September 16, 1998

Computerising and Coding the New Survey of London Life and Labour

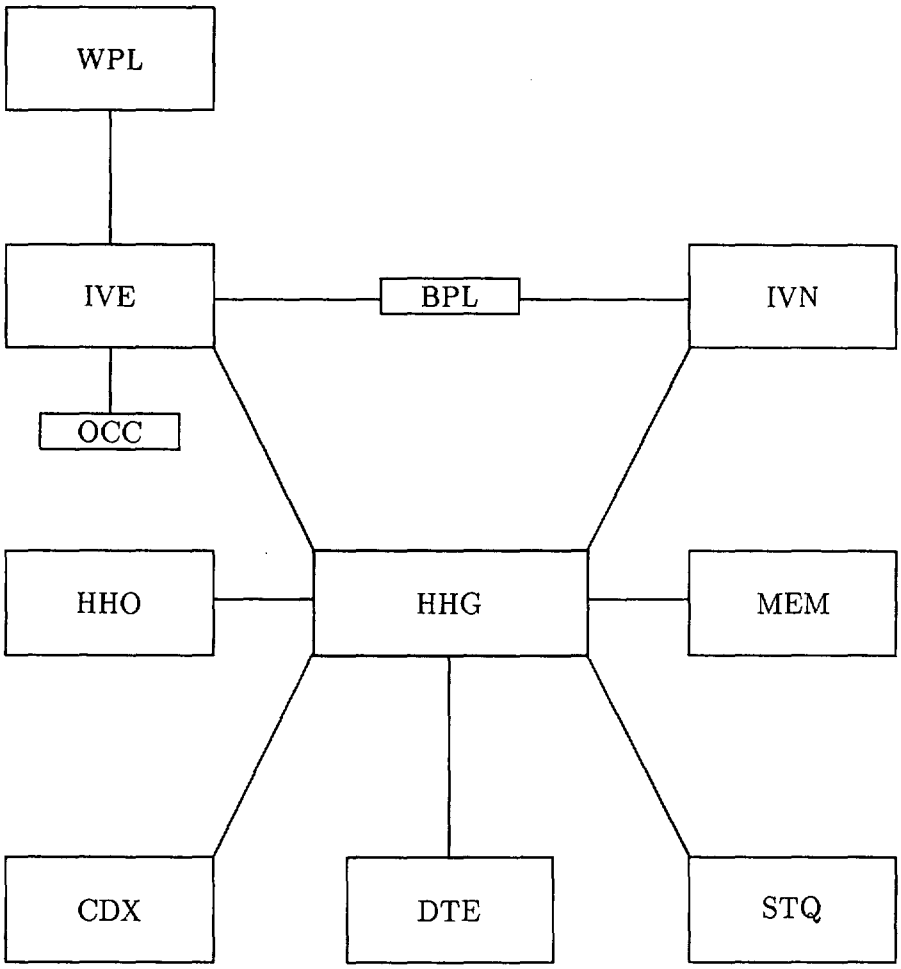
A Companion Paper for the Codebook (Release 2.0)

Roy E. Bailey
University of Essex

Anna Leith
London School of Economics

Acknowledgement: The computerisation, coding and analysis of the New Survey of London Life and Labour was funded under ESRC awards no. R000235697 and no. R000221981.

The *New Survey* Data Tables



New Survey Data Tables:	
HHG	General household information, one record per card.
HHO	Other sources of income, one record per source.
IVE	Earners' information, one record per earner.
IVN	Non-earners' information, one record per non-earner.
MEM	Free-format text fields.
STQ	Street Quality codes, at least one record per card.
CDX	Card numbers for supplementary cards, zero or more records per card.
WPL	Work place location codes, one record per earner.
DTE	Date of interview (modified), one record per card.
Look-up Tables:	
OCC	Occupation codes (related to IVE via a four character occupation code).
BPL	Birthplace codes (related to IVE and IVN via a seven character place code).

Each data table is related to the others via a unique, six-character code.

Contents

[1] Dr Simon Niziol, who carried out the workplace coding under the supervision of Anna Leith, kindly provided detailed notes for this section.

Section 1: Introduction

The aim of this paper is to provide users of the New Survey of London Life and Labour (hereafter the *New Survey*) data files with a more detailed discussion of the problems associated with the interpretation of some of the variables, than was suitable for the codebook (hereafter the *Codebook*) which accompanies the files. It should be regarded as a companion to the codebook and, while it can be read independently, reference to the *Codebook* may be necessary for detailed definitions of particular variables.

This paper also provides some brief background information about the *New Survey* itself, and about the general approach taken by the project team to the construction of the initial data files, both of which might be of interest to users.

Apart from the preliminary and design work, the computerisation project can be divided into three, overlapping, stages: (i) data entry from the questionnaire cards; (ii) checking and cleaning; (iii) coding. Data entry commenced shortly after the appointment of the project Research Officer, Anna Leith, in March 1995. Under her supervision and guidance, a team of Research Assistants completed the first stage by late November the same year. By that time, checking of the data had already commenced. Thereafter a wide range of tests for internal consistency and to identify oddities in the data was undertaken to uncover errors; consequently, amendments were made to the computer files. Where necessary, reference was made back to the original data cards to resolve problems of interpretation. Although the bulk of this work was completed by mid-1996, data cleaning continued in order to correct errors or obscurities that emerged during early substantive research using the data. By early 1997, the need for, and effort devoted to, such corrective work had become negligible. Even so, it is anticipated that minor corrections will continue to be needed and the opportunity remains open for these to be made as necessary.

The routine aspects of coding (e.g. converting money fields into a standard format) were carried out in early 1996. The process is designed in such a way that, following amendments to the data, the whole coded dataset can be re-created, thus automatically re-coding any element of the data that has been edited since the previous version of the dataset was produced. More importantly, several major special coding projects were also undertaken covering occupations, birthplaces and street quality. The bulk of the work on these projects began in the latter part of 1996 and was completed in May 1997.

Section 2 provides more detail about the *New Survey* itself and some aspects of the computerisation project. The remainder of the paper then falls into two parts. Sections 3 to 9 describe some of the areas for which the problems of interpretation were most acute and discuss how they were handled. Sections 10 to 12 explain the principles and practical details associated with the major coding projects that formed part of the project. Sections 13 (workplace coding) and 14 (interview dates) provide brief background for two extensions to the data added in Release 2.0 of the files.

Section 2: Background

The New Survey of London Life and Labour

The New Survey of London Life and Labour was one of a number of pre-World War II social surveys that sought to examine the extent and causes of working class poverty. Its main objective was to update the findings of Charles Booth's survey of the 1880's *Life and Labour of the People of London,* specifically to address the question of whether poverty in London was increasing or decreasing.

The *New Survey* area covered 38 London boroughs; the inner County (of London) as surveyed by Booth, and, in order to reflect population changes and preserve the economic character of the area under investigation, an additional 9 adjoining boroughs or urban districts: Barking, East Ham, West Ham, Leyton and Walthamstow in Essex; and Tottenham, Hornsey, Willesden and Acton in Middlesex.

More than one method of surveying was used: a street survey following Booth's practice; and the household sample survey from which the *New Survey* data files are constructed. The household survey was directed by Professor A. L. Bowley and used procedures he had developed earlier, in his Five Towns surveys. The most significant points are outlined briefly below:

- The survey was directed at working class households as defined by the occupation and earnings of the head, with an upper limit of £250p.a. Middle class households were excluded from the final sample.
- Addresses for investigation were selected using a variable sampling factor, depending on the size of the borough. The final sample represents approximately one in 50 working class households from the survey area.
- The information was gathered by a team of interviewers (there were at least 171 of them), who visited all the households at the selected addresses and recorded information on pre-prepared blank questionnaire cards, one household per card.
- There were precise instructions on how the cards were to be completed, and on the action to be taken if an address was unoccupied or information difficult to obtain. The completed cards were subsequently verified, and an attempt made to complete certain categories of missing information.

The data were gathered during the period 1928–March 1932. Complete sets of questionnaire cards survive for all but two of the boroughs (Tottenham and Walthamstow), providing detailed information for 26,915 working class households; their internal structure; income; the employment and birthplaces of individual members; and housing conditions.

The Data Files

The underlying aim during the construction of the *New Survey* input level data files was to create an accurate representation of the original data as it appeared on the cards and to preserve all the information collected. But in transcribing the data from the cards, two types of problem were experienced and mean that the entries in the computerised version of the dataset at level A (the textual version) are not an exact copy of the original in every case. Where this has occurred, it is because the intention is to make the computer record represent what the interviewer meant, rather than to reproduce the letter of each particular entry on a card.

The most common problem encountered was the difficulty in deciphering some of the entries. The handwriting of some interviewers is not easy to read and the amount of care they took in completing the cards varied considerably. In these cases detailed scrutiny, checking of external references sources and intelligent guesswork enabled the elimination of many of the queries with a reasonable degree of certainty. Of those that remained an attempt to make some sort of entry was possible in most cases; there are just a few instances in which data has been lost. Wherever there was any doubt about an entry, a note was made in the appropriate text field (REMI_M, REMA_M or REMG_M — see *Codebook* section 9, p. 35).

The second, and more contentious issue related to the cases where information had been recorded on the cards in an ambiguous, inconsistent or incorrect manner. This occurred for two main reasons: (a) because the interviewer had not followed the survey guidelines on how to complete the cards (see the *New Survey* vol. III, appendix I); and, (b) because the circumstances that the interviewer was attempting to record had not been envisaged in the survey design and, in consequence, were not provided for on the cards or covered in the instructions. The problem was compounded by the fact that there were so many interviewers (171), all with their own peculiarities. The approach in computerising the data was to seek to resolve as many of these cases as possible in a consistent manner that tries to preserve the intended meaning of the the card entries, and to signal these actions to the user via notes in the appropriate text fields of individual records or by comments in the *Codebook* where groups of records are affected.

As the process of data entry and checking progressed, it became clear that there were a number of areas that had caused particular difficulties for the interviewers and that these difficulties would have significance for users of the dataset. During the data cleaning stage, these areas were the subject of detailed scrutiny and much editing was done to try to make them reliable for the purposes of statistical analysis.

It must be acknowledged that the way in which these problems should be tackled by those constructing data files from historical source material is open to question, and not all will agree with the approach adopted by the project team. The purist, seeking to create an electronic edition of the source, might argue that no attempt should be made, in the first instance, to do other than reproduce the original entries as they appear on the cards; others like the project team, whose ultimate interest is in producing a structured dataset for the purposes of statistical analysis, would accept that, under properly controlled conditions, some interpretation and guesswork is necessary in order to produce a usable set of data files.

Coding

As already noted, the *New Survey* data files do not express exactly, and only, what was written on each survey card. In addition to the checking and cleaning already mentioned, the data were extended by coding much of the information obtained. The pattern of coding has three distinct aspects:

1. *Coding at the point of data entry.* While almost all of the effort in the initial, data entry, phase of the project was devoted to reproducing what appeared on the cards, a very limited amount of coding was undertaken by the inputters. This covers three topics: dwelling tenure status (RTEN_H); "other income" sources (SRCO_S); and, employment status (STAT_E). The inputters operated according to rules provided and, except where elsewhere noted, no significant issues of principle arose. In each case, the information entered was subjected to the same scrutiny and checking as if it had appeared on the original

questionnaire card.

2. *Routine coding.* All of the variables which have an obvious numerical value (e.g. rent, earnings, hours of work) were given coded values at level E. That this is necessary can be seen by casual inspection of the level A tables: all sorts of non-standard and non-numeric entries appear in these fields. On the whole, apart from the corrections mentioned above, the entries at level A reproduce what appears on the cards.

   Translation into standard code values took place by first identifying all of the unique entries in each of the fields in question. These were used to construct a set of "look-up tables" which record the unique entries together with the assigned code. Each look-up table was then applied, in turn, to the level A files in order to generate the associated level E values. (The look-up tables, themselves, do not form part of the *New Survey* data and have not been released into the public domain.) This process was applied to all the fields that appear in both level A and level E files.

3. *Special coding.* Separate projects were needed to handle the magnitude and complexity of three additional coding topics: street quality, occupations, and birthplaces. These coding exercises are explained in sections 10, 11 and 12, respectively, below.

   Two different approaches were adopted in the special coding projects. One approach, that used for street quality coding, involved identifying all the unique entries for the field concerned (in this case `ADDR_H`) and assigning codes to these entries. It was then a simple matter to apply the codes to the relevant records in the underlying data file. The second approach, that adopted for occupations and birthplaces, involved assigning a code to each of the records in the underlying data files. The appearance of identical entries (e.g. "Labourer" as an occupation) made replication necessary in this approach, although this was largely mitigated by the development of algorithms to code automatically repeated occurrences of identical entries. The second approach did, however, have the advantage of flexibility in that identical entries *could* be assigned different codes depending on the availability of information in addition to that which appeared in the field being coded. (The same effect could be achieved via the first approach by changing the underlying information that determines the codes, an artifice that was applied in a small number of cases in the street quality coding.) In practice, neither approach seemed better than the other for every coding exercise and that which was most suitable to the task in question was adopted.

Section 3: Rent and Housing Tenure

Characteristics of the Original Data

The difficulties presented by the *New Survey* household rent data arise from the inconsistencies that occur in it; in a significant number of cases it would appear that the interviewers either disregarded, or were unable to follow the instructions they were given in connection with the composition of the rent details to be obtained. As a result there is evidence on the cards of some considerable variation between households in the precise meaning of the information recorded under 'Rent', and, additionally, the same information is presented in different ways by different interviewers.

The instructions regarding the collection of rent information are reasonably explicit:

> *Rent* should include rates (including water rate if any). If the house is owned by [the] occupier, or he has it rent free, this circumstance should be stated and the approximate rental value should be entered. Special care in entry should be taken where the occupier sub-lets part of his accommodation. The total rent paid by him to the external landlord should be entered; the rent for the rooms sub-let will appear on the face of the card under "Income from Other Sources" and also on the back of the sub-tenant's card as the rent paid by him to the principal tenant. Where a house is occupied by two or more sub-tenants paying rent to an outside landlord, this case should be distinguished from the former. (*New Survey*, vol. III, p. 415)

In practice, cases of "two or more sub-tenants paying rent to an outside landlord" were often marked as Independent Tenants. Also, as reference to the specimen household investigation card shows, it was intended that the rent amount entered should be that paid *per week*.

Problems with the rent data recorded on the cards occur in relation to all of the areas covered by the above instructions:

1. There is evidence that interviewers did not always include rates in the rent amount. In some cases there is a specific qualification to the effect that the figure recorded for rent does not include rates and/or water rates. In many of these cases rates information is available elsewhere on the card, albeit on a different basis from the rent data, e.g. annual as opposed to weekly; for these households a total figure for rent plus rates can be calculated. An inconsistency arises for those cases in which rates are not included in rent and no information about rates is provided elsewhere on the card. Also, when there is no reference to rates, it is not evident whether rates are zero or have been included in rent, as required.

2. Where an occupier sublet part of the accommodation, interviewers were instructed to enter the gross rent paid to the external landlord. For the most part this instruction was followed, but in some cases there is evidence that the rent amount given is net of receipts from subtenants.

   This occurred on the cards for Bethnal Green, Shoreditch and West Ham, where the words "net" or "nett" sometimes appear on the card in connection with the amount entered for rent. The interviewer may also have recorded a corresponding gross figure in the accommodation or general remarks (`REMA_M` or `REMG_M`) but this is not always the case. The presence of a net figure implies that the household had an income from subtenants and there is usually an entry in the other income section on the face of the

card to indicate that the household was receiving rent payments from subtenants, but again this is not always the case. Additionally, there may be other evidence on the card to indicate that the household was the principal or main tenant at that address, thus supporting the implication that there were subtenants; however, in a few cases this evidence appears contradictory, and indicates that the household in question has itself the status of subtenant or is an independent tenant at the address.

It will be clear that the information on the cards in these "net rent" cases can vary considerably. In many of them, information is available to allow a gross rent figure to be computed but this is not always the case and, hence, inconsistencies can arise in comparisons among households. Also, there is the possibility of error in calculating aggregate household income when enough information is on the card to enable a gross rent payment to be calculated but there is no explicit source of income in the form of rent received. Finally, in some cases, the rent of a subtenant or independent tenant is recorded as net. This could imply that the subtenants were subletting or, possibly, that the interviewers misused the term "net rent".

3. In cases where it was stated that the accommodation was being occupied rent-free, it appears that one of three things could be meant: (a) that the household was not, in effect, paying any rent because it covered this cost entirely through subletting; (b) that the accommodation was provided free of charge by a relative; or, (c) it was being received as a perk connected with the employment of someone in the household, often described as part of earnings. (The last appears to hold for the majority of the rent-free households.)

   Contrary to instructions, a notional rent figure is not always given in these cases. Where it does appear, the notional rent is typically in parentheses. Finally, there is sometimes, but again not always, a corresponding entry recorded on the face of the card (under "Income from Other Sources") to the effect that the household is receiving accommodation as a perk (this is classified as a category of other income). Where this is absent, a question of imbalance in household income is raised.

4. The entries sometimes indicate that, rather than being a weekly amount, the period to which the rent figure relates is a month, quarter, six-months, a year or some other time interval. A problem arises in those cases where the amount given is clearly too large to be a weekly figure but the actual period referred to is not stated.

5. Where a dwelling is owner-occupied (1,128 households), the value of `RENT_H` should be treated with extreme caution. There appears to have been little consistent practice on the part of interviewers in handling these cases, presumably because they experienced difficulty in obtaining the required information; it is thought that the information about notional rent which was to be sought from owner occupiers, was probably not readily available to most of them and in its absence, each interviewer took whatever action he could. In about half of owner occupation cases there is no information at all beyond the entry of the words "own" or "owner"; in the remainder the meaning of what appears may be quite ambiguous. Where an explanation is given, there is considerable variation.

   Where a numerical entry is made, the "rent" figure may represent: mortgage payments; rates only, with or without water rates, taxes; rateable value; ground rent; or rent/rentable value, with or without rates.

INTERPRETATION

The priorities in constructing the dataset were to ensure that the entries in these different types of problem case had been correctly interpreted and consistently handled, and that their meaning would, as far as possible, be apparent to users.

It was decided that this would best be achieved through a combination of judicious changes to the original entries, and coding. The approach adopted was to make changes only when they could be supported by the evidence of other information on the cards; unsupported assumptions were avoided wherever possible. In all cases where the changes made are not obvious, the original entry has been recorded in a text field.

INTERPRETATION AND DATA ENTRY RULES

1. Wherever it was stated that the figure entered for Rent excluded rates, an inclusive figure was calculated from other information on the card, if it was available, and this amount was substituted. A record of the original figure was made in the accommodation remarks (`REMA_M`). If no rate figure was available, the words "rates excl" were inserted as the first line in accommodation remarks to ensure these records could be identified. Differential coding could then be applied (see the tenure status field, `RTEN_H`). Where the entry for Rent was described as "rates only", these words were inserted as an accommodation remark to facilitate coding. If, as in the vast majority of cases, there was no reference to rates, it was assumed they were included or were zero.

2. Where the interviewer indicated that the figure entered for Rent was a net amount, a gross figure was substituted whenever possible. This could be obtained in one of two ways: either (a) the interviewer may have recorded the gross rent elsewhere on the card; or (b) one or more entries for rent received from subtenants may have been made under other income which could be used to calculate a gross figure. A record was then made of the original net figure in accommodation remarks, together with the means used to arrive at a gross one if this was not already obvious. Again, if, as in the vast majority of cases, there was no reference to a net amount, it was assumed that a gross figure had been given. In a few cases in West Ham, where families were clearly occupying dwellings on the basis of independent tenancies, the net figure has been retained even when a gross one was given.

3. Where a possible imbalance in household income arose from the substitution of a gross figure for a net one in the rent field, because there did not appear to be any income from subtenants, it was decided that no special action should be taken, beyond the entry of a note in a text field to the effect that the household may have been making an amount of money from subletting, equivalent to the difference between the gross and net rent amounts. In no case was any entry for rent from subletting constructed from this information; such action could not be justified without evidence that subtenants existed. It is also argued that any attempt to identify these "imbalance" cases through coding would be misleading since it is not known how many others may exist among the vast majority of records that are assumed to contain a gross rent figure.

4. Where it was clear that the figure given for rent was net but there was no further information to allow the substitution of a gross figure, the words "net" or "nett" were added to the entry for rent to ensure these records could be identified. Again differential coding would be appropriate (using the tenure status field). This action was taken regardless of

whether the household appeared to be a principal, independent or subtenant, i.e. the use by the interviewer of the word "net" was accepted at face value and it is assumed that it always implies subletting.

5. Where the entry for rent was too large to be a weekly figure and the period referred to was not stated, an attempt was made to infer the intended period from (a) the rent amount, and (b) the practice of the interviewer in other cases. It seems that an annual figure was intended for most periods in excess of a week where the period is not explicit.

6. "Rent-free" accommodation was defined as meaning that rent was not payable, i.e those cases where the household was covering its rent costs through subletting were not treated as rent-free. In the majority of actual rent-free cases, there is no notional rent figure in the rent field, but where such a figure did appear, in cases where it is clear that the accommodation was being received as a perk, it has been entered in parentheses and a corresponding entry made under perquisites in the Income from Other Sources section, to avoid giving the appearance of an imbalance in household income. Rent-free cases are, of course, distinguished by the coding relating to tenure status (`RTEN_H`).

7. In contrast to the other categories of problem with respect to rent, no attempt at all has been made to interpret or clarify the individual entries recorded against rent in cases where the dwelling was owner occupied; this information has simply been entered into the database as it appeared on the cards. It was decided that the ambiguities and inconsistencies that existed in this data were too extensive, and the likelihood of being able to obtain sufficient information from the cards alone to provide comparable, or indeed useful, information across households, too remote, to make any further work on these cases worthwhile within the context of the existing project.

Implications

Despite the objectives underlying the interpretation rules, it is acknowledged that weaknesses remain in the *New Survey* rent data. As the rules outlined above indicate, these lie chiefly in the "net rent" and owner occupation categories. First and foremost, the entries recorded against rent in the latter category are clearly quite unreliable and do not provide any consistent information about this class of households as a whole. Second, the assumption that the appearance of a net rent figure on a card always implies that the household was subletting, may also be open to question; in those cases where other information on the card indicates that the household was an independent or subtenant and there is no evidence of subletting, it is possible that the interviewer has used the word "net" in a misleading way. Other information on the cards, in particular the occasional comments made by some interviewers, suggests that the cards for all the households dwelling at the same address may sometimes have completed at the same time; the use of "net" to describe the rent details recorded on one card, may therefore have taken place in the context of a gross figure recorded on another card in the same group.

Section 4: Accommodation

Characteristics of the Original Data

The problems experienced in handling the elements of the *New Survey* relating to accommodation characteristics, arise from the limitations imposed on those recording the original data by the standardised list of separate rooms or facilities that appears on the survey questionnaire, and from the absence of instructions as to how they should deal with the many cases that did not conform to this pattern.

The specimen investigation card shows that interviewers were required to ascertain the number of bedrooms available to the household, and to indicate whether or not any of the following facilities existed: parlour; kitchen; scullery; pantry or larder; bathroom; yard; garden; allotment. In practice, it appears that interviewers also commonly distinguished facilities that were shared with other households or which the household in question had the use of, but there is no evidence of any formal instructions to this effect. Space was also provided on the card for optional, additional remarks about the quality of the accommodation. The (published) instructions for completing the accommodation section of the card are as follows:

> *Rooms* A kitchen is distinguished from a scullery by having a coal or other cooking range and being usable as a living room. If a parlour is used for sleeping as well as for a living room it should be counted as a parlour only, but the fact of sleeping noted. (*New Survey*, vol. III, p. 415.)

No further guidelines appear to have been given.

When constructing the dataset, a wide variety of non-standard responses to the questions asked were encountered. These responses sometimes appeared beside the facility which they described, as an alternative to the standard "yes/no/shared" response, or were recorded in the accommodation remarks in as well. Examples are given below.

The most significant problem of interpretation arose in those cases where the entry on the card indicates that the facility which existed was not precisely the one listed but was an alternative which may or may not be regarded as equivalent. For example: should a "roof drying ground" or "veranda" be interpreted as the same sort of facility as a yard; a "tap on the landing" or a "washhouse" as a scullery; a "safe on the landing" or "coal cellar" as a larder/pantry? Another question of interpretation arose in cases where rooms were recorded as being combined or used for more than one purpose. Should a dwelling which has a combined kitchen and scullery be treated as having both of these facilities or just one of them? Lastly, how should cases be dealt with where the information given about a facility suggests that it might be shared with other households? For example, does the entry: "washhouse in the yard" imply a communal facility if the building is a multi-occupancy one?

Interpretation

Consistency in the interpretation of the data relating to accommodation characteristics was achieved by first identifying a number of distinct categories of non-standard response; then deciding which of the entries: "yes", "no" or "shared", is appropriate in each case, according to what was implied about the existence of a particular room or facility, or its equivalent; and, finally, assigning each individual response to one of the categories.

In dealing with cases in those categories where there was a question of whether or not a

facility might be shared, the general rule was to treat what was written on the card at face value and to apply a rule of interpretation (see below). But when checking and cleaning the data, an attempt was made, in particular cases, to take account of any obvious evidence that appeared to contradict the agreed entry for that particular response; for example, a "washhouse" or "drying ground" is likely to be a shared facility in a Peabody block, regardless of whether the card specifically states that it is shared, or is, respectively, in the yard or on the roof. However in most cases the response on the card has not been questioned.

When interpreting the data, precedence was given to what had been written in the accommodation remarks over the entry recorded against a specific room or facility, only where the latter was negative. But where there was already a "yes" or "shared" entry against the appropriate facility, that entry took precedence. For example, if the interviewer had recorded "no" against scullery but had indicated that there was a shared washhouse, the entry in the dataset for scullery would be "shared", but if instead, the interviewer had recorded "yes" against scullery, that entry would not be overridden by the information about the washhouse. In all cases, any additional information given on the card about the characteristics of the accommodation appears in the accommodation remarks field in the dataset.

EXAMPLES AND INTERPRETATION RULES

1. Comments describing the size, location, or other characteristic of the room or facility. Examples are: garden: "small"; garden: "back & front"; parlour: "1+dining room". As these do not affect the main issue (i.e. whether or not a room or facility is present) a letter Y (or a 2, in the second and third examples) was entered in these cases.

2. Comments indicating that a room is combined with another room or used for more than one purpose. Examples are: parlour, kitchen, scullery bracketed, or otherwise linked together as one room; scullery: "sink in kitchen"; bedroom: "bedsitter" or "one combined room"; parlour: "used for sleeping". These cases were interpreted as indicating the presence of just one room or facility; a letter Y was entered for the first room and N for subsequent rooms in the bracketed or linked cases. In the other multi-purpose cases, Y was entered for the room or facility that represented the originally designated purpose, e.g. Y for parlour in the last example.

   But it was decided that a positive entry on the card for bath should be interpreted in a broad sense as meaning that there is either a bath or bathroom, i.e. a facility other than the general sink for washing persons. Therefore, where bath occurred in combination with another facility, it was given a Y entry. Examples are: "scullery and bath combined", bath: "in scullery". Y was entered for both scullery and bath in these cases.

3. Comments which suggest that the facility may be shared with other households. Examples are: scullery "on landing "; bath: "in yard"; yard "all round house" or "on roof". Unless there was clear evidence that the dwelling was occupied by just one family, these were interpreted as meaning that the facility was shared; a letter S rather than Y was entered against the facility.

4. Comments which suggest that an alternative facility exists which is comparable to one on the list, and does not appear to be shared. Examples are: pantry: "cupboard"; pantry: "safe on landing"; scullery: "washhouse". These were interpreted as a comparable facility and Y entered against the appropriate one in the list.

5. Comments which suggest that an alternative facility exists which is comparable to one on

the list, but is probably shared. Examples are: scullery: "washhouse in yard", "communal washhouse", "laundry", "sink/tap/water on landing/in yard"; yard: "(roof) drying ground/platform", "playground", "square", "forecourt", "area"; kitchen: "gas stove on landing", "cooks on landing". These were interpreted as a comparable, shared facility and `S` entered against the appropriate one on the list.

6. Comments which suggest that although an alternative facility exists, it is not comparable to one on the list. Examples are: yard: "veranda", "balcony", "2ft forecourt"; parlour: "boxroom"; pantry: "coal cellar". These were interpreted as meaning that the facility does not exist and `N` entered in the appropriate place.

7. Comments which suggest that the facility exists but is not used by this household; is not available to it; or is used for a significantly different purpose from the one intended. Examples are: bath: "not used", "awaiting repair", "to look at"; yard: "upstairs tenants", "downstairs", "see card A"; parlour: "used as shop". These were interpreted as meaning that the facility does not exist and `N` entered.

Implications

The adoption of the principles detailed above has implications for users of the accommodation data. Most significantly, it means that certain facilities may have been interpreted more broadly than was intended by those who designed the *New Survey*; chiefly: "scullery" as meaning access to a water supply and/or some sort of household washing facilities — either for general purposes or for clothes, that is separate from the other rooms in the accommodation; "yard" as meaning access to an outside place for drying clothes and/or for children's play, or merely access to a reasonable sized open space designated for use by the household; "kitchen" as meaning access to cooking facilities that are separate from other rooms in the accommodation; and "bath" as meaning access to facilities for washing persons other than the general household sink or tap.

Also, a rough analysis of the post-interview coding recorded in the odd characters field, suggests that some 19,675 out of the total 26,915 households surveyed may have lived in multi-occupancy buildings, perhaps rather more than is otherwise suggested by the information on the cards. In view of this and the inevitable inconsistencies in the way the original data was recorded, users of the dataset may wish to consider how meaningful is the distinction between the entries: "shared" and "yes", in the context of certain facilities that were likely to have been communal in multi-occupancy situations, particularly yards and gardens.

Section 5: Household Income

Characteristics of the Original Data

The elements of the *New Survey* data that make up household income, i.e. the earnings of members plus other sources of income, are problematic not so much because of difficulties in interpreting what is meant by particular entries, but because without corrective action the information recorded in certain cases gives a false picture of the total income coming into the household when the completed dataset is analysed. This is not something which would have concerned those who conducted the original survey, for their data analysis appears to have involved a manual processing of each card that would allow for the examination of individual cases; it is an issue that assumes importance in the context of the computer analysis of the coded data.

According to the guidelines reproduced in the *New Survey* vol. III, Appendix I, interviewers were to record the full-time and last week earnings of every working member of the household and the full-time earnings in the last employment of those who were unemployed or incapacitated. This included lodgers and those "so occupied as to be usually living away from home" (p. 413). The latter were to be "specially marked so as to exclude them from some subsequent tabulations" but both their total earnings and the amount they contributed to the support of the family were to be recorded. Each lodger was to be described as such in the entry under relationship to head of household; as distinct from a subtenant, he/she was someone who shared a common commissariat with the family. However, "no one related by blood, marriage or adoption to the head of the household" (p. 414) was to be counted as a lodger. Lastly,

> *Income from other sources* is intended to include income from lodgers, rooms sublet, property, pensions, poor relief, insurance benefit, charitable sources, gifts from relations, and perquisites, etc., when they are known. ... It is realised that these items often cannot be completely ascertained, but where they can be stated they are of value in enabling the economic position of the family to be judged. (*New Survey* vol. III, p. 414)

The main problem, one of double counting or double recording of income, arises because in some cases an entry was made under sources of other income, as allowances from relatives, or board and lodging payments, that did not represent additional income coming into the household but was money already included in another source recorded on the card. This source is usually the earnings of the member of the household referred to in the allowance or board payment, but can also be his/her pension, social security benefit, poor relief, savings or other source. In recording both amounts in these cases, the interviewers were in part, acting according to instructions, but the fact that some of them included certain relatives (e.g. brother) within the definition of "lodger" increased the scale of the problem.

The potential for double counting to occur was found to exist in the following situations:

- Where both the earnings of an absent worker, e.g. someone away at sea or in domestic service, and an amount which he/she sent home to help the family, are recorded. The latter would usually be described as an allowance paid to a wife, or received from a son/daughter, under other income.
- Where both the earnings of a lodger and the amount paid by him/her to the family for

board and lodgings are recorded. Alternatively where another income source, e.g. pension, received by a lodger is recorded as well as the lodging payment.

- Where a member of the family or relative has been treated by the interviewer in the same way as a lodger, and both his/her earnings (or other source of income) and payments for board and lodging are recorded. This occurs most obviously in respect of adult working children who are living in the parental home and paying something for their keep, but is also found in connection with other relatives living with the family, e.g. brothers and sisters (in law), aunts, uncles and parents. In many cases the payments are described as "allowances, contributions, assistance or help" from a son/daughter *etc.*, but in others either the relative or his/her board payment is erroneously described on the card in terms of a lodger, and in yet others the payment is recorded as rent.

Some questions of treatment arise for these cases: should the information be recorded exactly as given and an attempt made to eliminate double-counting through the coding scheme? Alternatively, should the double-counting be eliminated from the dataset (and notes entered in a text field that would allow users to reconstruct the original entries)? If the latter course is chosen, it raises the question of which element of income to omit in each case.

INTERPRETATION

It was decided to give priority to providing as accurate a representation as possible of the income available to each household as an entity, rather than to concentrate on preserving the details relating to individual income in all cases. In consequence, an attempt has been made to eliminate the double counting (or double recording) of income in cases where the structure of the dataset as created at the input stage, would not allow its presence to be readily detected during subsequent work.

The first step in the process was to ensure that the source of income codes (SRCO_S) were correctly assigned with respect to: allowances from relatives (code letter A); payments by lodgers (code letter L); and rent received (code letter R). These were the three categories identified as being the most serious sources of ambiguity and inconsistency in the context of double counting.

Rent Income coded as "Rent" should always denote money received from a subtenant as payment for the use of a room or rooms. The important point here is that regardless of whether the subtenant is, or is not, described as a relative, he/she belongs by definition to a separate household. Hence, there should be no possibility of double counting: income coded as "Rent" always represents additional money coming into the household.

Lodgers Where the "Lodgers" code has been used, the income so described should always represent a payment for board and lodging made by an individual or individuals who are recorded in the Earners or Non-earners files as being resident in the household, but who are not members of the family. Here, double counting can occur: income coded as "Lodgers" may be duplicated by money recorded as earnings (or as benefits *etc.*) associated with the persons involved. Potential problem cases can be identified by the presence of the "Lodger" code in the relationship field, or of the word "Lodger" in the Description of source (SRCE_S), in the same household for which there is an entry coded "Lodger" in SRCO_S.

Allowances from relatives Unlike the other two categories, income coded as "Allowances from relatives" can have more than one meaning: it may mean financial assistance from

a relative or friend who is not part of the household (e.g. help with the rent from a son or daughter living elsewhere); it may be an allowance paid by a member of the family who is working away from home (e.g. a husband at sea); or it may represent a board and lodging payment made by a relative who is recorded as being resident in the household. The last two categories are the most problematic; double counting could occur in either the second or third situation when the income coded as an allowance was duplicated by money recorded as earnings or benefits, *etc.* of the individual in question. Such cases would not be obvious to the user from the structure of the input level data alone.

In view of this diagnosis of the potential problems, the next stage in resolving the issue of double counting was to examine all cases in the "Allowances from relatives" category where there was a possibility that the income thus coded might represent a duplication of money recorded elsewhere as coming into the household, and to eliminate the double recording. In order to achieve consistency, the following rules were applied when deciding which element of the income to delete in each case: if the person from whom the allowance derived was not normally resident in the household (i.e. was working away from home), the allowance was retained in the computerised record but the earnings of that individual were omitted; if the person from whom the allowance derived was resident in the household and was a relative, the earnings or benefits *etc.* of that individual were retained, and the allowance was omitted. Where information given on the card was omitted from the main part of the computerised record, it was reproduced in full in the text field, `REMI_M`.

Where the person was not normally resident in the household, but only earnings information for that individual appears on the card, the data is retained as it appears on the card; likewise, where only details of a board and lodging contribution appear on the card and no earnings information appears for an individual who is resident, the original data has been recorded in the computer files as written on the card. There is no double counting problem in these cases.

Implications

It is accepted that the solution outlined above in respect of the third type of "Allowances from relatives" cases does not address the issue of family income pooling, but is based on the simple assumption that all members of a household who were related and normally living together, always shared all of their income. There is clearly a question as to how far this would really have been the case where, for example, adult working children were living in the parental home, or relatives such as a brother or parent in law were living with the family. An alternative interpretation might therefore have involved an attempt to define the extent of "the family" in the context of income pooling. However it was felt that such an exercise would be beyond the scope of the computerisation project.

Section 6: Relationships to the Head of Household

Characteristics of the Original Data

Interviewers were instructed to write down the relationship to the person identified as head for all individuals in the household (wage-earners and non-wage-earners). The interviewers had little difficulty with the commonest nuclear families, comprising a man, wife and their children. Even here, though, there are a few peculiarities such as cases where the interviewer entered the name of each person rather than a relationship.

Most of the problems occur for the less common household structures or those which are complicated by the presence of several generations. The difficulties faced by the interviewers stem from two main sources: (a) identification of the head; and, (b) assigning relationships to other members of the household. In some ways, the degree of complication is an outcome of the survey procedures themselves: the broader the definition of a household, the greater the prospect of encountering awkward cases. This appears to be particularly true for the household survey of West Ham (see section 9, beginning page 25).

The identification of the head appears to have been especially awkward in cases where (i) there were two or more individuals of "equal status" (e.g. two spinsters); (ii) there were two (possibly three) generations of adults in the household; or (iii) the household appeared to be one component in a more complicated web of family relationships within the same dwelling. For each of these types, the interviewers often failed to identify a head, sometimes entering a status (such as, "Spinster" or "Widow"), a relationship to other members of the household, or a proper name.

At least for the multi-generation households, the instructions given in the *New Survey*, vol. III, p. 414 are clear:

> *The Head of the Household* is the husband in an ordinary family, or his widow; but where married sons or daughters are also living in the family, the married man of the younger generation is counted as head.

It is apparent that many interviewers, consciously or otherwise, resisted the application of this rule which would require "the married man of the younger generation" to be designated as head when their commonsense dictated otherwise. For instance, it must have seemed odd for an interviewer to designate as head a married man of the younger generation (perhaps in his early twenties) who was living, temporarily perhaps, in the home of his spouse's parents, the latter being clearly responsible for the domestic arrangements such as the payment of rent.

It appears that, by and large, the interviewers ignored the rule where it did not accord with their view of what was sensible. Instead of direct contravention, however, they often wrote down a set of relationships with no explicit identification of the head.

More frequent, but easier to handle, are the transparent cases for which the interviewer simply ignored the instruction to identify a head. For example, it is not uncommon to find the only two adults in a household being designated as "Husband" and "Wife", or as "Father" and "Wife". In these cases, the husband (or father) is obviously the head of household. More generally, the head (or husband) is not infrequently given as "Father".

Another circumstance is where the household is part of a larger domestic group. In these cases, the head is likely to be designated as a "subtenant" or even, in a few cases, as a "Lodger".

The identification of relationships to the head is obviously most problematical in cases where no head has been explicitly designated. In most, though not all, such cases it is possible to make an inference about the relationships from what is written on the card. Even for the more straightforward cases for which a head is designated, liberties were taken in the attribution of relationships. For example, the "Wife" of the head is sometimes entered as "Mother", not an uncommon working-class practice.

Interpretation and Coding

With respect to RELA_E and RELA_N (see *Codebook* pages 23 & 33), the database entries adhere closely to what was entered on the card. Translations, mainly to designate a head where none is given, were made in the clearest and uncontentious cases. No systematic attempt was made during computerisation to ensure consistency in the identification of relationships.

A head was not explicitly identified for over 11,000 households. Of these, some 2,900 are single-person households; hence, the individual in question can be interpreted as head. The bulk of the remainder straightforwardly yield a head by routine methods, leaving about 1,400 households in need of more careful consideration. A head can be identified with little ambiguity in most of these from a visual check of the family composition.

Given the perceived merit of explicitly identifying a head for each household, it was decided to create a new field to designate headship: RELN_E for wage-earners and RELN_N for non-wage-earners (see *Codebook* pages 26 & 33). The rules adopted for creating RELN_E and RELN_N are as follows:

1. Single person households: that person is head.

2. Where a head is identified, that person is head. If no head is found but a husband is present, that person is head.

3. Where no head or husband is found but a married couple plus members of an older generation (e.g. mother or father) are present, then the person of the senior generation is head (male in the case of a senior couple or the oldest individual) unless that person is 80+ and not an earner. When the person is 80+ and a non-earner, the male member of the (junior) married couple is head.

4. Where the senior generation comprises aunts or uncles, grandmothers or grandfathers or in-laws, the oldest male (or female) in the junior generation is head unless all the remaining members of the household are under 21 years old. In this case, the senior male (or female, if no male is present) is head.

5. Where no head is found under the previous rules but a widow or widower is present, that person is head unless he or she is 80+ and not an earner and there is someone else in the household 21+ years of age. Where there exist younger members aged 21+ (and the widower is 80+ and a non-earner) the oldest male or oldest female of the younger generation is head.

6. Where there is one generation of adults plus children less than 16 years old and no head is found under the previous rules, then the head is the male member of a married couple, or the person identified as wife or the father or mother of the children (in that order of priority).

7. When there is only one generation (e.g. brothers, sisters, cousins, *etc.*), the head is the eldest male earner aged 21+ and less than 80. If there is no male earner in this category, the head is the oldest female earner (21+ and less than 80); if neither, then the oldest male (21+ and less than 80); otherwise, the oldest female (21+ and less than 80).

8. If all of the previous rules fail to identify a head, then the head is the person who seems to be the householder (i.e. who owns the house or pays the rent).

For the person identified as head, `RELN_E` or `RELN_N` equals `01`. For all others, the entry is `00`.

A potentially useful coding exercise would be to recode all the non-heads according to their relationship to the head as given by `RELN_E` or `RELN_N`. For the great majority of individuals `RELN_E` would equal `RELA_E`, and `RELN_N` would equal `RELA_N`. There would be differences for some households with complicated family structures and a large number of cases would need to be checked even for simpler structures. This work has not yet been undertaken.

IMPLICATIONS

For all but about 100 out of the 26,915 households, the designation of headship is unproblematical — though it should be emphasized that the rules outlined differ in an important respect from the instructions quoted above from the *New Survey*. Most other conventional approaches to headship should allow a head to be identified without ambiguity and would probably yield the same outcome as here for the bulk of the households.

Caution should be exercised with respect to interpretation of relationships other than that which designates the head. How much caution is necessary depends largely on the purpose at hand. Clearly, the finer degree of detail needed and accuracy demanded, the greater the caution needed. For example, a "Wife" who is not also the head can safely be interpreted as the spouse of the head unless a "Husband" (who is not also head) is present. Similarly, sons and daughters under the age of 18 can reasonably (though not definitively) be interpreted to be the offspring of one or more adults in the household (normally the head and his spouse).

Even if a comprehensive revision is made of all relationships, it should be recognised that some ambiguity will remain given that the information available to, or recorded by, interviewers was incomplete. This being noted, the proportion of households for which ambiguity would remain is probably much smaller that one per cent of the total.

All of the comments made here, and work done in the construction of the database, assume that a "household" corresponds to the information written on a single card. In only a very few cases did the interviewer use more than one card and this was done for lack of space rather than to express a complex web of family relationships. If a different, or broader, view is taken of what constitutes a household, then the issue of relationships is more complicated than is addressed above.

Section 7: Birthplaces

Characteristics of the Original Data

The birthplaces are probably the most problematic of all the elements of the *New Survey* data, a feature that arises from two serious deficiencies in the original survey:

1. The instructions to interviewers for filling in the household investigation card in respect of birthplaces are imprecise and state merely that the "birthplace of adults should be named in sufficient detail. (Metropolitan borough, county or town, or country if foreign.)" (*New Survey*, vol. III, p. 415). No attempt was made to define what was meant by "adult". The result is considerable inconsistency between households as to which members' birthplaces are given. Evidently, the interviewers had no clear idea about how to interpret the instructions. It appears that some recorded the birthplaces of the head (or husband) and wife only; others included all members of the older generations (i.e. head, wife and their parents, if present) but excluded offspring; most appear to have applied various age limits (e.g. 21 years, 18, 16 or 14) and to have included, where possible, everyone in the household of that age and over; while the remainder tried to give birthplace details for all members.

2. The design of the survey questionnaire with respect to how birthplace details were recorded did not encourage clarity. Reference to the specimen household investigation card shows that this information appeared on the back of the card, quite separate from the records of earners and non-earners to whom it applies. The interviewer was not required by the design of the card to link birthplaces with particular household members; but merely to list them. Again the result is inconsistency between interviewers in the amount of detail given. G E Bartlett and A N Winter, who between them dealt with at least 5,845 households, tended simply to record the name of just one place with no indication as to whom it applies; others recorded full details, matching each individual with a birthplace; while yet others gave a list with only partial or vague attribution, or none at all, leaving the user of the data to guess at what was intended.

A further cause of inconsistency in the birthplace data is a common one: failure on the part of some of the interviewers to follow the instructions they were given; in this case to record the required amount of detail (see above). Hence the information given varies from a general entry such as "the country" or "abroad", to the name of a particular street. G E Bartlett, again, contrary to instructions, favoured a general entry of "London" in most cases with no further detail as to borough.

Additionally of course, no single interviewer was ever completely consistent in recording the birthplace details; this would depend as much on the extent to which the required information could be elicited from the household, as on the interviewer's own practices.

The problems described above presented clear problems of interpretation in constructing the dataset. Examples of the sort of ambiguities encountered are:

1. "All born in ...". This might apply to all members of the household; all members of the family, (i.e. excluding lodgers); or all adults (including or excluding lodgers), in which case, a definition of "adult" is required.

2. "Children born in ...". This might mean all children in the household, or just the adult children; again a definition of "adult" is required.

3. A single unspecified birthplace is given. This might apply to head; head and wife; all members of household; all members of the family; or all "adults" If the place is in London, it might be that the parents were born elsewhere. If the place is outside London, it might be that some or all children ("adults" or not) were born at the present address.

4. A list of places is given with no further details. Here there is a question about the order in which it should be applied; possibilities include: head, wife, children in order of age; head, wife, earning children, non earning children; earners, then non-earners. If the number of birthplaces in the list is less than the number of individuals in the household there is ambiguity about to whom the places apply, e.g. 3 places for a household of head, wife and children, might mean that the third place applies just to the eldest child, or to all the children.

INTERPRETATION

In view of the considerable variation in the way "adult" had been defined, it was decided that the aim should be to achieve consistency in the interpretation of the birthplace data for all persons in the dataset, aged 18 years and over. This age limit was chosen as the one which offered a reasonable compromise between all the possible definitions of adult: school leaving age at 14 years; state insurance payable at 16 years; age adjudged capable of having dependents for unemployment benefits assessment at 18 years; formal age of majority at 21 years. It was also felt to be the age at which birthplace information would be available for the majority of individuals; those age groups excluded would be those for whom coverage was only very patchy. (Of the 31,777 individuals below the age of 18 years, birthplace details are available for less than 30 per cent, as compared with about 93 per cent for all individuals aged 18 or over.)

The cases which were most vulnerable to misinterpretation during the initial data entry stage appear to have been the "multiple adult" households, excepting those where the adults comprised only husband and wife. Therefore, it was decided that the review at the data checking and cleaning stage should concentrate upon a significant subset, from which would be excluded the following types of household: single person households; households comprising only head and wife (or husband and wife); households comprising head and wife (or husband and wife) together with other persons, none of whom were aged 18 years or over. The decision to treat head and wife cases as reasonably uncontentious is based upon the assumption that there is a strong likelihood of the couple originating from the same place; thus even where just one birthplace was given, it would not be unreasonable to apply this to both persons.

Consistency in the interpretation of the birthplace data was achieved through the application of a fairly detailed set of rules; these are outlined below. But it should be noted that some adaptation was necessary to take account of the practices of individual interviewers, as far as these could be ascertained from an examination of the entries on the cards. For example, where an interviewer had obviously defined an adult as 21 years and over, it made sense to use this age limit, rather than 18, for those cases in which an interpretation of adult is necessary.

Interpretation rules in unspecified cases:

1. If a single, unspecified birthplace was given on the card and it was either "London" or the borough where the family currently resided, it was applied to the head/husband, wife, close relatives and all children aged 18 years and over; if it was not London or the current borough, it was applied to the head/husband and wife and close relatives only (close relative was interpreted as brother (in law) or sister (in law)).

2. If the household consisted of two or more unrelated people, e.g. Single, Spinster, Bachelor, and just one, unspecified birthplace was given, it was applied to all persons if it was "London", otherwise to the first person only.

3. If two unspecified birthplaces were given, they were applied to the head/husband and wife, in that order.

4. If two unspecified birthplaces were given, but the household consisted of only one parent and offspring; the first place was applied to the parent and the second to the eldest child aged 18 and over, or , if it was "London" or the current borough, to all the offspring aged 18 and over.

5. If three unspecified birthplaces were given, and the third birthplace was not the current borough or "London", these were applied to the head/husband, wife and eldest child in that order. But in cases where where the third birthplace was the current borough or "London", they were applied to husband, wife and all children aged 18 and over. Finally where the list contained birthplaces both in London and elsewhere, it is assumed that the latter applies to the older generation, regardless of the order in which they appeared.

6. If the birthplaces of the head/husband and wife are specified, and there is one other unspecified place, these were applied as in item 5, unless the card states that the third place applies to the children in general, in which case, it was applied to all children aged 18 and over, regardless of location.

7. If four or more, unspecified birthplaces or other variants are given, the above principles were adapted in the most reasonable way possible.

8. Where the household contained other relatives, the following assumptions were made:

   (a) A brother (in-law) or sister (in-law) was likely to have come from the same place as the head (or wife);

   (b) Where a head (or wife) was born in London, the parents (in law) may well have come originally from the provinces. But if the birthplace of the head (or wife) was in the provinces, Scotland, Ireland, or abroad, it was assumed that his or her parents came from the same place. The same would apply to aunts and uncles of the head and wife.

   (c) Step children related to one or other parent: if the birthplace was London or the current borough and was the same for both parents, this was applied to the step child. Otherwise it was deemed unsafe to make assumptions without further information.

   (d) It was assumed the birthplace of any grandchildren was the same as that of the parent(s), if known, and was London or the current borough. But if the parents' birthplace was unknown, it was not considered safe to assume it was the same as the grandparents.

9. Unspecified birthplaces were not applied to lodgers, adopted children, and sons/daughters in law.

10. It was assumed that an adult meant a person aged 18 years and over unless the interviewer clearly used 21 and over.

IMPLICATIONS

Users of the birthplace data should be aware that the adoption of the principles outlined above in interpreting the unspecified cases may have resulted in some errors. Firstly, certain groups may commonly have no birthplace where one was intended, i.e. lodgers and non-close relatives, the parents of a head or wife born in London, and the adult children of a head or wife born outside London. Secondly, it is acknowledged that the rules adopted are weighted in favour of the older generations being born outside London and the younger in London. To put some perspective on all this: a rough analysis of the dataset suggests about 30 per cent of interviewers did not specify which birthplace applied to a particular individual, but these interviewers account for about 49 per cent of the households surveyed and about 47 per cent of the individuals in the dataset.

Section 8: Odd Characters on Face of Card

Characteristics and Interpretation of the Data

During the initial analysis of the *New Survey* source data, it was observed that various sequences of letters and numbers had been recorded, either towards the top left or top right-hand corners of the front face of a large number of the survey cards. There is considerable variation in what appears, but it was thought that some of the letters may have had particular significance. Although at that stage there was little understanding of their meaning, the decision was taken to record them as part of the data entry. They were to be entered just as they appeared on the cards, in a separate field for odd characters, ODDC_H. (See the *Codebook*, p. 20.) A more detailed examination of the data subsequently, and of the survey report in *New Survey*, vols. III & VI, has indicated that some of these sequences do convey information about the household on the card. They appear to represent coding, applied after the interviews were completed and probably used in the original analysis of the survey data. Two separate coding schemes have been identified:

Letters on the Right of the Card:

The letters which appear on the right of the cards, in blue, categorise the employment and relief status of the household; this simple coding scheme is outlined in the *New Survey* vol. III, appendix I, part 2 and it is suggested that it may have been used for the analysis of the causes of poverty which appears in *New Survey*, vol. III, chap. VI & vol. VI, chap. IV. These letters have usually been entered into the ODDC_H field in the database, following those which appear on the left hand side. There are four codes which may appear in any combination: "U", meaning that one or more persons in the household is unemployed; "S", meaning that one or more workers is absent due to illness or accident; "R", meaning that public relief (as opposed to insurance payments) was being received; and "C", meaning that the principle wage-earner is in casual employment.

There are no further details about the circumstances in which these codes were to be applied; it is therefore unclear in a few cases where individual circumstances had changed very recently, whether the employment or relief status thus coded, applied to the last completed week before the interview, or to the situation on the actual date of the interview. This problem is reflected in the *New Survey*'s own employment status coding scheme, where the decision was taken that the status given should as far as was ascertainable, apply to the last completed or full week before the interview.

Letters on the Left of the Card:

As no documentation has yet been found referring to the letters which appear on the left of the cards, there can be no certainty about their precise meaning. The explanation presented here is a tentative and incomplete one which has emerged from experience with the data. It is suggested that these letters are connected with the status of the household in the dwelling which it occupied and that they may have been used to carry out the analysis of rent and housing by different classes of tenement which appears in *New Survey* vol. III, chap. III & vol. VI, chap. II.

It is known that the cards for all the individual households surveyed at each address were originally clipped together following the interviews. It would seem that the left hand coding was added at some stage before the cards were separated and sorted into the filing order by type of household in which they now largely appear (see *Codebook*, page 11 for information

about the "File Number", FILE_H). The following codes have been identified with a reasonable degree of confidence: "A"; "IT"; "Flat"; "B"; "C"; "D"; "E". Also found are: "F", "G", "H" and other letters but there is less confidence that these are part of the same pattern. At least 5,500 cards appear to have none of these codes. This may have been the result of processing errors, but the proportion affected (about 20 per cent of the cards) suggests that this too may have significance, i.e. that the absence of a code may itself represent a category. The left hand codes occur singly or in certain combinations.

It is suggested that the coding may denote the following categories of occupier status and/or tenement: "A" is usually found in cases where it appears that the household occupies a house as owner, as the sole tenant, or as the main or principal tenant among others, with responsibility for paying rent for the whole house to an external landlord; "IT" almost certainly stands for "Independent Tenant", and would appear to denote cases where two or more households were each paying rent for separate tenancies in a divided house direct to an external landlord; "Flat" appears to mean the household was the tenant of a flat in a block such as Peabody or Guiness Trust; and "B", "C", "D", "E" would seem to denote subtenants who are paying rent to another tenant in the house, presumably the main tenant, or to the owner-occupier. The cards where none of these codes appear are mostly cases where the household seems to be the sole occupier of the dwelling, either owner or tenant. "A", "B", "C", "D" and "E" may occur in combination with "IT", and also apparently with each other. The former combinations would seem to indicate principal and subsequent independent tenants but the latter combinations do not appear to make sense. It is thought probable that they are not significant at all, i.e. that one or other of the letters in the combination is spurious within the context of the scheme being suggested.

An examination of *New Survey* vol. III, chap. III & vol. VI, chap. II shows that the following classes of tenement were identified for the analysis of rent and housing: separate houses, i.e. buildings devised for and used by one family on one plot of land, or structurally separate buildings; divided houses, where originally or by alteration a building is arranged for two or sometimes three or more families; workmen's flats, constructed as such; lessors, i.e. a class containing houses where the principal occupier lets off part of the house to one or more other families; and subtenants of the principal occupier. These categories would seem to correspond fairly closely with the explanation of the left-hand codes given above; thus "IT" equates to "divided houses"; "Flat" to "workmen's flats"; "B", "C" *etc.* to "subtenants"; and "A" to "lessors". The absence of a left-hand code would then seem to indicate a "separate house". It is not clear how the combinations noted above, i.e. "A" or "B" *etc.*, with "IT" and with each other, were part of this, or whether they were simply the result of coding problems. It is possible that "A" or "B" *etc.*, on their own indicate respectively, lessors and subtenants living in a "separate house", i.e. one which has not been structurally divided up; and that "A" or "B" *etc.* in combination with "IT" means lessors or subtenants living in a house that has been structurally divided. However this is pure conjecture; there is no evidence of such categories in the survey analysis, and the problem of "A" with "B" *etc.*, still remains.

Implications

If the above interpretation of the odd characters found on the front of the cards is accepted, then it is clear that they do not represent original survey data, but are derived information. As such they do not have any immediate significance for users of the dataset. However it should be noted that reference has occasionally been made to these characters in arriving at decisions

about how to interpret problem cases; in particular the right-hand codes have been used to help inform the interpretation of employment status, earnings last week and income from social security benefits; while the left-hand codes can have significance in relation to questions of rent and housing tenure. It must be acknowledged that such evidence may be regarded as questionable, but it can be argued that in difficult cases, where there exists no other reliable evidence, these codes do at least give an indication of how researchers at the time of the survey interpreted the data. Finally, these characters may have significance for those interested in the conduct of the original survey, in that they appear to tell us something about how the data analysis was done.

Section 9: A Note About West Ham

General Characteristics of the Original Data

Although the most significant problems in the *New Survey* data are associated with individual variables, users of the data may also wish to note that the records for one particular borough, West Ham, have certain peculiar features which did pose problems of interpretation.

It is thought that West Ham may have been the first borough in which interviewing was carried out. According to the *New Survey*, vol. VI, p. 36, West Ham was certainly among the first group to be completed, in 1929; all of the boroughs in the group being in the east and north east of the survey area. However it is not possible to obtain more precise information about this from the data itself, because the questionnaire cards used in West Ham were exclusively of a design found in the early boroughs (Bethnal Green and Stepney, in particular) which does not specifically request the interviewer to record his or her name or the date of the interview. While the interviewers' names or initials are given on most West Ham cards, the date is not. Additional research in the future may throw more light on this question.

The collection of the West Ham data appears to have been handled by a completely different group of interviewers from those used in the rest of the survey area, since unusually, not one of them occurs in any other borough. The group was also the largest for any borough although West Ham was not the largest in terms of sample size; 23 different individuals were used for a sample of 1,263, many of them for a relatively small number of cases. Also unusually, two or three interviewers names are given on many of the West Ham cards, and it is often clear from the distribution of the different handwriting over these cards and from the notes that have been recorded, that each had a hand in the collection of the data, in some cases either completing or correcting information obtained earlier by another person. The result is not infrequently ambiguous and difficult to interpret. It suggests that the West Ham group of interviewers may have been relatively inexperienced, supporting the view, perhaps, that they were the first to try out the survey procedures and data collection rules.

It may also indicate that West Ham was used as a test case and that procedures followed later were improved in the light of the experiences of those who carried out the work here; in particular that some attempt may have been made to clarify certain aspects of the definitions and instructions for filling in the cards (see the *New Survey*, vol. III, appendix I) This could account for the peculiar features of the West Ham data.

The most noticeable feature of the West Ham cards is that they appear to contain a high proportion of unusually complex households. These are households in which two or more families are recorded as living together, typically married couples and their children living with the parents and siblings of one spouse, or families of the same generation related by marriage who are sharing a dwelling. While complex households do occur in other boroughs, they evidently posed a particular problem for the West Ham interviewers, for the relationships have more often been recorded here in a way that fails to establish a meaningful unified household structure, usually because no-one is identified as head. In some cases, as elsewhere, the individuals in such a household are described in terms of their relationship to another person, e.g. a mother; but in others the relationships indicate two separate families or are simply muddled. If two or more names are also recorded in the name field, there may be further ambiguity.

The difficulties with complex households may have been linked to another awkward problem: the distinction between lodgers and subtenants. Again, this is by no means peculiar to

West Ham, but the complicated, and sometimes incomplete, corrections to the data which appear on some of the cards where lodgers are recorded, suggest that it is an issue which caused particular problems there. It seems that the West Ham interviewers may have found it difficult to apply the survey definition of a household (namely, a family and other persons living as a separate economic unit) and of a lodgers (namely, persons paying for board-and-lodging, who were not relatives) to the situations that they encountered in practice, and this may have resulted in the artificial designation of complex households that should not have been treated as single households at all.

A further notable feature about the West Ham cards is that the interviewers seem to have been more than usually unsure about how to treat absent earners. The survey guidelines state that such persons should be listed on the front of the card, together with information about the amount they contribute to the support of the family and their total earnings, but that they should be excluded from the number of persons entered on the reverse. In West Ham, such persons, together with the details of their contributions and earnings, frequently appear only in the general remarks on the reverse of the card. In some cases they are included in the number of persons, and where the absent person is the individual who would normally be regarded as head, their omission from the front of the card has occasionally meant that the wife or eldest child is entered as head.

Interpretation Rules

As with all aspects of the data for which there are problems, the intention is to handle individual cases in West Ham as far as possible in a way that correctly interprets the meaning of the data and is consistent across households. It is acknowledged that this may not always have been achieved. Most of the inconsistencies and ambiguities found in the West Ham data also occur elsewhere and have been described in other sections of this paper or in the *Codebook*. The problems covered below are those which are either peculiar to West Ham or which occur more frequently there than elsewhere.

1. Where two or three interviewers' names appear on the card, an attempt was made to identify the one who had the "final say" and to record the information as given or verified by that interviewer. The earlier information has been recorded (in the appropriate text field) only if it there is a potential error of interpretation. The name of the final interviewer only has been given in most cases.

2. Where there are complex extended families in which no head has been identified, the relationship entries usually appear as they were written on the cards, i.e. no attempt has been made to identify a head or alter the relationships. Exceptions to this occur where failure to do so would have resulted in a non-sensical household structure, e.g. two wives, or where it has been possible to identify a head with some confidence. Occasionally a second interviewer has identified a head but has not altered all the other relationships. An attempt has been made to correct these so that the other individuals relate to the "new" head.

3. Where the interviewers appear to have experienced difficulty in deciding whether a person (or persons) were lodgers or subtenants, or how to record someone who was a lodger, a number of inconsistencies may appear on the card which leave doubt about what was intended. In the usual case, the "lodger" family was originally recorded on the front and back along with the main family, but while it has subsequently been crossed off the front

of the card, details remain in the birthplaces and/or no of persons fields on the back. In other cases the lodgers appear on the front of the card with the main family but have not been included in the number of persons, NOPN_H. In a third group, the lodgers were not listed on the front at all, but are referred to in General Remarks (REMG_H) and have been included in the no of persons. In any of these types of case, there may or may not be an entry in other income for board and lodging payments.

Each case has been judged on its merits and as a general rule it was assumed, in the first type of case, that the intention was to remove the "lodgers" from the card. These have been omitted from the face of the card (i.e. from the IVE and IVN files) and from the number of persons (NOPN_H). Any entry for board and lodging payments has been coded as rent from subtenants. But in the second and third types of case, it has been assumed that the intention was to include the lodgers on the card, and so they have been included, or added, to the face of card details and to the number of persons, as necessary. Notes have been made in the REMI_H field.

4. Where an absent earner has been omitted from the face of card information and is referred to only in General Remarks (REMG_H), a record for that individual has been added to the earners section (IVE) file. Details about total earnings and contributions to household income have been handled with due regard for the possibility of double counting, as outlined in this paper in section 5 on Household Income. Where necessary the number of persons has been altered to take account of the absence. Changes have been made to the relationships of other individuals in the household only when failure to do so would have resulted in a confused household structure, e.g. two heads.

IMPLICATIONS

The problems of interpretation inherent in the West Ham data do not have special importance for users of the dataset as a whole, but they obviously do have significance for anyone who is studying West Ham alone, and may also have implications for any analysis that is carried out by borough, particularly if it is concerned with lodgers and family relationships. Any results obtained should be treated with special caution.

Section 10: Street Quality Coding

Objectives

The purpose of the *New Survey* Street Quality Coding is to provide for each household in the dataset, an indicator of the overall social and economic condition of the street in which it was living. The coding uses the results of the *New Survey* Street Survey. Following the method of investigation used by Charles Booth in his Survey of London Life and Labour of the 1880's, this part of the New Survey represented an attempt to classify the population on a street by street, rather than individual household, basis according to a series of broad socioeconomic grades. Its results were published as a series of coloured street maps; one set for the eastern half of London, and another for the western half (*New Survey*, vols IV & VII).

The Street Quality Coding scheme uses the colour codings recorded on these maps to construct a file which links the data for individual households, gathered through the House Sample Survey, to the relevant socioeconomic classifications recorded in the Street Survey.

The Coding Process

The Street Quality Coding was carried out in three stages:

Stage 1 The starting point for the coding process was a list of all the different street names occurring in each borough in the computerised dataset. In stage 1, the names were standardised and verified as far as possible, so that each unique street appeared in the list in just one, correct form. The chief source of information used was the London County Council *List of streets and places within the administrative county, 1929-35,* and accompanying street maps. For those boroughs in the *New Survey* area which were outside the administrative County of London, a combination of *Stanford's Indexed Atlas of the County of London, 1911* and modern street atlases were employed.

Stage 2 Next, the socioeconomic classifications recorded for each street by means of different colour codings were extracted from the *New Survey* maps, borough by borough, and attached to the streets in the computerised list. Two-digit codes were used to represent the various classifications. Each street was given at least one code (QUAL_Q), but if it was found that different classifications applied in different parts of the street, a separate code was entered for each block of colour or classification present. There is also a single ranking order digit (PRED_Q) attached to each code entered, to show its order of predominance relative to other codes recorded for the same street. In addition, a match code (SQMC_H)was assigned to the street, to provide information about the type of match obtained between the street name entry in the dataset and the coding given. Lastly, in stage 2, any necessary street name standardisation and verification remaining from stage 1, was completed.

Stage 3 In the final stage, the street quality codes and corresponding ranking order digits were attached to individual households in the dataset using the street name as the intermediate match field. At the same time the street names, ADDR_H, in the dataset were amended according to the form of name recorded during the standardisation and verification process. The outcome of the coding is the STQxxxA file containing one or more Street Quality Code(s) for each individual household, as identified by its identification number, IDNO_Q, plus a predominance rank for each code given; together with a match code entry in the HHGxxxA file for each household.

Coding Scheme: Street Quality Codes

There are 13 street quality codes representing 11 different socioeconomic classifications as recorded on the *New Survey* maps, plus codes for two types of "unknown" situation. Each household has at least one code. Where more than one code is given, it indicates that different classifications applied in different parts of the street in which it was living, as shown by the colouring on the maps; but, as these maps do not show house numbers, it was not possible to identify the specific code for individual households, so all the codes that could apply are given. However, in these cases, the ranking order digit (PRED_Q, see below) allows the user to identify which code was predominant in the street.

The street quality codes which appear in the field, QUAL_Q, and the classifications on the *New Survey* maps which they represent, are given below. Codes 21–61 represent composite colour codings that were used on the *New Survey* maps, in addition to the single colour codings; where they appear in a street, a fairly complex mixture of social and economic circumstances is indicated, especially if other single or composite blocks of colour also appear in the street.

Code	Colour on *New Survey* map (classification)
01	black (criminals)
02	blue (poor)
03	purple (unskilled)
04	pink (skilled)
05	red (middle class)
21	blue with black stripe
31	purple with black stripe
41	pink with black stripe
32	purple with blue stripe
45	pink with red stripe
61	purple with blue and black stripes
98	street uncoded on *New Survey* map
99	street is unknown

The assignment of composite colours (codes 21–61) is described as follows:

> Where the majority of inhabitants of a street belong to one class, but there is also in the street a substantial number who belong to the lowest or highest grades of the classification, black blue or red stripes as the case may be are imposed on the predominating colour ... (*New Survey*, vols. IV & VII, map references.)

The composite colour codes should not be confused with the predominance order code, PRED_Q, discussed below.

Coding Scheme: Predominance Order

Associated with each value of QUAL_Q is a predominance order code, PRED_Q, which provides a ranking when two or more QUAL_Q codes are assigned to a street address. Examples of PRED_Q

sequences are: 1, 2, 3; or 1, 2, 2, 3; or 1, 1. In the first and second examples, the QUAL_Q code which is ranked "1" is predominant in the street in which the household lives, while that which is ranked "3" applies to the least area in both cases; however in the first example, there is a single code which is second in predominance, while in the second, there are two codes which occur second in equal proportions; lastly in the third example, there are just two codes which are equally predominant, probably one on each side of the street.

Although just one street quality code is found in most cases, the need for these rankings arose because of the complexity of the colour coding which was applied to some streets on the *New Survey* maps.

Coding Scheme: Street Quality Match Codes

For the majority of households the street quality coding that has been recorded, be it one code or many, may be taken to represent the outcome of a reasonably definite match between a single street name entry as given in ADDR_H and a coded or uncoded street that appears on the *New Survey* map. In these cases, the letter D has been entered in the street quality match code field, SQMC_H. A letter D also appears in cases where the street in which the address is located, is unknown.

However the street quality match code can also indicate that the coding represents one of three types of special case: *Adjacent streets*; *Combination streets*; and *Name changes.*

Adjacent streets Where the street in which the household lived was found in a directory, but could not be identified on the *New Survey* map, the coding given may be that which applies to an adjacent street. The name of this street appears in ADDR_H, after the actual street, and separated from it by a comma. This action was taken only in cases where there was a reasonable likelihood that the adjacent street would have had the same socioeconomic classification as the street in question, i.e. if the surrounding area had all been given the same colour coding on the map. The adjacent street used in these cases is the one which appears in the locality column of the London County Council *List of streets and places within the administrative county, 1929–35.* These cases are identified by the letter A in SQMC_H.

Combination streets Where the household is living in a block of workmen's flats that are known, for example, as "Peabody Buildings" or "Guiness Trust", the coding given may apply to a combination of two or more streets. This occurs in cases where the address information given on the card was incomplete. Because the block either fronted onto more than one street, or a block of this name occurred in more than one location in the borough, it was not possible to determine a precise entry for ADDR_H. The action taken instead was to enter all the possible street names into ADDR_H, separated by "and" or "or" respectively. Again the London County Council *List of streets and places within the administrative county, 1929–35,* was used as the reference source for identifying the missing street information.

A second type of case where the street quality coding given for a particular household may apply to a combination of streets, occurs where there were two or more streets with the same name in a single borough, e.g. High Street, and it was not possible to determine which applied to the household in question. The codes entered in these cases will be those that apply to both (or all) of the streets of that name in the borough

Combination cases can be identified by the letter C in SQMC_H. In all cases, the combination of streets involved will have been treated as if it were one single or continuous street for the purposes of assigning codes, and the ranking order digit given in PRED_Q will indicate the relative predominance of that code over the whole combination.

Name changes Finally, there is a very small number of cases where the name given in ADDR_H for the street in which the household lived, was found to be completely different from that which appeared for the same street on the *New Survey* map, i.e the street had obviously undergone a change of name at some point between 1928 and 1932. In these cases, the original entry on the card as given in ADDR_H was retained and the alternative name recorded in REMI_M. Cases in which there has been a change of name may be identified by a letter N in SQMC_H. Most of them can be verified in the London County Council *List of streets and places within the administrative county, 1929–35.*

Coding Rules

In addition to the special categories outlined above, there are three other types of problematical cases. None of these has been explicitly flagged in the data files. The way in which they have been resolved is noted below:

1. Where the boundary between two or more boroughs ran down the middle of the street in which the household is living, or crossed it, only those classifications that apply to that side or part of the street located in the relevant borough will appear in QUAL_Q for the household in question.

2. During the coding process, it emerged that some of the households surveyed in particular boroughs were living in streets that actually appeared from the maps to have been part of a neighbouring borough or a different section of the same borough, where it had been divided for the purposes of the survey. This seems to have occurred mostly in two places, Camberwell and Lambeth, although individual cases occur throughout the dataset. Firstly, about 60 streets from which addresses were selected for the Camberwell sample appear from contemporary maps to have lain within St. Pauls, Deptford. It is thought that these discrepancies may have arisen as a result of boundary changes around that time, but this has not been verified. Secondly, it is thought that when the cards for Lambeth were sorted into the separate sections for the North and South of the borough in which they now appear, some of the addresses were assigned to the wrong section. There are a number of households living in streets classified as being part of one section of the borough which, from the map, are clearly in the other part.

   No attempt has been made to reclassify addresses as a result of either type of discrepancy; the boroughs assigned in the *New Survey* have been accepted at face value. The street quality coding recorded in these cases is as given on the maps, regardless of the borough to which the street was attributed.

3. In some cases where the street in which the household lived could be identified in other sources but did not appear to be on the *New Survey* map, it was possible to identify a block of colour coding on the latter at the approximate place where the street should have appeared. Where it was felt that a reasonably definite match had been obtained with the street in the computerised dataset, this coding was applied to that street. It is thought that the scale of the *New Survey* maps was sometimes too small to allow every little side

street, "place" or "mews" to be individually marked; in some of these cases it would seem that the street surveyors nevertheless attempted to record the appropriate classifications.

Implications

Every household in the *New Survey* dataset has at least one street quality code associated with it. The code provides an indicator of the overall social and economic condition of the street in which the household was living, apart from those cases (codes 98 & 99) where it serves only to indicate that the street was not classified, or that it is unknown. Where more than one street quality code is given for a particular household, it implies that different conditions existed in different parts of the street in which it was living; the nature of the source precluded determination of the precise code to apply in these cases, so all the possible codes have been given. But for all households where there are multiple codes, the entries in the predominance order field allows the codes to be ranked in order of their prevalence in the street, so it is almost always possible to identify a single code which is most likely to apply on balance of probability.

In the majority of cases we are reasonably satisfied that the coding given is an accurate representation of the condition of the household in question as far as it could be determined from the information available to us, and bearing in mind the special and problematical cases outlined above; however the following caveat should be made: that owing to the quality of the printing and the smallness of the scale used, we could not always be sure that we had distinguished accurately between the different colour codings recorded on the maps, especially where these had been applied in very small blocks.

Section 11: Occupational Coding

Objectives

The *New Survey* record cards contain information about the occupations of some 42,500 individuals, a large proportion of the 49,445 individuals recorded in the IVE file. This information has been reproduced, as faithfully as possible from the interview questionnaire cards, in the OCCU_E field. It is highly variable, textual information. No control was imposed over the terminology used to describe the occupations; the interviewers simply recorded the details verbatim, as given to them by the individuals concerned or other members of their households. As might be expected, the information that appears is inconsistent and ambiguous; the same occupation may be described in several different ways (e.g. "Sailor", "Seaman", "Able-seaman", "At sea", "Navy", "Mercantile marine") and a particular term may not necessarily mean the same thing in every case. But occupation is widely regarded as the most useful indicator of economic or social status. In order to make this data usable in any form of quantitative analysis, the textual entry recorded for the occupation of each individual earner needs to be translated into a format (i.e. code) which standardises the way the activity being engaged in is represented, and provides a means of classification.

One of the criteria by which a coding scheme may be judged is its appropriateness for the data being coded. Appropriateness may be defined in terms both of time and of scope. The only general occupational classification available to the project that is contemporary with the *New Survey* is the scheme used in the *Census of England and Wales,* 1931. This is the scheme on which the *New Survey* Occcupational Coding is based. It is important to note, however, that the scheme has not been used here exactly as in the Census; some modifications have been made and these are explained below. The *Introduction to Census, 1931: classification of occupations,* HMSO, 1934, states that the classification is based on "material worked in and process carried out, and for the non-productive occupations, type of service individually rendered, while degree of skill is also taken into account." Although, it is described as an occupational classification, the scheme is more accurately defined as a combination of occupational status and industrial classification.

The initial purpose of the *New Survey* Occupational Coding, was to convert each individual's occupation into an appropriate numeric code at a level detailed enough to uniquely represent separate activities. But the ultimate objective was to be able to analyse earners, and the households to which they belonged, according to certain characteristics; for this it was necessary to group the coded occupational data into two broader analytical categories: socioeconomic (skill) class; and main industrial/occupational order.

The Coding Process

The occupational coding was carried out in three stages:

Stage 1 An attempt was made to match the occupation recorded for each individual in the earners (IVE) file to a code in the Census classification, using one of the following: the occupation as given; different combinations of the same words; or, synonyms. A limited amount of additional information was also used at this stage to make coding decisions; this included: other survey details about the individual in question (in particular the employer, EMPL_E); the entries for other family members in the IVE file; and, a small number of readily available reference sources, most importantly, *Kelly's Post Office London directory, 1968.* Each coded occupation was also given a single character match code (OCCM_E) to

indicate the quality of the match obtained between the actual occupation as given in the OCCU_E field and the occupation code, OCCC_E. Note that codes were assigned to individuals rather than to occupation titles. The same codes were almost invariably given to to the same occupations but this procedure allowed for the possibility that other information (e.g. about employer) could require a title to be interpreted differently for coding purposes in particular cases.

At this stage, efforts were concentrated on those cases where either a reasonably definite or an approximate match could be made; cases where this could not be achieved were flagged and left for later research. To ensure that the judgements applied across the whole dataset were as consistent as possible, a series of decision rules were followed.

Individuals in the IVE file for whom no occupation was given were identified at this stage and coded in one of three ways: those who were retired were coded as Y (preceded by three spaces), ␣␣␣Y (981 cases; the code is 9998 at level E); others who were not participating in the labour market (e.g. housewives) were coded as X (preceded by three spaces), ␣␣␣X (5,412 cases; code is 9997 at level E); and those for whom occupational information was missing were coded *␣␣␣ (547 cases, including a small number for whom the occupation proved to be unknown at stage 2; the code is 9999 at level E).

Stage 2 All those cases for which a particularly doubtful match, or no match at all had been obtained at stage 1, were reviewed and an attempt made to resolve them, either as definite or as approximate matches against classifications appearing in the coding scheme. Considerable use was made of additional sources of information; in particular, research was carried out into the employers, if given in these cases, to try to obtain clues as to the likely nature of the trade being engaged in. This involved examining the records of other individuals in the survey who were working for the same employer, as well contemporary directories listing businesses in the London area. Another approach involved reviewing groups of occupations that had been particularly problematic, most importantly, certain categories of the unskilled, engineering occupations, and those working for relatives. As a result, it proved possible to assign a meaningful code to almost all individuals for whom occupational information was given. The occupations for only 473 cases remain as doubtful, or tentative, matches.

Stage 3 At the final stage, each unique occupation code was assigned two additional codes; a single character code to associate it with one of five skill levels, and a two character code to indicate its membership of one of 31 main orders of occupations. Lastly, the these codes together with the occupation and match codes were added to the rest of the data for each individual in the IVE file.

Coding Scheme: Occupation Codes

The occupational coding scheme is based on a block code composed essentially of three-digit, non-significant, sequential numbers. At level A in the *New Survey* dataset, the individual codes are made up of three numeric digits and additionally, in some cases, a letter, E, F, R or U. Formally, OCCC_E defines a *four* character field, the fourth character being a space in most cases (At level E, a fourth numeric digit replaces the fourth character at level A; see below.) The letters E and F occur in the Census scheme, although here they are applied in a slightly different way; R and U do not appear in the Census. Each numeric code uniquely identifies a very closely related group of occupations or activities, as defined by a single occupational title. These are

as given in *Census, 1931: classification of occupations,* HMSO, 1934. The specific occupations that are included within each group may be found in the Classified list and Alphabetical index. The occupational groups are also assigned in the Census to one of 32 main industrial orders, and in some cases to a sub-order within the main order. Associated with each occupation code is one of five socioeconomic classes or skill levels. The numeric codes and the occupational groups that they represent have been applied essentially according to their definitions in the Census scheme, but with certain adjustments which are outlined in the remainder of this section.

To determine the appropriate code for each individual's occupation, the information recorded in the OCCU_E field of the dataset was matched against the list of occupations given in the Classified list or accompanying Alphabetical index, and the best possible match obtained.

The extra letters have been appended to the numeric codes, where required, to indicate certain additional categories of job within the occupational group. Hence, particular numeric codes may appear both with or without the extra letter, meaning something different in each case. The most significant departures from the original census classification scheme occur with respect to the use of these letters:

- *Employers and managers*: Where E appears as the fourth character ("1" at level E), the code denotes that the individual is an employer, proprietor or manager within the occupational grouping concerned. This is defined as someone who employs others, runs premises that are separate from the home, or manages other people. There may be some cases where a numeric code used without the E suffix also appears from the occupational title, to denote a grouping of employer or managerial occupations, e.g. $613_{\sqcup}$ (with no E) represents "haulage and cartage contractors and managers". In these cases the code denotes the occupational grouping as given, but indicates that the individual occupation thus coded did not fall within the "Employer" category as defined above. Continuing the above example, $613_{\sqcup}$ denotes a small contractor who was likely to be working alone from home, while 613E indicates one with a substantial business, probably employing others. Equally, there may be cases where a numeric code used with the E suffix does not appear at all from the occupational title to denote a grouping of employers, *etc.* Here again, the code represents the occupational grouping as given, but has been used with the suffix to indicate that the individual concerned additionally has employer status. For example, $261_{\sqcup}$ would mean an electrical engineer or fitter, or electrician, while 261E would indicate an employer, manager or proprietor in that field.

  In the Census scheme it would appear that the E suffix denoted a certain status or position within a trade or business. It has been used to indicate not only employer, proprietorial or managerial occupations, but also occupations of professional or independent status, i.e. those operating on their "own account". Its use is confined to certain prescribed classifications and it always appears when these codes are assigned. In most of these cases the occupational title does indicate a grouping that includes employer, proprietorial or managerial occupations (e.g. 480E identifies employers and managers in the building trades) so in practice the use of the E in the census and in the *New Survey* will often coincide. But this will not be the case where the E suffix has been used to indicate a professional status (e.g. 810E identifies a civil engineer in the census scheme); or someone likely to be working on his own account (e.g. 721E identifies costermongers and hawkers in the census scheme).

- *Foremen*: Where F appears as the fourth character ("2" at level E), the code denotes that

the individual is a foreman, overseer, charge hand, gaffer, superintendant or supervisor of other people, but not an assistant foreman or under-foreman. Again there may be cases where a numeric code used without the F suffix, also appears from the occupational title, to denote a grouping of foremen and overlookers, e.g. $251_{\sqcup}$ (with no F) denotes "foremen and overlookers – electric". In these cases, the classification is likely to have been selected for the individual in question, because the occupational grouping thus represented included activities which did not fall into the "foreman" category defined above. Where a numeric code used with the F suffix does not appear from the occupational title to denote a grouping of foremen, this again represents the occupational grouping as given, but indicates that the individual thus coded additionally has foreman status. For example, $215_{\sqcup}$ would mean a plumber, while 215F represents a foreman plumber.

It would appear that the F suffix also had a wider application in the census scheme, than its apparent meaning would imply. Again its use is confined to certain prescribed classifications and it always appears when these codes are assigned. In most of these cases the occupational title does indicate a grouping that includes foremen and overlookers, (e.g. 481F denotes foremen and overlookers in the building trades) so in practice the use of the F in the census and in the *New Survey* will often coincide. However this will not be the case where an F classification in the census represents an occupation or occupational grouping which does not fall within our definition of "foreman" (e.g. 594F denotes locomotive engine cleaners in the census scheme) but would appear as $594_{\sqcup}$ in the *New Survey.*

- *Assisting relatives*: Where R appears as the fourth character ("3" at level E), the code denotes that the occupation of the individual concerned involves assisting a relative; this is defined as someone who is employed by a member of his/her family. A particular numeric code may have been used with or without the R suffix; the occupation is the same in both cases, but in the first, the individual is working for a relative. Where the entry for occupation in the *New Survey* is of the form "assisting father", the appropriate classification will have been determined with reference to both the father's occupation and the likely status of the individual in question; hence if the father is a shopkeeper, the classification chosen will be that of a shop assistant. In cases where a female is assisting a male relative in a predominantly male occupation (e.g. building trades) an appropriate code denoting employers and managers will have been used, with the R suffix.

  It should be noted that this differs from the method by which relatives assisting in business are handled in the census; here the R suffix is not used and the individuals concerned are either coded to the appropriate occupational grouping, or in cases where a female is assisting in a predominantly male occupation, are classified within the appropriate employers and managers grouping, if aged 20 years and over, or as $889_{\sqcup}$ = other clerks, if otherwise. The census coding scheme does not therefore, allow individuals working for relatives to be distinguished from others engaged in the same occupations.

- *Unskilled occupations*: Where U appears as the fourth character ("4" at level E), the code denotes that the individual is engaged in an unskilled occupation. This suffix was used where a classification representing an appropriate "other" or "other skilled" occupational grouping was assigned to denote an unskilled occupation, in place of one of the general unskilled $920_{\sqcup}$, $930_{\sqcup}$, or $940_{\sqcup}$ codes which the census scheme provides. It is important to note that an unskilled occupation may be represented by one of three types of code: (i) a numeric code denoting an appropriate unskilled occupational grouping as defined

by the occupational title (e.g. $030_{\sqcup}$ represents "gardeners' labourers"); (ii) a numeric code plus the U suffix, in which the numeric on its own denotes a grouping of "other" or "other skilled" occupations in a particular main order or sub-order, but with the "U" represents an alternative grouping of unskilled occupations in that order (e.g. $528_{\sqcup}$ means "other skilled workers (rubber)", while 528U indicates unskilled occupations in the rubber industry, such as a labourer in a tyre company); (iii) a code representing an unskilled occupational grouping in the "other and undefined workers" main order, i.e. $920_{\sqcup}$ = "general labourers"; $930_{\sqcup}$ = "undefined labourers"; and $940_{\sqcup}$ = "other unskilled workers (class of work specified)".

The U suffix does not appear in the Census scheme. Where a specific numeric code existed in that scheme for an unskilled occupation, this was used in the *New Survey*; but in the majority of cases, the Census Classified list provides only the three general unskilled codes as a means of representing these occupations. The *New Survey* occupational coding scheme attempts to avoid using these general codes; wherever possible, an appropriate industry-specific main order or sub-order has been identified, and the classification for "other" or "other skilled" workers in that order used with the U suffix. However it was not possible to identify an appropriate alternative in every case, so a significant number of unskilled individuals (1,903) remain coded as $920_{\sqcup}$, $930_{\sqcup}$ or $940_{\sqcup}$, e.g. scavengers and other types of unskilled local authority workers.

Coding Scheme: Occupation Match Codes

The quality of the match obtained between each occupation recorded in the dataset and the occupation code selected, is indicated by one of four categories:

1. If the match is a definite one, it is identified by the value D in the OCCM_E field. This implies that the words used to describe the occupation are the same or similar in both the dataset and the classification scheme, or have the same meaning. Also cases where the occupation code is $\sqcup\sqcup\sqcup$X, (not participating in the labour market), and $\sqcup\sqcup\sqcup$Y (retired) have been treated as definite matches.

2. The match is approximate, identified by the value A, where the code used represents a fairly similar or related occupation, of equivalent status and main order to the one given, but is not quite identical. In these cases the code selected is the best possible choice; either there are no reasonable alternatives, or those that exist are in the same main order (or suborder) and/or require the same level of skill (e.g. a "sectional pipe maker" could be coded as 095 = earthenware pipes or 093 = zinc pipes). The actual choice in such a case is significant only for work at the most detailed occupational level; for most broader analytical purposes it is immaterial. A match will also be rated approximate where the coding represents a less specific occupational definition than that given in the dataset, but is otherwise equivalent (e.g. an "alaska furrier" would be coded as 282 = furrier (not employer)).

3. A tentative match, identified by the value T, indicates considerable doubt that the code chosen accurately represents the occupation of the individual concerned. All of these cases were reviewed at stage 2 of the coding, but remain doubtful because the information available proved insufficient to allow a classification to be chosen with any degree of confidence. The code provided, therefore, should be regarded only as a best guess. Tentative matches indicate in particular that there was a choice between two or more

classifications that were not closely related, because they were of different status and skill level and/or in different main orders; they are cases that collectively, could have significance at broader levels of analysis. An example is "seaman" or "sailor"; in the absence of any employer information, these occupations could mean $761_{\sqcup}$ = Royal Navy or $635_{\sqcup}$ = merchant seaman.

4. Where no match could be made at all, the value N appears in the match code field. These are cases for which the occupation code is $*_{\sqcup\sqcup\sqcup}$, defined as those for whom an occupation was not given or could not be matched to any code in the classification scheme.

Coding Scheme: Skill Level Codes

One of five socioeconomic classes or skill levels is associated with each unique occupation code:

Skill level 1 indicates higher professional, managerial and proprietorial occupations; it applies to very few individuals in the *New Survey* dataset (49), the largest group being engineers.

Skill level 2 denotes other professionals, employers and managers. Most of the 308 individuals to whom it applies are engaged in teaching or subordinate medical services, or fall within our definition of employers and managers. Wherever the suffix E appeared, the occupational grouping thus denoted was assigned this skill level.

Skill level 3 denotes skilled workers, including most foremen, and clerical workers; it also includes small proprietors such as shopkeepers, and almost all cases where the suffix F appeared. It applies to 20,562 individuals.

Skill level 4 denotes semi-skilled workers and includes domestic servants and others engaged in personal services, shop assistants and salesmen. It applies to 10,068 individuals.

Skill level 5 identifies 11,515 individuals who are engaged in unskilled occupations.

Where a numeric code appears with the R suffix, it has been interpreted as implying a lower level of skill than that which applies to the code without the R. A lower level of skill code is appropriate on the ground that working for a relative was in a sense more "sheltered" than working for a wage in the open market,

The assignment of skill levels to individual occupation codes in the *New Survey* largely corresponds to the classifications enumerated in the *Registrar-General's decennial report for 1931,* Table 1. But there are differences in the treatment of some occupational groupings: shopkeepers, shop assistants and salesmen, and domestic servants respectively were demoted by one class; the addition or subtraction of the E and F suffixes in appropriate groupings, as compared with the census scheme, served to raise or lower the skill level in some cases, although for the most part the appropriate underlying skill level remains unaffected; and, where an R suffix appears, the skill level assigned was demoted by one class compared with the numeric code used without the suffix.

Coding Scheme: Order Codes

The unique occupation codes and their associated groups of occupations were arranged in the Census classification scheme into 32 main orders, largely on the basis of the type of industry or the material used in production. These groupings are reproduced in the *New Survey* scheme to provide a broader analytical classification. The codes are listed in the *Codebook*, Table 27,

p. 41. With some exceptions they have been applied as in the Census scheme. One exception is that Order 32, "Retired or not gainfully occupied", has not been used: the codes assigned to this order appear in the *New Survey* classification under code 99. Of greater significance is the difference in the way many of the unskilled occupations have been handled. In the Census classification, the majority of these have been coded as 920, 930, or 940, hence they would appear under order 31, "Other and undefined workers". In the *New Survey* scheme, the unskilled occupations are assigned wherever possible to a suitable code within the most appropriate industry-specific main order. For comparisons with the Census, this decision has implications for the distribution of coded individuals across tha main orders but does not affect the result of any analysis by skill level. (Under the *New Survey* scheme 2,892 individuals fall within Order 31; if the scheme is applied as used in the Census, this number becomes 5,904; i.e. 3,012 unskilled individuals have been shifted to another main order) .

Coding Rules for Occupations

Inevitably a significant number of awkward cases were encountered during the process of coding the *New Survey* occupations, such that the occupational data recorded by the interviewer could not be readily matched to any occupational title in the classification scheme. In order to resolve these difficulties, a set of decision rules was adopted; these are outlined below. In many cases, additional information about the individual in question, other members of the household, or the person's employer, could be used to inform the decision making process, often allowing an approximate match to be made. However a relatively small number of cases remained for which no further reliable information could be brought to bear and a choice of code had to be made on the basis of criteria which may be regarded as questionable. These cases are identified by the assignment of the tentative match code, `T`, in `OCCM_E`. The coding rules are as follows:

1. Where the occupation given for an individual could be matched to several possible alternatives in the classification scheme that appeared to be similar or related, but not identical, the code representing the occupation which seemed closest in status was chosen; similarly, where no occupation at the level of specificity given for a particular individual was available in the scheme, the most appropriate broader classification was chosen with regard to likely status.

2. In some cases an occupation was recorded in the *New Survey* in terms that were too general to allow an appropriate classification to be selected from among several more specific alternatives, with any degree of confidence. If no suitable general class existed, the code which represented the "most likely" grouping was used, where "most likely" should be understood as the most obvious choice in the light of any other evidence available, or the most common occupation. Sometimes the choice lay between a small number of clear alternatives. An example referred to above is "seaman"; in this case, it was decided that, in the absence of further information, merchant seaman rather than Royal Navy was the most obvious choice. But in other cases the classification scheme provided alternative codes in relation to a number of specific trades or industries. Where the occupation was given as "shop assistant", the classification, 716 = "shop assistant (other retail)" was used as a last resort, but in some cases a choice had to be made between specific classifications on the basis of evidence which may be regarded as flimsy; e.g. in the absence of further information, female "machinists" were coded as 360 = "sewing machinists (not elsewhere enumerated)" on the basis of their sex.

3. Where either a specific or general occupation given in the *New Survey* could not be

matched to a code in the classification scheme, but it was possible to determine an appropriate industry-specific main or sub-order, the "other" or "other skilled" workers classification in that order was used, with or without the U suffix, as appropriate.

4. Where the occupation description is given as one of: "assistant", "hand", or "operative", it was assumed that an unskilled job was implied. If an appropriate industry-specific main or sub-order could not be identified, the general codes 930 or 940 were assigned. However, where the word "assistant" occurred in conjunction with other details , a distinction was made between cases where it appeared first (e.g. assistant printer) and those where it came second, (e.g. printer's assistant); the latter would be regarded as unskilled; the former would not.

IMPLICATIONS

An attempt has been made to provide a meaningful occupational coding for as many of the individuals in the earners file as possible for whom information about occupation is given. As a result a code has been assigned to 42,505 individuals out of a total of 49,445 persons recorded as earners. In the majority of cases a match has been obtained between the occupation as described by the interviewer and a classification which is reasonably definite or approximate. In just 473 cases the choice of code remains doubtful and, indeed, for these it could be argued that a positive match should not have been attempted at all.

Inevitably, in any large general occupational coding exercise such as this, many individual decisions have to be made of the basis of a personal judgment by the coder. It is acknowledged that different outcomes are possible for some of the cases, though the user can be confident that the most sensible decision has been made in the overwhelming majority of cases according to the rules set out above and that judgment has been applied reasonably consistently throughout.

Section 12: Birthplace Coding

Objectives

The purpose of the *New Survey* Birthplace Coding is to provide a code to represent the birthplace of each adult individual recorded in either the Earners (IVE) or Non-Earners (IVN) file. The scheme used is a structured one in which the codes serve both to standardise the way individual birthplaces are represented and to show how they are related or grouped geographically; hence the codes are made up of interdependent components that have significance both as a whole and in part. With some exceptions, each birthplace is uniquely identifiable by its code so that with additional processing, it may also be grouped in other meaningful ways depending on the requirements of the user. For example, in our scheme, the individual codes have been grouped, where applicable, into conurbations; another possible approach would have been to group the codes by actual population size.

The coding is restricted to adults, defined here as individuals aged 18 years or over. The basis for this decision lies in the characteristics of the original *New Survey* birthplace information, in that the remit of the interviewers was to collect details about the birthplaces of adults only. Although in come cases, the birthplaces of children were given, it was felt that the coverage of younger age groups was not sufficiently complete to make coding of these groups worthwhile. The adoption of 18 years and over as the definition of an adult for the birthplace coding is a decision made in the light of the problems associated with the *New Survey* birthplace data.

The net result is that of the 94,136 individuals in the dataset, birthplace coding is available at some level for 58,054. Birthplace codes were not assigned where (a) individuals are excluded from the exercise because they were not adults; (b) no information is available about the place of birth. The former group comprises 31,778 individuals which have been coded 9999997. The remaining 4,304 have been coded with 9999998. The code value 9999999 was reserved for cases in which a birthplace was given but no genuine code could be determined: this value has not been needed.

The Coding Process

The Birthplace coding was carried out in five stages:

Stage 1 At the preliminary stage, as many individuals as possible were automatically coded. This was achieved by identifying all those aged under 18 years and assigning 9999997; by identifying those for whom no birthplace information was available and coding as 9999998; and by assigning a precise place code to all those for whom an exact match could be obtained between the birthplace as given in the survey data and a formal place name title in the look-up file of birthplace codes (BPL). This left 16,269 individuals to be coded manually.

Stage 2 In stage 2, an attempt was made to match each of the uncoded birthplaces, to one of the place codes in the BPL file. At this stage, efforts were concentrated on those cases where a reasonably definite match could be made. Where necessary, additional information was used to make coding decisions. This may have been other survey details about the individuals in question, or the households to which they belonged. Alternatively reference was made to one of the following sources: the London County Council *List of streets and places within the administrative county, 1929–35* and accompanying maps; *Stanford's Indexed Atlas of the County of London, 1911*; *The survey atlas of England*

& Wales, 1939 (Bartholomew, J. G.); *Ordinance survey atlas of Scotland,* 1924; *Census of England and Wales, 1921* (Index of names of places. 1924); *Census of England and Wales 1951,* (Index of place names. 1955); world atlas.

Each birthplace was given a seven-digit place code (BPCC_E or BPCC_N)and in addition, a single character match code (BPCM_E or BPCM_N) to indicate the quality of the match obtained between the actual place and the code assigned. A query (not reported in the data files) was signalled in all cases where a definite match could not be made. Note that the codes were attached to *individual persons* so that it was possible to treat occurrences of the same place name differently if necessary. However, in most cases everyone with the same birthplace was assigned the same place code and match code.

Also at this stage, some standardisation and verification of the place names appearing in the birthplace field was carried out.

Stage 3 In the third stage, all those cases which had been flagged as queries at Stage 2 were reviewed and an attempt made to resolve them, either as a definite match against an entry in the file of place codes, or as an uncertain match against such an entry. In order to arrive at a consistent coding across the dataset, a series of decision rules were applied in these cases. Using these. it proved possible to assign a meaningful code at some level to every individual aged 18 years or over, for whom birthplace information exists.

Stage 4 Next, each entry in the file of birthplace codes (BPL) was assigned an additional single character code (CONU_B) to indicate its association, where appropriate, with one of eight conurbations. These were identified using the reference sources noted above.

Stage 5 In stage 5 the birthplace coding information was added to the rest of the data for each individual in either the Earners (IVE) or Non-Earners (IVN) files. In addition to the birthplace codes and the match codes, the conurbation codes were also added (as BPCN_E and BPCN_N, respectively). Lastly, the birthplace names were amended according to the form of name recorded during the standardisation and verification process.

Coding Scheme: Birthplace Codes

The coding scheme provides a numeric code to uniquely identify each country, county or specific place. The scheme is hierarchical. It is structured by country and, where applicable (see below), by county, and is based on the international boundaries and national administrative units that existed in 1931, or as close thereto as possible. Each code is described by a formal place name title. In the case of United Kingdom and Eire counties and places, the form of name used is that which was in official usage in 1931. It may be argued that 1931 is not an appropriate point in time on which to base the coding scheme, since the individuals whose birthplaces are being coded will have been born at least 18 years earlier, some of them under very different regimes to the ones in existence at the time of the *New Survey.* However, it is argued that this decision is appropriate for reasons of consistency; the *New Survey* information itself dates from that period, as do the sources upon which the street quality and occupational coding schemes rest. Also, with an age range of 18 years to 97 among the individuals being coded, it would have been difficult to devise a sensible alternative.

Each place code is made up of three parts. The first two characters represent the *country*; the third and fourth are the *county*; and the remaining three characters point to a *specific place name.* All seven characters are needed to uniquely identify each place name or geographical

area; the use of zeros for characters 5, 6 & 7 and for 3 & 4 allows an area, either county or country, as opposed to a specific place, to be represented; while the code 900 for characters 5, 6 & 7 indicates a "rural area" within the United Kingdom and Eire (see below). Examples are: 0100000 = England; 0100900 = English rural areas ("the country"); 0101000 = Bedfordshire; 0101002 = Bedford; 0101900 = Bedfordshire rural areas; 0125000 = County of London; 0125004 = Bethnal Green; 0126000 = Middlesex; 0126001 = Acton; 0600000 = Eire; 0612000 = Dublin County; 0612001 = Dublin City; 0000000 = Great Britain; 1800000 = France. Additionally there are a few codes with special meaning: 9000000 indicates "abroad" (no country specified); 9100000 indicates Jewish origin (no country specified); 9999998 indicates that there is no information about this particular individual's birthplace; and 9999997 indicates that there may be information but no code is being assigned because the individual is aged under 18 years.

Individual counties and place names are coded only for the countries of the United Kingdom, (i.e. England, Wales, Scotland and Northern Ireland) and Eire. However, rural districts in these countries (landward areas in Scotland) are not individually coded, but are represented by a single general code (900) within each county and country. Place names and area names outside these countries are covered simply by the appropriate country code. For birthplaces within the United Kingdom and Eire, the coding was always applied to individuals at the most detailed level possible in the light of the information available, i.e. a specific place code or general rural district code, as appropriate, was always preferred to a county or country level code, even though the match may be uncertain. A higher level code than that implied by the birthplace information was used only in the last resort. It should be pointed out, however, that in quite a significant number of cases, coding below county or county level was precluded by the lack of detail recorded by the interviewer.

Within the United Kingdom and Eire, the structure of the codes and the place name titles attached to each one are based on the administrative units in use in 1931 as identified from contemporary official sources:

1. For England and Wales, the source used was *The Registrar-General's statistical review of England and Wales for the year 1931: Tables. Part II. Civil.* HMSO, 1933: Table E. Within each county, a unique code is provided for each county borough, metropolitan borough and urban district. Where a county such as Yorkshire was administratively divided, there is also a general code for the county name.

2. For Scotland the source used was the *Seventy-seventh annual report of the Registrar-General for Scotland 1931,* HMSO, 1932: Tables 4, 5 & 6. Within each county, a unique code is provided for each large burgh and small burgh.

3. For Northern Ireland the source used was *Northern Ireland census of population 1926,* HMSO, 1929: Table 3; and *Northern Ireland census of population 1937,* HMSO, 1940: Tables 3 & 4. Within each county, a unique code is provided for each county borough, municipal borough, and urban district, and additionally for a small number of towns which do not have administrative status.

4. Lastly, for Eire, the source used was *Ireland census of population 1926: volume 1,* Stationery Office, 1928: Table 9; and *Ireland census of population 1936: volume 1,* Stationery Office, 1938: Table 9. Within each county, a unique code is provided for each county borough, municipal borough, and urban district. Additionally, there are codes for the four provinces into which Eire is divided, for a small number of towns which do not have ad-

ministrative status, and for Ireland as a whole. County codes have been designated in a way that allows them to be identified as part of the appropriate province.

Coding Scheme: Birthplace Match Codes

The quality of the match obtained between the birthplace recorded in the dataset for each individual and the place code selected, is indicated by one of three categories: a definite match, identified by the letter D in BPCM_E and BPCM_N; an uncertain match, identified by the letter U; and no match (cases coded 9999997 and 9999998), identified by the letter N. In the overwhelming majority of cases where a meaningful code has been entered, a reasonably definite match has been achieved; only 559 remain uncertain. The type of case which may fall within this category are discussed below, under coding rules.

Coding Scheme: Conurbations

This is an extra level of coding, added to the information about birthplaces for the purposes of this project. It allows individuals to be mapped, through the specific birthplace code, to one of eight conurbation codes. Only those individuals whose birthplace code represents a place that is geographically part of a conurbation, will have a meaningful conurbation code; the remainder are coded 0 (zero). The conurbations that are identified, together with their codes, are given below.

Code	Conurbation
1	Greater London
2	West Midlands (Birmingham, Wolverhampton, *etc.*)
3	Merseyside (Liverpool, Bootle, Birkenhead, *etc.*)
4	Manchester (including Salford, Stockport, Oldham, *etc.*)
5	Sheffield (including Rotherham)
6	Leeds and Bradford, *etc.*
7	Tyneside (Newcastle, Gateshead, North and South Shields, *etc.*)
8	Clydeside (Glasgow, Paisley, *etc.*)

A conurbation is defined as a single very large urban area, with a total population of 600,000 or over; it is usually identified by one particularly large county borough, and includes a number of county boroughs, metropolitan boroughs and urban districts which are contiguous, i.e. are without rural districts in between. The conurbation code allows these contiguous boroughs and urban districts to be linked. (In Scotland, the appropriate areas are large and small burghs.) The definition of the outer boundary of the conurbation may appear problematic, but contemporary maps showing the spread of built-up areas allowed a reasonable judgement to be made in each case. However, it is important to note that the degree of accuracy which could be applied to this process was necessarily limited by the underlying structure of the birthplace coding on which it was based. Because this rests on administrative units at the level of borough or district, compromises as to boundary were necessary in some cases, with the result that a few places are included within conurbations, that arguably should not be so. An example is North Shields and Tynemouth: the former was judged to be part of the Tyneside conurbation, but not the latter. However North Shields falls within the county borough of Tynemouth, so it was not possible to make a distinction; the whole county borough is included. The burgh of Motherwell

and Wishaw is a second example; this has been assigned to the Clydeside conurbation but, arguably, Wishaw only should have been included.

Coding Rules for Birthplaces

During the coding process, several types of problem cases were identified for which a reasonably definite coding could not be obtained, even after additional information had been sought from the survey data and other reference sources. These are cases where the original birthplace data for a particular individual was found to be ambiguous, incomplete or incorrect. The way in which they were resolved is noted below. With the exception of the first type of case, which is regarded as a definite match in most instances, they are made apparent to the user by the assignment of a letter U match code. The coding rules are as follows:

1. Where the birthplace was given as a region or area which could not be precisely coded, the most appropriate higher level code was chosen. Examples are: "East London" was coded as London (0125000); "The North" as England (0100000); "S. Wales" as Wales (0200000).

2. Where the birthplace of the individual was known to exist, but a place of that name occurred in more than one county, and county information was not given in the *New Survey* entry, one of three courses of action was chosen as appropriate: (a) reference was made to other information in the survey, e.g. the current place of residence of the person or the birthplace of another family member; (b) the place closest to London was chosen; (c) it was assumed to be the place which was largest or most well known. Examples of each type are: (a) a person born in "Ashford" was coded to Ashford, Middlesex if currently living in West London (e.g. Acton), but to Ashford, Kent if living in South-East London (e.g. Greenwich); (b) a birthplace given as "St Ives" was assumed to be St Ives, Huntingdonshire, rather than St Ives, Cornwall, as the former is nearer London; (c) an entry such as "Newcastle" would be interpreted as Newcastle-on-Tyne, not as Newcastle-under-Lyme, the former being the largest and best known place.

3. In some cases the birthplace information recorded for an individual was found to be incorrect, in that a place of the name given may have existed, but not in the county that the interviewer had entered with it. In these cases, it was assumed that the county information in the dataset was in error and the place was coded to the correct county. An example is "Towcester, Bucks": here the code applied was 0128900, reflecting the fact that Towcester is a rural district in Northants.

4. There were a number of cases where the birthplace was known, or could be identified from a map or the census index of place names, but was attributable to two or more contiguous administrative districts which might be of different status, or in different counties. Where the choice was between two administrative districts of different status, the one with the higher status was always chosen, i.e. a borough or urban district coding would have been used in preference to the rural district alternative. Examples are: "Aintree", which appeared to fall within the county borough of Liverpool and a rural district of Lancashire; and "Hendon", which was both an urban and a rural district of Middlesex. A more complex example is "Newmarket, Cambridgeshire"; Newmarket was an urban district in Suffolk, and also an adjacent rural district which fell within Cambridgeshire. Despite the ambiguity in this case, the urban district coding would have been selected. Where the choice was between two or more boroughs which fell within different but contiguous

counties, the one which was part of the County of London was selected, if applicable, or otherwise the one which was part of a conurbation. Failing this, the first borough alphabetically by name was chosen, on the grounds that it was unlikely to have a serious effect on the outcome of any analysis. Examples are: "Highgate", which was partly in the borough of St Pancras, and partly in Hornsey, Middlesex; and "Norwood", which was spread across Lambeth and Wandsworth, and the county borough of Croydon, Surrey. In the latter example, the code for Lambeth was selected, except in cases where a particular part of Norwood was specified.

5. Lastly, in a small number of cases where the birthplace could not be identified at the most detailed level of coding, or remained a completely uncertain choice between alternatives, the most appropriate code at a higher level, e.g. county or country, was applied. An example is "Colesham": in the absence of any information as to the location of this place, it has been coded as England.

Implications

Every individual earner or non-earner aged 18 years or over, for whom there is birthplace information in the *New Survey* dataset (58,054), has been assigned a meaningful code which represents that information at some level. In 37,965 cases, where the person was born in the United Kingdom or Eire, it has been possible to provide a specific place code which matches the birthplace given with a reasonable degree of certainty. In a further 17,466 cases, we have been able to give a code that indicates the county of origin with reasonable certainty; 772 of these individuals are known to be from rural districts which are not specifically coded. This leaves 2,621 individuals for whom country level coding only is available, some of them from outside the British Isles, and additionally a further 559 individuals at all levels for whom the coding is uncertain.

In the majority of cases, there is reason to believe that the birthplace coding given for each adult individual is an accurate representation, according to our coding scheme, of the data that is available; however, the deficiencies in the *New Survey* birthplace data which are outlined in section 7 (see p. 18), should not be disregarded.

Section 13: Workplace Coding[2]

Objectives

The *New Survey* record cards contain information about each earner's place of work, together with the employer's name, for most of the individuals who were gainfully occupied. This information enables inferences, with varying degrees of detail, to be made about the geographic location of workplaces. Although a more precise refinement of geographical coding would be feasible for some workplaces, it was decided to use London boroughs as the smallest units of reference. As a consequence, it was possible to apply the same coding scheme as for the birthplace codes. In addition, an adjacency code was assigned to identify how "close" the workplace was to the borough of residence.

The success of the workplace coding can be gauged by the proportion of the earners for whom a defined code was assigned. Some 38042 earners were assigned a workplace code out of a total of 49445 individuals in the `IVE` file, a success rate of about 77%. But a significant number of persons listed as earners were known not to be in the labour force or were without employment for one reason or another. With the 39798 individuals coded as being employed or self-employed in the denominator, the success rate increases to over 95%.

Coding Scheme

The criteria for assigning workplace codes were as follows:

1. Where an identifiable street or district was entered on the card, or an employer known to be located in such a place was entered, then the appropriate place code was assigned.

2. Where the place of work or employer was given as "local", "district" or "here", the place code corresponding to the borough of residence was assigned.

3. All street traders, hawkers, pedlars, *etc,* were coded to the borough of residence unless stated otherwise or unless high transport costs were entered on the card.

4. Where the place of work was given as a street which fell in more than one borough and where no employer was given, the workplace was coded as the borough of residence if this was one of the possible choices. Otherwise, the code is that corresponding to the borough containing the larger part of the street or area given as the workplace.

5. All individuals who claimed to work all over London or in a part of London which encompassed several boroughs (e.g. "North London" or "East London") were assigned the London code, `0125000`. This group includes: travellers, entertainers and others who worked all over London; all workers who claimed "various" or "many" employers; and, approximately 300 dockers (whose work encompassed as many as 10 possible boroughs). For those cases where a region corresponded closely to a borough, the relevant place code was entered, thus "West End" was coded as Westminster and "East End" as Stepney.

6. For nearly 200 individuals not currently in the labour force (being retired or unemployed), the borough of last employment could be identified and the workplace was coded accordingly.

[2] Dr Simon Niziol, who carried out the workplace coding under the supervision of Anna Leith, kindly provided detailed notes for this section.

The individuals assigned a code of 9999999 fall into the following broad, overlapping groups:

1. Where the individual was not in the labour force but, nonetheless, was included in the IVE file.
2. Where insufficient information was entered on the questionnaire card, including cases for which no workplace at all was entered, and, also, railway or tram workers for whom no specific depot or station was recorded.
3. Individuals for whom London was given as the workplace but for whom the employer could not be identified.
4. Casual workers, domestics and labourers for whom no other information was entered to enable identification of the workplace.

Adjacency Codes:

The adjacency code is one of six values assigned according to the geographical relationship between borough of residence and borough of employment. For details of the codes, see the *Codebook*, page 38.

Section 14: Interview Dates

Objectives

It is fortuitous that the *New Survey* was conducted at the onset of the Great Depression, the interviews being spread over the years 1929, 1930 and 1931.[3] Although the upheaval in London may have been less dramatic than in the northern industrial cities, it is of interest to identify the dates as accurately as possible. For instance, in March 1930 and October 1931 changes were made to the regulations governing national insurance contributions and benefits. Hence, studies in which these are of significance need access to an accurate rendering of the interview date for each household.

For 5110 cards no date was entered. These correspond largely to the boroughs the interviewing for which was completed early in the survey, i.e. in 1929. (See the *Codebook*, page 14.) For another 5211 cards, there is some information about the interview date (typically, day and month) but no year. The objective of completing the information on dates was, as far as possible, to identify the year in which the interview took place where this was not written on the card. A year was ascribed (in DATE_D) for every case, though with different degrees of reliability. A code variable, AMND_D, signals the confidence with with the year is assigned.

Coding Process

In the construction of DATE_D *no changes* were made to (a) the day and month for any card, or (b) the year when one was written on the card (even though the this may have been

[3]There are 387 cases, 1.4% of the total, for which the interview year is entered as 1932 (and none thereafter). Just 18 cards record an interview year before 1929. All but one of these carry the year 1928, the exception being 1926, surely an error.

later than the completion year noted in the *New Survey*). Where the year was entered on the card, the AMND_D takes the value "0".

The year 1929 was ascribed to all cards corresponding to boroughs listed in the *New Survey* (vol. VI, page 36) as completed in 1929. The boroughs are: Bethnal Green, Finsbury, Hackney, Islington, Poplar, Shoreditch, Stepney, Stoke Newington and West Ham. For three of these boroughs, Bethnal Green, Stepney and West Ham, there is virtually no information on interview dates. A value of "1" was assigned to AMND_D for every case corresponding to these boroughs, given the strong evidence that the interviews were completed early in the survey.

The bulk of the remaining cases for which an interview *year* was not entered can be associated with a small group of interviewers, in particular G. E. Bartlett, C. Etherington and A. N. Winter. Bartlett is pre-eminent in his[4] disregard for the interview year, a pre-eminence achieved no doubt because he signed by far the largest number of interview cards (some 4971, the nearest identifiable contender being J. Hopker with a mere 877). Bartlett did, however, almost always enter the month (as well as the day) of interview, information which is itself valuable in making inferences about the year.

It is possible to gain some insight, albeit incomplete, about Bartlett's activities from documents held in the archives at the BLPES. (The relevant library reference is: LSE CF 146/B-146/E; Organisation: Central Filing Registry; Section Heading: New Survey of Life and Labour.) These documents provide a partial account of payment requisitions made by A. L. Bowley to reimburse the interviewers. The interviewer's name, date of payment and borough are given on the payment requisitions. On the assumption that the payment for each interview was about 2 shillings, it is possible to obtain a rough guide of how many interviews had been undertaken in the weeks immediately preceding the payment. With very few exceptions, the information gleaned from the payment requisitions is compatible with other information about when the interviews were conducted. Some doubts remain, mainly because the surviving payment records are incomplete and, obviously, because inferences drawn from those which do survive cannot be entirely free from error.

On the basis of the payments to interviewers (particularly Bartlett) it has been possible to identify reliably the interview year for 5003 cards, such that AMND_D is assigned the value "1". (The remaining 4293 cards coded "1" correspond to the boroughs listed above for which the house survey was completed in 1929.) For 1024 cards, AMND_D is assigned the value "2" to reflect an element of doubt. This value was used where payment records have not survived or when the interview date seems to be at variance with the year in which the borough's interviews were supposedly completed. For just 5 cards, AMND_D is assigned the value "9" to indicate that the date may well be erroneous.

Among the puzzles which remain unresolved is Bartlett's prodigious interviewing capability. In no fewer than five separate months (September 1930, October 1930, December 1930, January 1931 and October 1931) he appears to have completed more than 400 interviews per month.[5] In October 1930, he seems to have completed well over 600 interviews. It is possible that these estimates are in error though it is unlikely that they are wildly misleading (Bartlett did, after all, record the month in nearly every case and the years are almost surely

[4]There is, strictly, no evidence that Barlett was male but this is assumed throughout.

[5]An interviewer working 20 days per month (five days per week for four weeks) would need to complete 20 interviews per day. For a working day of eight hours, this allows a little less than 25 minutes per interview, including the time taken in going between households.

limited to 1930 or 1931). One possibility is that Bartlett carried out the interviews over a longer span of time and subsequently signed and dated the cards in batches later on. Another is that he subcontracted some of the work. Yet another is that Bartlett sat at home and used his imagination to fill in the cards. The sheer variety of entries, carelessness in completion of the cards and awkwardness of non-standard responses suggest that this is unlikely — a most fertile imagination would have been required. It is possible that this enigma will eventually be disentangled though reliable evidence is hard to come by given the incomplete records and, by now, it is most unlikely that any person directly involved in organising the interviewing still survives.

*****