7 th WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

DATA PROCESSING AND DATA QUALITY

Madrid, Spain, 10 – 11 May 2012

E. Data quality: Quality issues on measurement of variables, issues and opportunities for standardisation between surveys or Internationally

  E3 – Calculating the eligibility rate of sampling units with unknown eligibility

Rita Ranaldi – Italy

Calculating the eligibility rate of sampling units with unknown eligibility[1]

Rita Lima (lima@istat.it)
Rita Ranaldi (ranaldi@istat.it)

ISTAT – Italian National Statistical Institute

1. Introduction

Response rate is considered one of the most important quality indicators for social surveys: the lower, the higher the potential non-response bias[2]. Response rate is used to compare survey quality between surveys, years and countries. The comparisons of the figures, however, may reflect not just different levels of survey participation, but also discrepancies in the formula applied for the response rate calculation and in the list of final disposition codes on which it is based.

Some attempts (AAPOR, 2011, Lynn et al., 2001) have been made for the standardisation of the final disposition codes and the response rate calculation in particular in the social surveys. In the recent years also EUROSTAT (2011a, 2011b) has undertaken a process aimed at developing a harmonised approach for the calculation of response indicators for the European Labour Force survey (EU-LFS). The main project foresees the harmonisation of different aspects concerning the topic: the list of final disposition codes by survey mode, the criteria to accept or refuse partial questionnaires, the main non-response indicators, the methods for estimating the share of eligible units among those of uncertain eligibility, etc. This paper, developed in this context, focuses on the latter topic, in particular with a review of the methods used in the calculation of the eligibility rate among units with unknown eligibility and with an application from Italian Labour Force Survey (IT LFS) micro data.

2. The classification of sampling units according to eligibility

In the literature on response rates, sampling units are divided into three main groups according to their eligibility: a. eligible units; b. units of unknown eligibility; c. not eligible units. The first group of the eligible units can be divided into two further subgroups: a.1. interviews and a.2. non–interviews (non-respondents). According to the AAPOR standard disposition codes, sampling units can be classified as the scheme reported in Figure 1.

Figure 1. Hierarchical level of the sampling units according to eligibility

---
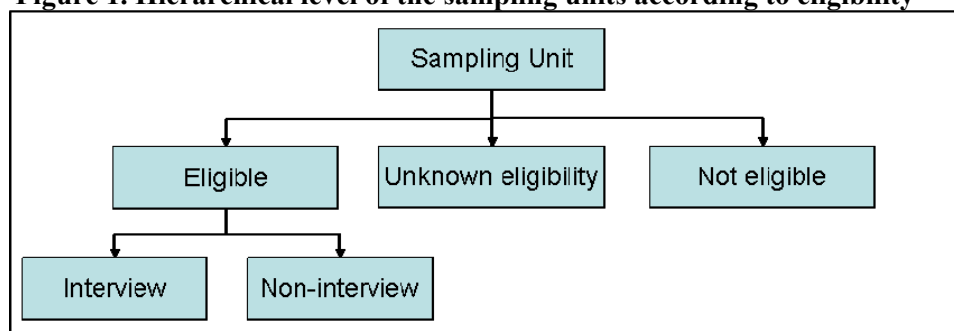
[1] The paper is the result of a joint research of the authors. As far it concerns this version of the work, Sects. 3, 4, 5 should be attributed to R. Lima and Sects. 1, 2 to R. Ranaldi.

[2] First of all, in case of low response survey results could be biased if the characteristics of non-respondents differ from those of the respondents; second, low response rate reduces the precision of estimates, because fewer cases are available for analysis.

A core issue is the definition of "*eligible unit*" as it affects the calculation of response rate:

$$RR = \frac{EI}{EI + EN + \alpha \cdot UN} \times 100 \quad [3] \qquad [1]$$

where:

EI   Number of eligible interviews
EN   Number of eligible non-interviews
UN   Number of units with unknown eligibility
NE   Number of not eligible units
α   Estimated proportion of units of unknown eligibility that are actually eligible

The results can visibly change, depending on the number of eligible units (*EI+EN*), but also on the number of units with uncertain eligibility (*UN*) and on the share of units of unknown eligibility which should be considered as eligible, represented by α.

The IT LFS sample design is a two stage sampling with stratification of the primary units (municipalities); final sampling units (households) are randomly selected from the registry offices in all the municipalities drawn at the first stage. The final unit is eligible if the name corresponds to the household selected and if it is a private household having usual residence in the municipality. The eligibility is unknown when it is not possible to collect sufficient information for a proper classification, for example in case of failing contacts.

The list of the final disposition codes by survey mode[4] associated to uncertain eligibility are reported in Figure 2.

Figure 2. List of final disposition codes associated to unknown eligibility by survey mode

CAPI	CATI
Maximum number of planned contact attempts reached, all with no contact	*Maximum number of planned contact attempts reached, all with no contact*
-- all no one at home or no other information is available	-- all no answer
-- all no one at home and unsuccessful telephone attempts	-- all busy
	-- all unswerving machine with no message
	-- mixed attempts but no contact
End of fieldwork period, with no contact	*End of fieldwork period, with no contact*
-- no attempt made	-- no attempt made
-- attempts made but no contact	-- attempts made but no contact
Unreachable due to wrong name/address	*Unreachable due to wrong telephone number[5]*

In general final disposition codes reflect the combination of the temporary disposition codes recorded in the different attempts of contact with the sampling units.

The "*maximum number of planned contact attempts reached*", in both the survey mode, concerns the cases where all attempts foreseen were made, but without managing to contact the household and thus to collect the necessary information to establish if the unit is eligible.

The "*end of fieldwork period*" covers all cases where the 5 weeks provided by EC Regulation n.577/98 are spent and no attempt was made or attempts were made but without a valid contact to determine if the unit is eligible or not.

Even the unreachable households because of wrong name/address/phone number are labelled with unknown eligibility.

---

[3] For simplicity we consider only the formula of un-weighted unit response rate.

[4] In the IT LFS the information is collected through a mixed mode data collection. Personal interviews (CAPI) are carried out in households of foreign people and in households with no telephone, by about 300 professional interviewers working throughout the national territory. Telephone interviews (CATI) are conducted in all other cases by professional interviewers working in a call center.

[5] The phone number is entered by the interviewer during the first interview. In the following waves, the households coded with "wrong telephone number" are labelled with "unknown eligibility", notwithstanding the household had been interviewed in the previous wave as some of the criteria to define a unit as eligible may not be met.

Figure 3 shows the distribution of the sampling units according to eligibility status by survey mode and wave[6] in the 1st quarter 2011 for the IT LFS.

Figure 3. Distribution of the IT LFS sampling units according to eligibility status by survey mode and wave - 1st quarter 2011
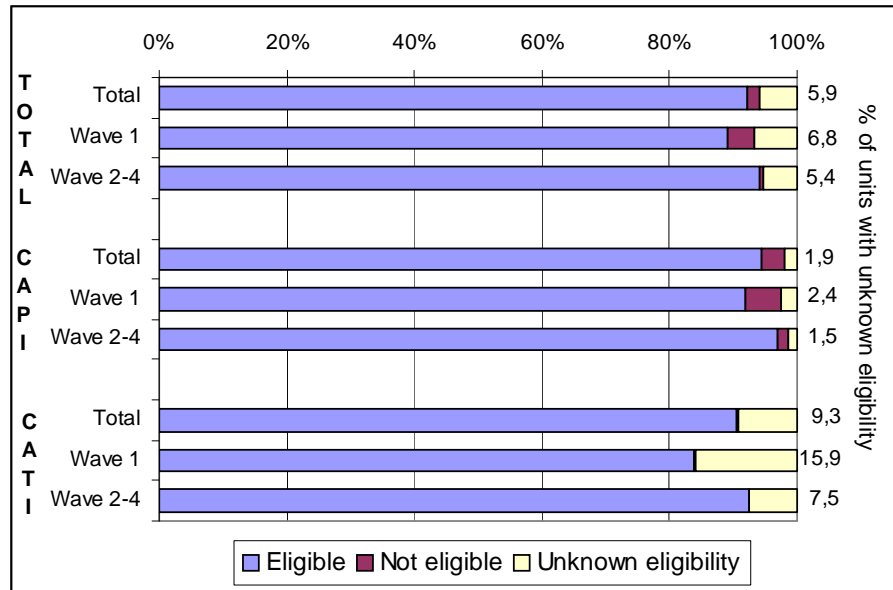
The share of the sampling units with unknown eligibility is 5.9%, but it is higher in the first wave and in particular for the CATI mode. In the first CATI wave this share reaches 15.9%. This result depends on the sampling frame: households are randomly drawn from the municipality registry offices, and their phone number is obtained by the matching with the list of national phone numbers for private users. At the same time in the waves following the first one the share of units with uncertain eligibility is lower because the probability of the eligible units increases.

3. Estimating Eligibility Rates: a brief review

Several different methods have been proposed in literature to estimate the share of eligibity sampling units among those of unknown eligibility, that is α in the formula [1] (Smith, 2009).
Here we compare the survival analysis method carried out by Brick, Montaquila and Scheuren (1997, 2002) with the minimum and maximum allocation method proposed by Montaquila et al. (1997), the proportional allocation method suggested by the Council of American Survey Research Organizations (CASRO, 1982) and the approach proposed by the American Association for Public Opinion Research (AAPOR, 2011).

The Minimum and Maximum Allocation method (MMA)
The MMA method defines the lower and upper bounds of the response rate assuming that all the units with unknown status are actually eligible and that all the cases of unknown status are actually not eligible. So it is possible define a range of response rates by setting:

$$\alpha_{MMA_{upper}} = 0 \text{ and } \alpha_{MMA_{lower}} = 1 \text{ [2]}$$

The Proportional allocation method (CASRO)
The CASRO method assumes that the proportion of eligible units amongst those whose eligibility is unknown is equal to the proportion of eligible units amongst the eligibility known sample units. The CASRO formula for estimating α is shown below:

---

[6] The sampling plan provides that final units are rotated according to a 2-(2)-2 rotation scheme: households are interviewed during two consecutive quarters; after a two-quarters break, they are again interviewed twice in the corresponding two quarters of the following year. As a result, each household is included in four waves of the survey.

$$\alpha_{CASRO} = \frac{EI + EN}{EI + EN + NE} \qquad [3]$$

The American Association for Public Opinion Research (AAPOR) approach

AAPOR states that the estimation of the eligibility rate is left to the discretion of the organization(s) and individual(s) undertaking the research, that the estimate for eligibility from unknown cases should be based on the best available scientific information, and that the basis of the estimate must be explicitly stated and explained. AAPOR suggests that all units of unknown eligibility are considered as eligible unless it can be shown that some of them are indeed ineligible. Then the AAPOR formula to calculate α is:

$$\alpha_{AAPOR} = 1 \qquad [4]$$

The Survival Analysis method (SAM)

The survival analysis approach, applied in particular for CATI surveys, obtains an estimate of eligibility status using the Kaplan-Meier estimate (Lawless, 1982) where the determination that a unit is *eligible* and the determination that a unit is *not eligible* may be thought of as the two "*causes of death*". The survival functions for these two causes of death are estimated by:

$$\hat{S}_{eligible}(t) = \Sigma_{i:t_i < t} \frac{d_{eligible,i}}{n_i} \hat{S}(t_i) \qquad \text{and} \qquad \hat{S}_{ineligible}(t) = \Sigma_{i:t_i < t} \frac{d_{ineligible,i}}{n_i} \hat{S}(t_i) \qquad [5]$$

where:

$\hat{S}(t)$ Kaplan-Meier estimate[7]

i Trial or call attempt at which there are non-censored cases resolved to be *eligible* or *not eligible*

n_i Number of cases "*at risk*" just prior to trial t_i (that it corresponds to still being contacted at the t_i^{th} attempt)

d_i Number of "*deaths*" or resolved cases at trial t_i

$d_{eligible,i}$ Number of cases determined to be eligible at trial t_i

$d_{ineligible,i}$ Number of cases determined to be not eligible at trial t_i

The summations are defined only at those trials t_i where $n_i > 0$.

The overall eligibility rate is then estimated as:

$$\hat{R}_\infty = \frac{\hat{S}_{eligible}(0)}{\hat{S}_{eligible}(0) + \hat{S}_{ineligible}(0)} \qquad [6]$$

With [6] and the resolved cases we come up with a distinct eligibility rate for the unresolved cases:

$$\alpha_{SAM} = \frac{(\hat{R}_\infty \times n_{tot}) - n_{eligible}}{n_{unknown}} \qquad [7]$$

where n_{tot} is the total sample size, $n_{eligible}$ is the number of sample cases resolved as eligible and $n_{unknown}$ is the number of cases with unknown eligibility. The numerator of this equation represents an estimate of the number of unresolved cases that are eligible (estimate of total eligible units minus known eligible units).

---

[7] $\hat{S}(t)$ is defined only for those t for which $n_i > 0$.

4. Results

We applied the MMA method to the IT LFS results for the 1st quarter of 2011, in order to determine the upper and the lower bounds for the response rates by survey mode and wave. The results are showed in Figure 4. The range of response rates is bigger when the survey mode is CATI, in particular in case of first wave; in this last case the range is 12.7 percentage points. Therefore the range is wider when the share of the uncertain eligibility is larger. For the IT LFS, while the estimate of the eligibility rate is almost irrelevant in terms of impact on the overall response rate for CAPI, it is crucial for CATI.

Figure 4. Ranges of the response rates according to MMA method in the IT LFS by survey mode and wave - 1st quarter 2011
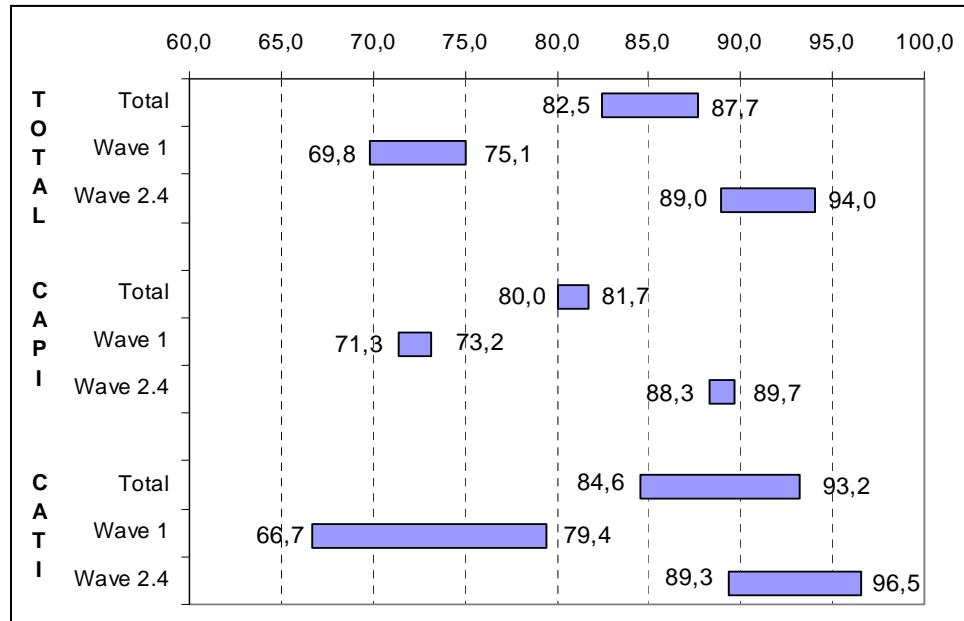

To test the effectiveness of the MMA, CASRO, AAPOR and SAM methods, we estimate α for the CATI component of the IT LFS. Because of the nature of sample and the likelihood that α would be quite different for the re-interviews (waves 2-4) and the first wave, we calculated α for these two groups when applying the different methods.

To use the survival method, we had to classify the units in addition to eligible-not eligible-unknown also by interim status codes (such as "*call-back*") and count the number of call attempts until a unit was resolved. For example, a case with 15 call attempts resulting in "*no answer*" or "*busy*" outcomes would be classified as unresolved after 15 attempts. In the case of a telephone number attempted 16 times with the first 15 attempts resulting in "*no answer*" or "*busy*" outcomes and the 16th call resulting in a refusal, this would be classified as eligible.

All the telephone numbers are dialed with a justifiable "*random censoring*" calling rules of call attempts for a given phone number that is the end of the fieldwork period. In figures 5 and 6 are graphed eligibility rates by wave. In the waves following the first one, we note that there are high numbers of call attempts and after 37 call attempts there are only 2% unresolved units (Figure 6).

Figure 5. Eligibility rate in the wave 1 of IT LFS by number of call attempts – 1st quarter 2011
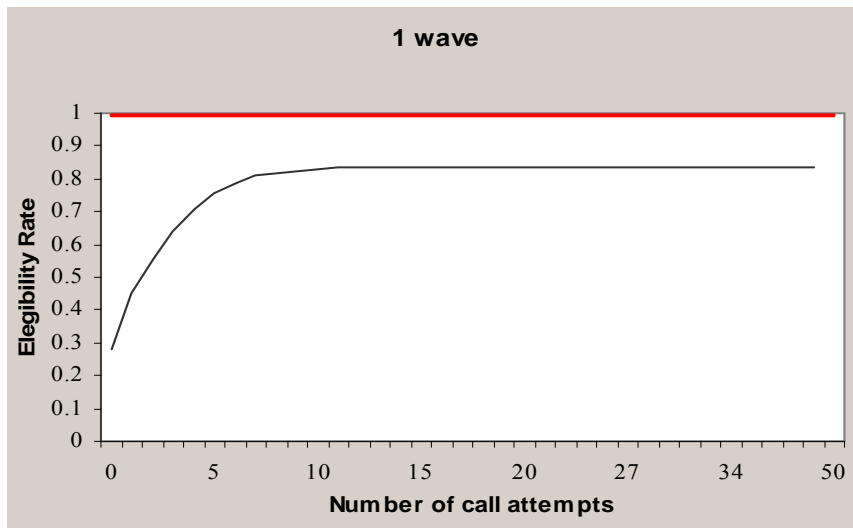
Figure 6. Eligibility rate in the waves 2-4 of IT LFS by number of call attempts – 1st quarter 2011
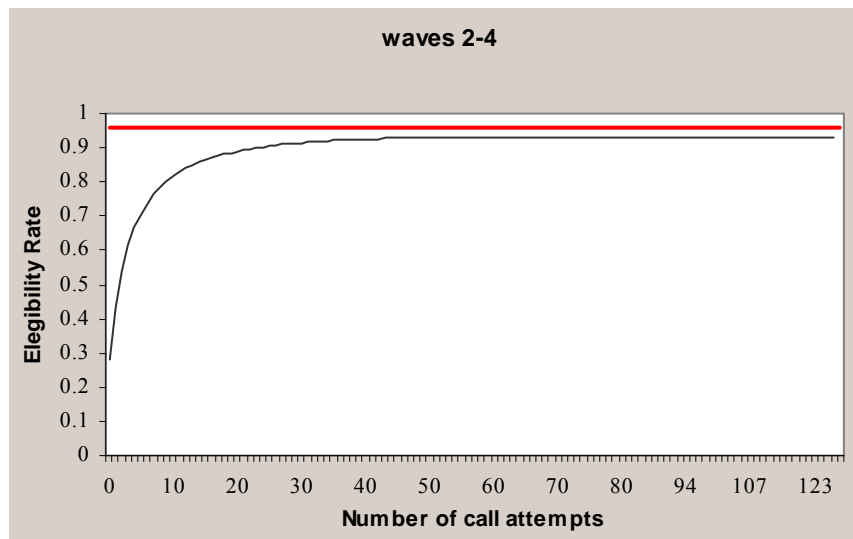
Note: The horizontal line is $\hat{R}_\infty$

Table 1 illustrates the estimates of the eligibility rates α and the response rates according to different methods. As the results show, there is no practical difference between the different methods, and the SAM eligibility rates appear to converge when there are a large number of call attempts, as there normally are for the IT LFS.

Table 1. Estimated eligibility rates and response rates according to different methods in the IT CATI LFS – 1st quarter 2011

	Eligibility rate (α)					Response rate (RR)				
	MMA lower	**MMA upper**	**CASRO**	**AAPOR**	**SAM**	**MMA lower**	**MMA upper**	**CASRO**	**AAPOR**	**SAM**
Total	1	0	0.9990	1	0.9990	84.558	93.210	84.566	84.558	84.566
Wave 1	1	0	0.9959	1	0.9967	66.728	79.439	66.771	66.728	66.763
Waves 2-4	1	0	0.9997	1	0.9996	89.313	96.545	89.315	89.313	89.316

5. Discussion and conclusions

The estimation of the share of eligible units amongst those of unknown eligibility is an important issue because it is a component in the denominator of the response rate formula.

While for some surveys, applying the different methods it may make a substantial difference in the estimated response rate, for the IT LFS we found that they produce similar results.

Although each method could be an alternative approach, we recommend these brief notes.

The MMA method implies the computation of two response rates, i.e. a range that has the advantage of the comparability of the results, but the drawback that two point estimates are difficult to manage with communication issues. Besides, as Smith (2009) stated, if the unknown eligibility level is large, as it often happens, the possible range of response rate is great.

The CASRO method has the advantage of being easy to apply and it does not inflate the response rate, as affirmed by Smith (2009). But the assumption on which is based, that is the units with uncertain eligibility have the attributes as the units with known eligibility, may be too strong.

Also the AAPOR approach has the advantage of being simple, but also in this case the assumption of considering all uncertain units as eligible may be strong.

Having sufficiently large sample of households that are contacted more times, the SAM method is essential for estimating the response rate accurately and for making sure that the number of attempts of contact is a reasonable measure of exposure. The main result is that, insisting endlessly to contact the units, at the end they are almost all eligible. But the SAM method has a lot of limitations as stated by Smith (2009), in particular, as we have experienced, it is a very complicated estimation method to carry out.

Under these conditions the CASRO method produces more similar results to the SAM method for the IT LFS. Moreover it seems the most appropriate one to estimate the eligibility rate in case there is evidence that α is less than 1.

References

Beerten R., Lynn P., Laiho J. and Martin J. (2001) "Response Rates as a Measure of Survey Quality" NTTS & ETK 2001 Conference

Brick, J. M. and Broene P. (1997) "Unit and Item Response Rates,Weighting and Imputation Procedures in the 1995 National Household Education Survey." U.S. Department of Education, Office of Educational Research and Improvement, NCES 97-06

Brick M., Montaquila J., Scheuren .(2002) "Estimating the Residency Rates for Undetermined Telephone Numbers." Public Opinion Quarterly, volume 66:18-39. American Association for Public Opinion Research.

Council of American Survey Research Organizations (1982), On the Definition of Response Rates, Port Jefferson, NY: CASRO

Eurostat (2011a) "Towards common guidelines for EU-LFS harmonised non-response indicators", January 2011.

Eurostat (2011b) "Towards common guidelines for EU-LFS harmonised non-response indicators – Follow-up of the June 2011" LAMAS WG Meeting, September 2011.

Ezzati-Rice T.M., Coronado V.G., Frankel M.R., Hoaglin D.C., Loft J.D. and Wright R.A. (1999) "Estimating Response Rates in Random-Digit-Dialing Surveys That Screen For Eligible Subpopulations", Accessed on 7/11/02 and retrieved from; ww.jpsm.umd.edu/icsn/papers/EzzatiRiceCoronado.htm

Giuliani G., Grassia M.G., Quattrociocchi L., Ranaldi R. (2004) "New methods for measuring quality indicators of ISTAT's new CAPI/CATI Labour Force Survey", CD-Rom Proceedings, European Conference on Quality and Methodology in Official Statistics (Q2004), Mainz, Germany, Castle of the Prince Elector, 24-26 May 2004, published by Statistisches Bundesamt (Federal Statistical Office Germany)

Keeter S., Miller C., Kohut A., Groves R.M. and Presser S. (2000) "Consequences of reducing non-response in a national telephone survey", Public Opinion Quarterly, Vol. 64, No. 2, pp 125-148

Lawless J. F. (1982) "Statistical models and methods for lifetime data", John Wiley & Sons, New York

Lynn P., Beerten R., Laito J. and Martin J. (2001) "Recommended Standard Final Outcome Categories and Standard Definitions of Response Rate for Social Surveys". Working Papers of the Institute for Social and Economic Research, ISER, paper 2001-23. Colchester: University of Essex

Minato H. and Luo L. (2004) "Towards a Better Estimation of Working Residential Number (WRN) Rate Among the Undetermined: An Application of Survival Analysis", Paper presented to the American Association for Public Opinion Research, Phoenix.

Montaquila, J. and Brick J.M. (1997) "Unit and Item Response, Weighting, and Imputation Procedures in the 1996 National Household Education Survey" Working Paper (NCES 97-40). U.S. Department of Education. Washington, DC: National Center for Education Statistics

Smith T.W. (2009) "A Revised Review of Methods to Estimate the Status of Cases with Unknown Eligibility," Report of the Standard Definitions Committee for the American Association for Public Opinion Research,

The American Association for Public Opinion Research (2011) *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 7th edition. AAPOR.