# Estimating the Eligibility Status of Cases with Unknown Eligibility

Ashley Amaya, Jenny Marlar, & Ned English

Based on Smith, T.W. (2009) A Revised Review of Methods to Estimate the Status of Cases with Unknown Eligibility. *Report prepared for AAPOR Standard Definitions, 5th Edition, 2008. Available online at https://aapor.org/wp-content/uploads/2022/11/ERATE09.pdf*

*Thank you to the numerous Standards Definitions Subcommittee members including Andrew Caporaso, Peter Lugtig, Cameron McPhee, and Amanda Nagle who brainstormed how to reenvision this document given changes in frames and appended data. Thank you to Andy Peytchev and Tom Smith for their review and feedback.*

## About this Report

This report serves as a companion to the tenth edition of the American Association for Public Opinion Research (AAPOR) *Standard Definitions*. The *Standards Definitions* provides multiple definitions for calculating response rates, refusal rates, and contact rates. Several of these include a term, *e*, an estimate of eligibility among cases with unknown eligibility. Unlike other inputs into the equations, *e* can be calculated in multiple ways and is at the discretion of the researcher. This report focuses on the most common methods to calculate *e*.

The primary difference between this report and its predecessor (Smith, 2009) is that this report broadens the scope of the conversation to reflect newer and more complex survey data collection methodologies, specifically multi-mode, multi-phase, and multi-frame designs where each component may have its own estimate of *e*. It also focuses on explicit calculation methods and offers recommendations on when each should be used. This change is consistent with the revisions made in the tenth edition of the *Standard Definitions*.

What does not differ between this report and its prior iteration is the ultimate take-away: in estimating e, one must be guided by the best available scientific information on what share eligible cases make up among the unknown cases. **One must not intentionally select a proportion in order to boost the response rate.**

## Introduction & Background

Most sample surveys are likely to have some proportion of sampled units for which eligibility has not been determined but who are, in fact, eligible. For example, a survey that uses an address-based sample (ABS) to mail survey invitations may get no response back for some sampled addresses. Some of those nonrespondents are likely eligible households. Or, a random digit dial (RDD) telephone survey might ring without connecting to a person or voicemail. Some of those phone numbers are likely working and connect to an eligible individual. Or, you've sampled a list of registered voters and emailed them to complete a web survey but get no response. Again, some of those emails are likely working and are used by a registered voter.

*Standards Definitions* offers multiple methods for calculating different key survey metrics. For the purposes of this report, the methods differ on whether they assume:

1) All unknown eligible sample units are eligible
2) Some unknown eligible sample units are eligible
3) None of the unknown eligible sample units are eligible.

This report focuses on the calculations in group two – assuming some of the unobserved sample units are eligible. This assumption is made in the following equations found in the *Standard Definitions*:

1) Response rate 3 (RR3)
2) Response rate 4 (RR4)
3) Refusal rate 2 (REF2), and
4) Contact rate 2 (CON2).

Each of these equations includes a term *e*, an estimated proportion of cases of unknown eligibility that are, in fact, eligible. However, there are many ways to estimate *e*, and different assumptions may have significant effects on the final rates. This document outlines the most common approaches.

Regardless of the approach used, one must do the following to compute response, refusal, and contact rates outlined in the *AAPOR Standard Definitions*:

1) Identify the number of cases of unknown eligibility (i.e., *UH*, *UR*, and *UO* in *Standard Definitions*), and
2) Use an appropriate method for estimating *e*. Three are listed in the rest of this document.

## Common Approaches to Compute *e*

This section presents three common methods for calculating *e*, but it is not an exhaustive list of methods for *e*. Ultimately, the method used is at the researcher's discretion. Regardless of the method used, it should be well documented and justified.

This section does not make any judgement about which method is superior, but it does outline strengths and weaknesses of each approach.

### Observed Eligibility Method

This approach assumes that the eligibility rate among cases for which eligibility is unknown can be inferred based on information known from the cases for which eligibility is observed. The AAPOR response rate spreadsheet uses the Council of American Survey Research Organization (CASRO) method to calculate *e*. The online AAPOR response rate calculator allows users to apply the CASRO method or input their own value for *e*.

The simplest way to use observed information to predict e for unobserved cases (i.e., those for which eligibility is unknown) is to use the CASRO method (Frankel, 1983).[1] This assumes that the eligibility rate among unknown eligible cases is equal to the eligibility rate among known cases. Therefore, to calculate *e*, one must calculate the eligibility rate among the known cases:

$$e = \frac{I + P + R + NC + O}{I + P + R + NC + O + NE}$$

Where:

*I* = completed interview (known to be eligible) (1.1)[2]

*P* = partial interview (known to be eligible) (1.2)

*R* = refusal (known to be eligible) (2.1)

*NC* = non-contact (known to be eligible) (2.2)

*O* = other non-interview (known to be eligible) (2.3, 2.9)

*NE* = not eligible (known to be ineligible) (4.0)

The main strengths of the CASRO method are that it uses the observed eligibility from the same sample to extrapolate to the rest of the sample and is relatively simple to calculate. However, the main limitation is that the eligibility rate for those with undetermined eligibility can be different. For example, the United States Postal Service (USPS) generally returns most undeliverable mail (i.e., unoccupied housing units) to the sender. As a result, the remaining addresses for which household status is unknown are likely occupied and eligible. Using the CASRO approach would likely underestimate eligibility.

Alternatively, one could apply a survival model on observed cases to estimate the eligibility rate among unknown eligible cases. Brick and his colleagues detail these equations and demonstrate this approach on the

[1] Since the creation of this method, CASRO merged with Marketing Research Association and became Insights Association.
[2] The codes found in parentheses are consistent with *AAPOR Standard Definitions* dispositions.

1999 National Household Education Survey and Round 2 of the National Survey of America's Families (Brick, Montaquila, & Scheuren, 2002).

While this approach is more statistically cumbersome than the CASRO approach, it still allows researchers to use sample information and does not require appending additional data from outside sources. It also relaxes the assumption that the observed eligibility rate is equivalent to the unobserved rate made by the CASRO method

## Global Auxiliary Information Method

In some instances, the eligibility rate of the population may be known. For example, let us draw a sample of residential addresses from the USPS's Computerized Delivery Sequence file with the goal of interviewing one adult in each occupied household. We conduct an in-person survey, but still have some sampled addresses for which eligibility is unknown. We look to the American Community Survey[3] to identify the number of housing units in the US (n=145,333,462) and the number of those that are occupied households (n=131,332,360). We may assume that the eligibility rate should be 90.4% (occupied households/total housing units).

This population estimate may be used to calculate *e* in two ways. First, one may opt to set *e* to the estimate itself. In our example, this would mean that *e* equals 0.904.

Second, one could adjust *e* so that that eligibility rate for the entire survey (known and unknowns) equals the population estimate. That is:

$$population\ estimate = \left(\overbrace{\frac{I+P+R+NC+O}{I+P+R+NC+O+NE}}^{\text{Eligibility rate of known sample}} * \overbrace{\frac{I+P+R+NC+O}{total\ sample}}^{\text{Proportion of sample known}}\right) + \left(e * \overbrace{\frac{UH+UR+UO}{total\ sample}}^{\text{Proportion of sample unknown}}\right)$$

Where:

UH = unknown if household (3.1)

UR = unknown if sampled unit is eligible/housing unit contains an eligible respondent (3.2)

UO = unknown other (3.9)

All inputs should be known other than *e*, and one may solve for *e*.

The main strength of this approach is that it is fairly simple to implement and that it does not require an assumption that those with and without established eligibility have the same eligibility rates. However, the eligibility rate is obtained from another source and not from the particular sample. It may even be based on another sampling frame, as in this example. This means that the eligibility rate can be incorrect for the particular sample, and may be biased.

[3] 2023 ACS 1-Year Table DP04.

### Case-level Auxiliary Information Method

Often, supplement information is available on the sampling frame or auxiliary information may be appended to individual cases. We may use this information to "guess" whether an individual sampled unit is eligible. The way in which we guess can take many forms and is at the discretion of the research.

For example, one may have a variable on the frame that is a sufficient proxy for eligibility. Let us assume a wholesale store would like to conduct a customer satisfaction survey among members who are new parents. The store has a list of members and their recent purchases, but it must rely on a survey question (Are you the parent to a child under the age of 1?) at the start of the survey to confirm eligibility. Eligibility is unknown for individuals who did not complete the screener question, but the store knows whether these members purchased diapers in the last 12 months. We may use this diaper indicator as a proxy for eligibility. In this case, *e* is a case level variable that is set to 0 (no diaper purchases) or 1 (diaper purchases), and the rate calculations would be slightly altered. For example, Response Rate 3 (RR3) would change from:

$$RR3 = \frac{I}{I + P + R + NC + O + e(UH + UR + UO)}$$

to:

$$RR3 = \frac{I}{I + P + R + NC + O + \sum e_i}$$

where $e_i$ is the proxy indicator for whether a given case ($i$) is eligible.

Alternatively, if a researcher has multiple frame or case-level auxiliary variables, he/she/they may build an imputation model among cases for which eligibility is known and then apply the model to the unknown cases. As noted above, *e* becomes a case level variable, and the rate calculations follow the revised equation.

The main strength of this approach is that an eligibility status may be obtained for every sample case. However, this approach relies on an accurate proxy measure for eligibility. In the example above, it is possible that some households always buy diapers from other stores, possibly biasing the eligibility rate and resulting response rate. It would be desirable to compare the direct and proxy eligibility for sample cases with known eligibility to evaluate this assumption.

## Frequently asked questions

### Which of the above methods should I use?

There is no one-size fits all method. Different approaches carry different strengths and weaknesses (as listed above). The decision is ultimately left to the researcher. However, in estimating e, one must be guided by the best available scientific information on what share eligible cases make up among the unknown cases. One must not select a proportion in order to boost the response rate. Regardless of the method chosen, the researcher should justify his/her/their choice and make it transparent in the documentation.

### I have multiple frames. Should I calculate *e* separately for each frame?

Yes. The available information or assumptions of eligibility for each frame may be quite different. For example, let's say you would like to measure presidential approval in Michigan using a voter file (i.e., named list) and random digit dial cell phone. We might assume that all people in the voter file are eligible and set *e*=1. However, the same assumption is not realistic for the cell sample. Cell cases for which eligibility is unknown

may not be working, belong to a person living outside of Michigan, or belong to a business. Given the lack of information available on the cell sampling frame, we may opt to use the CASRO method to calculate *e* for this frame.

## The above methods assume there is only one value for *e*, but the *AAPOR Standard Definitions* has the option for multiple *e*'s. Why are these documents different?

Under the simplest equations, a single value for *e* is applied to the sum of the count all unknown eligible respondents. For example, response rate 3 is calculated:

$$RR3 = \frac{I}{I + P + R + NC + O + e(UH + UR + UO)}$$

This can be an oversimplification in some instances, particularly when screening a household for the presence of a subpopulation. In this case, it may be more appropriate to use this equation:

$$RR3 = \frac{I}{I + P + R + NC + O + (e_{UH} * UH) + (e_{UR} * UR) + (e_{UO} * UO)}$$

where:

$e_{UH}$ = the probability that a sampled unit is a household[4]

$e_{UR}$ = the probability that someone in the household is eligible

$e_{UO}$ = the probability that a sampled unit that does not fit into the other unknown categories is eligible[5]

Any of the above methods for calculating *e* may still apply to its components. For example, let us assume that you are using an address-based sample to screen for adults over 65 years of age. You will likely have some addresses which you know nothing about (UH) and some that you know are households but do not know if an adult 65 or older lives in the household (UR). You're unlikely to have any other type of cases for which eligibility is unknown (UO). If you use CASRO assumptions, you may calculate the eligibility rate of your known households for which eligibility is unknown as:

$$e_{UR} = \frac{I + P + R + NC + O}{I + P + R + NC + O + NE}$$

You may similarly calculate the eligibility rate for addresses that may or may not be households. In this case, you will need to break out your "not eligible" cases by reason – not eligible because not a household and not eligible because they do not have eligible individuals living in the household. The equation may be written as:

$$e_{UH} = \frac{I + P + R + NC + O + UR + NE_{HH}}{I + P + R + NC + O + UR + NE} * e_{UR}$$

Where $NE_{HH}$ is the number of sampled addresses that are known household but did not contain an eligible adult.

[4] Note that if you are using a list of individuals (as opposed to households), then you should not have any cases classified as UH and do not need to calculate $e_{UH}$.

[5] It is extremely rare to categorize any cases into UO. If you do not have any cases in this category, then you will not need to determine $e_{UO}$.

## References:

AAPOR (2016). Standard definitions: Final dispositions of case codes and outcome rates for surveys, 9th ed. Research report. Available online at https://aapor.org/wp-content/uploads/2022/11/Standard-Definitions20169theditionfinal.pdf.

AAPOR (2023). Standard definitions: Final dispositions of case codes and outcome rates for surveys, 10th ed. Research report. Available online at https://aapor.org/wp-content/uploads/2024/03/Standards-Definitions-10th-edition.pdf.

Brick, J. M., Montequila, J., Scheuren, F. (2002). Estimting Residency Rates for Undetermined Telephone Numbers. *Public Opinion Quarterly*.

Frankel, L. (1983). Report of the CASRO Task Force on Response Rates. In *Improving Data Quality in a Sample Survey*, ed. F. Wiseman. Cambridge, MA: Marketing Science Institute.

Massey, J.T. (1995). Estimating the response rate in a telephone survey with screening. Proceedings of the Joint Statistical Meetings. Available online at http://www.asasrms.org/Proceedings/papers/1995_115.pdf.

Smith, T.W. (2009). A revised review of methods to estimate the status of cases with unknown eligibility. Research report. Available online at https://aapor.org/wp-content/uploads/2022/11/ERATE09.pdf.