GRADUATE OUTCOMES SURVEY METHODOLOGY STATEMENT PART TWO

SURVEY DESIGN AND IMPLEMENTATION
JULY 2020 VERSION 1.1 18/19 COLLECTION

HESA

HESA
95 Promenade
Cheltenham
GL50 1HZ
E Liaison@hesa.ac.uk
T +44 (0) 1242 211 144
W www.hesa.ac.uk

GRADUATE OUTCOMES SURVEY METHODOLOGY PART 2: SURVEY DESIGN AND IMPLEMENTATION

Contents

INTRODUCTION

Part two of the methodology statement contains details of the most important aspects of survey design, data collection, analysis and dissemination for Graduate Outcomes. It is aimed at the users of Graduate Outcomes survey data as well as those with an interest in survey methodology.

This part of the methodology statement will be a 'live' record of the methodology, at least in the first few years as we make improvements to our data collection, processing, analysis, and dissemination policies. We aim to revise it as changes to existing processes are made and new initiatives are introduced. Previous versions of the statement will still be available to users.

Part one of the methodology statement contains the history and background to the development of the Graduate Outcomes survey. It outlines the process HESA went through to review the need for a replacement to previous iterations, how we engaged with the sector on its design and the intended governance structure. View methodology statement part one: history and background.

Please refer to our glossary page for definitions of terms used in this document. View this document on HESA's website.

SURVEY COVERAGE

Graduate Outcomes is a population survey (meaning we aim to survey the whole population of interest, rather than a sample) of almost all graduates of higher education in the UK, in a given academic year. For the first time, we will have the opportunity to measure and understand graduate destinations in their entirety, across all Higher Education Providers (HEPs) in the UK and Further Education Colleges (FECs) in England, Wales and Northern Ireland.

Given the uses of Graduate Outcomes, it is important to try and collect information about graduates of every single HEP in the UK (large and small), to high standards of detail, completeness, accuracy, and consistency. With more than 400 providers in the coverage, a centralised population survey of all graduates was deemed the most suitable option, as opposed to a distributed or sample survey.

Student data on demographic and course characteristics from HEPs and FECs in Wales is collected by HESA. Similar data from FECs on their higher education provision are compiled by England and Northern Ireland. The Graduate Outcomes target population contains all students reported to these organisations as obtaining relevant higher education qualifications during the reporting period 1 August to 31 July and whose study was full-time or part-time (including sandwich students and those writing-up theses). This overall target population is then broken down into four cohorts, depending on when a graduate completed their course. For example, a graduate who completed their course between the months of May-July 2018 were surveyed in September 2019 cohort (circa. 15 months later). For each annual collection, the respective coverage definition is available on the HESA website.

As with any survey where there is no legal compulsion to respond, there will always be an element of non-response. We have described in later sections, the steps we are undertaking to maximise response rates, reduce non-response bias and make the achieved sample as representative of the population as possible.

SAMPLING FRAME

Although Graduate Outcomes is a population survey in practice, it is inevitable that there is some under-coverage: some graduates cannot be surveyed, perhaps because their contact details were unavailable, or because they are seriously ill, or have died. We call the list of all cases we can include, the “sampling frame”. Using the coverage criteria outlined above, a sampling frame is developed for each collection year. This contains contact details (email addresses, phone numbers and in rare cases residential addresses) for all graduates eligible to take part in the survey.

A bespoke data collection portal (the ‘Graduate Outcomes provider portal’) enables the collection of contact details to be used for the sampling frame. Contact details for graduates are supplied by providers using this portal. Using the eligibility criteria (as defined under survey coverage), a population file containing unique identifiers for all graduates in coverage is made available to each provider in the portal. For those providers who return student data to HESA, the population file is automatically generated from it. For other providers, i.e. FECs in England and Northern Ireland, the information for the population file will be taken from the appropriate student data collection (the Individualised Learner Record (ILR) in England and the Consolidated Data Return (CDR) from the Department for the Economy Northern Ireland (DfENI)). HESA supplies a separate report to provide the relevant derived fields which will support providers in identifying graduates for inclusion in this return.

Once the Graduate Outcomes data collection portal opens, providers begin uploading contact details for graduates, one cohort at a time, based on dates provided in data collection schedule. A quality assurance period runs from when the collection system opens until the ‘final updates’ deadline. This is the process in which a provider’s contact details submissions are validated on the provider portal to ensure they are of sufficient quality for use. This is facilitated by a set of quality rules developed by HESA, alongside providers’ own activity to ensure contact details are useable and accurate. Contact details should be continuously updated (if necessary) during the survey field period.

Once compiled and approved by the provider, the set of contact details for each cohort are exported to our survey supplier platform (Confirmit) through which we manage all aspects of survey design and data collection. Contact details for each cohort are compiled and processed individually, prior to data collection.

ROLE OF HIGHER EDUCATION PROVIDERS

Our model of open centralisation means that HESA is responsible for delivering the survey, including ensuring graduates are surveyed, with the support of a number of suppliers. HESA depends critically on providers to fulfil a number of roles to ensure the survey is a delivered successfully, response rates are maximised, and bias is not introduced.

The primary role of providers can be summarised in three activities:

- Collecting and maintaining contact details to support the creation of an accurate and comprehensive sampling frame
- Submission of contact details through a bespoke data collection portal
- Promotion of the survey to create brand awareness among prospective respondents (which we discuss in greater detail in a later section of this methodology)

Providers are instructed to refrain from contacting the graduates during the survey window to minimise risks of introducing bias.

These and other provider responsibilities are explained in more detail on our website.

HESA'S ROLE

Methodological decisions about the survey design and implementation are made by HESA, with input from a Steering Group comprising of regulatory and funding bodies, relevant sector bodies and providers.

All HESA collections have a coding manual, which contains the relevant operational information and guidance to help aid a provider in compiling and submitting appropriate data. For Graduate Outcomes, there are two relevant coding manuals for the different aspects of the collection:

- The contact details coding manual contains all the necessary information and guidance required to enable providers to submit their contact details correctly and on time. It contains several different areas of information, ranging from population coverage, data protection and the physical data structure of the collection, to the data items (required to be returned), the data quality rules (which must be adhered to), and the user guide (which sets out, step by step, the overall processes required during the collection and use of the provider portal).
- The survey results coding manual contains the necessary information for providers to download and analyse their raw survey results (during a cohort) and the final data delivery (when appropriate).

HESA has also provided a suite of communications materials for providers to use to build brand awareness for the survey. This includes logos (in various formats), engagement materials, social media content and much more.

SURVEY TARGETS

It is important to ensure that response rates do not fall significantly over time, especially if they are likely to reduce the usability of survey results (e.g. through the generation of less precise estimates for smaller groups of population). Target response rates could therefore be one possible measure of survey performance.

DLHE response rate targets

The DLHE survey was administered with target response rates, which HE providers were required to meet. Target response rates for 2016/17 were set at:

- 80% for UK-domiciled HE leavers who previously studied full-time.
- 70% for UK-domiciled HE leavers who studied part-time.
- 80% for Research Council-funded students.
- 50% for all other EU HE leavers.
- 20% for non-EU international HE leavers.

To set target response rates for Graduate Outcomes, response rates from DLHE and Longitudinal DLHE (LDHLE) were analysed; the former being a destinations survey six months after graduation and the latter being its longitudinal element conducted 42 months after graduation. In the absence of a precedent for a 15-month destination survey, the two existing surveys were deemed as the most suitable benchmarks.

As expected, the response rates achieved in the LDLHE survey were lower than those achieved in the DLHE survey. Although many factors may contribute to this decrease, time-lag is a major determinant.

The timing of the Graduate Outcomes survey is 15 months after graduates complete their studies. This is exactly one quarter of the way between the timing of the DLHE (six months) and the timing of the LDLHE (42 months). Assuming the drop in response rates over time is linear, expected response rates for Graduate Outcomes can be calculated by taking a quarter of the drop away from the response rate achieved in DLHE. This is adjusted to account for the fact that response rates may drop in a non-linear fashion and the aim to provide a more cost effective yet effective overall solution to the sector.

Graduate Outcomes response rate targets
The resulting targets for Graduate Outcomes response rates, by main groups are:

UK domiciled full-time: **60%**
UK domiciled part-time: **60%**
Research funded: **65%**
EU domiciled: **45%**
Non-EU domiciled: **25%**

These targets are mainly applicable for the entire set of survey results across a year but are also used to monitor performance of individual cohorts. This is to acknowledge the diverse composition of different cohorts in terms of graduate characteristics and how it impacts the response rates that can be realistically achieved.

The ‘headline’ response rates for the Graduate Outcomes survey are defined as:

- Numerator: Count of records with a valid response to a minimum[1] (pre-determined) set of core questions (classified with a status of ‘survey completed’).
- Denominator: Count of all records in the target population, excluding those marked as dead or seriously ill.

In addition to responses classified as ‘survey completed’, a status of ‘partially completed’ has been assigned where some of the core questions are missing but the first two questions have been answered. Although partially completed responses do not contribute to the ‘headline’ response rates, these are used alongside ‘survey completed’ responses in statistical outputs and form part of the base population for this purpose. Data from such responses will appear in published statistics with unknown values for questions that were not answered.

It should be noted that achieving a high response rate is not sufficient in itself to ensure good quality data has been collected. In later parts of this methodology statement, we will be discussing the other checks and processes we have undertaken in assessing and ensuring the quality of the data.

[1] Full details on mandatory and optional survey questions are available here https://www.hesa.ac.uk/innovation/outcomes/survey

SURVEY QUESTIONS

The Graduate Outcomes questionnaire design utilises a substantial proportion of established questions from DLHE and LDLHE, with some revision, where appropriate. Several additional questions have also been included following feedback from the sector, gathered through consultations. This includes enhanced questions on self-employed graduates, graduate reflections on their activity, and subjective wellbeing questions (see part one of this methodology statement for details).

A high level survey routing diagram of Graduate Outcomes is available, alongside the latest version of the survey.

In order to reduce measurement error (specifically respondent and interviewer error), the survey underwent cognitive testing with a sample of graduates who conformed to the criteria required for the actual survey (graduates who completed their course 15 months ago, across a range of provider types). It was also subject to numerous rounds of user acceptance testing by internal and external colleagues. The questionnaire was finalised in consultation with funding and regulatory bodies. View the cognitive testing technical report and outcomes report.

The survey has been updated for each cohort since it was first launched in December 2018. These changes have focused on minor amendments to routing and changes to text that was deemed necessary to improve data quality and likely to improve response rates. Particular attention has been paid to making sure the changes do not have a material impact on the meaning of the questions and do not bias the survey results in any way or make the data incomparable.

The survey results coding manual is available to users, which contains necessary information and guidance on the data that is returned from the survey.

ONLINE SURVEY DESIGN

The survey questionnaire is hosted on an online platform using specialist survey software. All questions are programmed using the software’s coding language. The system is widely used to conduct surveys by leading sector bodies. It conducts the management of survey contact with graduates both online and for telephone interviewers, meaning there is live interaction between the different channels.

The same online version is accessed by telephone interviewers ensuring the same survey is used across both modes of data collection at any given point in time. This is aimed at reducing survey measurement error. The system also includes sophisticated technology which is instantly smartphone compatible, making the survey more accessible by the target audience.

The online survey is accessed through a URL (link) which is unique to each graduate and sent to them via email or SMS (text message). The survey can be conducted on multiple devices (desktop and mobile) with in-built compatibility functions that enable seamless transfer from one device to another. Respondents are provided with data validation prompts to help them with specific questions as they go through the survey. This minimises the risk of respondent error, particularly in self-administered surveys. The following are examples of questions that use validation checks:

1. Activity – if a respondent indicates they are in employment as well as retired, they see a prompt that requires them to check their answer and correct if necessary.

2. Salary – if a respondent's salary seems too low or too high, based on currency and intensity of employment, they are asked to check their answers. Typical salary ranges are obtained from ONS' Annual Survey of Hours and Earnings[2].

When completing the survey online, respondents are unable to go back and change their answers to previous questions. This is done for data protection reasons. We are unable to confirm whether the contact details submitted by providers are unique to the graduate because, in several cases, the email addresses used to contact the graduate were previously supplied to them for a different purpose. We therefore consider the data protection risks this may cause.

There is a possibility that a graduate has two email addresses, one that is only available to the graduate in question and the other available to another individual. In such instances, personal data entered by a graduate using a link from one of these emails could be accessed by the person with access to the other email address. By removing the 'back' button on the online survey, it is not possible for anyone other than the respondent to have access to information already entered into the survey.

This issue does not affect surveys conducted over the telephone as interviewers confirm the name of the respondent before the survey begins.

PROVIDER PERSONALISATION OF SURVEY AND EMAILS

Graduate Outcomes depends upon strong collaboration with providers. While HESA manages the planning, delivery and data capture elements of the survey centrally, providers fulfil equally important roles in collating and submitting good contact details and publicising the survey to their graduate communities. It is important that graduates understand the official status of the survey and although few will have heard of HESA, most are likely to feel more confident about the credentials of the survey when it is visibly supported by their 'home' providers.

To support this recognition by respondents, we collect providers' logos for co-branding the survey and a link appears at the end of the survey to a relevant area of each provider's website (e.g. their careers service). Email invitations and reminders to complete the survey are sent under the name of the provider (but from the central Graduate Outcomes email address). We also include the name of the provider in interviewers' scripts. All of these approaches are intended to convey the collaborative approach underpinning this survey and reassure graduates about its legitimacy. The principle of survey customisation was agreed with providers during survey design consultations.

To minimise risking the introduction of bias, we ask providers to refrain from any attempts to drive up response rates through direct engagement with graduates during the live survey period. We also strongly discourage the use of incentives for the same reason. Providers can, however, make full use of non-direct channels for promotion, for example social media.

TELEPHONE SURVEY DESIGN

As mentioned previously, our contact centre uses the same survey platform and questionnaire design as used in the online mode. Additionally, they also use a pre-determined script for interviewers to guide them through interviews. This is designed to complement the survey by

[2] https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/bulletins/annualsurveyofhoursandearnings/previousReleases

providing useful prompts that aid interaction between interviewers and respondents. While the survey itself is identical across the two modes (online and telephone), the interviewer script contains additional prompts and guidance for the interviewers to support direct interaction with respondents.

Prior to launch, the survey and platform were rigorously tested by the contact centre. This led to the identification of areas that required further guidance and recommendations from this exercise were incorporated into the original script.

The script is constantly reviewed by HESA, especially considering any issues encountered by interviewers during live interviews. Additional support in the form of guidance, prompts and reassurance texts are offered to mitigate the risk of respondents disengaging part way through the survey.

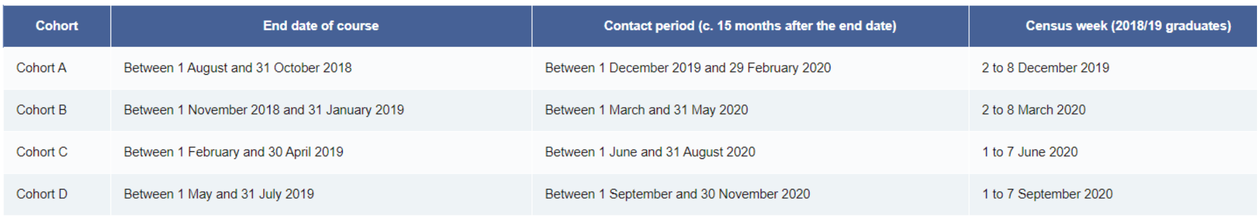
The entire interviewer script is also available in Welsh, to allow interviews to be carried out in that language should a respondent choose to do so. Further information on the handling of Welsh language requirements can be found in the section on data collection.

Read more about our contact centre methodology on our website.

DATA COLLECTION

Graduate Outcomes data, for a given academic year, is collected in four instalments, known as cohorts. Each cohort represents a group of graduates who completed their course during a certain period 15 months prior to start of data collection. Figure 1 outlines the data collection plan for 2018/19 collection year:

Figure 1: Data collection plan for the 18/19 collection



Cohort	End date of course	Contact period (c. 15 months after the end date)	Census week (2018/19 graduates)
Cohort A	Between 1 August and 31 October 2018	Between 1 December 2019 and 29 February 2020	2 to 8 December 2019
Cohort B	Between 1 November 2018 and 31 January 2019	Between 1 March and 31 May 2020	2 to 8 March 2020
Cohort C	Between 1 February and 30 April 2019	Between 1 June and 31 August 2020	1 to 7 June 2020
Cohort D	Between 1 May and 31 July 2019	Between 1 September and 30 November 2020	1 to 7 September 2020

As not all graduates will have access to the internet (or a telephone), the survey adopts a mixed mode design to maximise contact with respondents. The primary modes of data collection in every cohort are web and telephone, with several strategies (outlined below) that look to maximise response rates. Postal surveys are also used for a small number of graduates with no other contact details except a residential address.

The two main modes of data collection interact with each other seamlessly in that respondents starting the survey on one mode could easily finish it on another, without having to start at the beginning. They are also able to access the survey online, multiple times, until they reach the end and submit all of their responses. Respondents can choose not to complete the survey over the phone and in such instances, interviewers can transfer a respondent to the online survey by sending a link to the survey via email instantaneously.

ONLINE DATA COLLECTION

About

Data collection commences at the start of a cohort with an invitation email that is sent to all graduates, using email addresses submitted by providers. The email contains a survey link that is unique to every graduate. This is followed by an SMS (usually the following day but it can take longer for larger cohorts) to UK mobile numbers only. All graduates therefore receive a form of invitation in the first week of data collection. Telephone follow-ups with all non-respondents commence in the second week. Respondents who only partially complete the survey online are given a few weeks to complete it online before they are contacted by telephone.

In year one, providers were asked to submit up to a maximum of 10 email addresses and mobile numbers per graduate. This requirement is being revised for future years in light of evidence that most graduates only have one email address and mobile number and having more contact details does not have a significant impact on response rates. Every single contact detail submitted and approved by providers is used to send emails and SMS messages.

During the entire 13-week field period in each cohort, up to five emails and SMS messages are sent to all non-responding graduates and those partially completing the survey. The exact timing of these reminders varies slightly from one cohort to another and is communicated on the engagement plan which is published for each cohort on our website.

Enhancements

The first year of Graduate Outcomes has seen the implementation of several enhancements during and in between cohorts. The objective of these enhancements has always been the improvement of data quality and/or effectiveness of the data collection instrument which in turn leads to higher response rates. Some of the enhancements include:

- Trialling email and SMS delivery on different days and time of day. Using paradata to inform future deliveries.
- Recognising respondents who may have partially completed the survey, through targeted emails and SMS messages.
- Using SMS messages flexibly as a prompt or to encourage a direct response.

One of the main changes to our data collection strategy is the use of pre-notification or “warm up” emails to prospective respondents, before the start of data collection. This was implemented for the first time in cohort D in the 17/18 collection year. All graduates with approved contact details in this cohort received a pre-notification (warm-up) email at least a week before they received the first invitation. The purpose of this exercise was to improve the take up of online data completion and to ‘warm-up’ our IP addresses, to raise their recognition as legitimate by the information security utilised by the service providers that respondents receive notifications on e.g. gmail and microsoft.

We have taken steps to risk-assess these improvements prior to implementation to minimise any likely impact on bias in the survey. Balancing the potential improvements in response rates and data quality with assessed risk of bias has been a key consideration, but in the case of all improvements implemented, we believe the balance of benefits has been compelling.

View the emails used in the engagement strategy and survey materials on the HESA website.

TELEPHONE DATA COLLECTION

About

Telephone interviewing usually commences in week two of field work. For the larger cohorts, graduates with no email addresses but a valid phone number are called in the first week as that is the only mode of data collection available for them.

Calls are handled using an auto-dialler that randomly selects respondents from the entire sample and connects them to an available interviewer. Depending on the outcome of the call, it is marked as a complete, incomplete or refusal. An incomplete status is further classified into the nature of the call and its outcome, for example, 'no reply', 'busy', 'answer phone' etc. To try and maximise response rates, interviewers are also able to book appointments with respondents if they wish to be contacted on certain days or time of day.

As with email addresses and mobile numbers, a graduate can have up to 10 UK landline and international numbers, although this is being reviewed for future years (see online data collection for more information). All numbers are used to contact respondents and collect a valid response. Once a number has been used to make direct contact with a graduate, it is marked as 'successful' and used in all subsequent attempts. As advised by our contact centre, mobile numbers are likely to be more unique to the graduates, therefore they are used before landline and international numbers.

Geo-dialling

The contact centre operates using a geo-dialling system, whereby the area code of the telephone number displayed to graduates matches that of the location of their university. Graduates are presented with a telephone number that is more familiar to them, increasing the likelihood of them answering a call rather than ignoring or rejecting it as they might from an unknown/unrecognisable number. This approach is supported more generally by existing best practice within the Market Research sector. As well as increasing the likelihood of graduates picking up the phone, it also dilutes the risk of a single number becoming backlisted.

Despite the benefits of a geo-dialling system, the use of phone numbers that are visible but unknown to respondents does increase the likelihood that they will repeatedly ignore or even bar the calls, especially where they are called multiple times from the same number. It was therefore vital to consider any steps that could be taken to reduce this behaviour, with a view to increasing levels of response. Therefore, during the second half of year one, the approach to geo-dialling was further enhanced by changing the telephone numbers used during fieldwork, once or multiple times, whilst retaining the geographical link to the area of each HEP.

Third-party interviewing

During the second half of the field period, interviewers are advised to collect responses from third parties, where possible, and where a suitable proxy respondent (defined as a partner, relative, carer or close friend) is available. Only the mandatory questions are asked and subjective questions are excluded.

Interviewer training and development

To minimise interviewer error, the contact centre undertakes an extensive training exercise to train their interviewers on Graduate Outcomes. HESA worked with them to compile a set of guidance notes and training materials on every question in the survey. The training covers practical, theoretical and technical aspects of the job requirements. For quality control purposes, team leaders provide ongoing support throughout, enhancing interviewer skills and coaching around areas for improvement. This is carried out through top-up sessions, structured de-briefs and shorter knowledge sharing initiatives about “what works”.

For Graduate Outcomes, all interviewers receive a detailed briefing upon commencing interviewing, covering the purpose of the survey, data requirements (for example level of detail needed in certain free-text questions), running through each survey question, and pointing out areas of potential difficulty so objections and questions can be handled appropriately and sensitively.

Making calls and scripting

Interviewers are randomly allocated to respondents by the telephone dialler. This reduces the risk of creating interviewer-respondent clusters based on common characteristics. The only exception to this rule is the employment of Welsh speaking interviewers who are allocated to Welsh speaking respondents only.

Interviewers introduce the Graduate Outcomes survey as the reason for the call and state they are calling on behalf of the provider for the particular graduate. If asked for further information, they will explain that they are from a research agency that has been appointed by HESA to carry out this work. If required, the interviewer can also advise that the survey has been commissioned by the UK higher education funding and regulatory bodies.

All interviews are recorded digitally to keep an accurate record of interviews. A minimum of 5% of each interviewers’ calls are reviewed in full by a team leader. Quality control reviews are all documented using a series of scores. Should an interviewer have below acceptable scores, this will be discussed with them along with the issue raised, an action plan agreed and signed, and their work further quality controlled. Team leaders rigorously check for tone/technique, data quality and conduct around data protection and information security.

Recontacting graduates

Some of the data collected on the survey is coded by an external supplier, using national industry and occupational coding frameworks. Where they are unable to code verbatim responses, these are returned to the contact centre who try and supply more detailed responses by listening back to the interview and where necessary calling the graduate again.

HESA collects regular feedback from interviewers on the handling of different questions and respondents with the aim of identifying survey or script modifications.

POSTAL DATA COLLECTION

A third and final mode of data collection used in Graduate Outcomes is postal. Under exceptional circumstances, where a higher education provider is unable to supply email addresses or phone numbers for graduates, survey questionnaires are sent by post to the residential address supplied by the provider. The number of records with only residential addresses is not permitted to exceed 5% of a provider’s population in a given cohort.

The postal survey is a much shorter questionnaire, containing only a subset of the core survey questions that are required as a minimum to produce the main outputs. This is largely done to keep the survey short and minimise the level of navigation required due to routing. So far, the requirement for postal surveys has been minimal across all cohorts and approximately 10% of recipients have returned a completed questionnaire. Data from completed surveys is manually entered into the system by HESA.

OPT-OUTS

Graduates are able to opt out from the survey and any further communication through a number of different channels. The email invitations and online survey instrument provide direct access to information on how to opt-out. Respondents can contact HESA at any point to request an opt-out or deletion of their survey data or contact details as per their rights under GDPR (this extends to after the survey closes up to a fixed point which is outlined on the privacy notice).

Respondents can also refuse to take part in the survey over the phone and interviewers are trained to handle such requests.

Graduates can also get in touch with their providers to request an opt-out. Such requests are redirected to HESA for a formal action. Respondents who opt-out are marked as such on the survey data collection system and all future communications cease within five working days from receipt of the request.

CASE PRIORITISATION

While achieving a higher response rate can improve the precision of estimates, the impact this will have on bias is ambiguous. The reason for this is that non-response bias depends not only on the level of response, but also the discrepancy between respondent and non-respondent values. As the latter component can continue to widen as more individuals complete the survey, a better response rate will not necessarily solve the problem of bias. It has generally been the case that the post-collection procedure of weighting is applied as a solution to this issue. However, rather than simply relying on this technique on its own, it was concluded that trying to additionally address bias during the data gathering phase could bring supplementary benefits (e.g. less variable weights).

Consequently, for cohort C and D in year one, a case prioritisation approach was introduced (due to competing operational commitments necessary for firmly establishing Graduate Outcomes as a data collection service, case prioritisation could not be introduced until cohort C). This involved developing a response propensity model around halfway through the collection period for a cohort. While the dependent variable was a binary indicator highlighting whether the individual had responded to the survey, the independent variables all related to demographic (e.g. sex, disability, age etc) and course characteristics (e.g. level and mode of study) available through the HESA student record.

Following the creation of the logit model[3], each individual was assigned their predicted probability of responding and these were ranked into order. Among those who hadn't submitted the survey, the quartile with the lowest propensity scores were selected to be given extra priority.

The priority sample was identified on the Computer Assisted Telephone Interviewing (CATI) system and allocated to a group of interviewers for a few weeks, towards the end of field period. This was done to enable a more concentrated effort to contact non-respondents who are least likely to respond to the survey. In theory, this would not only result in more calls per graduate for this group but also a higher response rate than what would be achieved if they were part of the

[3] A statistical technique that is used to investigate the relationship between the probability of an indicator and a few explanatory variables.

main sample. This approach was first used in cohort C with the aim of identifying operational improvements that were subsequently implemented in cohort D and continue in year two of the survey.

WELSH LANGUAGE REQUIREMENTS

HESA is committed to providing access to Graduate Outcomes in Welsh, recognising the importance of ensuring Welsh speakers are not treated disadvantageously in comparison to English speaking graduates. Working alongside the Welsh funding and regulatory body, we have contracted with a partner organisation to undertake all English to Welsh translation work for Graduate Outcomes. This includes the logo, Graduate Outcomes website, the survey, script, results, email and SMS text.

Following feedback from Welsh providers, HESA undertook a review of the approach to communication with graduates and it was agreed to adopt a nuanced approach based on Welsh language proficiency. We now offer all communications in Welsh, English or bilingual modes, depending on a graduate’s ability to speak fluently in Welsh.

DATA PROCESSING

DATA CAPTURE

All data collected across the three modes (online, telephone and postal) are captured in a single location provided by the same software used to administer the survey. Every day, all completed survey results are transferred to HESA’s internal databases. These are processed overnight, ready for dissemination through the provider portal the following day.

Data captured in the internal databases are also used for quality assurance and output production.

DATA QUALITY CHECKING

A series of data quality checks were carried out on the data collected in year one. Most of these checks will also be relevant in future years except those that relate to fixed survey components such as routing and completion logic. The main areas of consideration included:

- Survey completion logic - assessment of the coding of data fields which informed whether the survey was completed or not. Where coding issues were identified, fixes were implemented in a timely manner.
- Survey routing - capturing any errors around survey routing, such as incorrect questions being answered given the activities selected and compulsory questions skipped, allowing graduates to proceed and answer questions from following sections. An example of this occurrence is a small number of cases (less than 50 graduates) that were identified and attributed to the existence of the “back” button in the survey, allowing graduates to go back to earlier questions and delete answers. This issue was virtually eliminated mid-way through cohort A (17/19 collection) and in the following cohorts when this feature was disabled from the online survey. This is only retained in the telephone survey to maintain a good interviewer-respondent relationship. Those completing the survey online can contact HESA by email if they wish to request a change to their survey answers or completely reset the record.

- Free text fields - analysis of the data captured in the free text fields and proposals for future modifications to help improve quality. This included identifying trends in responses from graduates who were unable to select the appropriate response from a drop-down menu and opted to select “other” and complete the free text box.
- Salary - analysis of salaries returned by graduates, including percentage of known salary for those graduates paid in UK pounds and in full-time employment or self-employed or running own business and known salaries split by currency. This also included comparison of minimum, maximum, average (mean and median) and missing values against previously published material in DLHE and other national sources.
- Standard Industrial Classification (SIC) and Standard Occupational Classification (SOC) - analysis and trends found within the free text fields for those graduates for which SIC or SOC could not be coded. Further information on the SIC/SOC coding process is covered below.
- Partial responses - analysis of what could constitute as a sufficient response in order to be included within published material.

SIC/SOC DATA CODING

Where we have received sufficient data (more than one alpha-numeric character in one of the four employment fields) in the employment and/or self-employment sections of the survey, responses are passed on to Oblong, our supplier for coding of Standard Industrial Classifications (SIC) and Standard Occupational Classifications (SOC). Surveys completed in Welsh are first translated and then sent to the coding supplier.

The SOC2020 framework is being adopted for the 18/19 collection.

The fields used for SIC coding are:

- Company Name
- Company Town/City
- Company Postcode
- Country
- Company Description
- Job Title (to help with School/Healthcare classifications)
- Course Title
- Self-Employed or Own Business

The fields used for SOC coding are:

- Company Name
- SIC Code
- Job Title
- Job Duties Description
- Most Important Activity
- Self-Employed
- Own Business
- Portfolio
- Qualification Required
- Course Title
- For business owners, whether they have employees
- Whether they Supervise Others

- Company Description

The Company Description can help in some cases to clarify the SOC code. A combination of the Course Title studied and the Qualification Required question, where appropriate, help to inform and give confidence to the coding.

Ideally, all of the above variables are needed to obtain the most relevant SOC code for a given record. In some instances, it may be possible to obtain a code even when all the information is not provided. However, as previously noted, at least one of the four employment fields must be provided as a minimum.

SIC/SOC coding process

Over the years, our supplier has developed self-learning software to deal with the classification and matching of company data. This software has been re-written and trained to work with HESA data, and utilises fuzzy logic, knows of common typos and uses spelling error algorithms to deal with the free text in the data. The software is underpinned by our supplier's own database of UK companies and uses machine learning on both SIC and SOC from historic data to improve coding. They also employ a dedicated team of manual coders who check all codes and fill gaps where the software could not apply a code.

Our supplier first loads the data into their systems and pre-processes it, tidying it up, addressing common issues and putting it into the right format ready for further automated processing. Each field has its own set of unique pre-processing tasks, which can range from keyword replacement, keyword removal and character substitution.

Next, industry classifications (SIC codes) are automatically added to companies that employ graduates. The manual coding team then complete an initial check of the data and fill the gaps where the system cannot apply a SIC code. The codes are then checked again by a quality control team and amended where necessary.

The data are then automatically SOC coded, and the system uses various methods to apply a SOC code to a record. It looks for keywords in both the job title and job duties fields. The system learns from data that have already been coded (including previous manually SOC coded records), so if it sees a record with similar details to one that was seen before, it can be assigned the same SOC code.

Of the responses collected through telephone interviewing, any uncodable records identified by Oblong are sent back to IFF for a follow-up interview where there is a reasonable case for going back to the respondents.

Supplier-led quality assurance process

The SOC codes are manually reviewed, and the gaps filled where the automated systems could not apply a code. All records are then sent to the SOC quality checking team to be checked before being released back to HESA.

The manual coders are in constant contact with each other and the quality team, and any new/different occupations encountered are discussed with the quality team, who will then research an occupation if necessary, or discuss with HESA or the ONS if required.

For most of the job titles, the coding index (list of job titles in the SOC framework) contain the job titles and records can be coded from them. Where the job title is not in the indexes detail in the job duties is used to ascertain what the job involves and code accordingly. Due to the international element of the data, jobs which do not appear in the indexes are also encountered. Coders are adept at assessing the job duties and placing the job with the appropriate code, and this is all subsequently checked by a quality checker. If a coder still cannot code then they raise a query with the quality checkers, who will discuss with other team members, research the role if necessary, and advise on coding. Research is done online using reputable sources (for example the company website where the person works, NHS websites, large well-known job sites, where one can see what qualifications are required and what a job involves). Where appropriate, the documentation which the coders use is subsequently amended for future reference.

Doing this exercise over multiple years, and given the volume of data, allows Oblong to refresh their databases with new jobs that did not exist before. When new jobs are encountered, a decision is made on an appropriate code and this information is disseminated to all coders via their coding indexes for future reference.

A final consistency check is completed at the end of each cohort and for many records a final data consistency/quality review takes place at the end of the collection. This involves consistency checks across employers, job titles and all cohorts to make sure no single cohort within the collection looks different to the rest. By the end of the process, every SOC and SIC code will have been manually checked multiple times.

At the end of year one (17/18) data collection, providers had the opportunity to review their data including the draft SOC (occupation) coding and submit feedback to HESA. All of the provider feedback received was individually reviewed and tracked. Outcomes from this review have been published on the website, alongside a description of the review process itself and next steps. The provider feedback allowed Oblong to correct and improve on coding for year one, and the learning has been fed into the systems to enhance future SOC coding.

With the introduction of SOC2020 for year two, our supplier has taken the opportunity to review all logic and associated reference data (over 50,000 sets of keywords and SOC associations) within the SOC coding automated systems, to ensure that provider feedback has been embedded in the software, and has also refined and added to the guidance documentation used to manually classify the responses. The manual coders have been retrained on the new SOC2020 taxonomy, and also continue to be re-briefed ongoing following changes based on provider feedback.

Oblong also provide a standardised company name, improved business postcode, Companies House registration numbers and employee size information in the final, returned data, in order to aid analysis.

For more information about the coding process for year two (18/19), please visit the operational survey information page.

Usage of salary
Graduate Outcomes collects data on annual salary from all respondents in employment. We have reviewed the case for using this data in the coding process.

Salary is an optional question as respondents can skip it to move onto the next question. As such, there is a level of missing data for this question. Furthermore, salary is one of the sensitive questions in any personal survey and is not likely to yield accurate responses all the time. In the absence of a mechanism to validate this information, using administrative data for example, we are unable to comment on the accuracy of data in this field.

The use of salary data in SOC coding poses unique challenges. Given the geographic, industry and demographic variations in earnings, it is not possible to identify a set of principles that could be applied to the coding of different occupations, consistently for all graduates. For example, there is not a single rule of thumb that dictates the salary of people working in highly skilled roles across all industries and sectors. The case of part time roles and those working under other forms of contractual arrangements makes this task even more difficult. A system that is based on a set of standard principles that are consistently applied across thousands of records using a combination of automated and manual tools cannot accommodate a variable with a high degree of variance. After careful consideration it has been decided that salary will not be utilised in SOC coding of Graduate Outcomes data. HESA will, however, consider its use in the quality assurance of coded data.

FREE TEXT FIELD CLEANING

At the end of the collection process, data returned for questions that permit a free-text response goes through a cleansing process, in order to improve data quality. This is usually where the respondent has not chosen a value from the drop-down list provided but has instead selected “other” and typed their own answer. This process also runs for questions seeking postcode, city/area and country of employment, or self-employment / running own business; country in which graduate is living and of further study; provider of further study, and salary currency. Where possible, the free text maps to an appropriate value in the drop-down menu or the appropriate country or region.

DERIVED FIELDS

Further aggregation of some key fields is carried out to produce standard derived breakdowns used across HESA’s published material. Key areas of derivation include minimum response for inclusion in publication, method of response, activity (including most important activity), location of activity; grouping of standard industrial classification (SIC), standard occupational classification (SOC) and salary; employment and study undertaken after graduate and prior to survey activity. Details of these derivations will be published within the survey results coding manual.

DATA ANALYSIS

APPROACH TO WEIGHTING

In summary
HESA has agreed with the Graduate Outcomes Steering Group that weighting will **not** be applied to all statistics published by HESA for this first year (17/18) of survey data. Our analysis of the survey data has not identified any evidence of bias relating to mis-match between the achieved sample and graduate population characteristics in any direction at sector level. Indeed across a range of demographic and course variables, we see a high level of similarity between the sample and population distributions.

The weighting approaches we have developed as part of this analysis have shown little if any divergence between weighted and unweighted estimates at sector level. At a more granular level, some effects have been seen when applying particular weighting methodologies, especially for small sample sizes (e.g. statistics disaggregated by provider and subject with very small numbers of responses) but these are subject to high variability. When considered together, we have been unable to determine any weighting approach which consistently and materially improves the quality of estimates.

The following assessment further explains our conclusion on this issue. For those with greater expertise in this field, a more comprehensive technical description of the analysis and conclusions on the weighting methodology will be published shortly.

Background

As described in previous sections, Graduate Outcomes aims to survey all (with a small number of necessary exceptions) individuals who qualified from higher education in each academic year. With participation being voluntary, non-response is recognised as one of the factors that could impact on the quality of the collected data, both in terms of potential non-response bias and precision of the resulting estimates. By non-response bias, we mean that any estimates generated from the sample will not accurately reflect the outcomes of the wider population. In Graduate Outcomes, this is likely to occur if the composition of the sample differs to that of the population.

Greater precision meanwhile would mean that we can be confident that an estimate we derive from our achieved sample of respondents, with a small margin for error, fairly reflects the true statistic for the whole population. For example, if we could estimate that the whole population percentage in employment for a given university was 81.5% (with a margin of error plus or minus 0.5%) based on the sample of responses, that would be a relatively precise estimate. If, on the other hand, that 81.5% was subject to plus or minus 10% then it would not be precise.

Higher response rates provide the advantage of generating more precise estimates; this can be interpreted (loosely) as more ‘reliable’ estimates at granular levels such as by HE provider and subject, mainly because higher response rates provide greater numbers of graduate responses at these levels. Non-response bias, however, is determined by a combination of the response rate and the difference between respondents and non-respondents for any given statistic of interest. Consequently, a larger response rate does not always guarantee a reduction in non-response bias, as it is possible that it is the most hardened of non-respondents who are most different from those who respond. It is entirely feasible for unbiased statistics to be derived from survey data based on relatively low response rates if appropriate survey design and operation have been deployed and, where required, approaches such as weighting have been applied.

As previously described, as part of the response-chasing operation, HESA has utilised a case prioritisation process to try to balance response rates across a range of groups. This technique involves identifying those considered least likely to respond and giving such individuals a higher priority as part of our engagement strategy in the latter stages of the field work. Such an approach aims to mitigate possible bias resulting from non-response, rather than simply ensuring high response rates are achieved.

Use of weighting

One of the techniques commonly deployed in surveys post-collection is the use of weighting. This involves the use of ‘scaling factors’ (e.g. a factor of 0.75 applied to a response would reduce its relative weight) applied to each survey response in an attempt to make the sample more representative of the population. Weighting seeks to mitigate the impact of non-response bias and, under certain conditions, can also improve the precision of estimates.

Survey weighting is almost always used when the survey is designed around a ‘structured sample’ (a specific subset of a population designed to conform to certain characteristics) but Graduate Outcomes is not designed in this way – it is a census survey. Even with a census survey, once the resulting responses are analysed, they can be found to show materially different characteristics from the population and the application of weighting can ‘correct’ for this imbalance.

HESA analysts have worked in collaboration with analysts from the Office for Students and with advice from experts at the Office for National Statistics to undertake extensive analysis of the first year of Graduate Outcomes survey data (for academic year 2017/18). This work has focused on assessing the extent to which the achieved sample for the survey shows similar characteristics to the population of all graduates, deriving and implementing a number of different statistical models for weighting and then testing to assess the impact of each weighting model through comparing weighted and unweighted data.

It is important to note here that HESA holds data on the population of graduates through the HESA Student Record and associated census records for HE taught in Further Education Colleges, so it is possible to compare demographics, study, qualifications and HE provider characteristics between the achieved sample of respondents in the survey and the entire population. HESA does not hold population data on outcome characteristics, such as nature of employment or other outcome activities (though future work might provide some insight into this missing element, such as use of Longitudinal Educational Outcomes (LEO) data). Consequently, we have only been able to make inferences about bias using the data that we hold. We cannot definitively know anything about the responses to the survey that would have been provided by those who chose not to respond (without identifying alternative sources of these data).

Our findings

Notwithstanding the above caveat on our analysis, our findings are that there was little observed difference between the achieved sample and the population across a range of demographic, study, qualification and HE provider characteristics that were examined. Having applied a variety of different weighting models, the weighted and unweighted estimates (such as percentages of those in employment or study) were very similar and this also often applied at sub-sample level too (for example, by subject and/or provider). The largest differences we observed between weighted and unweighted estimates were most commonly found in instances of small sample sizes, which are estimated less precisely in any case. We note that, in general, weighting led to less precise estimates than unweighted data.

The above findings have led to HESA agreeing with the Graduate Outcomes Steering Group that weighting will not be applied to all statistics published by HESA for this first year of survey data.

The position regarding use of weighting in future years of the survey remains under review. HESA is planning some additional exploration of more nuanced approaches to weighting through the remainder of 2020 (and early 2021), which will utilise data from the second year of the survey once available. If an approach to weighting can be identified at that time that can be shown to improve the quality of statistics derived from the surveys, then this will be applied from 2018/19 Graduate Outcomes data onwards and will also be retrospectively applied to key data outputs from the 2017/18 survey to enable comparisons across years.

RELIABILITY

Some statistics published from the Graduate Outcomes survey will be at a very granular level, e.g. employment rates by HE provider and subject. In some cases, the sample of respondents for such statistics may be small and/or the response rate for that sample may be lower than the overall survey response rate. In these cases, the statistics may be subject to high levels of variability and a lack of statistical precision. HESA intends to publish confidence intervals on these statistics (ranges within which we have a high level of confidence that the equivalent whole-population statistic would fall, where a narrow range indicates greater precision and a wide range indicates less precision).

In addition, for some statistics, it may be necessary to introduce publication thresholds whereby statistics based on very small sample sizes and/or lower response rates are suppressed. The actual decisions on use of these techniques will be clearly explained in each HESA statistical release.

DISSEMINATION

A data dissemination policy provides a high-level description of the range of statistical outputs as well as outlining the key aspects HESA of policy and practice in publishing and disseminating Graduate Outcomes survey data.

View the Graduate Outcomes dissemination policy

SECTOR ENGAGEMENT

COHORT REPORTS AND REGULAR COMMUNICATIONS

As part of our communications strategy, HESA releases regular communications to providers to ensure important operational details about the survey are shared, as well as a range of best practise and additional support. This is to ensure the sector learns as we learn. These communications are sent via email to the appropriate provider representatives and frequently shared on the Graduate Outcomes Jiscmail forum.

In addition to this, and to share insight with the sector as a whole, formal reviews are also created and shared at key survey milestones. At present, these are at the mid-point and end of each cohort. They summarise key operational information from the cohort in review and set out the changes made for the following cohort. As the sector has an appetite for additional statistics, each end of cohort review (for at least the first collection) includes an infographic which shares both response rates and engagement statistics in a graphical way. These reviews are added to the HESA website for anyone interested in Graduate Outcomes.

DATA DISSEMINATION COMMUNICATIONS APPROACH

HESA has created a data dissemination communication plan which encompasses all potential users of the data. The plan contains information to be communicated in the build up to release in order to build understanding and awareness of the distinctive characteristics of the Graduate

Outcomes survey and the methodology that has been employed. Content includes blogs and news items from key HESA experts and clear guidance on the releases themselves.

As part of the delivery of this plan, we will seek support from key sector agencies in reinforcing key messages.

Like the cohort reports, these key communications will be disseminated to relevant IDS roles, shared on the Jiscmail, added to the HESA website, shared (where relevant via HESA social channels (twitter and LinkedIn) and in the weekly update.

EVALUATION

Continuous evaluation and improvement are among the main features of the design and delivery of Graduate Outcomes, as evidenced throughout this methodology statement and in our regular communications with providers. We are keen to build on the foundations from the first year and enhance our data collection, processing and dissemination systems as we move into the next year and beyond. This is partly being achieved through a series of post-implementation reviews with our key stakeholders, including suppliers.

The following are a number of potential survey enhancements that HESA will consider for future collections of the survey. These will be described once they have been fully explored in future iterations of this methodology statement:

DATA COLLECTION

- Understand what encourages individuals to engage with emails and SMS messages and use it to inform the content of our communication / graduate engagement materials.
- Enhance the survey experience for users of mobile devices, considering recent technological developments and increasing use of such devices.
- Explore the use of web survey design features such as progress bars and information buttons.
- Explore the use of a single number for telephone interviews versus geo-referenced phone numbers to increase uptake of telephone calls.
- Assess the costs and benefits of various approaches to incentives.
- Investigate and make recommendations for closer alignment with UK and international labour market data standards.

DATA QUALITY ANALYSIS

- Measure the quality of graduate contact details supplied by providers and its impact on response rates.
- Identify factors contributing to variability in response rates and, where possible, determine ways of reducing it.
- Determine the extent of measurement error introduced by questionnaire design, mode effects, interviewer bias and respondent bias.
- Identify ways of improving the quality of free text responses.
- Obtain access to raw LEO data, in order to quality assure collected salary data, and to contribute to our understanding of non-response bias in the survey.

DATA PROCESSING AND DISSEMINATION

- Solicit feedback from users on our first Statistical Bulletin and Open Data release, to help HESA refine and develop our outputs.

- Undertake an investigation into linking Graduate Outcomes data to subsequent years of the HESA student record, in order to quality assure further study outcomes data, and enhance our understanding of undergraduate to postgraduate transitions.
- Undertake further research into weighting methodologies incorporating the second year of survey results (18/19), when available, to ascertain whether more nuanced approaches can be identified that improve quality of statistics derived from the survey.

HESA
95 Promenade
Cheltenham
GL50 1HZ
E Liaison@hesa.ac.uk
T +44 (0) 1242 211 144
W www.hesa.ac.uk