OVERVIEW OF AGS DATA PROCESSING AT GCA

About this document

This document details procedures followed at GCA for processing Australian Graduate Survey (AGS) responses and converting these to data files, focusing on quality control processes. Data processing at GCA includes three main types of data: (1) hard-copy forms, (2) data files downloaded from online instruments, and (3) data files which are the result of processing procedures at institutions.

Auditing is used to try and make the data as accurate as possible by eliminating mistakes as they arise, and also by recognizing consistent errors and correcting procedures to stop those errors from happening in the future.

PRIVACY AND CONFIDENTIAILITY

All forms and data received by GCA are treated as confidential and as such are handled only by GCA full-time staff or, in the case of forms and online response data files, temporary staff who are required to sign a confidentiality agreement before doing so (refer to Apprendix A).

RECEIPT OF HARD-COPY FORMS AND DATA FILES AT GCA

A consignment is the total delivery from one institution on a particular day. Consignments are prioritised for processing by their time of arrival at GCA.

When a consignment arrives:

- It is assigned a tracking number and logged in GCA records
- The *packing slip* (refer to Appendix B) is checked against the details of the consignment (e.g. number of October GDS+CEQ forms, partially or full coded, etc.)
- Exact requirements for processing are ascertained and logged, and in the case of forms, recorded on a cover sheet attached to each batch of approximately 100 forms
- If there is a discrepancy between the details on the packing slip and the details recorded at GCA, the relevant Survey Manager will be notified immediately.

RESPONSES FROM AN EARLIER COHORT

From time to time, GCA receives responses where it is apparent that the respondent completed prior to the appropriate time for the current survey year, and is not part of the relevant survey populatioin. These cases will be excluded from processing but retained and returned to the institution with the consignment and identified in a covering note.

CHECKING INSTITUTIONAL CODING

- If fields have been coded at the institution, a sample of cases from each institution will be checked for presence of codes for each of the following variables: maj1-4, permnat, ceqmaj1-2, industry, empnat2, duties, furinst and furmaj1-2.
- A sample of forms will also be checked for coding accuracy and coding consistency, for the following variables: maj1, furinst, furmaj1, industry, empnat2 and duties. Maj1 provides an indictor of the quality of ASCED coding of major fields of education offered internally in the variables maj1-4 and ceqmaj1-2, while furmaj1 indicates the quality of ASCED coding of a mix of internally offered majors and majors offered by other institutions, in maj2 as well.
- Persistent or frequent errors will result in further checking. If, after checking, the error rate remains at an unacceptable level, the institution will be notified immediately. Each code will then be checked and corrected where necessary at GCA, and this will result in an additional processing charge to cover the costs of the work involved.

PROCESSING STAFF AT GCA

- All processing staff working at GCA undergo training, regardless of experience.
- Any errors identified during processing are recorded in a log and linked to a specific individual and institution, with details about the nature of the error recorded (whether this was a coding error, data entry error, what the error was etc). Cases which have not been covered with the staff previously are discussed with all relevant staff and notes are provided as a permanent record.

CODING

- Coders at GCA specialise in coding one of two sets of codes only: (1) ASCED-coded variables, furinst (E306), and SACC-coded variables or (2) industry (ANZSIC) and duties (ANZSCO) variables.
- Training includes:
    - an overview of the importance of the work and the end use of the data
    - an introduction to the form
    - an introduction to hierarchical code frames
    - the principals underlying the coding of specific AGS variables
    - the meaning and importance of accuracy and consistency in coding
    - how to self-verify a choice of code
    - an overview of the checking processes in place
    - best practice in number writing
    - the very few cases where respondent error may be corrected on hard-copy forms (e.g. when a response mark is outside the relevant box)
    - practice coding sessions.
- Practice coding involves all coders independently classifying the same raw responses (within their respective area of coding specialisation). Responses are then checked for consistency and accuracy, and the basis of each choice is discussed. Coding errors are discussed in detail before a second round of practice and checking ensues. The sample of responses used for practice includes those chosen because they frequently occur, others so that examples from all broad areas of the code frames are covered, and others because they are known to be difficult.
- Approximately the first 100 cases handled by each coder are independently scrutinised by two permanent GCA staff who then compare results.

CODING QUALITY CONTROL AT GCA

- A sample from each batch of cases (approximately 100) coded are auditied by GCA permanent staff or designated senior coders who have been recognised for a high level of performance. A sample of these audited forms are further checked by another senior coder and any discrepancies are resolved among GCA permanent staff and senior coders, before covering these cases with remaining coding staff.
- Any errors identified during coding are added to a record linked to the specific coder and institution, with details about the nature of the error recorded. The level of error with regards to the level of detail in the code is recorded to determine how inaccuracte the code selected was. For example, for a given code of 083612, if the code deemed correct by the auditor is 083611 (i.e. only the last digit is wrong) this will be recorded as a level-6 error. In contrast, if the code deemed correct by the auditor is 183611 (i.e. the first digit is wrong) this will be recorded as a level-1 error. Cases which have not been covered with the coding staff already are discussed with all coding staff and notes are provided to coding staff. The error record is used to feed back into both internal and institutional training and coding documentation for the following year, as well as recording the error rate and which errors are most common.
- After data has been coded (and in the case of forms, scanned and verified), each value in coded variables is checked against a relevant list of valid codes. These lists are either the full-set of the relevant codes or, where available in the case of maj1-4 and ceqmaj1-2, sub-sets of these lists which are relevant to a particular institution. In the case of any invlaid codes identified, the original response can then be retrieved and the value changed to the correct code.

DATA ENTRY FOR HARD-COPY FORMS AT GCA

The AGS questionnaires (forms), are scannable documents. This means that once completed they are able to be run through a document scanner to collect the data.

At Graduate Careers Australia we use a software package called Teleform to capture scannable data.

The scanner essentially takes a photograph (image) of the form and then the document reader module of Teleform converts the image to data. The data is created using a code frame that has been programmed into the document reader.

The scanning audit process can be broken into four parts, they are: audit whilst scanning, audit whilst reading, verification and the post scanning audit.

Audit whilst scanning

Whilst scanning the AGS forms, any forms that are not identified correctly by the scanner are picked up by a couple of methods.

The scanner uses recognition marks and the bar code to identify each form. The recognition marks are the squares in the corner of each page. If the recognition marks or bar codes are in any way damaged or altered, e.g. torn off, folded, scribbled on, then the reader will not recognize the form.

Whilst the forms are scanning the image of the form comes up for review on the monitor. The scanning operator watches the forms scan on screen so that they can tell if there are any forms that have not gone through the scanner correctly (e.g. a form may be stuck to another form, a form may not be laid flat, or forms may have obstructions over the alignment boxes).

If only one side of a form is unable to be read it will warn that there is a page missing. The scanning operator must then determine where the page is missing, correct the error (e.g. white out scribble on the barcode) and rescan the whole batch. Any forms that are not scanned correctly will be rescanned.

Audit whilst reading

Following scanning, the forms are 'read' by Teleform. Following the reading, any other forms that have been unreadable will be highlighted as 'non-forms' any non-forms must be rescanned once they have been identified and corrected if possible.

Verification

Verification is perhaps the most important step in auditing scanned data. Once a batch is read successfully, the verification process begins. Each form is scanned and read by Teleforms software, before a digital image of the form is presented to a keyboard operator who may modify the interpretation Teleforms has made of the form. The Verifier operator may also review the hard copy form itself if necessary, and then correct as required.

The verifier module in Teleform will automatically bring up any data that the reader has either misread or is undecided about what a piece of data may be. This may be for a variety of reasons but the most common are because of poor handwriting, or because a box is not completely ticked.

- Teleforms software converts selection marks, digits (and characters in specified office use only areas) into values in a data file. Cases where the system can not verifiy whether a legitimate response has been provided (for example, where there is very little ink in the relevant response option tick box) are brought to the attention of a keyboard operator who will view an image of the actual response before inputing the correct value. The system has been adjusted to very conservative settings, so that the vast majority of cases which are presented to a keyboard operator have actually been correctly interpreted by the Teleforms system.
- Cases were multiple responses have been entered for a single response variable, are presented to a keyboard operator for verification of which response, if either, is valid.
- Cases where no value is identified in the key variables maj1, level, sex, permres, and working, are always presented to a keyboard operator.
- Teleforms presents important text responses such as employer name, occupation title, contact details and long-term email to a keyboard operator for key punching.
- The coded variables that use large code frames (e.g. ANZSCO, ANZSIC) may be brought up for verification if the response does not match a valid code.

- 'Verifiers' are given instruction on using Teleforms verifier software, and on standard procedures for dealing with various cases of respondent error (e.g. deleting CEQ responses where the respondent has selected two values for the same item, but not indicated one of these is an error).
- Verifiers are closely supervised during the handling of their first approximately 200 forms, where after close supervision may be extended as required.
- Ten per cent of cases from the resulting data are checked for errors, and the errors are recorded in a log, along with information about the source and nature of the error.

Post scanning audit

Once the data is committed to the final data file a post scan audit is conducted on 10% of all forms.

Each batch is reviewed and 10% of the forms are randomly selected to be audited.

The auditing process is a manual task which requires the auditor to check every variable on the hard copy form (or an image of this) against every column in the output data file. Images of both sides of the form are captured as GIF files and saved with the GCAID.

Any errors that are found are recorded and then corrected by the auditor.

CHECKING COMBINED FORM AND ONLINE DATA FILES

Once an institution's form and online data has been assembled into a single data file, the file is checked for whether the number of October CEQ+GDS cases, October PREQ+GDS cases, April CEQ+GDS cases, and April PREQ+GDS cases match what was received.

Each value in coded variables is checked against a relevant list of valid codes. These lists are either the full-set of the relevant codes or, where available in the case of maj1-4 and ceqmaj1-2, sub-sets of these lists which are relevant to a particular institution. In the case of any invalid codes identified, the original response can then be retrieved and the value changed to the correct code.

PROCESSING DATA IN SPSS

1. Institutional data files are saved as a tab delimited file which helps to find characters that are invalid in SPSS. An increasing problem with data files that come from online forms is that some respondents copy and paste notable amounts of text into text response fields (e.g. their occupational description - *dutyraw*) often adding line returns, bullets, etc. as a result. These can be read into an Excel cell but cannot be read by SPSS and need to be cleaned out. This is a problem whether the file comes from the oAGS or an institution's own online form and can add some time to the processing of an institution's data.

2. Data is then read into SPSS and saved as an SPSS file using the SPSS syntax file <**1 200X AGS read data.SPS**>. This reads the data into an SPSS file, allocating variable names and value labels.

3. Data is then cleaned and a series of frequency tables are produced for all variables using the file <**2 200X AGS clean data.SPS**>. This cleans mis-codes and out-of-range codes for variables and assigns missing values. The output file is a set of cleaned frequencies that forms one of the files that goes back to Survey Managers for checking. Survey Managers are asked to check that their frequencies do not indicate any notable or systematic errors in their data file.

4. Cleaned frequency tables for KEY variables are then produced using the file <**3 200X AGS key variable check.SPS**>. The output file is a set of cleaned key frequencies that forms one of the files that goes back to Survey Managers for checking. This file is intended to make it simpler for survey managers to check important variables that will affect data quality.

5. Cleaned frequency tables for CEQ items are then produced using the file **<4 200X AGS CEQPREQ variable check.SPS>.** The output file is a set of diagnostic tables to allow Survey Managers to checking their CEQ and PREQ response numbers. Again, this file is intended to make it simpler for survey managers to check important variables that will affect data quality.

6. Initial Table 200X A is then produced using file <**5 200X AGS Table 200X A INITIAL.SPS**> and initial Table 200X B using file <**6 200X AGS Table 200X B INITIAL.SPS**>. These are the initial runs of the main destination table sets and are also sent to Survey managers so that they can check that the figures are as expected and can be related to the previous year's figures (which are also sent to them).

7. The output files are sent to Survey Managers for checking and confirmation or changes. Any problems noted with the data are fixed before repeating steps 1 through 7. If at this time there are no problems noted, step 8 is then undertaken. At this time GCA also works through a list of checks for the institution's current output and for some broad comparisons against the institution's previous year's output.

8. A final Table 200X A is then produced using file <**5 200X AGS Table 200X A FINAL.SPS**> *(output file <inst_year_FINAL_AGS_Table_A>)* and a final Table 200X B using file <**6 200X AGS Table 200X B FINAL.SPS**> *(output file <inst_year_FINAL_AGS_Table_B>).*

9. These are then sent to Survey Managers with the cleaned data file and a response rate summary.

10. At this point, hardcopy forms can be returned to an institution (if applicable).