# Notes on the Large Data Set OCR Mathematics B (MEI) (H630/H640)

**LDS6 used for H630/02 in June 2023 and H640/02 in June 2024**

These notes outline the requirements for OCR’s large data set for specification B(MEI), include some notes on interesting features of the large data set and some links to some useful resources.

Please note that the approach outlined here and the suggested activities are suggestions only; you are free to deliver this in any way that suits you, your students and your setting.

If you have any comments on the contents of this guide, including suggestions for other activities, please get in touch with the Mathematics Subject Advisors at Maths@ocr.org.uk .

In the ‘Overview’ section we take a look at the *requirements* in terms of teaching from the Department for Education and in terms of assessment from Ofqual, and introduce one of the large data sets chosen for OCR’s Maths B (MEI) specification. We also suggest some software you will find useful when working with the LDS.

In the ‘Key features’ section we will take a look at some important aspects of the structure and set up of the data. We will also explore some aspects of data cleaning in this section, including a list of things to look out for in the way that the data were presented by the London Datastore, along with some ways to explore these issues and to set up your students to stumble across them.

In ‘Online resources’ we highlight some useful places to go for more contextual information, maps, more data, tools and so on.

You can find teaching activities using the LDS within MEI’s Integral resources <https://integralmaths.org/statistics.php>

***DISCLAIMER***

This resource was designed using the most up to date information from the specification at the time it was published. Specifications are updated over time, which means there may be contradictions between the resource and the specification, therefore please use the information on the latest specification at all times.If you do notice a discrepancy please contact us on the following email address: resources.feedback@ocr.org.uk

### Contents

[Notes on the Large Data Set OCR Mathematics B (MEI) (H630/H640) 1](#_Toc66531256)

[Contents 2](#_Toc66531257)

[1 Overview 3](#_Toc66531258)

[1.1 Introduction 3](#_Toc66531259)

[1.2 MEI Large Data Set 6 (for AS in June 2023 and A level in June 2024) 4](#_Toc66531260)

[1.3 Why three large data sets for the MEI specification? 5](#_Toc66531261)

[2 Key features 7](#_Toc66531262)

[2.1 Population 7](#_Toc66531263)

[2.2 Data cleaning and manipulation 7](#_Toc66531264)

[3 Online resources 12](#_Toc66531265)

[3.1 Useful additional information 12](#_Toc66531266)

[3.2 Other useful websites 12](#_Toc66531267)

# 1 Overview

### 1.1 Introduction

All AS and A Level Mathematics for first teaching from September 2017 include the requirement to work with one or more given large data sets (LDS). The criteria from the Department from Education says that

AS and A Level Mathematics specification must require students to:

* become familiar with one or more specific large data set(s) in advance of the final assessment (the data must be real and sufficiently rich to enable the concepts and skills of data presentation and interpretation in the specification to be explored)
* use technology such as spreadsheets or specialist statitstical packages to explore the data set(s)
* interpret real data presented in summary or graphical form
* use data to investigate questions arising in real contexts.

Specifications should require students to explore the data set(s) and associated contexts, during their course of study to enable them to perform tasks that assume familiarity with the contexts, the main features of the data and the ways in which technology can help explore the data. Specifications should also require students to deomnstrate the ability to analyse a subset of features of the data using a calculator with standard statistical functions.

The prerelease LDS is primarily a resource for the classroom, to encourage the use of real, LDSs when learning statistics. *Some* questions in the assessment, on the statistics content, will be set in the context of the LDS, in such a way as to provide an advantage to students who have spent time exploring the data. However, this is only a small part of the exam and students will not have access to the LDS, or to a computer, in the assessment. The focus of this resource is very much on teaching and learning.

Note that the first bullet point above includes the phrase ‘to enable the concepts and skills of data presentation and interpretation in the specification to be explored’. This refers to that particular section of the content document, i.e. Section D of the OCR Mathematics B specification. You are welcome to explore the full range of concepts and skills, but the focus of the requirement, and therefore of the assessment, is on data presentation and interpretation.

### 1.2 MEI Large Data Set 6 (for AS in June 2023 and A level in June 2024)

OCR’s MEI Large Data Set 6 consists of data about individuals who took part in the American National Health and Nutrition Examination Survey (NHANES) in 2003-4.

This was originally published as LDS\_3 and used for the AS Mathematics B (MEI) – H630 Autumn 2020 paper H630/02, and A Level Mathematics B (MEI) – H640 Autumn 2021 paper H640/02.

The data were downloaded on 27/03/17 from http://www.eeps.com/zoo/nhanes/source/choose.php Further data are available from that source.

The National Health and Nutrition Examination Survey has been taking place since the 1960s in the United States. It combines interview and physical examination to assess the health and nutritional status of adults and children in the United States. More information about the survey can be found at http://www.cdc.gov/nchs/nhanes/about\_nhanes.htm .

The data set includes an information sheet which describes the various terms used. Rather than repeat this information here, please refer to that sheet. The terminology and information on the metadata sheet of the LDS is part of the data set, so students are assumed to be at least familiar with that terminology and information. That’s part of the point of putting it there, i.e. that you then know that these are the words students should understand. *However*, the assessment is not a test of memorisation of the details of the data, only that students have worked with the data enough to have some familiarity with the key features. We will take a more in-depth look at *some* features in section 2.

Please note that further commentary on specific aspects of the LDS can be found within the resources on the OCR website for this qualification, on the cdc.gov website and in resources in Integral.

###

### 1.3 Why three large data sets for the MEI specification?

The large data sets associated with AS and A Levels in Mathematics should serve two purposes: they are a teaching resource and they provide a context for setting examination questions. Our hope is that teachers will use all three for teaching, but for each cohort of students just one will be the focus of some of the questions in the exam. Each data set will be clearly labelled as to when it is used.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | June 2020\* | June 2021\* | June 2022 | June 2023 | June 2024 | June 2025 |
| **AS** | 3  | 4  | 5  | 6  | 7 |  8 |
| **A Level** | 2  | 3  | 4  | 5  | 6 |  7 |

\* Assessments held in Autumn.

So if you teach A Level Maths over two years, then the class you start teaching in September 2022 will see some questions on LDS\_6 in their AS exams in 2023 (if they sit AS) and their A Level exams in 2024, as the following table demonstrates.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Publish** | June 2017  | 2019  | 2020  | 2021 | 2022 |
| **Start Teaching** | Sept 2019  | 2020  | 2021  | 2022 | 2023 |
| **AS Exam (if sat)** | June 2020  | 2021  | 2022  | 2023 | 2024 |
| **A Level Exam** | June 2021  | 2022  | 2023  | 2024 | 2025 |
| **Large Data Set** | 3  | 4  | 5  | 6 | 7 |

MEI and OCR have some experience of pre-release data for [Core Maths B – H869](https://www.ocr.org.uk/qualifications/core-maths/b-mei-level-3-certificate-h869/). The CIA World Factbook data set that forms the current pre-release for that qualification became the basis for our thinking and development for the reformed AS and A Level (LDS 1). We tried to write different types of questions using that data set, based on A Level content. When doing this, we realised that things in some countries have changed quite a lot during the lifetime of the legacy mathematics specifications so the data set would need to be updated from time to time - we didn’t want students learning about how things used to be in the world 15 years ago if that no longer reflected the current position.

We were aware that some students (and maybe teachers) did not enjoy the statistics in the legacy Mathematics A Levels. We think that this may be because in mathematics the focus has traditionally been on learning statistical techniques without much focus on why you might want to use them. The large data sets provide a context to use the techniques and interpret the results.

The use of large data sets in teaching and examining A Level Mathematics is new – it is an opportunity to make the statistics students learn more similar to the ways they will use statistics in future study and work. We thought it was important to review the data sets used and to make sure they continued to be suitable for examining. This needs a three-year cycle – two years for using the data set in teaching and a year to review and update if necessary. LDS 4 ia a refreshment of the data from LDS 1. LDS 5 may be a refreshment of the data from LDS 2, or a new data set dependent upon the post assessment review of the questions set in the live assessment. The information in LDS 3 does not vary as much with time as the other two sets so the decision has been taken to republish this set as LDS 6. The decision to replace or refresh the LDS is also dependent upon the post-assessment review of the questions set in the live assessment so LDS 7 may be a refreshment or replacement of LDS 1/ LDS 4.The data in the CIA World Factbook is grouped by country; we realised that data based on individuals would allow better teaching of distributions. There aren’t many publicly available data sets which contain ungrouped data on individuals. The NHANES data set, from American health surveys, is often used in statistics courses and it contains a wealth of data so we decided to use that as one data set.

Having got data about countries and data about (American) individuals, we thought it would be good to have some England-based data – the London Datastore is a good place to find suitable data and so we ended up with the following three data sets which we hope will appeal to students with different interests in terms of other subjects they are taking.

* [LDS\_4 Data about countries](https://www.ocr.org.uk/Images/400305-large-data-set-lds_1-h630-2018-h640-2018-2019-pre-release-data-mei.xlsx)
* [LDS\_5 Data about boroughs of London and the regions of England](https://www.ocr.org.uk/Images/400306-large-data-set-lds_2-h630-2019-h640-2020-pre-release-data-mei.xlsx)
* [LDS\_6 Health data about individuals](https://www.ocr.org.uk/Images/400307-large-data-set-lds_3-h630-2020-h640-2021-pre-release-data-mei.xlsx)

We wanted to make the process of working with data manageable for teachers, educationally valuable for students and workable for examining. We decided that three data sets – one per cohort – updated on a rotating cycle would do the trick. In the first year of teaching the new specifications, teachers might choose to work with one data set. The next year, they could still use the lessons that had gone well as well as introducing the next data set and so on.

# 2 Key features

### 2.1 Population

The NHANES uses a representative sample of the United States population; the data are from 2003/4. Students will find it helpful to understand that the US population is not the same as the UK population in all respects and that 2003/4 was some years ago so the sample may not be representative of the US population now. Data from more recent surveys and information on how to access the data is available at <https://wwwn.cdc.gov/nchs/nhanes/sasviewer.aspx>

### 2.2 Data cleaning and manipulation

This list is a few details covering data cleaning issues, along with a couple of possible stumbling blocks to watch out for.

1. Data fields. The Information Sheet in the LDS gives details of what the data mean.
2. Marital status. The NHANES collected data on marital status for persons in the sample aged 14 and over; the LDS only includes individual aged 16 and over. Note that it is possible for people to marry from the age of 14 in some states in some circumstances <https://en.wikipedia.org/wiki/Marriage_age_in_the_United_States>

Data on marital status for the whole 2003/4 NHANES can be found at <https://wwwn.cdc.gov/Nchs/Nhanes/2003-2004/DEMO_C.htm#DMDMARTL> and is reproduced below.

| **Code or Value** | **Value Description** | **Count** | **Cumulative** |
| --- | --- | --- | --- |
| 1 | Married | 2721 | 2721 |
| 2 | Widowed | 588 | 3309 |
| 3 | Divorced | 473 | 3782 |
| 4 | Separated | 137 | 3919 |
| 5 | Never married | 2502 | 6421 |
| 6 | Living with partner | 345 | 6766 |
| 77 | Refused | 0 | 6766 |
| 99 | Don't know | 0 | 6766 |
| . | Missing | 3356 | 10122 |

Students may find it interesting to compare the proportions in the whole survey with those in the LDS and with data from later years.

1. BMI. BMI combines height and weight data to produce a single measure. Information about BMI is available on the NHS website <https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/> Some students will be sensitive or worried about their weights but this measure is often used in health screening. You may find articles which question the use of BMI are of interest <https://www.npr.org/templates/story/story.php?storyId=106268439&t=1582623820055>
2. Blood pressure. The data include measurements of systolic and diastolic blood pressure. Systolic blood pressure is the pressure when the heart is pushing blood out round the body; diastolic blood pressure is when the heart is resting between beats. There is more information about blood pressure on the NHS website <https://www.nhs.uk/common-health-questions/lifestyle/what-is-blood-pressure/>

The systolic blood pressure is higher than the diastolic blood pressure; note that drawing a scatter diagram of systolic against diastolic blood pressure shows that there is an individual with a recorded diastolic blood pressure of zero; this is clearly an error in the data.

Drawing histograms of systolic and diastolic blood pressure separately will show that the Normal distribution is a good model for one of these but not the other.

|  |  |
| --- | --- |
| Systolichistogram about systolic blood pressure | Diastolic histogram about diastolic blood pressure |

1. Drawing graphs and charts. The LDS includes both categorical and numerical data so a variety of different charts can be drawn; the article at <https://mei.org.uk/files/pdf/03-MEI-Insights-November-15.pdf> gives some ideas. You may want to draw the following types of graphs and charts. In each case you will probably find it easier to copy the data you are working with to another Excel spreadsheet or to software such as GeoGebra or Autograph; this will make it easier to select the data you want.
	* bar charts for marital status or other categorical variables
	* scatter diagram using data from two different variables, looking for correlation
	* histograms, looking for Normal and skewed distributions – this will be easier in GeoGebra or Autograph (Excel is not good at drawing histograms)
	* box plots comparing male and female data – this is easy to do in GeoGebra or Autograph. Similary for those who have had food in the last 30 minutes and those who have not.
2. Missing data. Data values which are missing are indicated by #N/A in the LDS. Different software treats this code in different ways. You might want to draw scatter diagrams between different data fields. For example, a scatter diagram comparing pulse rate to systolic blood pressure.

Excel draws the scatter diagram correctly as follows.



Copying the two data columns into GeoGebra gives the following graph. It may not be obvious that it is not the same but notice that the Excel graph has two points at about (70, 180) but the GeoGebra graph only has one. Looking carefully at the LDS shows that there should, indeed, be two points. GeoGebra has not interpreted the #N/A symbols correctly. If you try to find the correlation coefficient in Excel, the existence of #N/A in some cells will prevent the automatic correlation function from calculating.



Different software uses different methods of showing that a data value is missing. #N/A ensures that Excel draws graphs correctly but will prevent use of formulas on the data. You may want to use the filter function in Excel to filter out rows which have #N/A in the fields you are working with before copying into other software or when doing a calculation. Some fields have all or nearly all data so you won’t want to delete all rows which include #N/A anywhere.

# 3 Online resources

### 3.1 Useful additional information

Please note, these links are offerred as additional informaion which students may find interesting and useful; it is not expected that you would use all of these with students.

**NHANES**

The NHANES website offers a great deal of information about the survey including access to data and methodology from different years.

### 3.2 Other useful websites

As mentioned earlier, a classic spreadsheet is not the only way to interact with the LDS. These links are a starting point for exploring other tools.

**Tinkerplots**

TinkerPlots is a simple, but powerful, data visualization and modelling tool developed for use by schools.

<https://www.tinkerplots.com/>

**CODAP**

Online data visualisation based on TinkerPlots and Fathom – just drag a CSV file in and off you go – free to use and needs no download.

<https://codap.concord.org/>

**Desmos**

Desmos is a free dynamic mathematics tool, including graphing, geometry and (most importantly) a spreadsheet. the website also hosts a collection of materials.

<https://www.desmos.com/>

**Geogebra**

Geogebra is a free dynamic mathematics tool, including graphing, 3D graphing, geometry, CAS and (most importantly) a spreadsheet. the website also hosts a vast collection of materials.

<https://www.geogebra.org/>

**JASP**

This is free statistical software which has been developed with the support of the University of Amsterdam. It is fairly intuitive to use for people who can use spreadsheets, comes with online support materials (in English) and has some features which are not available in either Excel or GeoGebra. There are more features than are needed for A level but this software is a good starting place if you want to try specialist statistical software. For Windows, MAC or Linux.

<https://jasp-stats.org/>

**PAST**

This is free statistical software which has been developed by a member of staff at the University of Oslo. It is fairly intuitive to use for people who can use spreadsheets, comes with a manual (in English) and has some features which are not available in either Excel or GeoGebra. There are more features than are needed for A level but this software is a good starting place if you want to try specialist statistical software. For Windows or MAC.

[folk.uio.no/ohammer/past/](http://folk.uio.no/ohammer/past/)

**R**

Finally, R is a free software environment for statistical computing and graphics. It is readily available on a wide variety of operating systems. The interface and language may take some getting used to, but the flexibility and power rewards the effort.

<https://www.r-project.org/>

We’d like to know your view on the resources we produce. Click ‘Like’ or ‘Dislike’ to send us an auto generated email about this resource. Add comments if you want to. Let us know how we can improve this resource or what else you need. Your email will not be used or shared for any marketing purposes.

Looking for another resource? There is now a quick and easy search [tool to help find free resources](http://www.ocr.org.uk/i-want-to/find-resources/) for your qualification.

**Resources**: the small print

OCR provides resources to help you deliver our qualifications. These resources do not represent any particular teaching method we expect you to use. We update our resources regularly and aim to make sure content is accurate but please check the OCR website so that you have the most up to date version. OCR cannot be held responsible for any errors or omissions in these resources.

Though we make every effort to check our resources, there may be contradictions between published support and the specification, so it is important that you always use information in the latest specification. We indicate any specification changes within the document itself, change the version number and provide a summary of the changes. If you do notice a discrepancy between the specification and a resource, please contact us.

© OCR 2021 - You can copy and distribute this resource freely if you keep the OCR logo and this small print intact and you acknowledge OCR as the originator of the resource.

OCR acknowledges the use of the following content: N/A

Whether you already offer OCR qualifications, are new to OCR or are thinking about switching, you can request more information using our [Expression of Interest form](https://www.ocr.org.uk/qualifications/expression-of-interest/).

Please get in touch if you want to discuss the accessibility of resources we offer to support you in delivering our qualifications.