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

**LDS10 used for H630/02 in June 2027 and H640/02 in June 2028**

These notes outline the requirements for OCR’s large data set for specification B (MEI). They 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 within the activities to be found on the website 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 Mathematics 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 CIA World Factbook, 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://my.integralmaths.org/course/view.php?id=287&sectionid=11304>

***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](#_Toc195624297)

[Contents 2](#_Toc195624298)

[1 Overview 3](#_Toc195624299)

[1.1 Introduction 3](#_Toc195624300)

[1.2 MEI Large Data Set 1 4](#_Toc195624301)

[1.3 MEI Large Data Set 4 4](#_Toc195624302)

[1.4 MEI Large Data Set 7 4](#_Toc195624303)

[1.5 MEI Large Data Set 10 5](#_Toc195624304)

[1.6 Why three large data sets for the MEI specification? 6](#_Toc195624305)

[2 Key features 8](#_Toc195624306)

[2.1 Geography 8](#_Toc195624307)

[2.2 Data cleaning and manipulation 8](#_Toc195624308)

[3 Online resources 12](#_Toc195624309)

[3.1 Useful maps 12](#_Toc195624310)

[3.2 Other useful websites 12](#_Toc195624311)

# 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 for Education say that:

AS and A Level Mathematics specifications 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 statistical 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 demonstrate the ability to analyse a subset of features of the data using a calculator with standard statistical functions.

The pre-release 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 1

OCR’s MEI Large Data Set 1 consisted of data about countries taken from the CIA World Factbook and was used for H630/02 and H640/02 in June 2018 and H640/02 (A level only) in June 2019.

### 1.3 MEI Large Data Set 4

OCR’s MEI Large Data Set 4 was a refreshed version of LDS1 and intended for use for H630/02 in June 2021 (actually used in Autumn 2021 series) and H640/02 in June 2022 (actually used in Autumn 2022 series). Data were refreshed to the latest available and additional changes were made as listed below.

* The list of countries has been checked to update names of countries and remove rows which are not usually considered to be countries in their own right. Svalbard has been removed as a country as it is part of Norway. The area and population for Norway are as given in the CIA World Factbook and so do not include Svalbard.
* The Sub region data in LDS1 is no longer a part of the CIA World Factbook. This has been replaced by the region field as detailed in the information sheet in the LDS.
* Time series data on life expectancy has been added from the World Bank DataBank. To ensure that the LDS is of similar size to LDS1, the following fields which existed in LDS1 have been removed; Land area, water area and dependency status.

### 1.4 MEI Large Data Set 7

OCR’s MEI Large Data Set 7 was a refreshed version of LDS4 and was used for H630/02 in June 2024 and H640/02 in June 2025.

Data were mostly been refreshed to the latest available and additional changes made as listed below.

* The list of countries has been checked to update names of countries and remove rows which are not usually considered to be countries in their own right. Swaziland (Eswatini) is now just given the name Eswatini. Western Sahara has been removed as it is a disputed territory.
* The land borders field has been removed.
* The labor force data has not been updated from LDS4 as the data in the latest CIA World Factbook included some older estimates of labor force so the data in LDS4 is, overall, more recent. Note that labor is an American spelling as the CIA World Factbook is an American source.
* The Current Health Expenditure data field has been updated to that currently used in the CIA World Factbook. Note that this is calculated in a different way to the Health Expenditure data field in LDS4.
* Data on mobile phone users have been included; mobile phone subscribers is the same field as in the LDS for the sample assessments but with more up to date data.
* The 2020 life expectancy data from the World Bank DataBank have been included where these figures are available.

### 1.5 MEI Large Data Set 10

OCR’s MEI Large Data Set 10 is a refreshed version of LDS7.

Data have mostly been refreshed to the latest available and additional changes made as listed below.

* The list of countries has been checked to update names of countries and remove rows which are not usually considered to be countries in their own right. Turkey has been renamed as Turkey (Turkiye) to reflect the country’s official name while recognising that it is still oftern referred to as Turkey in common usage.
* Two new fields have been added: carbon dioxide emissions and percentage of electricity generated from fossil fuels.
* To ensure that the data set remains a similar size, two fields have been removed: labor force and umemployment (%).
* Note that for some data fields, the values for the Gaza Strip and the West Bank are combined; in these cases the same value apears for both of these.

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, in the CIA World Factbook and in resources in Integral.

### 1.6 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Publish** | June 2023 | 2024 | 2025 | 2026 |
| **Start teaching** | 2024 | 2025 | 2026 | 2027 |
| **AS Level exam (if sat)** | 2025 | 2026 | 2027 | 2028 |
| **A Level exam** | 2026 | 2027 | 2028 | 2029 |
| **Data set** | 8 | 9 | 10 | 11 |

So if you teach A Level Mathematics over two years, then the class you start teaching in September 2026 will have large data sets 9, 10 and 11 for use in lessons, but will only see some questions on large data set 10 in their H630/02 AS exam in 2027 (if they sit AS) and their H640/02 A Level exam in 2028.

MEI and OCR have some experience of pre-release data from our Quantitative Problem Solving Core Mathematics qualification. 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. 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 in ways 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.

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 initial data sets which we hope will appeal to students with different interests in terms of other subjects they are taking.

We wanted to make the process of working with data manageable for teachers, educationally valuable for students and workable for examining. We decided to have three data sets – one per cohort – updated on a rotating cycle. 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.

LDS 10 is a refreshment of the data from LDS1, LDS4 and LDS7. LDS9 is not such a time dependant data set and as such is the same set as LDS3 and LDS6, but may be refreshed or replaced (dependent upon the post assessment review of the questions set in the live assessment in 2025) and then published as LDS12 in 2027. Similarly LDS8 is a refreshment of the data from LDS2 and LDS5, and will be refreshed or replaced as LDS11.

# 2 Key features

### 2.1 Geography

Each country has been put into a region as detailed on the information sheet in the LDS; this should make it easier to find the countries on a map. The list of countries is taken from the CIA World Factbook but those which are not generally regarded as independent countries have not been included. If you look at a different list of countries, such as the one found on The World Bank website <https://data.worldbank.org/country> or in Gapminder <https://www.gapminder.org/> you will find that most of the countries are in both lists but there are some which are not. There is no internationally agreed list of countries. You will also find that there can be differences in the ways that countries are classified into regions. You may also find that the same country is called by different names in different lists.

### 2.2 Data cleaning and manipulation

This list contains the details covering data cleaning issues, and any data manipulation which was done to create a single data set from all the fields of data, along with a couple of possible stumbling blocks to watch out for.

1. Data fields. An easy way to get the latest data from the CIA World Factbook is to use the country comparisons <https://www.cia.gov/the-world-factbook/references/guide-to-country-comparisons/>. This will give the data values for a particular field in order of size. These lists were used to construct the data set by reordering in order of country name and inserting #N/A into the LDS where the data value was missing. Note that many of the data values are estimates and that they do not all date from the same time. When comparing data from different countries, it is common for data values not to be collected by exactly the same method or at the same time so all comparisons between countries are approximate.
2. Population. Look at the population data. It is given to the nearest person. No country ever knows its population to that level of accuracy. Even when there is a census, the data are out of date by the time they are put together. You might want to think about what a suitable level of accuracy would be.
3. Missing data. You might want to draw scatter diagrams between different data fields using Excel, for example, a scatter diagram of median age against life expectancy in 2020. 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 the point at about (57, 78) in the GeoGebra graph does not exist in the Excel graph or in the LDS. 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 formulae 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.

1. The Holy See. Notice that the area of the Holy See is listed as 0 but it has a population of 1000. The Holy See has Ambassadors to other nations. It is not the same as Vatican City State which is in Rome <https://whc.unesco.org/en/list/286/>. Vatican City has an area of 0.44 square kilometres so to the nearest square kilometre, the area is zero.
2. France. The CIA World Factbook lists France as composed of 6 areas:
	* metropolitan France: Western Europe, bordering the Bay of Biscay and English Channel, between Belgium and Spain, southeast of the UK; bordering the Mediterranean Sea, between Italy and Spain
	* French Guiana: Northern South America, bordering the North Atlantic Ocean, between Brazil and Suriname
	* Guadeloupe: Caribbean, islands between the Caribbean Sea and the North Atlantic Ocean, southeast of Puerto Rico
	* Martinique: Caribbean, island between the Caribbean Sea and North Atlantic Ocean, north of Trinidad and Tobago
	* Mayotte: Southern Indian Ocean, island in the Mozambique Channel, about halfway between northern Madagascar and northern Mozambique
	* Reunion: Southern Africa, island in the Indian Ocean, east of Madagascar.

None of the five overseas areas are listed as separate countries because they are part of France and not self-governing. As part of France, the overseas areas listed above belong to the EU even though they are not in Europe. See <https://ec.europa.eu/regional_policy/en/information/videos/the-outermost-regions-of-europe> for parts of the EU which are not in Europe. It is not always clear whether data values for France include the overseas regions.

1. North, Central and South America. Drawing box plots of median age for North, Central and South America shows more overlap between Central and South America than between either of these and North America.



Geographically, the countries of Central America are at the Southern tip of North America but students may find it interesting to consider whether data for these countries are more similar to data for South America.

# 3 Online resources

### 3.1 Useful maps

Please note, these maps are offered as useful tools to get a sense of where places are and what sort of places they are.

**CIA World Factbook Maps**

<https://www.cia.gov/the-world-factbook/maps/> The maps are available as JPG.

**Eurostat statistics and maps**

<https://ec.europa.eu/eurostat/cache/RCI/#?vis=nuts2.labourmarket&lang=en> Life expectancy, unemployment and more broken down by regions of EU countries.

**Whole world map**

<https://geology.com/world/world-map.shtml> You can click on a country to make it bigger.

**Visualisation of country population by area**

<https://ourworldindata.org/world-population-cartogram>

### 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 visualisation 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/>

**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. By clicking on ‘Like’ or ‘Dislike’ you can help us to ensure that our resources work for you. When the email template pops up please add additional comments if you wish and then just click ‘Send’. Thank you.

Whether you already offer OCR qualifications, are new to OCR, or are considering switching from your current provider/awarding organisation, you can request more information by completing the Expression of Interest form which can be found here: [www.ocr.org.uk/expression-of-interest](http://www.ocr.org.uk/expression-of-interest)

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

**OCR Resources**: *the small print*OCR’s resources are provided to support the delivery of OCR qualifications, but in no way constitute an endorsed teaching method that is required by the Board, and the decision to use them lies with the individual teacher. Whilst every effort is made to ensure the accuracy of the content, OCR cannot be held responsible for any errors or omissions within these resources.

Our documents are updated over time. Whilst every effort is made to check all documents, there may be contradictions between published support and the specification, therefore please use the information on the latest specification at all times. Where changes are made to specifications these will be indicated within the document, there will be a new version number indicated, and a summary of the changes. If you do notice a discrepancy between the specification and a resource please contact us at:

[resources.feedback@ocr.org.uk](file:///%5C%5Cfilestorage%5COCR%5CPD%5CProdSup%5CDesign%5CStudio%5CVisual%20Style%20Guidelines%5C2016_Templates%5Cresources.feedback%40ocr.org.uk).

© OCR 2025 - This resource may be freely copied and distributed, as long as the OCR logo and this message remain intact and OCR is acknowledged as the originator of this work. OCR acknowledges the use of the following content: n/a

Please get in touch if you want to discuss the accessibility of resources we offer to support delivery of our qualifications: resources.feedback@ocr.org.uk