The Development of the Questionmark Data Literacy Test by Cambridge Assessment

Introduction

The test was developed to measure data literacy skills. In today's world of digital record keeping generating vast quantities of data, the importance of data literacy for organizations is growing.

Improved performance and competitive advantage depend on the ability to make the best use of the increasing amount of data available. This ability helps understand business processes and customer needs, inform decision making and evaluate effectiveness.

This test aims to allow organizations to better understand the data literacy of their staff to support decision-making in recruitment, understand training needs and better allocate people to roles. The test is aimed at people in generalist roles, rather than specialist data analysts. It is relevant to the many roles where employees are expected to collate statistics on performance or output, communicate the performance of themselves or their department to others, understand the implications of internal or external data for their own decision making or use data in other ways to support their main working role.

Development of the Test

Our understanding of data literacy was informed by the work of Cambridge Assessment in developing related training programs. This was furthered by research on several organizations regarding their current need for data literacy in different roles and how it impacts performance at the individual and business-wide levels.

Based on this work we developed an operational definition of the skill to inform the development of the test. We considered data literacy as the ability to:

- understand data – this might be tables of figures, charts, and graphs
- analyze data to understand its implications – getting to grips with the meaning of the data
- communicate effectively with data – being able to provide others with a clear picture of what the data is telling you, whether in words or visualizations

www.questionmark.com

USA
35 Nutmeg Drive
Suite 330
Trumbull, CT 06611
+1 (203) 425 2400
(800) 863 3950

UK
New Kings Beam House
22 Upper Ground
South Bank
London SE1 9PD
+44 (0) 20 7263 7575
0800 731 5895

Germany
Hopfenstr. 8,
80335, Munich
+49 (0) 89 220 61272

Australia
RSM Australia, Level 13
60 Castlereagh Street
Sydney NSW 2000
GPO Box 5138
Sydney NSW 2001
+61 2 8073 0527

- understand the quality of data – check whether the data is appropriate for the use to be made of it.

Writing Items

Experts with knowledge of both data literacy and item writing developed a bank of items to measure the different elements of data literacy included in the definition. Items were designed to capture data literacy skills in real-world contexts without the need to understand specialist vocabulary, advanced mathematics, or technical statistical terms.

The items were reviewed by several experts in data literacy, measurement, and question writing, including US and UK English speakers. They are designed to be at an appropriate reading level for those with high school education and with a fluent, but not necessarily native, knowledge of English. The language and topics are appropriate for most English-speaking countries. The items are accessible to those with basic numerical skills such as understanding multiplication, division, and percentages – but advanced skills are not required.

Item Trial

Of the 52 items written and reviewed, 42 were chosen as appropriate to progress to the trial that would ensure their psychometric effectiveness. The items were loaded on the Questionmark OnDemand systems and completed by a range of individuals working in different organizations in various roles.

The sample included some people recruited via Amazon Mechanical Turk to broaden out the range of roles included.

The trial and the final test allow a generous time so that someone working with reasonable efficiency can easily complete all the questions on time.

In total the pilot achieved 169 full completions of the test. Based on the pilot data, 32 items were selected for inclusion in the final test form. Items were selected which had good psychometric properties and to reflect the full range of different skills included in the test.

These 32 items constitute the final test. The sections below summarize the psychometric properties of the test.

Score Distribution

A test must show a good spread of scores so that it is possible to differentiate those with different levels of ability. The scores on the test ranged from 6 to 31 points with a mean score of just under 20 and a standard deviation of 6.2. The difficulty of the test is set at an appropriate level with the average

person answering about two-thirds of the questions correctly. There is a good spread of scores allowing differentiation into at least three bands of scores.

Reliability

The accuracy of a test is measured using the reliability coefficient. This indicates how likely it is that two people of the same ability will achieve similar scores on the test. The desirable level of reliability depends on how the test is used. Greater reliability is required for selection decisions than when using a test for developmental purposes only.

There are several approaches to assessing reliability looking at accuracy in different ways.

The most used measure of reliability is internal consistency reliability, typically measured using Cronbach's Alpha coefficient. It reflects the consistency with which individual items contribute to the accuracy of the scale, including how closely they relate to the construct measured and how well-written items are. Cronbach's Alpha ranges from 0 to 1 with higher values reflecting greater reliability. Values 0.70 and above are appropriate for development purposes.

Internal consistency reliability can be used to estimate confidence intervals around scores, i.e. how accurate a score can be considered. This is done by calculating the Standard Error of Measurement statistic (SEM). One SEM on either side of a score provides a 67% confidence interval for the score and two SEMs give a 96% confidence interval.

The reliability of the test (Chronbach's Alpha) was 0.87 which shows a good level of accuracy of scores.

The standard error of measurement is just over 2 – suggesting that scores for people with the same level of ability will typically differ by only 1 - 2 points. This is around 7% in percentage score terms.

When interpreting scores, the SEM should be taken into account and differences of this order of magnitude should be discounted.

Validity

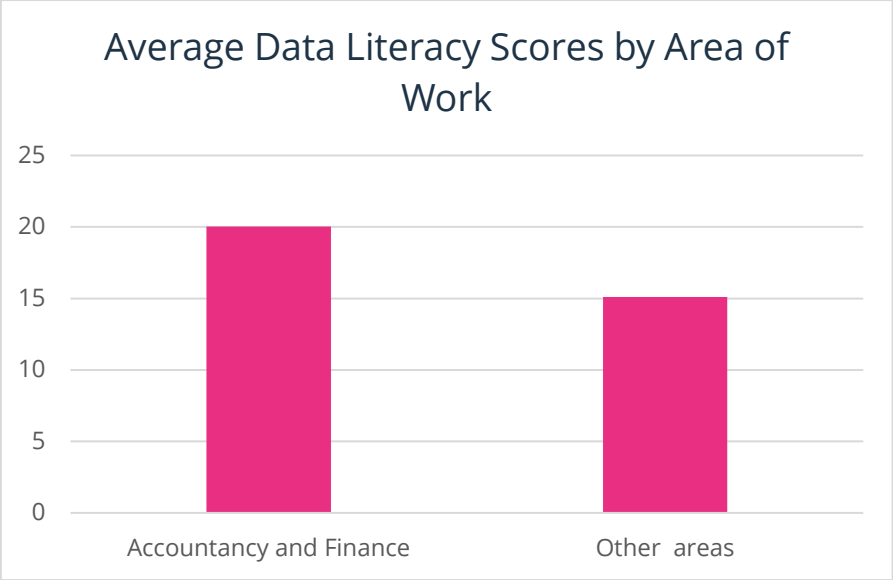
In developing the test, it is important to collect evidence of how well the test is meeting the measurement need to be identified. In this case, we consider the evidence that the test is measuring data literacy.

An initial consideration is face validity – that is the extent to which the test is acceptable to those who use and take it as a measure of data literacy. Does it seem to them an appropriate measure? The feedback we have received from those who have seen the test suggests that it does and indeed the transparent relationship between the questions and what is measured supports this.

A more important aspect of validity is content validity – that is the extent to which the content captures the construct of interest. The content validity of the test is evidenced by comparing the item content to the operational definition of data literacy we used. The questions clearly sample the different skills discussed at varying levels of difficulty. The use of subject matter experts in the design and development of the test helps to support this.

From an empirical perspective, during the trial, we asked respondents to self-assess their competence in areas related to data literacy. We were able to compare their data literacy scores with their average self-rating. The result was a correlation of 0.20 which is a statistically significant result ($p<0.01$ one-tailed, $df=158$). This is evidence for the criterion-related validity of the test.

For construct validity, we hypothesized that those who consistently worked in data-rich environments would show better data literacy scores than those who did not. This is because the work itself is likely to help develop skills and those with better skills are more likely to gravitate toward working in data-rich environments. The first comparison was based on the respondents recruited via Amazon Mechanical Turk. Those who identified their area of work as accountancy or finance (n=36) were compared with those working in other areas (n=43). The chart below shows that the accountancy and finance workers performed better on the test on average. The difference was 0.8 of a standard deviation, which is a large difference.



Another study compared the frequency of engaging in several data-related activities and scores on the test. We hypothesized that those who work frequently with data will develop better skills and this type of role will attract people with better number skills. Three of the comparisons showed statistically significant relationships which support our hypotheses. The table below shows the statistically significant correlations between the frequency of engaging in various activities and data literacy scores.

www.questionmark.com

USA
35 Nutmeg Drive
Suite 330
Trumbull, CT 06611
+1 (203) 425 2400
(800) 863 3950

UK
New Kings Beam House
22 Upper Ground
South Bank
London SE1 9PD
+44 (0) 20 7263 7575
0800 731 5895

Germany
Hopfenstr. 8,
80335, Munich
+49 (0) 89 220 61272

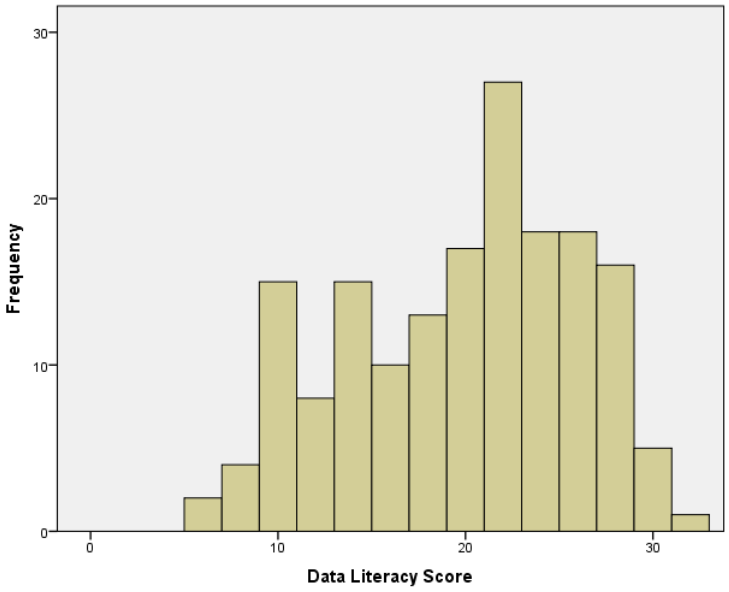
Australia
RSM Australia, Level 13
60 Castlereagh Street
Sydney NSW 2000
GPO Box 5138
Sydney NSW 2001
+61 2 8073 0527

Activity	Correlation with data literacy score	Sample size	Statistical significance level (one tailed)
How often do you use spreadsheets	0.16	156	P<0.05
How often do you look up data in spreadsheets	0.24	153	P<0.01
How often do you read reports with tables of data, graphs or charts	0.24	154	P<0.01

Norms and Standardization

The data from the trial sample was used to provide the standardization data for interpreting test scores. The chart below shows the distribution of scores in the sample ranging from 6 to 31.

The average score was 19.4 and the standard deviation was 6.2.



The standard error of measurement for the score is 2.2. The standard error of measurement can be used to provide a confidence interval around a score to take into account the inevitable error of measurement. For 67% confidence and band of one standard error around the score should be taken. For 96% confidence and band of two standard errors should be used.

For interpretation, scores were divided into three bands. Bands support the interpretation of scores and help address the error of measurement by collecting similar scores together. The table shows the allocation of scores to bands and the interpretation of the band scores.

	Raw Score Range	Percent correct	% of sample scoring in range	Interpretation of score
Band 1	1-15	0-49%	28%	Scores in this range suggest that the person has only a very basic understanding of numerical information and data. While they may be able to deal with simple data, they are likely to be confused by more complex tables and graphs and make errors when interpreting data or use data inappropriately. This may lead to poorer decision-making. Training should focus on understanding how to identify appropriate data, handling and reporting data in different contexts, using different types of displays as well as how to draw appropriate inferences from data.
Band 2	16-23	50%-74%	45%	People with scores in this range can understand and work well with raw data and tables, charts, and graphs. This will usually include evaluating the quality and appropriateness of data and communicating the meaning of the data to others. However, they are more likely to make errors or misunderstand more complex data without support and this could lead to mistaken interpretation or less than optimal decision making. Training should focus on developing an understanding of dealing with data to more complex data structures and more difficult contexts. A broader range of options for displaying and interpreting data could be introduced and the implications for effective communication of results should be studied as well as an understanding of the limitations of data.
Band 3	24-32	75%-100%	27%	People in this range have scored better than 75% of the comparison sample. This suggests they have a good understanding of working with a range of data, tables, charts, and graphs. They tend to make correct inferences from data and should be able to identify the best data to use and select appropriate displays to communicate results to others. Even with a good level of skills, most people could benefit from becoming familiar with new ways of interpreting and displaying data. This might also include topics such as sampling and statistical inferencing.