Quantum: tests worth teaching to

Status report April 2017

1 Background

Project Quantum aims to help computing teachers check their students' understanding, and support their progress, by providing free access to an online assessment system. Distinctive features of Quantum (do read the white paper here) are:

- **Formative**. Quantum is focused on frequent, low-stakes, formative, diagnostic assessment to support learning (in contrast to high-stakes summative assessment).
- **School-led, crowd-sourced.** Teachers both use the corpus of questions on the system and upload questions of their own.
- **Open**. Quantum uses a free, online platform, Diagnostic Questions. Moreover, the questions themselves can be re-used by other platforms; and anonymised data will be available to researchers.
- **Evidence-driven, research-led.** Quantum partners include two of the leading assessment experts in the nation, Tim Oates (Cambridge Assessment) and Robert Coe (Durham Centre for Evaluation and Monitoring). The CEM contribution will be to provide quality control for the crowd-sourced questions, by analysing the data from thousands of students doing thousands of questions. No one has ever done this before.
- Research and reality**.** The project combines two goals
  - **Reality**: being immediately useful to practising computing teachers. They have a crying need for high-quality assessment material, and Quantum will produce this, quickly. We aim to cover both primary and secondary.
  - **Research**: no one has tried to crowd-source assessment items, and then use data to evaluated and improve their quality. If we can make this work, the results will be useful for all subjects in any country. We aim to change the world!

The initial project is generously funded by Google, Microsoft, and ARM, over two years, starting April 2016. The main project partners are:

- Computing at School (CAS)
- Durham Centre for Evaluation and Monitoring (CEM)
- Cambridge Assessment
- Eedi / Diagnostic Questions

This status report summarises our progress over the last six months, and current status.

2 Headlines

Quantum is proceeding apace, on three fronts:

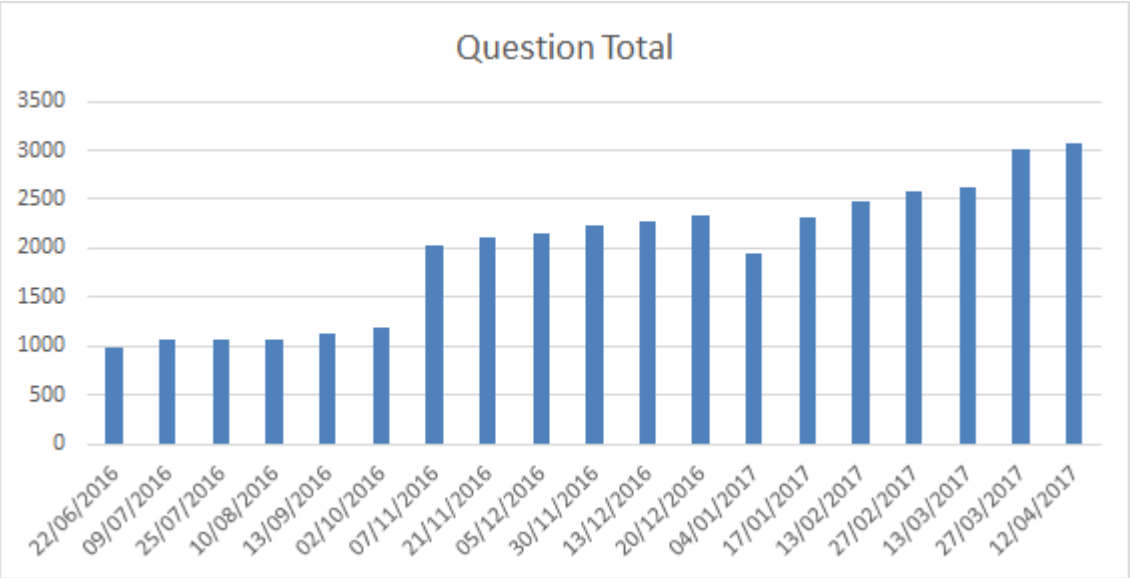
- **Content**. The computing content is developing rapidly (Section 3). We now have over 3,000 questions in the system, and that is enough to be genuinely useful. Alongside developing more content, we will now focus on increasing usage.
- **Analysis and quality control**. The unique feature of Quantum is our ability to analyse data from thousands of students answering thousands of questions, to provide quantitative, data-driven feedback to authors about the quality and effectiveness of their questions. A key issue is how to present this information to our authors, who are not assessment experts. We have made real progress here, described in Section 4.
- **Platform**. We are designing and implementing changes to the Diagnostic Questions platform itself, in direct response to the needs of the first two strands; Section 5 elaborates.

3 Developing computing content (CAS)

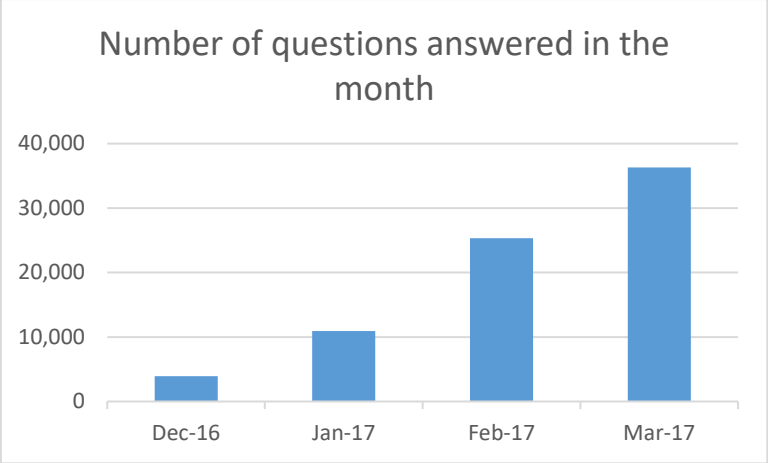
3.1 Usage

The number of Computing questions in Quantum is rising steadily. We have just exceeded 3,000 questions, and have at least another 1,000 in the pipeline. It is already a usable and useful resource for computing teachers wishing to avoid reinventing the wheel for low-stakes, formative assessment of their pupils' knowledge and understanding in computing.

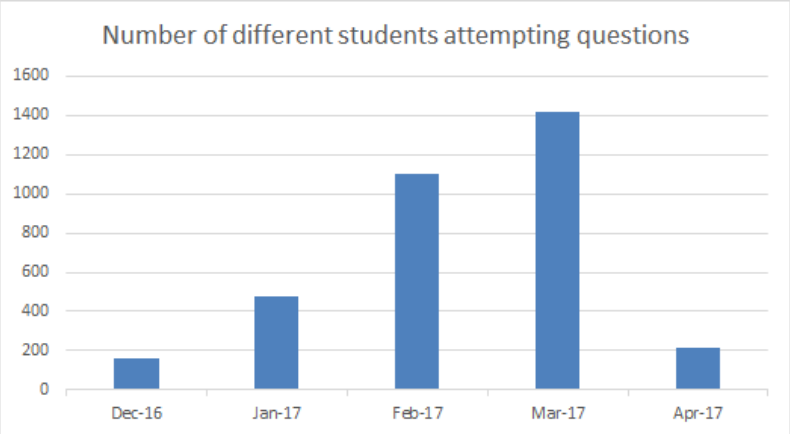
This chart shows our progress over time. (NB: the data before Jan 2017 includes items that had not yet passed moderation; the data after does not; hence the apparent dip in Jan.)



The number of questions actually answered by students is also rising rapidly. The total number of questions attempted varies from week to week, but since October 2016 to the end of April 2017, approximately 96,000 questions have been answered. A low number of questions in a week is still around 2,000, while a big week might be 11,000.



It is not easy to generate the number of different individuals answering questions, as the number is reset in the statistics each week. However, from the totals for each month, it can be calculated that, on average, 241 pupils are answering questions each week. (Data for April is incomplete.)



3.2 Coverage

Although the Quantum project aims to cover the whole of school-level computing, based on the English national curriculum plus GCSE and A Level, the current collection of questions is focussed more heavily on computer science (CS) than either information technology (IT) or digital literacy (DL), although the coverage of these latter two elements is certainly improving. Of the 3,066 questions,

- 2,158 are tagged as CS/foundations, including
  - 1,699 relating to systems (which includes the popular categories of hardware, binary representation and binary arithmetic
  - 606 are on programming
  - 283 on computational thinking
- 677 are on IT/applications
- 341 on DL/implications.

The questions cover GCSE computer science and Key Stage 3 more extensively than A Level or primary. Our view is that primary computing will remain something of a challenge, as assessment of foundation subjects is rather downplayed in most primary schools given the seemingly relentless focus on English and mathematics, and, where computing is assessed, this tends to be through summative assessment of the projects produced by pupils: that said, CAS Master Teacher Iain Davis has had much success with using Quantum in his school, as he discussed in an article for the January 2017 edition of Hello World.

3.3 Taxonomy

We have established a reasonably robust 'taxonomy' for computing, taking our lead from the Royal Society's Shutdown or Restart report in thinking of computing in terms of computer science, information technology and digital literacy, but framing these as the foundations, applications and implications of the discipline respectively. The categories of our initial taxonomy seemed to be rather too broad to make it easy for teachers to find questions on the topics they were teaching, so we've now added another layer to the tree, providing a finer grained approach to classifying questions, and indeed the scope of computing as a subject.

Initial conversations with colleagues from the USA's Computer Science Teachers Association (CSTA) suggest that our taxonomy could be used for the topics in the CSTA draft framework for K-12 CS, opening up further possibilities for the use of Quantum questions for assessment of computing in other English speaking countries.

3.4 Sources

3.4.1 Crowd sourcing

So far we have 75 authors of Quantum questions, of whom six are prolific.

Many of the questions now available as part of Quantum have been provided by class teachers, using the DQ platform with their classes and sharing their computing questions with other users, in line with the crowd-sourced vision for the project.

3.4.2 Commissioning authors

To address the balance of coverage, Quantum has recruited, trained and deployed a small team of CAS question authors, who create sets of three, ten-question quizzes on assigned topics in return for a small payment, following an approach adopted by DQ for developing additional mathematics content. The three-quiz format allows for similar, although not identical, quizzes to be used as pre-test, in-lesson diagnostic assessment and post-test measurement of progress on particular topics.

This team of authors have been effective in 'filling the gaps' in content coverage, and have focussed on topics from the taxonomy tree where coverage has been identified as sparse, with a focus on Key Stage 3. There are now over 800 live questions written by this team. These questions receive rather more scrutiny than others, with quizzes reviewed by members of the Quantum content group. At present a further 80 quizzes (i.e. another 800 questions) are awaiting review and upload to the platform.

3.4.3 Existing collections of questions

Quantum also draws on questions developed by others:

- Exemplar materials from OCR, code.org and Code Club (qv evaluation report via https://www.raspberrypi.org/research-and-insights/).
- The Bebras computational thinking questions, and have negotiated access to further questions from this source.
- The A Level CS wikibooks project supported by CAS #include
- The Canterbury Question Bank developed by Raymond Lister et al at ITICSE13 for undergraduate CS1 courses
- Questions on computer networking for the CISCO Academy course developed by Duncan Maidens of the CAS West Midlands regional centre.

There is no easy way to automatically import questions from other sources into Quantum; we continue to explore automating this process, but in the meantime one of our commissioned authors has agreed, in return for a small payment, to process this body of material for the project.

3.5 Quality

There are, as with any crowd-sourcing, inevitable issues around the quality of some of these questions: we have a low bar moderation system, in which new questions must be checked to ensure that they contain nothing inappropriate and that the answer indicated as correct is indeed correct. We have added a facility to allow users of Quantum to provide feedback to authors on their questions: this doesn't seem to be used particularly extensively at present, as far as we can tell.

The best indication of quality seems to come through the usage of particular questions - the DQ platform allows questions to be sorted

- by use,
- by "likes",
- by inclusion in quizzes, and
- by the most wrong answers.

The first of these three measures provide some indication of how useful a question is. The last is particularly interesting, highlighting where those answering questions have particular misconceptions: already we're seeing Quantum highlighting issues around the teaching of variables, as well as the interpretation of technical vocabulary such as 'selection'; wrong answers can also indicate particular issues with questions themselves, such as poorly worded stems or distractors, ambiguity, or issues with low quality images making it hard for pupils to actually read what the question is asking.

Our hope is that the data analysis undertaken by CEM will provide much better insights into the quality of individual items; see Section 4.

3.6 Curation and search

The number of questions available now makes it quite challenging for teachers to pick good questions to use with their class. The taxonomy tree is useful, although we have a number of questions written in the early phases of the project which now need to be retagged using the finer-grained tree we've developed. There's no free-text tagging of questions, nor does the platform allow the text of a question to be searched (as questions are essentially treated as images on the platform). Similarly questions are not tagged by age, school year or key stage, although CEM's analysis might subsequently provide a mechanism for determining age appropriateness.

The DQ platform does, however, provide a mechanism for grouping individual questions together into quizzes. These can kept private by the quiz creator or shared publicly. At present we have 94 public quizzes, 77 of which have been developed by the CAS Quantum authors.

A priority for the next phase of the project will be to gather together questions into key stage / topic based quizzes, to make it easier for teachers to make use of the materials developed to date. Quizzes themselves can be curated into 'Collections', which might be used to draw together material appropriate for a year group or key stage, particularly where questions can be linked to a scheme of work (such as Switched On Computing or Barefoot Computing) or exam specification.

3.7 Publicity and uptake

As indicated above, Quantum now has lots of good assessment items, and the DQ platform allows questions to be used for a range of low-stakes formative assessment purposes by teachers immediately.

Usage, however, is not yet at the level that we would hope for. Whilst this can be partially explained by teachers' and schools' reluctance to move to a different approach to or system for assessment, we could now be rather more active in promoting Quantum as a source of questions and quizzes for teachers to use.

We've had coverage in Hello World and the Independent Schools Portal site, and an item is planned for CSTA Voice. Miles has spoken about Quantum at a number of events, including the BETT Show and the Education Show, and further presentations are planned, particularly through the CAS Regional Centre network but also for ITTE, the association for information technology in teacher education. Cynthia has produced good training material on writing questions and this will now be extended to include presentations and screencasts for using questions and curating and sharing quizzes, so that CAS hubs and Master Teachers could help to promote Quantum as a source for questions in their training and professional development work.

4 Data analysis (Durham CEM)

During the last six months we have focused on the question of **how to provide meaningful feedback to the author of an item**, based on Rasch analysis student responses. We have developed three possible approaches to establishing the quality of a question, triangulating the evidence to establish the appropriateness of the automatically generated quality rating scale.

4.1 Data set

A data set was used from November 2016. This was filtered down to those students with greater than 50 responses. The data set was then further restricted to items with greater than 500 responses. Over 2000 items remained for more detailed analysis.

4.2 Determining an initial quality measure

4.2.1 Using statistical analysis

The quality rating was calculated based on the following statistical item parameters;

- Actual point measure correlation (PTME)
- Fit
- Discrimination
- Deviation of empirical values from the theoretical estimates or how well the data fits the model (PTME.E)

Typical published values for the acceptable limits of each parameter (for 'low stakes' testing) were used as benchmarks for item performance. The weighting of each parameter's contribution was determined by expert judgement, taking into account the consequences of exceeding the published parameters on the valid interpretation of student performance.

An initial quality rating scale was developed that ranged from a value of 0 to infinity. A value of 0 indicated that the item was performing well from a statistical perspective. A statistically poorly performing item returned a rating greater than 0. In practice, the scale reached a maximum of approximately 12. This rating is dependent on the ability profile of the students.

An initial review of the items using the rating scale as a benchmark indicated that a rating of 1.6 was a reasonable cut score for item screening. Items with a rating greater than 1.6 would form the basis of further investigation, with the intention of explicating the poor statistical performance.

4.2.2 Using distractor analysis

An alternative means of establishing item quality is using distractor analysis. The number of functioning distractors was analysed. A functioning distractor was defined as one that more than 5% of students who were presented with the item went on to select. The number of functioning distractors ranged from zero (no distractors were selected by greater than 5% of pupils; the correct response was attracted by almost all pupils) to three (three distractors were each selected by greater than 5% of pupils).

Based on a review of the literature, the following benchmarks were identified as a means of exploring the performance of items.

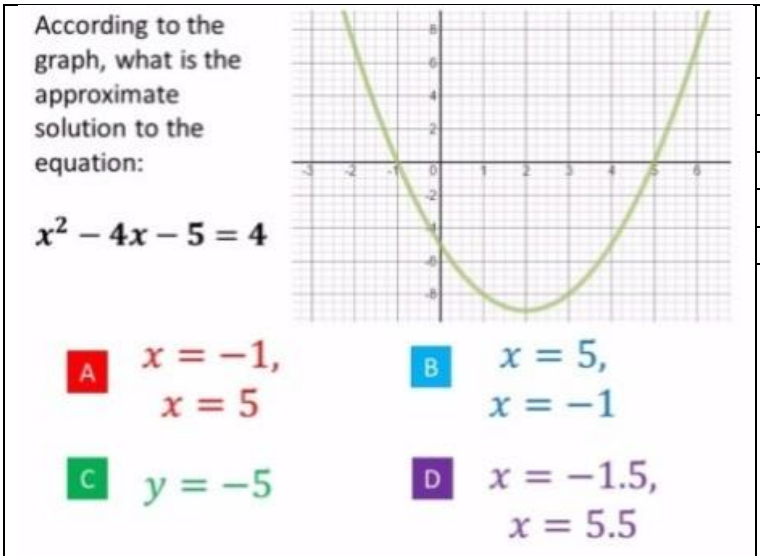
- Distractors which attract less than 5% of all student responses are classed as non-functioning.
- The correct answer curve should be monotonically increasing. Therefore, the distractor curves when combined should be monotonically decreasing.
- Individual distractor curves may not be monotonically increasing – a peak may tentatively indicate a misconception at a given ability range.
- Individual distractor curves may be monotonically increasing or uni-modal. They should not show a multi-modal distribution.

4.2.3 Expert item review

A further means of establishing item quality was using expert judgement. The review covered (not exclusively):

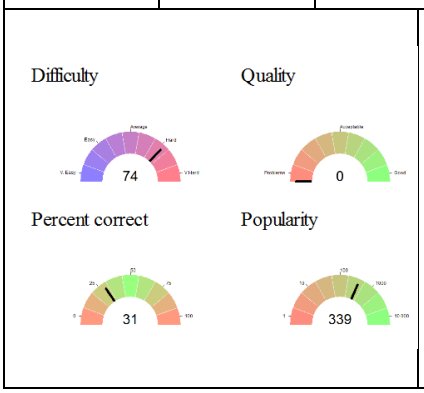
- The content of the stem.
- The content of the distractors.
- Whether the distractors identified typical misconceptions.
- The overall visual presentation of the material.

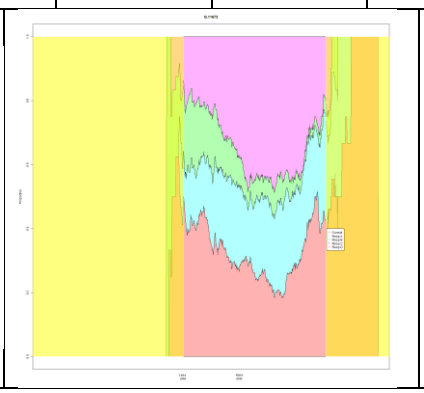
4.3 Q11672

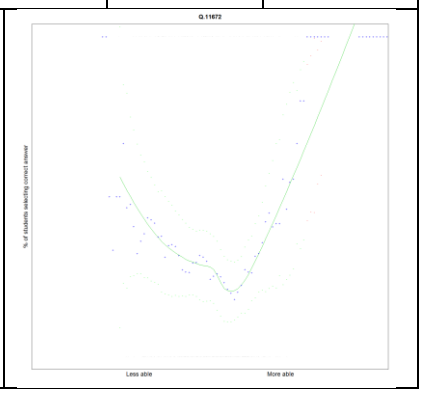


Option	Response count	%
A	124	37%
B	79	23%
C	31	9%
D	105	31%
Totals	339	100%

Measure	MODLSE	Infit	Outfit	PTMA	Discrim.	PTMA (E)	PTMA (diff)
2.40	0.13	1.4132	2.1191	0.0025	0.009	0.4164	0.4139

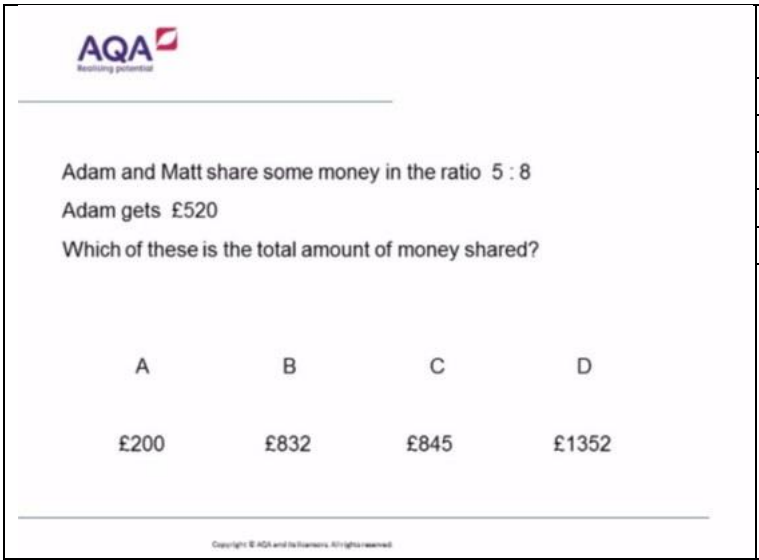






Distractor analysis	**Quality measure (based on statistical analysis)**
There are three functioning distractors. The correct answer curve is not monotonically increasing. The individual distractor curves are monotonically increasing/uni-modal.	Quality rating of 11.6 Statistical performance: POOR
Expert item review 105 students are correct. More students are selecting option A (124) than the correct response. Option A gives the solutions to the equation $x^2 - 4x - 5 = 0$. Option D gives the solutions to the $x^2 - 4x - 5 = 4$. More clarity in the question stem is suggested so that more able students are more likely to get the correct answer.	
Has the analysis recognised the quality of the item? Yes; both the distractor analysis and quality measure has recognised an item that would benefit from revision.	

4.4 Q15260



Adam and Matt share some money in the ratio 5 : 8

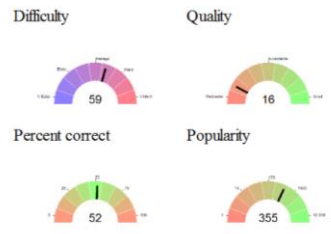
Adam gets £520

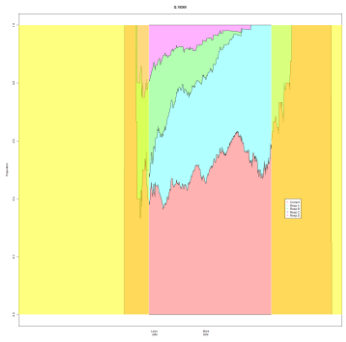
Which of these is the total amount of money shared?

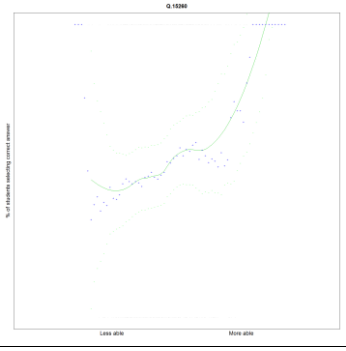
A	B	C	D
£200	£832	£845	£1352

Option	Response count	%
A	19	5%
B	116	33%
C	36	10%
D	184	52%
Totals	355	100%

Measure	MODLSE	Infit	Outfit	PTME	Discrim.	PTME.E	PTME-PTME.E
0.90	0.12	1.3443	1.4567	0.1549	0.0235	0.444	0.2891







Distractor analysis	**Quality measure (based on statistical analysis)**
There are three functioning distractors. The correct answer curve is broadly monotonically increasing. The individual distractor curves are broadly monotonically increasing/uni-modal.	Quality rating of 8.31 Statistical performance: POOR

Qualitative item review

A multi-step problem for which many students (832) are only completing the first step of the problem by calculating how much money Matt would get.

Has the analysis recognised the quality of the item?

No; neither the distractor analysis nor the quality measure has recognised a good item.

4.4.1 Q148

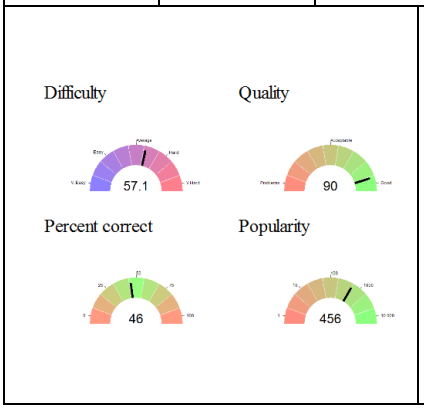
$$\frac{3}{4} + \frac{5}{6}$$

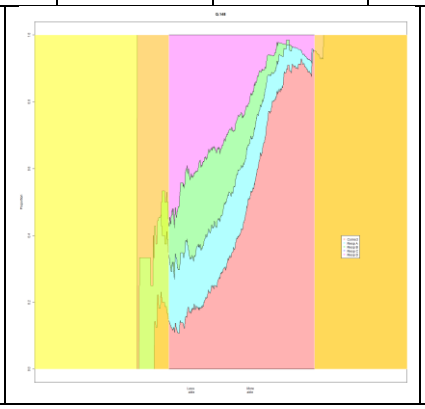
A $\frac{8}{10}$ B $\frac{19}{24}$

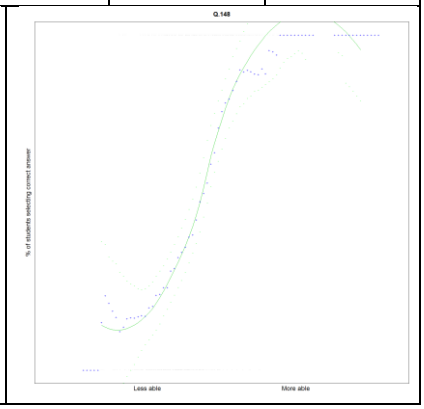
C $\frac{8}{12} = \frac{2}{3}$ D $\frac{19}{12}$

Option	Response count	%
A	110	24%
B	70	15%
C	67	15%
D	209	46%
Totals	456	100%

Measure	MODLSE	Infit	Outfit	PTME	Discrim.	PTME.E	PTME-PTME.E
0.71	0.1	0.8922	0.871	0.5112	1.317	0.4208	-0.0904







Distractor analysis

There are three functioning distractors.
The correct answer curve is broadly monotonically increasing.
The individual distractor curves are broadly monotonically increasing/uni-modal.

Quality measure (based on statistical analysis)

Quality rating of 0.
Statistical performance: GOOD

Qualitative item review

A good quality question which recognises common misconceptions in the distractors.

Has the analysis recognised the quality of the item?

Yes; both the distractor analysis and quality measure has recognised a good quality item.

4.5 Recommendations for future study

4.5.1 Refining the calculation of quality scale

- Change the infit and outfit parameters to be acceptable to a maximum of 1.5 (1.3 currently).
- Review the weighting and combination of variables in the model which are likely to be overcompensating for poor fit due to the interdependency between PTMA, fit and discrimination.

4.5.2 Seeking evidence to validate the scale

- Focus reviews on items with three functioning distractors as these return the most useful information on student performance.
- Qualitatively review those items which are returning the worst quality ratings.
- Determine if the quality scale is recognising the majority of poorly performing items by comparing with expert review and distractor analysis, determine how many false positives are returned (good items which flag as poor on the rating scale), and how many items are of low quality but which are not flagged. This should begin on a sample of items in the first instance drawn from across the rating scale.
- Independent expert review of items sampled across the full quality scale.

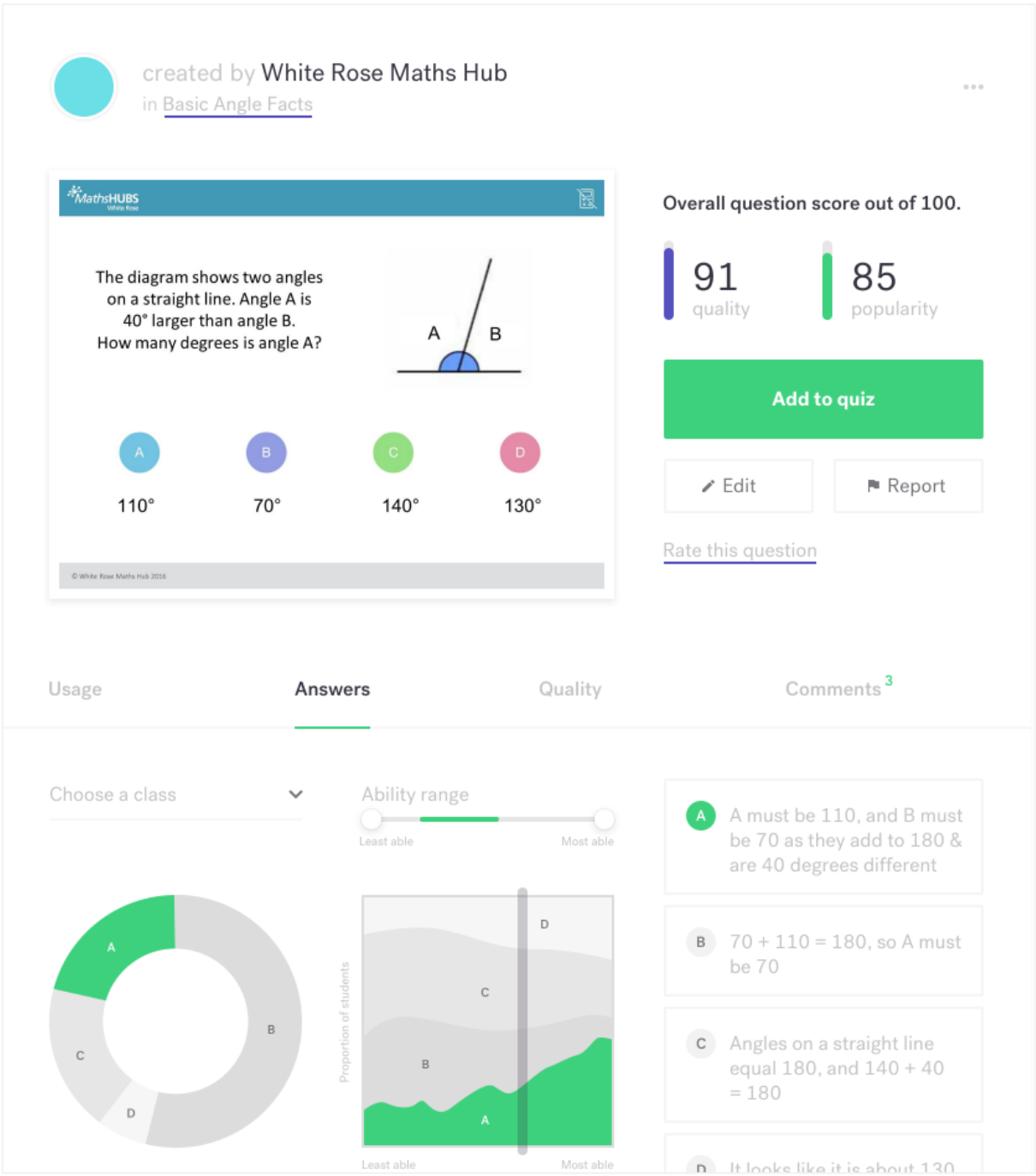
5 Developing the platform (Eedi)

We are working on a range of changes to the Diagnostics Questions platform itself, in direct response to the needs of the Quantum project.

5.1 WP1: Quality Assurance Processes

On 6th March 2017 we started a week-long design sprint with designers, developers and data scientists from Eedi and assessment experts and statisticians from CEM. The focus for the sprint was to identify how to use CEM's research into question quality to enhance teachers' creation and use of assessments.

By the end of the week we had a design prototype illustrating how we could enhance the search for questions with qualitative and quantitative measures of quality. We identified that it was important to show these statistical measures alongside comments from other teachers and explanations given by students. This was particularly interesting in making the distribution of answers by student ability more approachable (see screenshot below). Example student explanations are shown for the selected ability range.



The whole prototype can be viewed here: https://invis.io/FYASHHWDA

After the design sprint, Sam conducted 4 interviews to test our solution with users.

Interview 1: https://youtu.be/LJVcJZQ5tjo
Interview 2: https://youtu.be/5lhADpRzsqs
Interview 3: https://youtu.be/PzjeHQkIiqQ
Interview 4: https://youtu.be/_Qq3V3xsfDc

5.2 WP2: Framework for Computing Curriculum

When a student answers a question incorrectly we want to be able to recommend an appropriate follow-up question, one which tests the same thing as the original question. We currently assume

that questions tagged with the same leaf subject are testing the same thing, this assumption is not valid and we require a solution that suggests appropriate follow-up questions reliably.

We have added support for tagging by “construct”. Questions should test a single construct but may be categorized with multiple subjects. We have built tools to make it easy to assign constructs in bulk to existing questions.

5.3 WP3: Author Interface

The site navigation was completely replaced following feedback from the Quantum team (and quite a few teachers!).

We recorded a new set of videos for helping users through the site. We also integrated Intercom, a knowledgebase and chat support system.

Question and quiz creation is being redesigned as part of a new resources section which we plan to release in September 2017.

In the meantime, we have added support for explanations on individual question creation and for displaying these question explanations on the question insights page. For adding explanations to questions Cynthia and Miles have already uploaded we have written a batch explanation uploader.

5.4 WP8: Integration and Permissions

We are currently building a complete API for Diagnostic Questions which will make integration with CEM and other partners possible.