The Development and Piloting of an Online IQ Test

Examination number:

Sidney Sussex College, University of Cambridge
Abstract

29 items for an online IQ test were designed and piloted on an IQ test run on Facebook. The items were designed to set problems that could be solved by identifying rules (similar to those used in the Raven’s Progressive Matrices). Once designed, the items were published along with items designed by other researchers on an IQ test on the ‘My IQ Intelligence’ application on Facebook. The responses to the items designed for this study were analysed, and where required the items were modified to make them more suitable for inclusion in the IQ test. The modified items were then re-published on the application, and analysed to confirm suitability for inclusion in the IQ test. The study found 22 out of the 29 designed items to be suitable. The items were designed with the intention of eventually compiling an online IQ test that could be administered quickly and inexpensively, yielding a large database of scores that could be used for research into intelligence.
The Development and Piloting of an Online IQ Test

Psychometrics is the science of psychological assessment, and is becoming increasingly important in today’s society (Rust & Golombok, 2009). Psychological assessments have a widespread audience; they are used not only by educational, clinical, occupational and legal psychologists, but also by public servants and the government for use in policy developments in childcare, education, healthcare, criminology and social policy. They can also have direct influences on individuals’ lives, directing the paths that their occupations and education may take. Since the 1980s the use of psychometric testing for recruitment purposes has increased and is now widespread among medium and large-sized organisations (Jenkins, 2001), while some tests (e.g. the Binet tests) have been used since the early 1900s to diagnose mental retardation in children (Rust & Golombok, 2009).

Psychometrics is based upon a specific view of intelligence; that it is not education but educability, and hence is necessarily genetic in origin (Rust & Golombok, 2009). An intelligent person is someone who can make the most of their education, not one who merely receives a good education. This was the view of 19th century scientists, and still holds today. However, this view has since been built upon. More recent models of intelligence suggest that intelligence is not a single entity, but that there are several forms of it. For example, Howard Gardner (1983) asserts that multiple intelligences exist, each with separate systems that operate independently – linguistic, logical-mathematical, spatial, musical, bodily-kinaesthetic, interpersonal and intrapersonal. Robert Sternberg (1990)’s triarchic model suggests three forms of intelligence: analytic (doing well on IQ tests, solving problems, doing well at school), creative (identifying new questions, generating ideas), and practical (wanting to understand, questioning things). The view that there are multiple types of intelligence is compatible with psychometrics, as all these types can be tested psychometrically using tests which are similar to those that are in existence today (Rust & Golombok, 2009).
While Gardner (1983) and Sternberg (1990) assert that multiple intelligences exist independently of each other, others assert that a general mental ability underlies all forms of intelligence. Spearman (1904) was a fierce advocate of this view; he believed that ‘g’ (a general factor of mental ability) explained all variations in “academic”-related intelligence test scores. This view has formed the basis for the development of many psychometric tests. One of the most well-known examples is Raven’s Progressive Matrices (RPM; 1938), developed to assess eductive ability, one of the two components of ‘g’. Eductive ability is the ability to ‘squeeze’ new information and insights out of what is perceived by first forming a holistic impression of the information presented, and then discerning relationships in order to recognise variables (Raven, 1998). The RPM is a multiple-choice nonverbal test consisting of 60 patterns from each of which a section has been removed. The task is to select the missing part on the basis of the principle underlying the pattern. It is designed to "test a person's present capacity to form comparisons, reason by analogy, and develop a logical method of thinking, regardless of previously acquired information" (Raven, 1948).

Scores achieved on the RPM correlate highly with scores on other intelligence tests, for example the Wechsler Adult Intelligence Scale and the Otis Gamma Test (Vincent and Cox, 1974). For this reason research into the relationship between RPM scores and environmental influences often equates RPM scores to ‘intelligence’ or ‘mental ability’. The RPM was originally developed for use in research into the genetic and environmental origins of cognitive ability, and has continued to fulfil this role ever since its publication in 1938 (Raven, 1998). For instance, studies into differences between scores on the RPM have found that in all countries children from less privileged socio-economic backgrounds and from rural areas have lower average scores than others (Raven, 1998). A lot of research into the relative influences of social class, race, culture, gender, age and education on differences in scores on the RPM has also been undertaken (e.g. Tulkin & Newbrough, 1968; Guttman, 1981).
Research into the changes in IQ scores that have taken place over time has also been influenced by the RPM - there is a suggestion that increases in IQ scores are actually indicative of increases in abstract problem solving (deductive ability) rather than ‘intelligence’ (Flynn, 1987).

Despite the capacity of the RPM to be used as a vehicle for research into intelligence (or otherwise), it also appears that the RPM somewhat limits this research. For instance, many studies into the relationship between RPM scores and factors such as race and gender use only a limited sample of participants from a specific group (e.g. 356 in Tulkin & Newbrough (1968); 408 from just 100 families in Guttman (1981); 142 students in Brown & Day (2006)), which can make the generalisation of results difficult. Both the time and cost of collecting RPM score data could limit the size of the samples used. Research into changing IQ scores is also somewhat limited by the RPM, as up-to-date norms are required. This means that frequent recirculation and redesigning of the test is required, which can only be done by the publishers.

With the RPM’s limitations of cost, time and publishers’ rights in mind, this study was designed to develop and test new items for an online IQ test that could be circulated to a large audience, free of charge. It was planned that the items would follow similar rules as those followed by the RPM items. The items were then to be piloted to Facebook users, and responses to them analysed to identify which were suitable for inclusion in the IQ test. A ‘good’ item is one for which different respondents give different responses, is able to discriminate respondents according to whatever the questionnaire is measuring, and whose distractors are endorsed a similar number of times (Rust & Golombok, 2009). Hence the items were to be tested to ensure that they a) have a difficulty index between 0.25 and 0.75, b) discriminate respondents according to what the other items in the test measure, and c) have distractor options which are endorsed by a similar proportion of respondents. It was intended
that any items that do not fulfil these requirements would be modified, re-piloted, and re-analysed.

It was hoped that the study would result in around 20 items being deemed suitable for inclusion in an online IQ test, which in the future could yield a large database of scores. The similarity of the new items to the RPM items would hopefully mean that these scores would be a measure of a facet of intelligence analogous to that measured by the RPM, and could be used for further research into intelligence.

**Method**

**Design**

The study took place in six broad stages:

1. Designing items that followed similar rules to those followed by the Raven’s Progressive Matrices items.
2. Piloting the items to participants.
3. Analysing the responses to the items.
4. Modifying some items as a result of the outcomes of the analyses.
5. Re-piloting the modified items to more participants.
6. Analysing the responses to the modified items.

**Participants**

**Pilot test 1.** 7087 participants took the online test during the first piloting stage. Their responses were analysed to test the newly designed items.

**Pilot test 2.** 8329 participants took the online test during the second piloting stage. 15416 participants’ responses in total were hence available for analysis to test the modified items.
For both tests, the participants were Facebook users who had used the ‘My IQ Intelligence’ application. The participants found the application independently, were told about it by others, or were invited to take the test by their Facebook friends.

Materials

‘My IQ Intelligence’ application. The items were published on this Facebook application conceived and designed by Michal Kosinski, Bartosz Kielczewski and Professor John Rust (University of Cambridge), and David Stillwell (University of Nottingham). The application comprises a 20-question long IQ test based on an item bank consisting of 120 items designed by these researchers, and validated against the RPM. A few RPM items were also included in the item bank. In order to pilot the items that this study is concerned with, the 20-question long test was reduced to 15 questions. A random selection of 5 of the 24 newly-designed items was then added to complete the 20-question test. Each time the test was taken, the 5 newly-designed items were randomly selected. Hence any respondent who took the test once would come across only 5 of the newly-designed items, and no item was responded to by all the respondents (see appendices A and B for exact figures). The items were published in this way because in order to calculate an accurate IQ score, the application’s IQ test needed to consist mainly of items already known to reliably measure intelligence. As the newly-designed items were not yet tested, respondents’ answers to these were not taken into account when calculating their IQ score.

SPSS. The responses to the items were analysed using this statistics programme.

Raven’s Progressive Matrices. The rules followed by RPM items are similar to those followed by the newly-designed items. See Figure 1 for an example of an RPM item.

1 http://applications.facebook.com/myiq-test/
Figure 1

*An example of a Raven’s Progressive Matrices item: A12*

Note. *Correct response = 6*
Procedure

**Designing items.** Items were designed and organised into 5 ‘sets’, A, B, C, D and E. The correct response for each item is found by recognising the rule that the item’s pattern follows. Items in set A follow one type of rule, while items in set B follow another (and so on). The general rules for the items in each set (similar to those followed by items in the RPM) are summarised as follows:

*Set A.* The missing piece completes the continuous pattern of the rectangle.

*Set B.* The bottom right cell of a 2 x 2 matrix is missing. The piece that fits will complete the matrix so that any differences between two cells are mimicked in the other two.

*Set C.* The bottom right cell of a 3 x 3 matrix is missing. The piece that fits will complete the matrix so that the type of difference between the shapes in two adjacent cells is mimicked in any other two. The rule applies both horizontally and vertically.

*Set D.* The bottom right cell of a 3 x 3 matrix is missing. The piece that fits will complete the matrix so that in any row or column of 3 cells, the following rule applies: shapes/lines that are present in two of the cells are not present in the third, shapes/lines that are present in one of the two cells are present in the third, and shapes/lines that are present in neither of the two cells are not present in the third.

*Set E.* The bottom right cell of a 3 x 3 matrix is missing. The piece that fits will complete the matrix so that combinations of the rules above are fulfilled. The specific rule is likely to be unique to an item.

It was planned that the upcoming IQ test would consist of around 20 items in total. This is because in order for a test to yield a large database of scores, it needs to be quick to complete. A test with fewer items encourages respondents to complete it fully and perhaps enjoy it enough to encourage others to complete it. In order to construct a test with around 20
items, 29 items were initially designed. They were organised into sets A, B, C, D and E, following the rules above. For each item, 8 responses were also designed – one correct response, and seven distractors. The more distractors there are, the lower the probability of a respondent getting a correct answer by chance. By designing 7 distractors, there was just a 12.5% probability of a respondent guessing a correct answer. See Figure 2 for an example of a designed item.
Figure 2

An example of an item designed according to the rules described

C3

Note. Correct response = 7. The others are distractors (incorrect responses).
**Piloting items (1).** 24 out of the 29 designed items were published on the ‘My IQ Intelligence’ application in the manner described under the ‘Materials’ section. Items A1-5 were not published as they were considered too easy by the creators of the application (see appendix C for images). Participants who took the online test were given simple information and instructions on how to do so (see appendix D for exact wording).

**Item analysis (1).** Item analysis was then conducted using the responses chosen by the participants in order to assess the suitability of the items for use in an IQ test. This was done by assessing the difficulty, discrimination and distractors of each item.

The difficulty of an item is important to analyse, as each item should help the tester to discriminate between respondents. If an item is answered correctly by too many respondents, it is too easy; most respondents’ total scores will increase by one point, hence the item will not help to discriminate between respondents of differing ability. Similarly, if an item is answered incorrectly by too many respondents, it is too difficult; most respondents’ total scores will decrease by one point, hence the item will not help to discriminate between respondents of differing ability. The difficulty of the items was assessed by finding the difficulty index (the number of respondents who got the item correct/total number of respondents). The ideal difficulty index for an item is between 0.25 and 0.75 (Rust & Golombok, 2009).

Assessing the discrimination of an item involves assessing the item’s ability to discriminate respondents according to whatever the questionnaire is measuring. Only items that measure the same ability as the other items in the test should be selected for the final version. In this study, discrimination was measured by finding the correlation coefficient between the total scores for each designed item and total scores for all the other items in the application. The higher the correlation coefficient, the more discriminating the item is. Rust
and Golombok (2009) suggest that a minimum r-value of 0.2 is usually required for an item to be deemed satisfactorily discriminating.

Distractor analysis involves looking at the number of times each response option is chosen. For good items with clear rules, the correct response should be chosen by the greatest proportion of respondents, and each distractor (incorrect response) should be chosen by a similar proportion of respondents as other distractors (Rust & Golombok, 2009). If a distractor was chosen by more respondents than others, it would suggest that the item’s rule was unclear, and that there could be more than one feasible correct answer. If a distractor was not chosen at all or by very few, it would show that it was obviously incorrect and hence not useful. In this study, the responses given by the respondents were examined graphically using bar charts to assess the functioning of the distractors. Those items for which a higher proportion of respondents chose a distractor than the correct response were judged to be in need of modification, as were those for which each distractor was chosen by a different proportion of respondents (see Figure 3 for an example of improper functioning distractors).
Figure 3

The improper functioning distractors of item E3

Note. Response 6 is the correct response; the others are distractors. This item’s distractors are not functioning properly because the proportion of respondents who chose each distractor is different. Response 7 in particular has been chosen by too high a proportion of the respondents.
Modification of items. After item analysis was carried out on the 24 items, items that did not show satisfactory difficulty indices, satisfactory r-values, and well-functioning distractors were removed from the online test and modified. Modification usually involved changing the distractors to increase the likelihood of an equal number of respondents choosing each distractor and the correct response being chosen most frequently. In cases where it seemed that the rule was unclear, the item was re-designed entirely (e.g. E5). See appendix E for comparisons between the original and modified items.

Piloting items (2). The modified items were re-published on the Facebook application in the same way as before.

Item analysis (2). Item analysis using the responses to the modified items was carried out in the same way as before. Any further modifications to be made were noted, and a record of those items suitable for inclusion in an IQ test was made.

Results

Item analysis (1)

24 of the 29 initially designed items were published on the Facebook application. The responses of 7087 participants were analysed to assess the suitability of each item for inclusion in an IQ test. Each participant answered 5 of the newly-designed items (i.e. the 5 that were randomly selected to form part of the 20-question long test). Table 1 summarises the results of these analyses, and shows the following.

Difficulty index. The difficulty index of the items analysed lay between 0.051 and 0.876. Two items had a difficulty index that was lower than 0.25, and seven had a difficulty index that was greater than 0.75.

Discrimination. The r-value of the items analysed lay between -0.021 and 0.584. Two items had an r-value that was lower than 0.2.
**Distractors.** For each item, the number of times each distractor was chosen was analysed graphically. Six items had improperly functioning distractors (i.e. not all the distractors were chosen by a similar number of respondents and/or some distractors were chosen more often than the correct response). Table 1 shows the bar charts on which graphical analysis of distractors was based (for items where distractors functioned improperly). See appendix F for larger images.

Items that had a difficulty index of below 0.25 and/or an r-value of below 0.2 and/or improperly functioning distractors (marked by * on the table) were modified accordingly. The item names of the 8 items that were modified are indicated by bold case.
Table 1

Results of item analysis for 24 published items

<table>
<thead>
<tr>
<th>Item name</th>
<th>Difficulty index</th>
<th>Correlation coefficient (r value)</th>
<th>Evidence of improper functioning of distractors</th>
</tr>
</thead>
<tbody>
<tr>
<td>A6</td>
<td>0.836</td>
<td>0.253</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>0.597</td>
<td>0.254</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>0.868</td>
<td>0.192</td>
<td></td>
</tr>
<tr>
<td>B3</td>
<td>0.836</td>
<td>0.321</td>
<td></td>
</tr>
<tr>
<td>B4</td>
<td>0.856</td>
<td>0.390</td>
<td></td>
</tr>
<tr>
<td><strong>B5</strong></td>
<td>0.425</td>
<td>0.351</td>
<td></td>
</tr>
<tr>
<td>B6</td>
<td>0.876</td>
<td>0.130</td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>0.866</td>
<td>0.288</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>0.708</td>
<td>0.382</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>0.703</td>
<td>0.415</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C4</td>
<td>0.629</td>
<td>0.584</td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>C5</td>
<td>0.350</td>
<td></td>
<td>0.523</td>
</tr>
<tr>
<td>D6</td>
<td>0.661</td>
<td></td>
<td>0.520</td>
</tr>
<tr>
<td>D2</td>
<td>0.718</td>
<td></td>
<td>0.478</td>
</tr>
<tr>
<td>D4</td>
<td>0.448</td>
<td></td>
<td>0.417</td>
</tr>
<tr>
<td>D1</td>
<td>0.812</td>
<td></td>
<td>0.482</td>
</tr>
</tbody>
</table>
The development and

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>0.780</td>
<td>0.388</td>
</tr>
<tr>
<td>E2</td>
<td>0.596</td>
<td>0.533</td>
</tr>
<tr>
<td>E3</td>
<td>0.314</td>
<td>0.229</td>
</tr>
<tr>
<td>E4</td>
<td>0.437</td>
<td>0.398</td>
</tr>
<tr>
<td>E5</td>
<td>0.188*</td>
<td>0.263</td>
</tr>
<tr>
<td>E6</td>
<td>0.263</td>
<td>0.162*</td>
</tr>
</tbody>
</table>

Note. Correct response in bar chart is shown in bold case. C6: response 6 was chosen by no respondents hence is not shown on the bar chart.
**Item analysis (2)**

The modified items were then re-published on the Facebook application. The responses of 15416 participants (including the 7087 respondents from analysis (2)) to these were then analysed in the same way as before. Again, each participant answered the 5 newly-designed items that were randomly selected to be part of their IQ test. Table 2 summarises these results, and shows the following.

**Difficulty index.** The items analysed had difficulty indexes that lay between 0.355 and 0.785. None had a difficulty index below 0.25. E5 and D3, previously too difficult, were now slightly easier, showing that the modifications were successful.

**Discrimination.** The items analysed had r-values that lay between 0.274 and 0.538. None had an r-value below 0.2. E6 and D3, whose r-values were previously too low, now appeared to measure what they were designed to measure more effectively, showing that the modifications were successful.

**Distractors.** The modifications made to the 4 of the items with improperly functioning distractors were successful, as the correct response was chosen by the highest proportion of respondents, and each distractor was chosen by a relatively similar proportion of respondents. See Figure 4 for an example of a successful distractor modification. C6 and E6 may require further modification, as some of their distractors were chosen by quite a high proportion of the respondents (see bold case items in Table 2).
Table 2

*Comparison of results of item analysis for items before and after modification*

<table>
<thead>
<tr>
<th>Item name</th>
<th>Difficulty index</th>
<th>Correlation coefficient (r-value)</th>
<th>Evidence of improper functioning of distractors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
<td>Before</td>
</tr>
<tr>
<td>D3</td>
<td>0.051*</td>
<td>0.761</td>
<td>0.021*</td>
</tr>
<tr>
<td>E5</td>
<td>0.188*</td>
<td>0.367</td>
<td>0.263</td>
</tr>
<tr>
<td>B5</td>
<td>0.425</td>
<td>0.735</td>
<td>0.351</td>
</tr>
<tr>
<td>C4</td>
<td>0.629</td>
<td>0.693</td>
<td>0.584</td>
</tr>
<tr>
<td>C6</td>
<td>0.627</td>
<td>0.648</td>
<td>0.427</td>
</tr>
</tbody>
</table>
Note. Only the bar charts of items whose distractors functioned improperly before modification are shown.
Figure 4

Distractor functioning of item B5 before and after modification

Before – Response 1 is correct. Response 6 was chosen by too high a proportion of the respondents, so was modified to appear less like the correct option.

After – Note. Response 1 is correct. The distractors were chosen by a similar proportion of respondents.
Using the results of item analyses (1) and (2), the study found 22 items (some modified, some original) to be suitable for the inclusion in an IQ test. These items were: A6, B1-6, C1-5, D1-4, D6, and E1-5.

**Discussion**

24 items were piloted on Facebook, and after some were modified, 22 were deemed suitable for inclusion in an IQ test.

**Difficulty index.** Despite Rust and Golombok (2009)’s suggestion that the ideal difficulty index for an item is between 0.25 and 0.75, only those items with a difficulty index of below 0.25 were considered in need of modification. A difficulty index greater than 0.75 suggests that the item is too easy, but the results of the item analysis showed that in most cases, items with a difficulty index greater than 0.75 had r-values of greater than 0.2 and properly functioning distractors. The process of choosing which items to include in any questionnaire involves balancing factors against each other based on the need of the questionnaire (Rust & Golombok, 2009), so it seemed appropriate to retain easy items as long as they had reasonable discrimination. Item analysis (1) showed that B2 and B6 were the only items that had a difficulty index greater than 0.75 and an r-value lower than 0.2. These perhaps should have been modified, but at the time it seemed appropriate to retain some very easy items to ensure the test could be used by people with a range of abilities.

**Discrimination.** The usual procedure for analysing the discrimination of items involves piloting all the items to be tested together as one test, and finding the correlation between the total scores on each item with the total scores for the whole test. This establishes the extent to which any given item measures what the test as a whole measures. However, this was impossible to do in this study, as the total score for each respondent was given based on the answering of 20 items, only a few of which were the items that needed to be assessed. In such cases, it is generally considered appropriate to assess discrimination using Item
The development and Response Theory (Birnbaum, 1968). However, due to time constraints it was decided that discrimination would be assessed based on the other items in the 20-question test regardless; this meant assessing the extent to which each item measured what all items in the application measured, rather than the extent to which each item measured what all the newly-designed items measured. This was considered appropriate as many of the other items were designed to be similar to the RPM items (by other researchers), or actually were items taken from the RPM; as a result, most of the items that the discrimination analysis was based on most likely measured a similar facet of intelligence.

**Distractors.** The functioning of distractors was assessed visually, by looking at each item’s bar chart showing the frequency of responses for each distractor. Assessing the functioning of the distractors in this way is obviously relatively subjective, but was sufficient for the purposes of this study as the differences tended to be either very significant (in which case the item was judged to require modification), or very subtle (in which case the item was judged to have well-functioning distractors).

**Sample size.** The study comprised two separate rounds of item analyses based on the responses of those who took the online ‘My IQ Intelligence’ test within the two two-week periods of testing. Hence there was no control to be had over the sample size. Item analysis (1) was based on a much smaller sample’s responses than item analysis (2) (7087 compared to 15416 respondents), which could mean that item analysis (1) is less accurate than the latter.

However, the change in sample size does highlight a major strength of administering tests online – participants can be found with very little effort at all, and very quickly. Both analyses were based on the responses to the items after just two weeks of the items being published on the Facebook application; in each two week period 7000-8000 participants took the test. As the development of the items tested in this study was inspired by the need for
intelligence tests that can inexpensively yield large databases for use in research, it is useful to find that online administration is a fast method for this purpose.

**Future work.** Before the items tested in this study can be used for research into intelligence, some further work is required. Firstly, the items deemed suitable according to this study should be compiled into one test, independent to the 20-question ‘My IQ Intelligence’ application test. This will allow further work on the development of the test as a whole (rather than of individual items) to be undertaken.

This includes measuring its reliability; i.e. estimating the accuracy of the test by checking to see whether the same respondents answer it in the same way at different times. The validity of the test as a whole could also be assessed to check the extent to which it measures what it is intended to measure (the same facet of intelligence measured by the RPM). This has been done in this study for the individual items to an extent, but it will be necessary to do it for the whole test. Various forms of validity (face, content, criterion-related and predictive) can be tested (Rust & Golombok, 2009).

Once this test is shown to be reliable and valid, it could then begin to be used for research – for example to study the relationship between intelligence and racial differences. The test could be administered online in a similar fashion to how the items were piloted in this study, and hopefully a large database of scores could be produced for and used in more accurate and representative research studies.

**Conclusion**

The study comprised the designing, piloting and analysis of 29 items for an online IQ test. The items were designed to set problems that could be solved by identifying rules that are similar to those used in the Raven’s Progressive Matrices. Once designed, 24 items were published along with items designed by other researchers on an IQ test on the ‘My IQ Intelligence’ application on Facebook. Item analyses based on the piloting of the items
showed that some needed modification; during the course of the study the necessary modifications were made to 8 of them to make them more suitable for inclusion in the IQ test. The modified items were then re-published on the application, and analysed to confirm suitability for inclusion in the IQ test. After the modification stage, 22 items in total were found to be suitable. This study is merely one stage in the process of constructing a test, but is essential, allowing suitable items to be identified. It also suggests that online administration of tests is an efficient way of yielding large databases of test scores, which may come in useful in the future when preparing to administer the test for research purposes.
References


Appendix A

Table 3

*Number of participants whose responses were accounted for in item analysis (1)*

<table>
<thead>
<tr>
<th>Item name</th>
<th>Item code</th>
<th>Number of respondents who answered this item</th>
<th>Number of respondents who did not answer this item</th>
</tr>
</thead>
<tbody>
<tr>
<td>A6</td>
<td>q39</td>
<td>293</td>
<td>6794</td>
</tr>
<tr>
<td>B1</td>
<td>q40</td>
<td>263</td>
<td>6824</td>
</tr>
<tr>
<td>B2</td>
<td>q41</td>
<td>235</td>
<td>6852</td>
</tr>
<tr>
<td>B3</td>
<td>q42</td>
<td>219</td>
<td>6868</td>
</tr>
<tr>
<td>B4</td>
<td>q43</td>
<td>216</td>
<td>6871</td>
</tr>
<tr>
<td>B5</td>
<td>q44</td>
<td>212</td>
<td>6875</td>
</tr>
<tr>
<td>B6</td>
<td>q45</td>
<td>209</td>
<td>6878</td>
</tr>
<tr>
<td>C1</td>
<td>q46</td>
<td>224</td>
<td>6863</td>
</tr>
<tr>
<td>C2</td>
<td>q47</td>
<td>216</td>
<td>6871</td>
</tr>
<tr>
<td>C3</td>
<td>q48</td>
<td>229</td>
<td>6858</td>
</tr>
<tr>
<td>C4</td>
<td>q49</td>
<td>221</td>
<td>6866</td>
</tr>
<tr>
<td>C5</td>
<td>q50</td>
<td>223</td>
<td>6864</td>
</tr>
<tr>
<td>D6</td>
<td>q52</td>
<td>242</td>
<td>6845</td>
</tr>
<tr>
<td>D2</td>
<td>q53</td>
<td>238</td>
<td>6849</td>
</tr>
<tr>
<td>D4</td>
<td>q55</td>
<td>252</td>
<td>6835</td>
</tr>
<tr>
<td>D1</td>
<td>q57</td>
<td>245</td>
<td>6842</td>
</tr>
<tr>
<td>E1</td>
<td>q58</td>
<td>227</td>
<td>6860</td>
</tr>
<tr>
<td>E2</td>
<td>q59</td>
<td>255</td>
<td>6832</td>
</tr>
<tr>
<td>E3</td>
<td>q60</td>
<td>223</td>
<td>6864</td>
</tr>
<tr>
<td>E4</td>
<td>q61</td>
<td>238</td>
<td>6849</td>
</tr>
<tr>
<td>E5</td>
<td>q62</td>
<td>245</td>
<td>6842</td>
</tr>
<tr>
<td>E6</td>
<td>q64</td>
<td>137</td>
<td>6950</td>
</tr>
<tr>
<td>D3</td>
<td>q65</td>
<td>137</td>
<td>6950</td>
</tr>
<tr>
<td>C6</td>
<td>q66</td>
<td>158</td>
<td>6929</td>
</tr>
</tbody>
</table>
Appendix B

Table 4

Number of participants whose responses were accounted for in item analysis (2)

<table>
<thead>
<tr>
<th>Item name</th>
<th>Item code</th>
<th>Number of respondents who answered this item</th>
<th>Number of respondents who did not answer this item</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3</td>
<td>128</td>
<td>677</td>
<td>14739</td>
</tr>
<tr>
<td>E5</td>
<td>130</td>
<td>651</td>
<td>14765</td>
</tr>
<tr>
<td>B5</td>
<td>131</td>
<td>683</td>
<td>14733</td>
</tr>
<tr>
<td>C4</td>
<td>132</td>
<td>685</td>
<td>14731</td>
</tr>
<tr>
<td>C6</td>
<td>133</td>
<td>701</td>
<td>14715</td>
</tr>
<tr>
<td>D4</td>
<td>134</td>
<td>646</td>
<td>14770</td>
</tr>
<tr>
<td>D6</td>
<td>135</td>
<td>669</td>
<td>14747</td>
</tr>
<tr>
<td>E6</td>
<td>137</td>
<td>667</td>
<td>14749</td>
</tr>
</tbody>
</table>
Appendix C

*Items A1-5, considered too easy for publication*

Answer: 6
The development and

Answer: 4
Answer: 8
Answer: 2
The development and

Answer: 5
Instructions - read carefully

- MyIQ is a **20-question test** and it should take between **20-40 minutes** to complete.

- **Don't leave questions unanswered!** If you are not sure - use your intuition. You still have a good chance of getting it right.

- Make sure that **you have enough time** and that you are not going to be distracted. To get the best score you should complete the test in a quiet environment and also avoid taking the test while others are watching.

- Remember that your true score may be lower or higher than one achieved on this test.

- Please see our terms of use page including privacy policy for full details. By starting the test, you agree that you have read all of the above, including the terms of use page, and that you agree to abide by its recommendations.
Appendix E

Images of items before and after modification (if modification was required)
The development and

B4 before

B5 before

B5 after
The development and
The development and C4 before C4 after

C5
The development and

D6 before

D6 after

D2
The development and

D4 before

D4 after

D1
The development and
The development and
The development and

E5 before

E5 after

No image available

E6 before

E6 after
The development and
Appendix F

Bar charts showing frequency of participants who chose each response for items where the distractors were judged to be functioning improperly

B5

![Bar chart for q44](image)

C4

![Bar chart for q49](image)
Note. The response option on the very right of each bar chart is the correct option. The others are distractors.