Research and Analysis

The impact of qualification reform on the practical skills of A level science students

Paper 4: An analysis of the functioning of examination items that indirectly assess A level science practical skills

Stuart Cadwallader

ofqual

Contents

Executive summary

This report describes one study from a programme of research that Ofqual is conducting to evaluate the impact of qualification reform on the practical skills of A level science students (see Annex A). The post-reform qualifications assess practical skills in 2 ways. First, each student's practical work must be observed and assessed by their teacher throughout their studies, during which they must complete a minimum of 12 'hands-on' practical assignments. Students are assessed against criteria which reflect the broad competencies that A level science students are expected to develop. They receive a separate grade for their performance in this 'practical endorsement' (either 'Pass' or 'Not Classified'). This constitutes *direct* assessment of practical skills[1]. Second, at least 15% of all available marks in written examinations must be allocated to questions that *indirectly* assess practical skills and should draw on the experience of students across a range of activities.

It is the second way of assessing practical skills, through examination questions, which is the focus of this study. Data from the summer 2017 A level science examinations was collected for 16 separate specifications across the 3 subjects (biology, chemistry and physics). The data was provided by 4 examination boards and comprised the mark achieved by each student on each question (item). There were three stages to the analysis. First, the overall performance of students on the 'practical skills' examination items (as identified by the examination boards) was compared with their overall performance on the other (non-practical skills) items. Second, the performance of those students who received a 'Pass' grade for the practical endorsement was compared with those who did not. Finally, a principal components analysis was conducted to establish whether the practical skills items were measuring a statistically unique underlying construct (factor).

The findings suggest that, on average, the practical skills items were more difficult than the other items. However, there was a high degree of variability, with some practical skills items proving to be relatively hard and others relatively easy. There are many variables which combine to dictate the difficulty of an item and it is not possible from this analysis to establish the extent to which the 'practical element' of the practical skills items affected their level of difficulty. That caveat aside, the higher difficulty may be in part a result of practical skills items being relatively new to teachers and students. The effect may dissipate in future years, once teachers and students become more familiar with the style of question.

In terms of their performance on practical skills items, there was no difference between those students who received a 'Pass' grade for the practical endorsement and those who did not, once their overall exam grade was taken into account. This was true across all subjects, even when those who did not pass were compared to a 'matched' group of those who did (eg. a group of equal size that had taken the same specifications and achieved a similar total mark). This finding is less surprising than it may sound. The students who did not pass the endorsement are a relatively very small group who failed to achieve the endorsement for a wide variety of reasons (eg. for personal reasons or administrative reasons).

[1] The distinction between *direct* and *indirect* assessment of practical skills (see Abrahams & Reiss, 2015) is further discussed in the main body of this report.

Finally, principal component analysis (PCA) did not suggest that the practical skills items were measuring a statistically unique factor (ie. the patterns of correlations did not suggest that practical skills items tended to cluster together to assess the same underlying construct). This may be because practical skills items tend to also assess knowledge of subject content (to at least some degree) and are therefore not a 'pure' measure of practical skills. Alternatively, it may simply be that a student's performance on practical skills items is closely related to their performance on the other items in the test – if the student does relatively well on the practical skills items they also do relatively well on the other items.

In conclusion, this analysis provides an insight into how the practical skills items functioned in relation to other items in the 2017 A level science examinations. From a statistical perspective, though the practical skills items are often more difficult, they do not appear to be measuring a unique underlying construct or differentiating those who passed the practical endorsement from those who did not. However, as we have established, there are many reasons why one may not necessarily expect them to, the main one being that they are not intended to assess the exact same skills as those assessed by the endorsement. The findings are therefore not necessarily surprising because it may be desirable for practical skills items to assess practical skills *in the context of* scientific knowledge and understanding.

Introduction

Reformed science A levels were introduced for first teaching in September 2015, with the first full cohort of students certificating in summer 2017. Ofqual is undertaking a programme of evaluative work regarding the impact of this qualification reform on the practical skills of students. This report describes one study from this programme, a statistical exploration of the examination items that are intended to indirectly assess science practical skills. For a detailed discussion of the new assessment arrangements we recommend that you refer to the companion reports that are listed in Annex A (eg Ofqual, 2017b, 2018).

Assessing practical skills in the reformed A levels

The assessment of practical skills in A level science is now achieved in 2 ways: through questions in the written examinations and via the 'practical endorsement'. For the endorsement, each student's practical work is observed and assessed by their teacher throughout their course, during which they must complete a minimum of 12 'hands-on' practical assignments. The teacher assesses the student against 5 criteria (the common practical assessment criteria, or CPAC). These criteria reflect the broad competencies that A level science students are expected to have developed by the end of their course (see Department for Education, 2014). If the student has evidenced that they are competent against all 5 criteria they receive a 'Pass' grade for the endorsement (if they do not demonstrate competence they receive the grade of 'Not Classified'). The endorsement grade is separate to the primary A level grade (A*-E) but is reported alongside it.

As expected, the pass rate for the endorsement is high. In 2017, 99% of all entries for an A level science qualification attained a 'Pass' grade (Ofqual, 2017a). Those students who received a lower A level grade were less likely to attain a 'Pass' for the endorsement, though the pass rate for those receiving grade U was over 90%. Table 1 shows the overall pass rates for the endorsement (across biology, chemistry and physics) by each A level grade in 2017. This high pass rate is not surprising given the 2 year time period over which the practical assessment is completed. Teachers and students have the opportunity to work towards success against the CPAC, developing the required competencies over time and over multiple practical activities.

Table 1. *A level science practical endorsement outcomes by grade (Ofqual, 2017a)*

Grade	Not classified %	Pass %	Total Entry
A*	0.09	99.91	10,860
A	0.08	99.92	25,724
B	0.18	99.82	27,551
C	0.61	99.39	25,670
D	1.19	98.81	20,446
E	2.72	97.28	10,901
U	8.53	91.47	4,324
Total	0.91	99.09	125,476

The second way in which practical skills are assessed in post-reform science A levels is *indirectly* through the written examinations (the distinction between the *direct* and *indirect* assessment of practical skills is discussed in the next section). At least 15% of all available marks in the terminal examinations must be allocated to items (questions) that indirectly assess practical skills. These items can cover a broad range of skills and knowledge in relation to practical work. The subject content provided by the Department for Education (2014, p. 19) uses the following 4 headings to summarise the skills that have been identified for indirect assessment:

- Independent thinking
- Use and application of scientific methods and practices
- Numeracy and the application of mathematical concepts in a practical context
- Instruments and equipment (with an emphasis on knowledge and understanding of the instruments and related techniques for their use)

A practical skills item might therefore assess a student's ability to design an experiment, to interpret or draw a graph, or to understand the function of a scientific instrument in an experimental context. It is the indirect assessment of practical skills in the examinations which is the focus of this report.

Indirect assessment of practical skills

'Indirect assessment of practical skills' (IAPS) refers to any form of assessment in which a student's competency in practical skills is inferred from some secondary artefact of their work (Abrahams, Reiss, & Sharpe, 2013). This may be data that the student has produced, reports of the practical work they have undertaken, or, in the case of the reformed science A levels, their response to a question in a written examination paper (Abrahams & Reiss, 2015).

The use of a standard written examination format precludes the student from physically undertaking practical work during the assessment. However, strong performance on IAPS examination items should theoretically be at least partly dependent on the degree to which the student has had experience of 'hands on' practical work. For the reformed A levels, the intention is that a student should have a significant advantage in their written examinations if they have conducted a broad range of practical activities as part of their practical endorsement. In this way, it is intended that the two assessment methods (the endorsement and the examinations) are complementary.

The first thing to consider is precisely what a written examination can achieve with regard to the assessment of practical skills. Much depends on the precise definition of 'practical skills' that is adopted. The term is sometimes reserved only for describing the skills associated with the physical manipulation of objects and apparatus (Abrahams et al., 2013; Gott & Duggan, 2002), but the definition may also incorporate other knowledge, skills and understanding necessary for planning and presenting practical work (SCORE, 2014).

Gott & Duggan (2002) provide some useful terminology for discussing the various elements of practical work. They essentially group the skills and knowledge necessary for practical work into three broad categories: 'conceptual understanding', 'process skills' and 'practical skills'. Conceptual understanding refers to knowledge of the substantive scientific concepts which underpin or give purpose to the practical activity (eg thermodynamics). Process skills are defined as generic skills that are transferable between practical activities, such as following instructions and applying principles for ensuring reliable measurement (eg understanding error). Practical skills are defined purely in terms of the physical performance of specific practical techniques (eg titration).

It is important to state that Gott & Duggan (2002) define these 3 categories to allow for conceptual clarity when discussing practical work. They do not state that practical skills, process skills and conceptual understanding can easily be isolated from one another for the purposes of teaching or assessment. Arguably, such categories are inseparable in practice and, therefore, in terms of valid assessment (Harlen, 1999).

To give an example, an IAPS exam question might describe a scenario in which an electrical circuit is constructed to measure the potential difference across a resistor. The question may require the student to explain how they would set up the circuit (indirectly assessing a practical skill), then to explain how they would account for measurement error (assessing a process skill) and finally to demonstrate some conceptual understanding of potential difference and current.

Arguably, it would be difficult to write a question which focussed purely on the practical or process skills because the context (measuring potential difference) is a necessary part of the question. Without any such context a question would be rather abstract and perhaps, therefore, not a particularly valid way to assess the targeted skills (which, in the real world, are only ever applied in context). Equally, the use of a real world context for framing a question about potential difference may be a good way of allowing a student to engage their conceptual knowledge on the topic.

In general, despite fairly wide use in science qualifications (Abrahams et al., 2013), there is a dearth of evidence regarding how best to conduct IAPS with regard to written examinations. Brown, Pacini & Taylor (1992) compared two different methods of assessing practical skills in biology using correlational and principal components analysis. They found some evidence that teacher assessment was more valid (in terms of construct validity) than a terminal practical examination. However their study related only to a set of examinations from a single year and focussed largely on a narrow definition of practical skills. A follow up study (Brown & Moore, 1994) found that the teacher assessment method and the written paper were assessing somewhat different things and the exact skills assessed in the examination were dependent on how the questions were framed (eg the context in which the problem was presented).

Given the complexity in defining what exactly it is they should assess, it is probably not surprising that there are challenges in writing (and identifying) IAPS items. Even with relatively well defined criteria, classifying whether or not an examination question is assessing a practical skill is difficult because most questions require some degree of supporting knowledge to answer and some degree of contextualisation to ask.

Research objectives

The subject content provided by the Department for Education (2014) suggest a fairly broad and inclusive definition of practical skills for the new A levels, one which, arguably, incorporates practical skills, process skills and conceptual knowledge. The IAPS examination items are clearly intended to cover a wide range of skills and knowledge relating to practical work, not solely the assessment of competency with apparatus and instruments. For example, assessing the ‘application of mathematical concepts in a practical context’ is unlikely to require the student to have had ‘hands on’ experience of practical work, but such experience is likely be helpful for them in understanding the context of the question and formulating their response. The physical performance of such practical skills can only be validly assessed directly, which is the role of the endorsement.

The IAPS items are likely to be at least somewhat different in style and content to other items on the assessment. These differences may cause IAPS items to function in a manner that is statistically distinct, meaning that students perform better or worse on these items compared to other items on the test. Though this is a possibility, there are many things that affect the performance of students on a given item (Stiller et al., 2016) and it is notoriously difficult to predict the ‘difficulty’ of examination questions (Crisp & Green, 2013; El Masri, Ferrara, Foltz, & Baird, 2017; Pollitt, Ahmed, & Crisp, 2008). In other words, there are many variables which cause differences in the functioning of items which may make it difficult to discern between non-IAPS and IAPS items.

Despite this complexity, it is important to consider how these new IAPS items have functioned in the first year of their introduction. The analysis which follows compares the overall student performance on IAPS examination items with the overall performance on the other (Non-IAPS) items in the summer 2017 examinations. The intention is to engage with the following research questions regarding how the IAPS items have functioned:

1. How did the IAPS items generally function in relation to other items in the reformed A level examinations with regard to their difficulty (eg item facility)?

2. What was the relationship between students' practical endorsement grade and their performance on the IAPS items?

3. Do the IAPS items measure something which is statistically distinct (a unique underlying construct) within the A level science examinations?

In terms of the above research questions, this report will provide an overview of the functioning of the IAPS items across subjects and specifications, allowing us to consider and discuss the (initial) impact of the new assessment arrangements at a systemic level. Note that the intention of this report is not to explore the extent to which individual IAPS items are suitable for assessing practical skills, nor is it to evaluate the extent to which examination boards are complying with the current conditions and guidance[2]. This is something Ofqual will continue to monitor on an ongoing basis, but it is not the focus for this report. For this reason, the names of examination boards and their specifications are anonymised. The focus of this report is the 'big picture'.

[2] Indeed, for this study we use the examination boards' classification of whether or not an item is IAPS.

Analysis

Data and overview

Item level data and supporting information for 16 science A level specifications were gathered from 4 examination boards: AQA, OCR, Pearson and Eduqas. Assessment materials (assessment grids, mark schemes and specifications) were used to identify all items which were specified by the exam boards for the IAPS.

For a minority of items, only a percentage of the total marks were identified for IAPS (eg. 2 marks out of 5). For these items, it is not usually possible to ascertain whether a student has attained credit for the IAPS element of their response or some other element (e.g. if a student were to attain 3 out of 5 we would not know whether they had attained 0, 1 or 2 marks for IAPS). In such cases, items were flagged as IAPS only if 50% or more of the available marks were allocated to IAPS in the examination board's assessment grid.

The final dataset comprised 2,430 assessment items from 65 components of 16 specifications. Across all of the components, data from 84,475 unique students (taking at least one science A level) were analysed3. The full dataset included 120,447 unique entries to a science exam (this counts any students who took more than one science A level multiple times). The maximum available mark for each item (IAPS and non-IAPS) ranged between 1 and 25 but approximately 69% of items had a maximum mark of either 1 or 2.

A total of 459 (around 19%) of the items that were analysed were classified as IAPS (note this is the number of *items* that were IAPS, not the number of marks that were allocated to IAPS). The mean maximum available mark for non-IAPS items was 2.22, while for IAPS items it was 2.76, suggesting that IAPS items tended to have higher mark tariffs.

Item difficulty (facility)

The facility index is essentially a measure of how difficult the item was for the cohort of students who took the test. It is the mean item score divided by the total number of marks available for the item. Values therefore range between 0 and 1, with a value of 1 indicating that all students achieved full marks and a value of 0 indicating that none of the students achieved any credit (ie the entire cohort scored 0). For a 5 mark item, a facility of 0.4 would indicate that the mean mark was 2 out of 5.

Overall, the mean facility value was slightly lower for IAPS items (mean = 0.48, SD = 0.21) than non-IAPS items (mean = 0.55, SD = 0.21), suggesting that students found these items to be slightly more difficult on average ($t = 6.90$, $df = 690$, $p < 0.01$, Cohen's $D = 0.36$). The difference is statistically significant but small. As can be seen from the standard deviations, there is considerable variability in the item facilities for both IAPS and non-IAPS items. The boxplots in Figure 1 show the

[3] Note that for one of the exam boards, it was necessary to exclude a small proportion of the students (approximately 10%) from the analysis because of missing data for some of the items. This data appeared to be missing at random (eg the overall performance of the excluded participants was similar to that of the cohort as a whole).

spread of facility values for both item types (each boxplot shows the median value, inter-quartile range and maximum and minimum values).

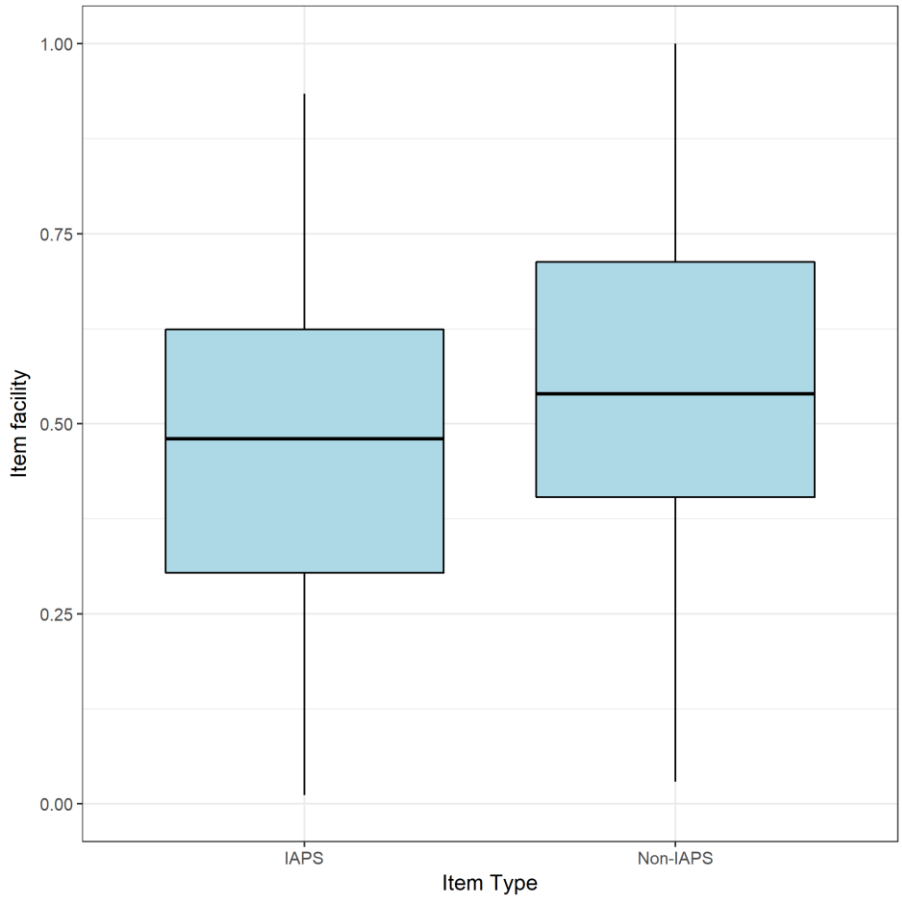

Figure 1. *Boxplot of item facility for IAPS & non-IAPS*

This overall difference does not tell us whether there was variation between subjects. The graphs below summarise the item facilities for IAPS and non-IAPS by subject (Figure 2) and then for each specification (Figures 3 to 5). The mean for each subject/specification and question type is plotted with the range of the data (which is 1 standard deviation either side of the mean). Annex B displays, for each subject and specification, whether these differences in facility between the IAPS and non-IAPS items are statistically significant.

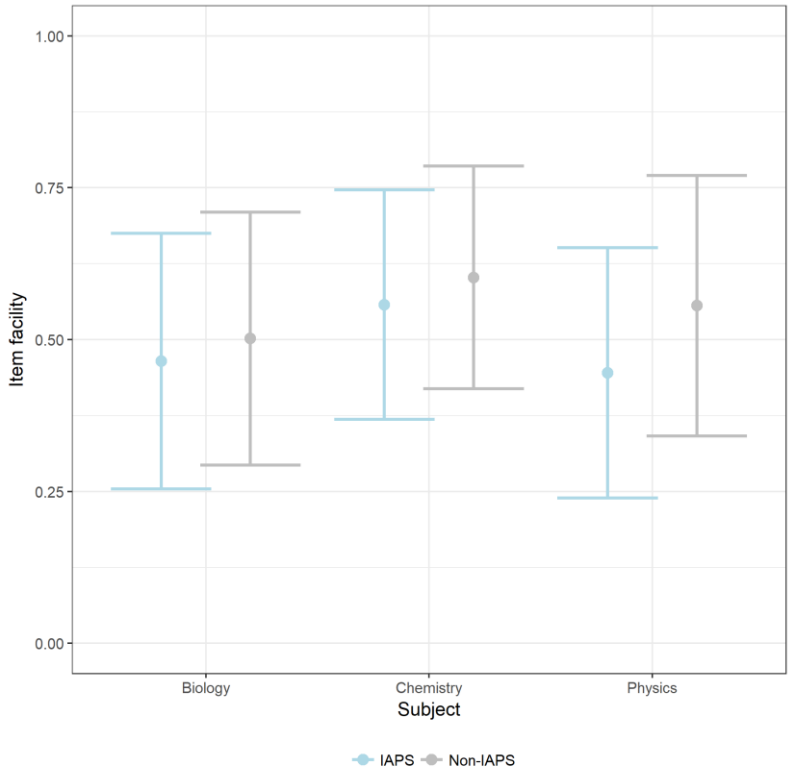

Figure 2. *Mean and standard deviation of facility index for IAPS and non-IAPS items by subject*

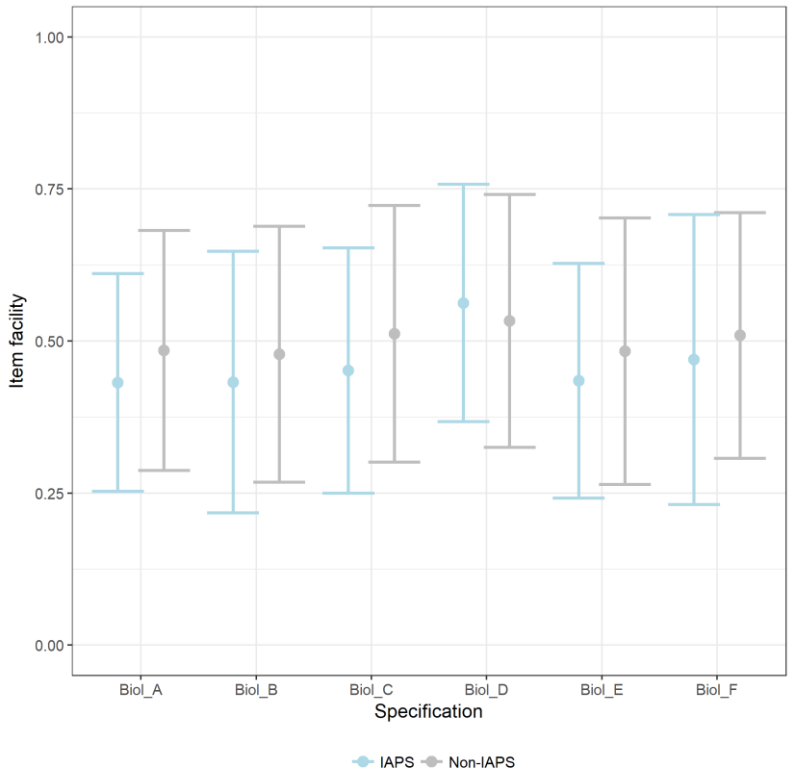

Figure 3. *Mean and standard deviation of facility index for IAPS & non-IAPS items by specification for biology*

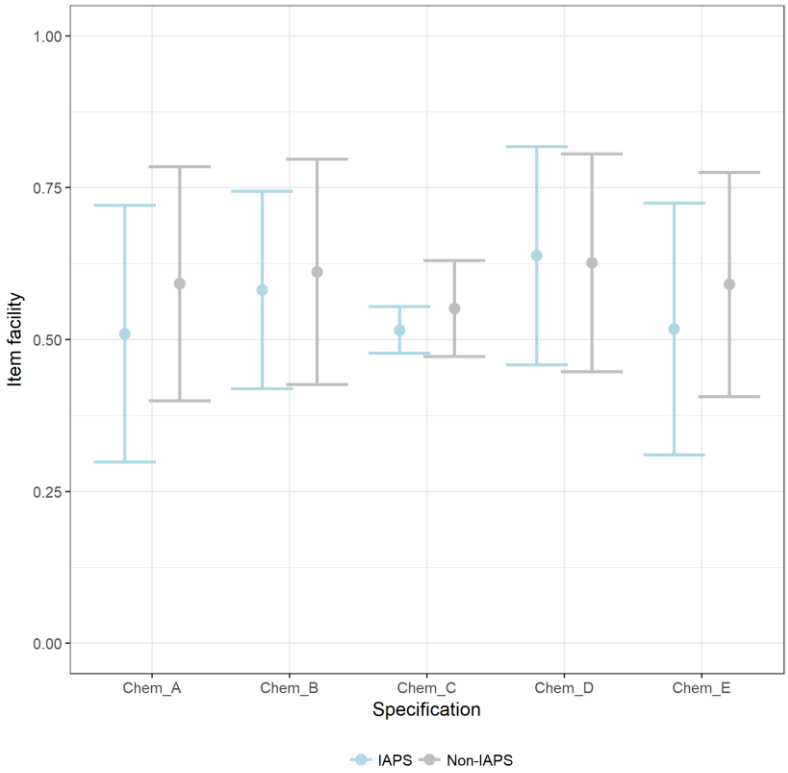

Figure 4. *Mean and standard deviation of facility index for IAPS & non-IAPS item by specification for chemistry*

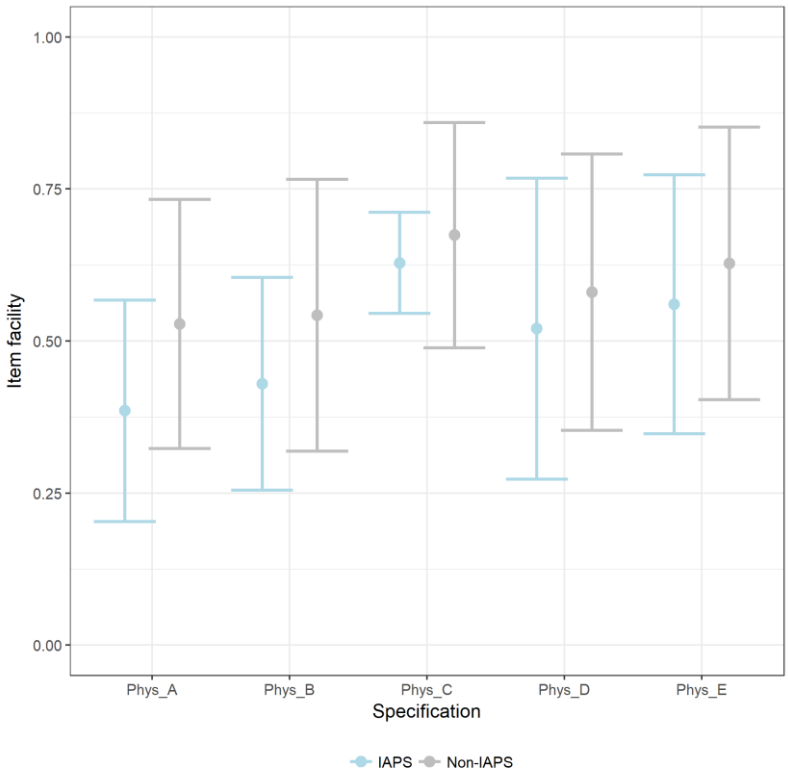

Figure 5. *Mean and standard deviation of facility index for IAPS & non-IAPS item by specification for physics*

To summarise, across all subjects and most specifications, there is a tendency for the IAPS items to have a slightly lower mean facility, indicating that the students found these items to be somewhat more difficult. At the subject level, the difference is negligible for biology, small but statistically significant for chemistry, and of a moderate size (and statistically significant) for physics (see Annex B).

At the specification level, differences of statistical significance are less likely given the smaller number of items (and therefore smaller sample size). For nearly all specifications, the mean facility value was higher for non-IAPS items than for IAPS items but the difference was only statistically significant for two of the physics specifications (Phys A and Phys B), both of which exhibited a medium effect size (and are therefore the main cause of the difference at subject level).

The only cases where IAPS items had a higher facility than non-IAPS items were Biol D and Chem D, and in neither case was the difference statistically significant. It is beyond the scope of this paper to explore exactly why these two specifications differed, or why the effect was significant (and stronger) for the two aforementioned physics specifications, but an exploration of how the IAPS items differ qualitatively in the context of these specifications would be worthwhile.

Before moving on, a word of caution. Item facility generally exhibits a high level of variance, regardless of whether or not an item is indirectly assessing practical skills (the facility values across all items in this study ranged from 0.01 to 1.00). This is perhaps to be expected because there are likely to be many variables that influence the difficulty of an item, not just whether or not it is indirectly assessing practical skills. Although there is a tendency for the IAPS items to be more difficult, it is not clear whether this is driven by the level of experience candidates have with practical work or some other feature that is more common in IAPS items than non-IAPS items.

The practical endorsement

Students who received a ‘Pass’ (P) grade for their practical endorsement were compared to those who had received a grade of ‘Not Classified’ (N) with regard to their performance on IAPS and non-IAPS items. Figure 6 compares the overall difference[4] in the mean performance of the two groups across subjects (the error bars represent 1 standard deviation either side of the mean). Those who pass the endorsement get a larger percentage of the available IAPS and non-IAPS marks across all subjects.

[4] Note that endorsement data was not available for one of the biology specifications (Biol C).

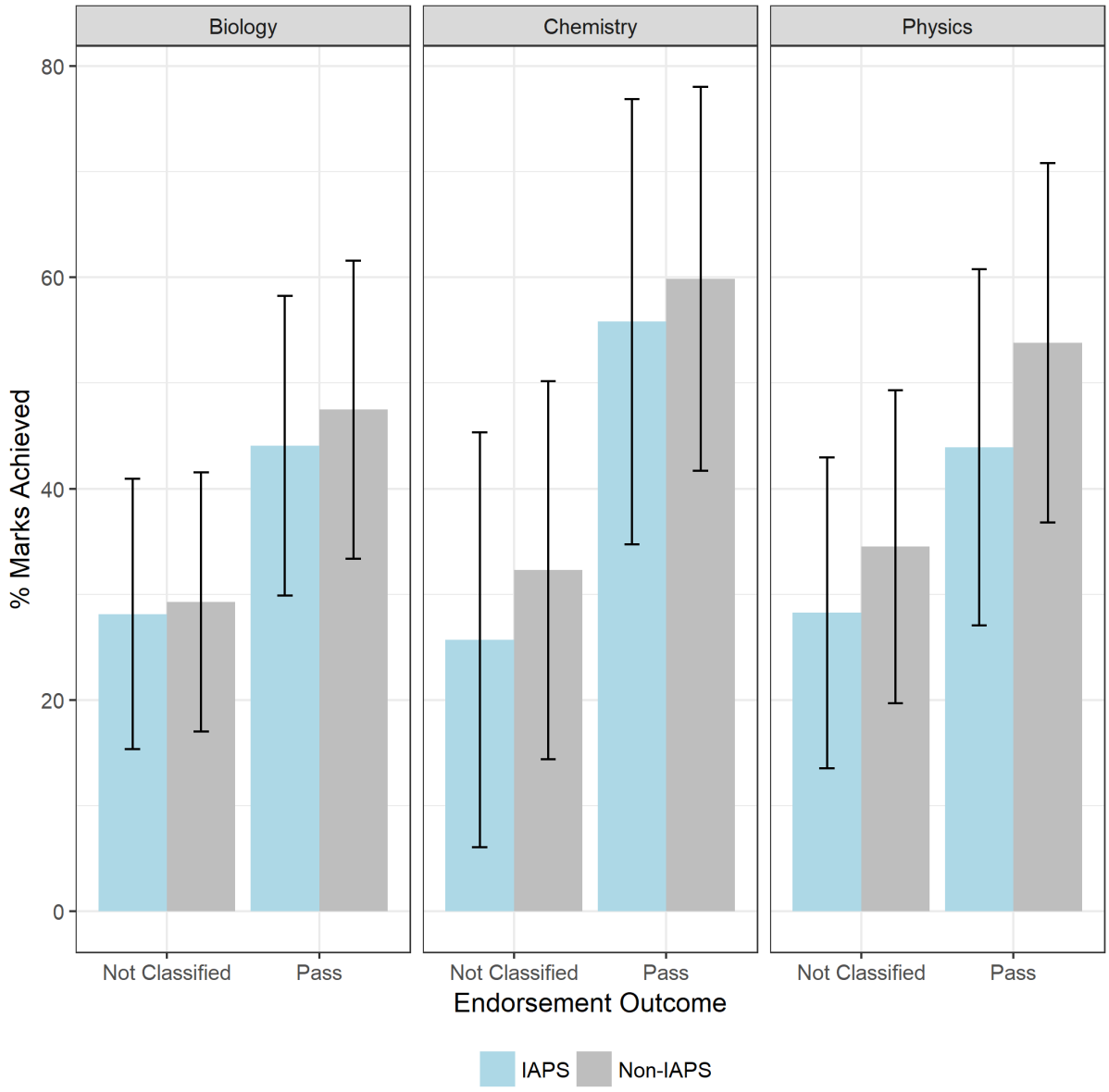

Figure 6. *Mean percentage of marks achieved for IAPS & non-IAPS items by practical endorsement grade*

As we saw in Table 1, those students who did not pass the endorsement were more likely to receive lower grades at A level than those who passed, hence their relatively low performance across both IAPS and non-IAPS item types. Figure 7 is therefore more helpful as its breaks performance down by grade. When the data is explored in this way, there appears to be little difference in how well those who received a P performed on the IAPS items relative to those who received a N. The only statistically significant differences were at grades C and U, where those who received a P outperformed those who achieved a N, though in both cases the size of the effect was small (see Annex C for a summary of statistical comparisons).

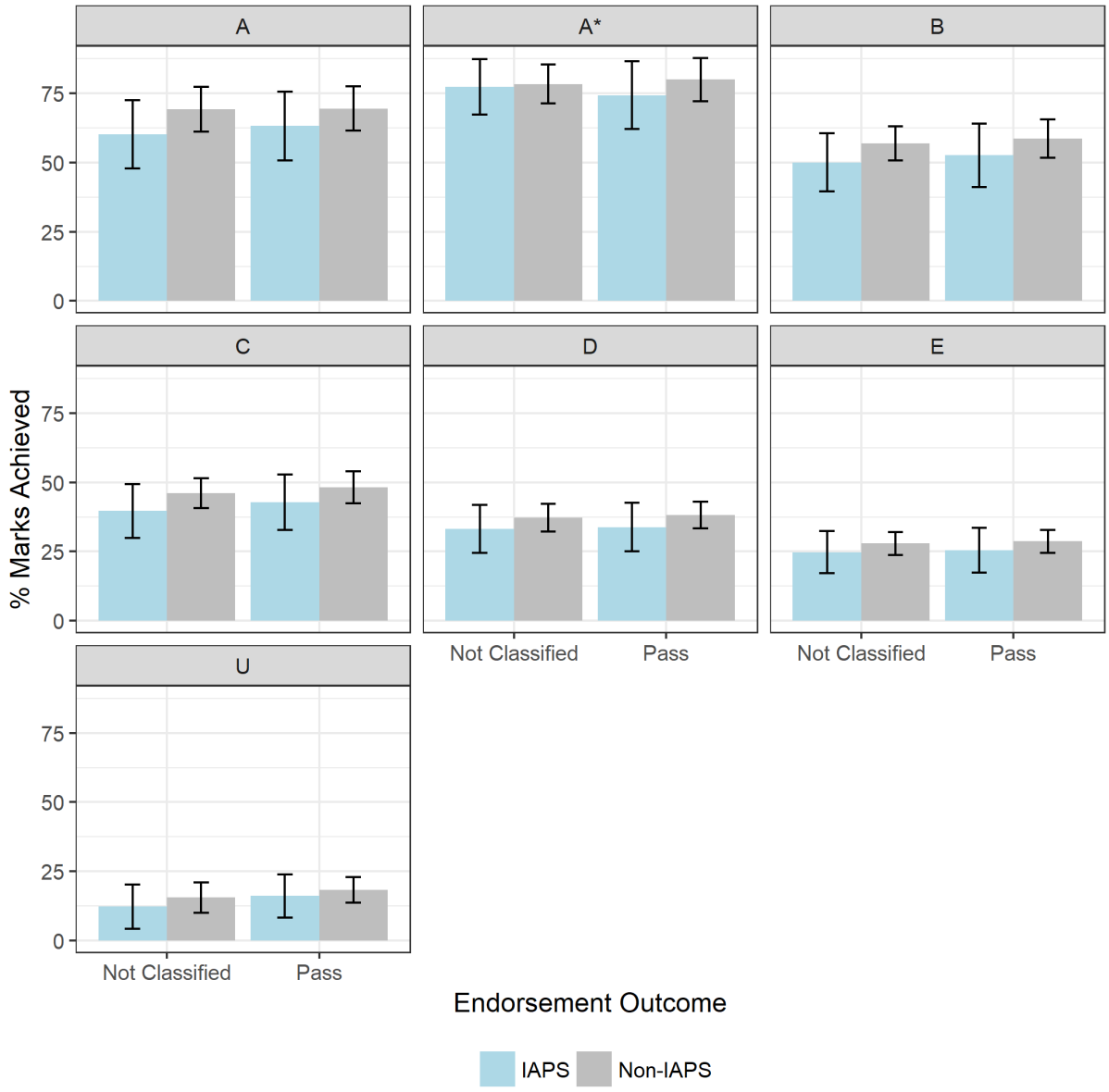

Figure 7. *Mean percentage of marks achieved for IAPS & non-IAPS items by practical endorsement grade for each A level grade*

The comparison is complicated by the asymmetry of the two groups – less than 1% of candidates received an N for the endorsement, making them very much the minority cohort. In an effort to account for this, propensity score matching was used to create a comparable sample of students who had attained a P but had received similar overall marks and grades (across the same specifications). Only specifications where at least 50 students had received a N for the endorsement were included in this analysis. This left 2 specifications for each subject in the analysis (6 in total). Following this matching process, the characteristics of the two groups can be seen in Table 2.

Table 2. *Characteristics of 'Not Classified' group and matched 'Pass' group*

	Grade N	**Matched Grade P**
Mean mark (% total)	30.45	30.48
Number candidates Biol A	260	263
Number candidates Biol E	182	184
Number candidates Chem A	135	135
Number candidates Chem D	118	120
Number candidates Phys A	199	200
Number candidates Phys D	104	96

Those who received a P and those who received a N in the endorsement did not differ to a statistically significant degree in their performance on the IAPS items, as can be seen in Figure 8 and Figure 9 (and in the table in Annex D).

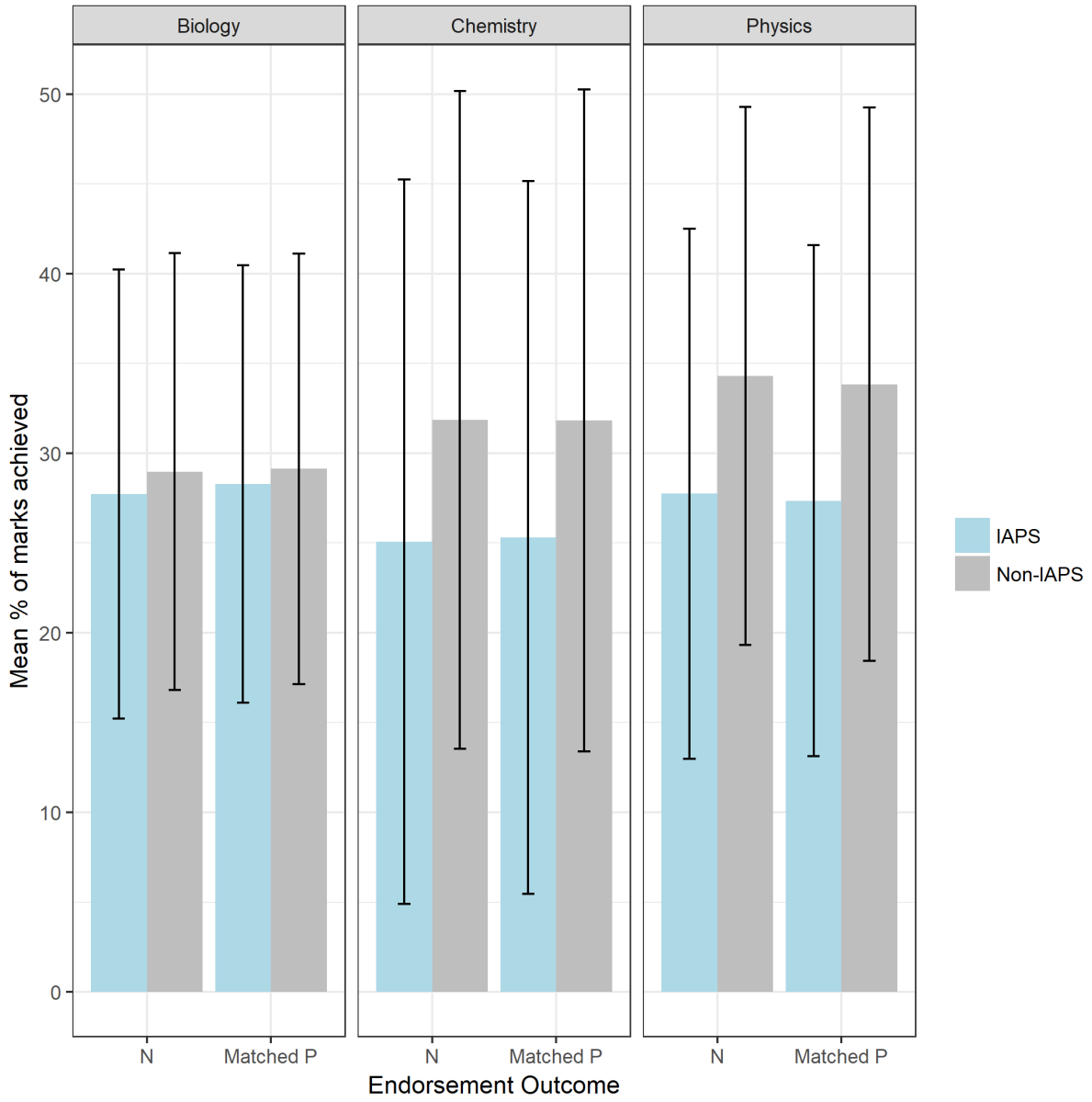

Figure 8. *Mean percentage of marks achieved for IAPS & non-IAPS items by practical endorsement grade for each subject (for propensity matched candidates)*

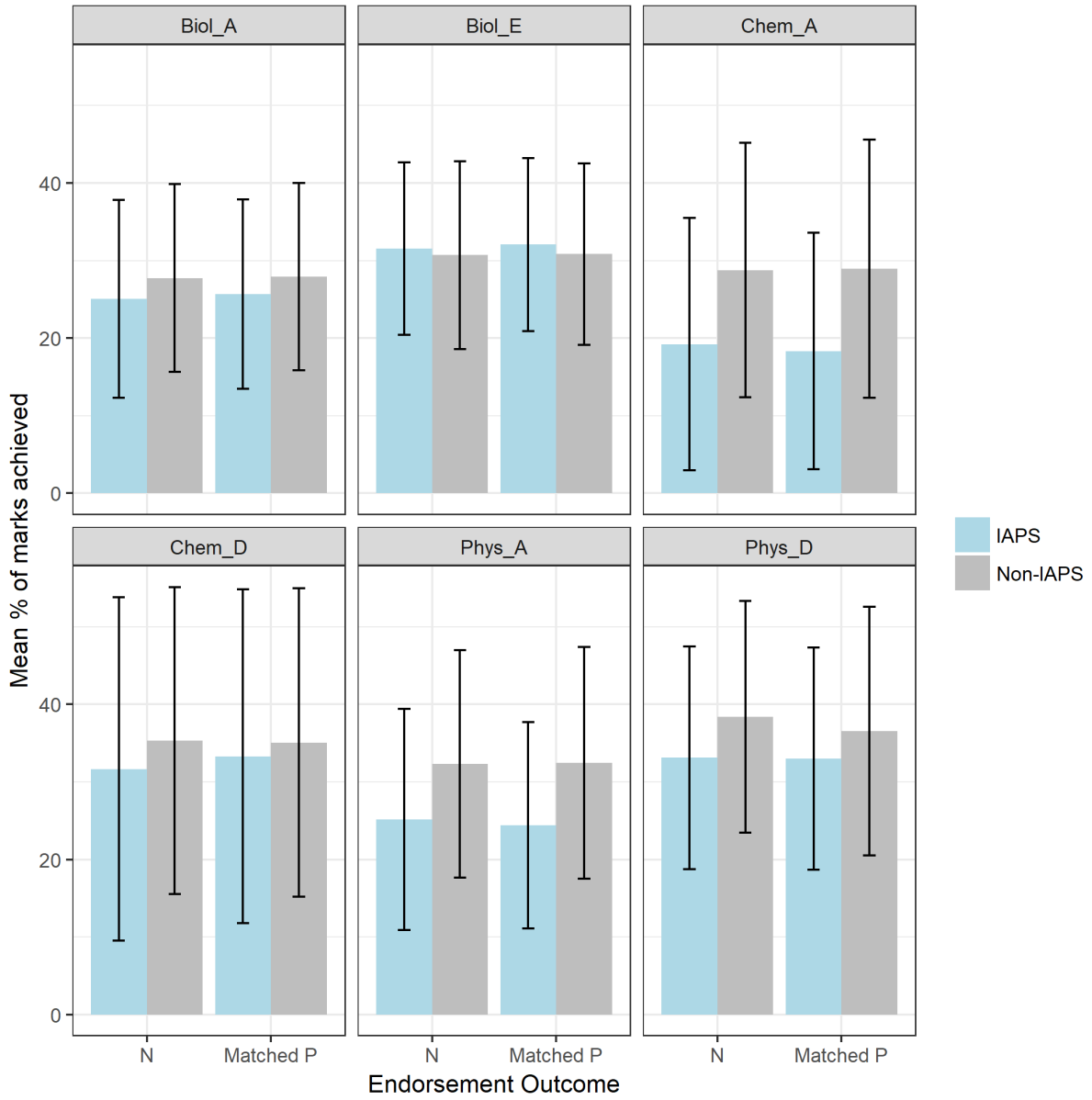

Figure 9. *Mean percentage of marks achieved for IAPS & non-IAPS items by practical endorsement grade for each specification (for propensity matched candidates)*

Again, caution is necessary when interpreting this result. As discussed earlier in this report, the vast majority of students for each A level science qualification receive a 'Pass' for the practical endorsement (99%), suggesting that in normal circumstances students were able to work towards the required standard by the end of the course. It is not possible to identify *why* a student did not achieve a 'Pass' for the practical endorsement, but reasons may include:

- The student failing to demonstrate the required standard with regard to one or more of the five assessment criteria (a genuine failure to meet the standard)

- The school failing to provide sufficient opportunity for the student to complete the required practical activities or else failing to gather sufficient evidence of the student's performance

- The student being unable to complete the required practical activities due to regular absence, disability or some other impairment

In some of the cases above, the student may have carried out quite a lot of practical work but not achieved a Pass, either for administrative reasons or for failing to meet just one of the 5 assessment criteria. In such cases their performance on IAPS items may not be significantly impaired. In summary, those who passed the endorsement did not do substantially better on the IAPS items than comparable students who did not pass, but there are a variety of possible reasons for this.

Principal component analysis

Principal component analysis (PCA) was conducted for each of the specifications. The purpose of PCA is to identify patterns of correlation in a dataset. In this case PCA was conducted to ascertain whether sets of items cluster together into groups that represent underlying constructs in student performance (these groups are known as principal components or *factors*). If the IAPS items were to cluster in a separate factor to the non-IAPS items, it might suggest that they were assessing a unique underlying construct (ie practical skills). It is important to point out that this is not necessarily the outcome one would expect for the reformed A level qualifications. As discussed in the introduction section, it may be desirable for IAPS items to assess practical skills *in the context of* scientific knowledge and understanding. This would mean that IAPS items would be less likely to cluster to a single factor as they would be more closely integrated with the rest of the assessment.

To prepare for the analysis of each specification, data from those students who had not completed all of the assessment components were removed, as were data from any optional items (eg items which were not completed by all of the students). Pairwise exclusion of missing data points was employed to account for data that was missing at random (eg where a student was missing data for a specific item). Each PCA was conducted using oblique rotation (oblimin) to account for the expectation that performance on extracted factors was likely to correlate (ie one might expect a student who did well on one set of items to also do well on another).

To decide on how many factors to extract for each specification, the eigenvalues for each potential factor and the scree plot were considered in tandem. In addition to these bespoke models, a model where only 2 factors were extracted was also produced. The intention behind these '2 factor' models was to explore the hypothesis that IAPS items are assessing a different underlying construct to non-IAPS items (ie IAPS items generally load on to one of the two factors while non-IAPS items generally load on to the other). The 2 factor models are likely to over-simplify the true component structure of the data but they will identify if there is a clear distinction between IAPS and non-IAPS items in a given specification.

This report will not present each of the PCAs - this would not be in service of the high level research questions. What follows is a summary which seeks to provide an insight into the degree to which IAPS items are measuring a statistically distinct

construct within A level science specifications, so the focus is on the '2 factor models' described above. For each subject there are a pair of tables. The first provides an overview of the PCA for each specification, providing:

- the number of IAPS and non-IAPS items[5]
- the Kaiser-Meyer-Olkin (KMO) value, a measure of sampling adequacy for the model that ranges between 0 and 1 (sampling is generally considered to be adequate where the value is greater than 0.8)
- the number of components with an eigenvalue greater than 1.0 (this is generally considered to be the threshold for extracting a component, though Field, Miles & Field (2012) suggest using the scree plot where the number of items in the analysis is high)
- the number of components to extract based on the scree plot[6]
- the proportion of the total variance across items which is explained by the 2 factor model

The second table for each subject shows the proportion of IAPS and non-IAPS items that loaded on to each of the two extracted factors (F1 and F2) for the 2 factor model for each specification[7]. By looking across the rows it is possible to judge the extent to which IAPS and non-IAPS items are loading on to different factors for each specification. It is important to note that an item can load on to one factor, either factors, or both. Table 3 and Table 4 show this information for biology.

Table 3. *Overview of principal component analyses for A level biology specifications*

Spec.	No. IAPS items	No. Non-IAPS items	KMO	No. factors eigenvalue >1.0	No. factors suggested by scree plot	Total variance explained by 2 factor model
Biol A	19	89	0.99	12	3	0.20
Biol B	28	93	0.96	35	4	0.19
Biol C	19	97	0.97	27	4	0.17
Biol D	21	93	0.95	33	5	0.16
Biol E	32	95	0.99	21	5	0.16
Biol F	39	90	0.70	45	4	0.16

[5] Note that this shows the number of items identified as IAPS, not the number of marks allocated to IAPS.
[6] This indicates the number of statistically distinct components which underpin each specification, based on the eigenvalues and the scree plot. These models are not discussed in depth because specification specific analysis
[7] Items with a loading of 0.3 or more were considered to have loaded to the factor (see Field et al., 2012).

Table 4. *Proportion of IAPS and non-IAPS items loading (after oblimin rotation) on to each factor by biology specification (2 factor model)*

Specification	Proportion IAPS items with loading greater than 0.3 on each principal component		Proportion Non-IAPS items with loading greater than 0.3 on each principal component	
	F1	F2	F1	F2
Biol A	0.31	0.37	0.49	0.22
Biol B	0.19	0.57	0.48	0.11
Biol C	0.11	0.53	0.42	0.21
Biol D	0.14	0.29	0.42	0.17
Biol E	0.25	0.22	0.44	0.14
Biol F	0.10	0.36	0.41	0.19
Mean biol.	**0.18**	**0.39**	**0.44**	**0.17**

In the case of biology, there does appear to be some tendency for IAPS items to cluster onto a different factor to the non-IAPS items in at least 4 of the 6 specifications. Taking Biol B as an example, 57% of IAPS items loaded on to the second factor F2 compared to only 19% on F1. For the non-IAPS items, the pattern is the reverse – nearly half of non-IAPS items load to F1 while only 11% load to F2.

This result should be interpreted very cautiously given that the model only explains 19% of the total variance in student performance across the specification. This suggests that the 2 factor model is not doing a particularly good job of explaining variance in student performance across all the items. This is not simply a result of too few factors being extracted to build the model. For example, even if a 5 factor model is constructed (ie including all factors with an eigenvalue greater than 1.5) only 22% of the variance is explained.

Table 5 and Table 6 display the same information for the chemistry specifications[8]. There does not seem to be an obvious pattern regarding the factor loadings. All items, regardless of whether or not they are IAPS, are more likely to load to F1. This suggests that the students' performance on IAPS items do not correlate strongly together and do not therefore appear to be assessing a statistically distinct construct. In other words, an IAPS item is as likely to group with non-IAPS items as it is with other IAPS items.

Table 5. *Overview of principal component analyses for A level chemistry specifications*

Spec.	No. IAPS items	No. Non-IAPS items	KMO	No. factors eigenvalue >1.0	No. factors suggested by scree plot	Total variance explained by 2 factor model
Chem A	25	116	0.99	14	3	0.25
Chem B	28	97	0.99	20	3	0.27
Chem D	19	103	0.99	12	3	0.25
Chem E	20	112	0.98	28	4	0.22

Table 6. *Proportion of IAPS and non-IAPS items loading (after oblimin rotation) on to each factor by chemistry specification (2 factor model)*

Specification	**Proportion IAPS items with loading greater than 0.3 on each principal component**		**Proportion Non-IAPS items with loading greater than 0.3 on each principal component**	
	F1	F2	F1	F2
Chem A	0.52	0.36	0.47	0.42
Chem B	0.43	0.46	0.43	0.36
Chem D	0.74	0.21	0.54	0.28
Chem E	0.60	0.25	0.60	0.13
Mean chem.	**0.57**	**0.32**	**0.51**	**0.30**

Finally, Table 7 and Table 8 display the PCA information for the 5 physics specifications.

Table 7. *Overview of principal component analyses for A level physics specifications*

Spec.	**No. IAPS items**	**No. Non-IAPS items**	**KMO**	**No. factors eigenvalue >1.0**	**No. factors suggested by scree plot**	**Total variance explained by 2 factor model**
Phys A	23	103	0.99	9	4	0.19
Phys B	20	85	0.98	19	4	0.21
Phys C	11	70	0.98	19	4	0.39
Phys D	23	104	0.99	18	3	0.21
Phys E	30	95	0.96	35	5	0.18

Table 8. *Proportion of IAPS and non-IAPS items loading (after oblimin rotation) on to each factor by physics specification (2 factor model)*

Specification	Proportion IAPS items with loading greater than 0.3 on each principal component		Proportion Non-IAPS items with loading greater than 0.3 on each principal component	
	F1	F2	F1	F2
Phys A	0.09	0.65	0.39	0.24
Phys B	0.30	0.20	0.40	0.32
Phys C	1.00	0.00	0.58	0.16
Phys D	0.39	0.35	0.53	0.21
Phys E	0.50	0.10	0.63	0.16
Mean phys.	**0.46**	**0.26**	**0.51**	**0.22**

The picture for physics is similar to that of chemistry – there is not strong evidence for IAPS items clustering together on one of the two factors in the model. Phys A is possibly the exception, as 65% of IAPS items loaded on to the second factor F2 compared to only 24% on F1 (for the non-IAPS items, 30% load to F1 while 24% load to F2), but this is not clear cut. Again, the model only explains a relatively small proportion (24%) of the variance.

Overall, it is notable that for all of the PCAs there were a large number of factors with high eigenvalues (for example, the number of factors with eigenvalues greater than 1.0 ranged between 9 and 45). This is partly a result of the large sample size for each specification. However, this might also be taken to suggest that the exams were assessing a variety of constructs and the PCA was not able to adequately disentangle them. This theory is given extra weight by the fact that the 2 factor PCA models generally explain little of the overall variance (between 13% and 39%).

It is also important to note that, although the focus of this section has been on the 2 factor models, the bespoke models for each specification (whereby the scree plot suggested the extraction of between three and five factors) produced similar results. None of the bespoke models suggested that IAPS items were likely to load on to a particular factor, mirroring the findings for the 2 factor models.

Discussion

Summary of findings

For the written A level science examinations in 2017, the IAPS items were, on average, slightly more difficult for the cohort than the other (non-IAPS) items. This is caveated by the fact that both IAPS and non-IAPS items were highly variable in terms of their difficulty – IAPS items exhibited a range of difficulties, with students doing very well on some and very poorly on others. The difference between the mean difficulty of IAPS and non-IAPS items at the subject level was generally either small or medium[9] (in terms of effect size, see Annex B) and, with a few exceptions, there was no statistically significant difference at the specification level. The effect seemed to be more prominent for physics, perhaps owing to differences observed for 2 of the 5 specifications in particular. It is beyond the scope of this report to explore why the relative difficulty varied between specifications. The differences may reflect the varying approaches taken by exam boards when writing IAPS items, differing approaches to linking the practical endorsement to the examinations, different characteristics in the students who are taking each specification, or perhaps something else entirely.

At the system level, it is perhaps not surprising that the IAPS items appear to have been relatively difficult in the 2017 exams. The reformed science A levels are new qualifications and both teachers and students are encountering IAPS items of this type for the first time. It may be that performance on IAPS items will change in future years as teachers become more familiar with the new courses and examinations. Such future improvement in the relative performance of IAPS items could be driven by a number of factors: an increase in the type and frequency of (relevant) practical work, increased familiarity with regard to IAPS items, or a greater breadth of practice materials (eg past papers).

An alternative theory is that IAPS items tend to be naturally more difficult. Indeed, the mean maximum available mark was higher for IAPS items than for non-IAPS items, suggesting that they tend to have a higher tariff. This means that IAPS items are likely to place a different type of demand on the student, perhaps having a tendency to require a more detailed, complex or synoptic response. There is no expectation that IAPS items should be of equal difficulty to other items in the assessment and it may be valid and appropriate that they have a tendency to be relatively stretching for students.

Despite these theories, care should be taken when interpreting the item facility data. Though the results tell us is that there was a tendency for IAPS items to be somewhat more difficult than non-IAPS items in 2017, they do not tell us anything about the source of that difficulty.

The second part of the analysis in this report suggests that those who did not achieve a ‘Pass’ for their practical endorsement did not generally perform either better or worse on the IAPS items compared to those who did pass. Though an interesting finding, it is difficult to establish a hypothesis about the extent to which

[9] For guidance, a Cohen’s D value of less than 0.3 is considered to be a small effect, between 0.3 and 0.6 is considered to be a medium effect, and a value of greater than 0.7 is considered a large effect (Field et al., 2012).

one would expect these two assessment outcomes to be associated. On the one hand, we might assume that those students who did not pass the endorsement missed out on high-quality training in practical skills across a broad range of topics and therefore we would expect them to perform relatively poorly on IAPS items. On the other hand, IAPS items are intended to assess somewhat different skills and subject content to the endorsement and also employ a substantially different method of assessment.

The Department for Education (2014) delineates skills that are to be assessed directly by the endorsement and those that are to be assessed indirectly in the examination. There is a clear intention for the 2 assessment methods to cover somewhat different ground. It is therefore worth further considering what is expected from IAPS items. If the intention is for them to indirectly assess practical skills alongside other subject content and in the context of other knowledge and skills, it is perhaps unsurprising that performance on the endorsement does not appear to strongly relate to performance on IAPS items.

For example, one of the criteria for the endorsement is ‘Safely uses a range of practical equipment and materials’. This is a competency that cannot be assessed validly in a written examination because it requires the performance of hands-on practical work. The intention is to assess whether the student has achieved a particular level of proficiency (competence) by the end of their course and the outcome is binary; the student has either demonstrated their competency or they have not. Reflecting this competency based approach, the pass rate for the endorsement in 2017 was high (99%). This means that, by the end of their course, most students had been able to demonstrate and evidence the required level of competency.

By way of contrast, the IAPS examination items are allocated multiple marks within the framework of a wider examination and are intended to differentiate across a broader range of performance. There is no minimum expectation of competence for these items because examinations are compensatory. The student can accumulate the marks required to reach a particular grade boundary through sufficient performance on any combination of items within the assessment, they do not necessarily need to perform strongly on the IAPS items (though, arguably, they may need to if they wish to accumulate enough marks for the highest available grades).

A student may therefore do relatively well on the IAPS items despite separately failing to meet one or more of the five assessment criteria (CPAC) required for the endorsement. Those who fail to pass the endorsement may actually have experienced all of the required practical work ahead of the exam. They may have good overall practical skills but fallen over one of the hurdles that all students are required to clear in order to pass the endorsement. Equally, the opposite may be true. A student may have passed the endorsement by meeting all of the criteria but may perform relatively poorly when responding to IAPS items in the context of a written examination.

Another issue is that some of the students may have failed to pass the endorsement for administrative reasons rather than due to a lack of aptitude for practical work. Perhaps their school or college did not provide all of the necessary evidence against one or more assessment criteria. As discussed earlier in this paper, the potential reasons for failing to achieve a pass for the endorsement extend beyond failing to achieve the required standard across all of the 5 CPAC assessment criteria.

Finally, the principal component analyses that were conducted on each of the specifications found that, with a few exceptions, IAPS items were not clustering together to assess a single underlying construct. Though it may be appealing to consider IAPS items to be assessing a common unidimensional construct (eg practical skills) this, in reality, would be an oversimplification. IAPS items are not indirectly assessing practical skills in isolation from other knowledge and skills and the variety of content that IAPS items cover probably further serves to reduce the inter-correlation of the items. Even if IAPS items were indirectly assessing practical skills and nothing else, there are a variety of specific skills and techniques that are encompassed in the subject content and it is easy to imagine students having varying levels of proficiency across these.

Alternatively, it may simply be that a student's performance on IAPS items is closely related to their performance on the other items in the test; if the student does relatively well on practical skill items they also do relatively well on other items.

Limitations

There are two significant limitations to these findings. First, as discussed in the method section, this analysis relies on the exam boards' identification of IAPS items. Though this is unlikely to be a significant issue, it is possible that there may be some cases where items have been incorrectly identified as IAPS, or items that should have been identified were not. There is also the issue of coding items for which only some of the available marks are allocated to IAPS. In short, the coding of items as either IAPS or non-IAPS is in some ways subjective, which is likely to undermine the precision of the analysis.

Second, our analysis is necessarily very broad. It seeks to explore how IAPS items have performed at a system level but it does not engage with the reasons for differences between subjects and specifications in depth, nor does it explore the qualitative features of IAPS items. The PCA does not seek to investigate why particular factor structures emerge for each subject and specification because this would require a far more detailed understanding of the individual items.

Conclusions

For the reformed A level science qualifications, the introduction of indirect assessment of practical skills through written examinations was intended to help facilitate and encourage the teaching of a broad range of practical skills and to embed practical work more firmly within the course content. The IAPS items should therefore be more challenging for a student if they have not experienced a wide range of relevant practical work during their course.

IAPS items cover a diverse range of skills and knowledge – they do not assess practical skills in isolation because they are contextualised by real world scenarios and scientific knowledge. Unlike the direct assessment of practical skills, indirect assessment does not (and cannot validly) assess 'hands-on' practical skills. Instead, the focus is on the process skills and conceptual understanding that one may acquire through conducting a range of practical work. The direct assessment of practical skills which is targeted by the endorsement necessarily covers different ground, which makes the comparison of student outcomes across the two assessment approaches problematic.

With this in mind, and the very small proportion of candidates who did not achieve the practical endorsement, it is probably not surprising that there was not a strong relationship between candidates' endorsement outcomes and their performance on IAPS items, or that the IAPS items were generally too diverse to statistically represent a single unidimensional construct. The IAPS items cover a range of skills, are underpinned by a breadth of scientific concepts, and are presented across a range of contexts. They tend to be a bit more difficult than non-IAPS items, which may reflect their synoptic nature, or perhaps that the style and format of the items is new and unfamiliar in these first examinations of the reformed qualifications.

This work serves as reminder that the endorsement and the examination questions are in fact assessing 2 different aspects of practical skills, rather than the same thing in different ways. The endorsement seeks to assess hands-on practical skills in a laboratory environment in the most valid way possible. The IAPS items in the examinations seek to assess the process skills that underpin practical skills and the application of knowledge and experience gained through practical work. This is something which, arguably, can be validly achieved in a written examination.

How effectively this is being achieved, and the washback effects that these items are having on teaching and learning, is something which should continue to be monitored and considered. Exam boards are pioneering these items and are likely to refine them over time. Ofqual will continue to monitor the new qualifications against the current conditions, and will keep an open mind with regard to the fitness for purpose of those conditions.

References

Abrahams, I., & Reiss, M. J. (2015). The assessment of practical skills. *School Science Review*, *96*(June), 40–44.

Abrahams, I., Reiss, M. J., & Sharpe, R. M. (2013). The assessment of practical work in school science. *Studies in Science Education*, *49*(2), 209–251.

Brown, C. R., & Moore, J. L. (1994). Construct Validity and Context Dependency of the Assessment of Practical Skills in an Advanced Level Biology Examination. *Research in Science & Technological Education*, *12*(1), 53–61. http://doi.org/10.1080/0263514940120107

Brown, C. R., Pacini, D. J., & Taylor, D. J. (1992). Two Different Methods of Assessing Practical Skills at an Advanced Level Examination in Biology: demonstration of construct validity or the appraisal of non-events? *Research in Science & Technological Education*, *10*(1), 23–35. http://doi.org/10.1080/0263514920100103

Crisp, V., & Green, S. (2013). Teacher views on the effects of the change from coursework to controlled assessment in GCSEs. *Educational Research and Evaluation*, *19*(8), 680–699. http://doi.org/10.1080/13803611.2013.840244

Department for Education. (2014). GCE AS and A level subject content for biology, chemistry, physics and psychology. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/593849/Science_AS_and_level_formatted.pdf

El Masri, Y. H., Ferrara, S., Foltz, P. W., & Baird, J.-A. (2017). Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *The Curriculum Journal*, *28*(1), 59–82. http://doi.org/10.1080/09585176.2016.1232201

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London: Sage Publications.

Gott, R., & Duggan, S. (2002). Problems with the Assessment of Performance in Practical Science: Which way now? *Cambridge Journal of Education*, *32*(2), 183–201.

Gove, M. (2013). Letter from the Secretary of State for Education to Glenys Stacey at Ofqual. Retrieved from https://www.gov.uk/government/publications/letter-from-the-secretary-of-state-for-education-to-glenys-stacey-at-ofqual

Harlen, W. (1999). Purposes and Procedures for Assessing Science Process Skills. *Assessment in Education: Principles, Policy & Practice*, *6*(1), 129–144.

Ofqual. (2016). *GCE subject level conditions and requirements for science (Biology, Chemistry, Physics) and certificate requirements*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/526286/gce-subject-level-conditions-and-requirements-for-science.pdf

Ofqual. (2017a). Guide to AS and A level results for England, 2017. Retrieved May 25, 2018, from https://www.gov.uk/government/news/guide-to-as-and-a-level-results-for-england-2017

Ofqual. (2017b). *The impact of qualification reform on A level science practical work*

- Paper 1: Teacher perspectives after one year'. Retrieved from https://www.gov.uk/government/news/the-impact-of-qualification-reform-on-a-level-science-practical-work

Ofqual. (2018). *The impact of qualification reform on the practical skills of A-level science students. Study 2: Pre and Post reform evaluation of practical skills*. Retrieved from https://www.gov.uk/government/publications/evaluation-of-qualifications-reform

Pollitt, A., Ahmed, A., & Crisp, V. (2008). The demands of examination syllabuses and question papers. In *Techniques for monitoring the comparability of examination standards* (pp. 166–206). London: Qualifications and Curriculum Authority.

SCORE. (2014). *SCORE principles: the assessment of practical work*. Retrieved from http://www.score-education.org/reports-and-resources/publications-research-policy

Stiller, J., Hartmann, S., Mathesius, S., Straube, P., Tiemann, R., Nordmeier, V., … Upmeier zu Belzen, A. (2016). Assessing scientific reasoning: a comprehensive evaluation of item features that affect item difficulty. *Assessment & Evaluation in Higher Education*, *41*(5), 721–732. http://doi.org/10.1080/02602938.2016.1164830

Annex A: Ofqual's A level science research programme

Reformed A level qualifications in science were introduced for first teaching in September 2015 (Gove, 2013). The reform led to significant changes to the assessment arrangements for practical skills (Ofqual, 2016). Ofqual is conducting a programme of research to evaluate the impact of A level qualification reform on the teaching and learning of science practical skills.

The programme is comprised of four main studies, of which this report is Study 4:

- Study 1: Teacher interviews – Perspectives on A level reform after one year
- Study 2: Pre and Post reform evaluation of practical ability – A comparison of science practical skills in pre and post reform cohorts of undergraduate students
- Study 3: Valid discrimination in practical skills assessment – An exploration of classification reliability when assessing the performance of practical skills
- Study 4: Technical functioning of assessment – An analysis of A level examination items that assess science practical skills

Annex B: Statistical comparison of mean facility values for IAPS and non-IAPS items across subjects and specifications

Subject	Spec	IAPS		Non-IAPS		t	df	Sig.	D
		mean	SD	mean	SD				
Biology	*All*	0.46	0.21	0.50	0.21	**-1.99**	**258**	**0.05***	**0.18**
	Biol A	0.43	0.18	0.48	0.20	-1.14	28	0.26	0.29
	Biol B	0.43	0.21	0.48	0.21	-1.00	44	0.32	0.21
	Biol C	0.45	0.20	0.51	0.21	-1.18	26	0.25	0.30
	Biol D	0.56	0.20	0.53	0.21	0.69	38	0.49	-0.15
	Biol E	0.43	0.19	0.48	0.22	-1.19	60	0.24	0.24
	Biol F	0.47	0.24	0.51	0.20	-0.91	63	0.36	-0.19
Chemistry	*All*	0.56	0.19	0.60	0.18	**-2.14**	**138**	**0.03***	**0.24**
	Chem A	0.51	0.21	0.59	0.19	-1.79	33	0.08	0.39
	Chem B	0.58	0.16	0.61	0.19	-0.84	49	0.41	0.18
	Chem C	0.52	0.04	0.55	0.08	-1.35	15	0.20	0.71
	Chem D	0.64	0.18	0.63	0.18	0.26	25	0.80	-0.07
	Chem E	0.52	0.21	0.59	0.18	-1.48	25	0.15	0.36
Physics	*All*	0.45	0.21	0.56	0.21	**-6.83**	**296**	**<0.01***	**0.53**
	Phys A	0.39	0.18	0.53	0.20	**-7.51**	**175**	**<0.01***	**0.77**
	Phys B	0.43	0.17	0.54	0.22	**-2.44**	**35**	**0.02***	**0.61**
	Phys C	0.63	0.08	0.67	0.19	-1.37	29	0.18	0.49
	Phys D	0.52	0.25	0.58	0.23	-1.07	31	0.29	0.25
	Phys E	0.56	0.21	0.63	0.22	-1.49	51	0.14	0.31

*p < .05

Annex C: Statistical comparison of mean performance on IAPS items for N and P candidates

Subject Grade	Grade NC		Grade P		t	df	Sig.	D
	mean	SD	mean	SD				
Overall	27.52	15.45	48.25	18.48	**-43.38**	**1085**	**<0.01***	**1.12**
Biology	28.14	12.77	44.06	14.17	**-26.68**	**475**	**<0.01***	**1.12**
Chemistry	25.70	19.63	55.80	21.06	**-25.70**	**286**	**<0.01***	**1.43**
Physics	28.26	14.70	43.90	16.85	**-18.67**	**320**	**<0.01***	**0.93**
A*	77.40	10.03	74.30	12.19	0.98	9	0.35	-0.25
A	60.22	12.28	63.24	12.46	1.12	20	0.27	0.24
B	50.02	10.50	52.61	11.44	1.74	49	0.09	0.23
C	39.72	9.72	42.87	10.03	**3.94**	**150**	**<0.01***	**0.31**
D	33.17	8.68	33.82	8.77	1.15	240	0.25	0.07
E	24.80	7.65	25.48	8.12	1.49	304	0.14	0.08
U	12.30	7.97	16.12	7.80	**8.08**	**357**	**<0.01***	**0.49**

*p < .05

Annex D: Statistical comparison of mean performance on IAPS items for N and propensity score matched P candidates

Subject Grade	Grade NC		Grade P		t	df	Sig.	D
	mean	SD	mean	SD				
Overall	27.05	15.48	27.12	15.30	-0.11	1994	0.91	0.00
Biology	27.72	12.52	28.08	12.39	-0.43	887	0.67	-0.03
Chemistry	25.07	20.18	25.45	19.85	-0.22	506	0.83	-0.02
Physics	27.74	14.75	27.13	14.71	0.50	597	0.61	-0.04
Biol A	25.06	12.76	25.50	12.44	0.40	521	0.69	0.03
Biol E	31.53	11.13	31.76	11.36	0.20	364	0.84	-0.02
Chem A	19.22	16.24	19.21	15.79	-0.01	268	0.99	0.00
Chem D	31.64	22.13	32.59	21.60	0.33	236	0.74	-0.04
Phys A	25.16	14.26	24.25	13.91	-0.65	397	0.52	-0.06
Phys D	33.12	14.35	32.65	14.69	-0.23	197	0.82	-0.03

*p < .05

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

October 2018 **Ofqual/18/6433**