Item response theory, computer adaptive testing and the risk of self-deception

Tom Benton (Research Division)

Introduction

For more than a century, the vast majority of high-stakes exams in England have been paper based. Moreover, aside from occasional differentiation of students into tiers, all students taking an assessment are presented with exactly the same set of questions at the same time.[1] This has obvious advantages in terms of transparency. If one student is ranked ahead of another it is simply because, given the same set of questions, one answered more of them correctly than another. All students answering the same questions within a given exam ensures there can be no argument of one student being given an easier assessment than another.

However, as more and more activities in modern life move from the physical to the online realm it is natural for people to consider what benefits might be achieved if high-stakes exams became computer based.[2] The switch to computer-based testing has already begun in other countries such as Israel, Finland and New Zealand (Meadows, 2021). Among the potential benefits that are considered is whether a computer-based format would make it easier to tailor assessments to each individual student through computer adaptive testing.

Computer adaptive testing involves selecting which items to present to a student on-the-fly as the test progresses. In particular, if a student answers an item (or a set of items) correctly then the next item (or set of items) presented to them will be more difficult. Conversely, if a student is struggling, they will tend to be presented with easier items. As such, as the student progresses through the test, the items they are presented with are tailored to match their ability level.

Clearly, computer adaptive testing cannot rely on simply counting how many items students have answered correctly as, by design, some students have been presented with more difficult items than others. To address this, item response theory (IRT) is used. IRT is an overarching theory describing how students respond

[1] Although, in a minority of assessments, students may choose which items they answer.

[2] Risks are also considered by some authors. See for example Bramley (2021).

to individual test questions (items). In its most common form, it assumes that the probability that any student will answer any item correctly is defined by just two things: a single number describing the ability of the student (unidimensionality) and a small set of numbers (item parameters) describing the key characteristics of the item such as its difficulty and how discriminating it is. For further details, see Harris (1989).

To use IRT within the context of computer adaptive testing, the parameters of all items, such as how discriminating and (most crucially) how difficult they are, must be estimated. This requires some form of trialling before items are used in a high-stakes setting. Then, after students have taken a test, IRT is used to calculate the score that should be assigned to each student while properly accounting for the difficulties of the items they have been presented with (see Wainer et al., 2000).

In theory, computer adaptive testing should make the test more engaging for students as the items they are presented with are more appropriate for their ability. For example, if a student is struggling, rather than being repeatedly presented with questions that are too hard for them to answer, they will find that the test automatically adapts to present them with items more appropriate to their current performance level. Computer adaptive testing should also allow more accurate assessment of each student's ability. For example, ensuring that high ability students are presented with lots of challenging tasks should make it easier to distinguish their relative abilities than if they also had to answer many easy questions.

The potential improvement in measurement precision that can be achieved by computer adaptive testing is normally presented in terms of the extent to which testing time can be shortened without any loss in reliability. That is, rather than keeping the length of exams the same and reducing the level of uncertainty around the score assigned to each student, the benefit of a computer adaptive test (CAT) is usually realised in terms of reducing the length of time students are required to spend taking an exam. According to Straetmans and Eggen (1998, p.51) "on average CATs require about 60 percent of the number of items needed in traditional paper-based test". Other authors suggest that in their specific contexts CATs allow test lengths to be halved with no loss of measurement precision (Kreiter et al., 1999; Weiss, 1982).

The aim of this article is to explore the potential gains from a switch to adaptive testing in the context of large qualifications such as GCSEs and A Levels in England. This context is potentially different from some typical applications of CATs such as general intelligence testing (e.g., non-verbal reasoning tests) or tests of foreign language fluency. In particular, GCSEs and A Levels require students to learn a range of knowledge and skills from a broad range of topics within a subject. As such, the design of examinations is intended to cover numerous topics and skills rather than tightly focus on a single concept.

More specifically, the aim of this article is to better understand whether apparent gains in reliability coefficients from a switch to CATs are likely to translate into real world improvements of the validity of our assessments. The interest in this topic stems from previous research (described in the next section) showing instances where improvements in reliability do not translate to concomitant changes in predictive validity. The logic for being concerned about a potential gap between supposed (reliability) and actual (validity) benefits of CATs would run as follows:

- In order to even estimate reliability of scores from a CAT we are forced to use IRT. As such, most existing estimates of the improved efficiency through using CATs are based either directly on the output from IRT models or on simulation studies run on the assumption that they are correct.
- The famous aphorism that "all models are wrong, but some are useful" (George Box) clearly applies to IRT. While IRT models function as a very good approximation to our data in many situations, they are not true in an absolute sense.
- Thus, given that, to some extent, the model we are using to estimate scores must be "wrong", how far should we trust estimated improvements in reliability when these are estimated on the assumption that the model is completely correct?
- The real risk here is one of self-deception—thinking that a move to computer adaptive testing is a bigger improvement (in terms of reliability and validity) than it really is. This article will provide an evaluation of the potential size of this risk using real data, that is, not based purely on simulations that assume IRT models fit perfectly.

Previous examples of self-deception risks

This article is by no means the first to draw attention to the risk of self-deception through an over-reliance on IRT models. Three examples are listed below. The first relates specifically to computer adaptive testing and the following two to large-scale empirical analysis of the impact of relying on IRT in other contexts.

Capitalisation on chance

This issue was explored by Veldkamp (2013) and van der Linden & Glas (2000). The issue is that in order to work, computer adaptive testing requires an initial estimate of the difficulty and discrimination of each item. These initial estimates are usually based on relatively small samples of students (perhaps a few hundred), and hence have non-negligible levels of uncertainty attached to them. The result is that, when a CAT selects the next item for a student, it may believe it is selecting a highly discriminating item targeted at just the right ability level, when in fact it is not. Furthermore, since some CATs are designed to try and pick the most discriminating items more frequently, they are liable to tend to select items where the discrimination has been overestimated. As a result, according to Veldkamp and Verschoor (2019, p.293), "the measurement precision of the CATs might be vastly over-estimated".

In other words, the reliability measures generated by a CAT may produce an over-optimistic picture of test quality that is not reflected in reality.

Rescoring using item weights or IRT

Benton (2018) compared various alternatives to simply summing item scores to create overall test scores. The alternatives were intended to help optimise reliability and included both classical methods such as the one suggested by Guilford (1941), and IRT methods such as using a graded response model to produce pupil ability estimates. On average, across analyses of more than 500 assessments, these methods increased the reliability indices (on scales from 0 to 1) from about 0.88 to about 0.89. This may appear a very minor improvement but is actually equivalent to the increase in reliability we might get from lengthening our assessments by 10 per cent[3] (without any of the associated cost). In these terms, it is also very similar to the reported improvement in reliability from transferring the reading literacy tasks in the Programme for International Student Assessment (PISA) to a multi-stage adaptive format in 2018 (OECD, 2019, p.27).

According to IRT, increases in reliability should relate to a reduction in the influence of random error on test scores. This should reasonably be expected to in turn lead to increased correlations with other measures of student ability. However, for Benton's 2018 study the supposed increases in reliability were associated with absolutely no improvement in the predictive value of test scores.

This example illustrates how, if we were entirely reliant on the numbers coming out of an IRT analysis, we might convince ourselves that reweighting items provides an easy way of improving the reliability of test scores at no cost. In fact, the lack of any concomitant improvement in predictive value suggests that, as has been suggested many times in previous research, reweighting items is "futile" (Wang & Stanley, 1970, p.688).

Optimal (fixed) test construction using IRT

Similarly, Benton (2018) compared various approaches to optimal construction of fixed tests. Again, this analysis was based upon real data from more than 500 separate assessments. The research compared the predictive value of half-length tests constructed out of real full-length tests so as to optimise various classical and IRT measures of test reliability. Unlike the research on simply rescoring tests, optimised approaches to item selection did indeed lead to improvements in predictive value when compared to simply selecting items at random. However, the scale of improvement was not as high as might be expected based upon the associated reliability values.

This example again reinforces the possible risks of self-deception from relying entirely upon reliability statistics from IRT analyses. That is, gains in reliability may not necessarily translate into validity. However, the example also accentuates the fact that, while like all models they are "wrong", IRT models are nonetheless "useful". The application of IRT in the study did indeed identify selections of items with greater predictive value on average–just not to the extent that might be hoped given the reliability coefficients.

[3] This is easily seen using the Spearman-Brown formula. $0.89 \approx 1.10*0.88/(1+0.10*0.88)$.

The present study

The current article builds on the item selection research in Benton (2018). In particular, it extends the research to include an examination of the possible gains from allowing the items to be assigned to each student to be selected in an adaptive way, rather than using the same set of items for all students.

The present study makes use of real data throughout. This includes using responses of real students to real items to mimic how they might perform in a CAT, as well as making use of assessment data beyond the tests being studied to give some idea of how different approaches to assessment affect validity. The use of real data for evaluation is crucial. While it is easy to estimate the likely impact of using CATs through simulation, such simulations tend to rely on the assumption that the underlying IRT model is absolutely correct. As such, the use of simulations would entirely undermine the purpose of the research. In order to make some inferences about validity we will look at the correlation of test scores (derived in various ways) with external measures of academic achievement. We will refer to these correlations as “predictive value”.

Having said this, the use of real data does have some limitations. All of the data used in the present study is drawn from tests that were originally delivered in a fixed (paper-based) format. This means that the analysis (presented next) cannot entirely mimic the way in which a genuine CAT would operate. In particular, a real CAT would start with a large bank of items that could be presented to students. By carefully selecting which items are presented to each student, the idea would be to either improve test reliability while maintaining test length or to maintain reliability relative to a fixed format while reducing testing time. Neither of these two aims can be tested directly using our real data from fixed format tests. In particular, our methodology will necessarily involve imagining a CAT that assigns each student a subset of items from the original full-length test. Since the imagined CAT is only a subset of the original full-length test it is likely that we will lose rather than gain reliability. As such, the focus will be on which approaches to selecting the subset of items (CAT or fixed form) lead to the smallest losses in terms of reliability and validity. That is, although our real interest is in whether CATs improve test quality, with our data we can only test whether they lead to smaller reductions in reliability and predictive power than other approaches.

A second drawback of using real data from fixed format tests is that they tend to be presented in terms of question stems with a number of subsequent sub-questions. Although, for the purposes of analysis, it is necessary to treat sub-questions as separate items (or else there are too few items to work with) they may not be quite as independent of one another as would generally be the case for distinct items within an item bank underlying a CAT. Although some effort has been made to mitigate the impact of this issue (particularly through checking data for unidimensionality—see below), it remains a caveat against the results presented here.

Method

The data for the present study comes from 159 assessments that were completed as part of GCSEs, A Levels or equivalent international qualifications between 2013 and 2017. The assessments were chosen to meet the following criteria:

- Taken by at least 5000 students. This ensured the accuracy of any item parameters estimated via IRT thereby avoiding issues of capitalisation on chance (see earlier discussion). The median entry size for selected assessments was just under 9000.
- No optional questions. In other words, all questions were compulsory so that whichever items were selected for retention for each student, an item score would be available.
- At least 20 items. This criterion ensured that there would be a reasonable number of items to choose for each student. The median number of items in selected assessments was 32.
- No items worth more than five marks and at least one item worth only one mark. Since the focus of this article is on computer adaptive tests, and such tests rarely (if ever) incorporate items with long mark scales, it seemed reasonable to restrict attention to assessments consisting of relatively low tariff items. Having said that, none of the assessments included in analysis consisted entirely of one-mark items.
- The assessment was deemed to be unidimensional. Unidimensionality was important for the analysis as the intention was to focus upon CATs based on unidimensional IRT models. Unidimensionality was confirmed for each of the assessments using Velicer's MAP criterion (Velicer, 1976) as evaluated by the R package *psych* (Revelle, 2020).[4]

The principle of analysis is as follows. For each assessment we apply some method to select items for each student, calculate a score for each of them based only on data from the selected items, and then calculate the correlation[5] between the resulting scores and a measure of the students' achievement more widely. We label these correlations "predictive value". We also calculate estimates of the reliability of scores from the item selection method. Finally, we compare both predictive value and reliability from the selected items against the original value based on retaining the whole full-length test. The idea is that methods that are more effective at selecting the most appropriate items for each student will retain a greater amount of reliability and predictive value from the full-length tests.

For the purposes of calculating predictive value, the wider achievement of each student was measured via each candidate's external ISAWG[6] (Benton, 2017). The

[4] Note that this criterion led to the removal of several hundred assessments from those available for inclusion in the study.

[5] To avoid potential issues with outliers, and also the impact of the scales used for different scoring systems, Spearman's rank-order correlations were used.

[6] ISAWG stands for Instant Summary of Achievement Without Grades.

external ISAWG is a measure of each candidate's performance across all of the tests that they have taken in a particular examination session, excluding the one being analysed. It is derived using a form of principal components analysis and can be interpreted as a very general form of ability across different subjects. It was used in this analysis as it was easily available for nearly all the candidates included in analysis.

Analysis focused on three methods that selected a single, optimal set of items for all students and two CAT-like approaches where the selected items could vary across students. The three single-form methods were: to select items at random; to select items that maximise expected test information (a concept from IRT relating to the likely reliability of a test) based on a Rasch partial-credit model (PCM); and to select items that maximise expected test information based upon a graded response model (GRM). The CAT-like approaches each attempted to maximise the expected test information for each student individually using either a PCM or a GRM.

The difference between the PCM and the GRM is that the former requires estimation of item difficulty only whereas the latter also estimates the discrimination of each item. In theory, the GRM approach should be superior in that it can ensure that the most discriminating items are selected in addition to ensuring that they are at the most appropriate level of difficulty for the students. In contrast, the PCM model assumes that items worth the same number of marks have the same discrimination parameters and focuses purely on ensuring that items of the most appropriate difficulty are selected. Evaluating whether extra focus of the GRM on how well each item discriminates between students of different abilities actually translates into improvements in predictive value was a key question within this research.

Each item selection method was designed to select items worth half the total number of marks available on the original full-length test.[7] Furthermore, the selected items were intended to reflect as closely as possible the distribution of item tariffs (i.e., the maximum available marks on each item) in the original test.

To further illustrate the procedure that was applied for each assessment, we consider a 40-mark Biology test that was included in analysis. The test consisted of two 4-mark items, two 3-mark items, eight 2-mark items and ten 1-mark items. In this particular instance, each method was designed to select one 4-mark item, one 3-mark item, four 2-marks items and five 1-mark items. Further details on each method are below:

- **Fixed test with random selection of items.** The required number of items with each tariff were simply selected at random. The scores on these same items were retained for all students.
- **Fixed test with item selection relying on the GRM.** First, we fitted a GRM model to the full data set and calculated the item information functions for

[7] Real CATs may use more complex stopping criteria such as whether the estimated error of measurement for each student is below some threshold.

each item. These provide an estimate of how much information each item is expected to provide about students at each ability level. For estimation of the GRM the distribution of ability was assumed to follow a normal distribution with a mean of 0 and a standard deviation of 1. Using this fact, we then calculated the expected information we expect each item to supply across students (i.e., averaging the item information functions across the ability distribution). For each item tariff we then selected the items with the highest expected information for retention. The scores on these same items were retained for all students.

- **Fixed test with item selection relying on the Rasch PCM.** The same process as for selecting a fixed test using the GRM was followed. The only difference was that a different IRT model was fitted as a starting point. For estimation of this model, it was assumed that the ability distribution was normal with a mean of 0. However, because, in contrast to the GRM, discrimination parameters are fixed, the model estimates the standard deviation of abilities and this estimate was used in the subsequent calculation of the expected information from each item.
- **CAT-like test with item selection relying on the GRM.** The initial steps for this approach were the same as for the fixed test based on GRM in terms of model fitting and calculation of item information functions. After this, the following procedure was followed separately for each individual student.

1. Initially set the distribution of the student's ability to be normal with a mean of 0 and a standard deviation of 1.
2. From the items with the highest tariff still required, select an item with the highest expected information given the individual student's ability distribution.[8] That is, if we still need a 4-mark item we select from among these, if we have already selected sufficient 4-mark items we select from among 3-mark items and so on. Starting with items with the highest tariff makes sense as these are most likely to provide useful information about candidates across a range of different abilities. Choosing items with the highest expected information close to each student's estimated ability will tend to mean more difficult items are assigned to high performing students and easier ones are assigned to lower achievers.
3. Based on the student's (known) response to the item, update their ability distribution. For example, if they have answered an item fully correctly the mean of their ability distribution will be adjusted upwards whereas if they have answered incorrectly, it will adjusted downwards. The uncertainty around their ability estimate (i.e., the standard error) will also be adjusted.
4. Unless we have selected all of the items we require of the various tariffs return to step 2 until complete.

[8] The first item selected for each student is the same. Subsequent questions will differ across students.

5. The final IRT ability estimate of each student based on their individually selected items is used as their final score. Note that these ability estimates will adjust for the difficulty of the items that were assigned to each student and also give more weight to performance on items estimated to have a higher discrimination.

- Note that simulating a CAT using the above procedure assumes that we would be able to automatically mark all items regardless of their tariff or format. That is, we are assuming that all technological barriers to computer-based testing and auto-marking have been overcome so that we could run a CAT using the same style of items currently used in qualifications in England. This is a fairly large assumption but is used here to allow us to explore the potential of computer adaptive testing in a best-case scenario. From a technical perspective note that all ability estimations made use of expected a posteriori (EAP) estimation and that the item selection approach reflects the posterior-weighted information criterion described by van der Linden and Pashley (2010).
- **CAT-like test with item selection relying on the Rasch PCM**. The procedure was exactly the same as above but with all calculations, including calculating information function and assigning ability estimates (including final scores) to students, based upon the Rasch PCM model. Crucially, the Rasch PCM model assumes the same discrimination parameters for items with the same tariff. This means that the model will not give additional weight to performance on items estimated to be highly discriminating.

Having calculated the scores that would be assigned to each student by each method all that remained was to calculate predictive value and reliability. Predictive value was calculated as the Spearman correlation between final scores and the external ISAWG (i.e., performance more widely beyond the assessment of interest). Note that for fixed form tests, final scores were always simply the sum of the item scores on the selected items. For CAT-like tests, the final scores were based on EAP ability estimates as described above.

There are many ways to calculate test reliability. However, in order to enable the best possible comparability between different techniques, an IRT method of estimating test reliability was calculated for each test score. For the CAT-like methods, this was simply provided by the reliability indices associated with their final set of IRT ability estimates. As noted earlier, for the fixed form methods, each student's score was simply a sum of scores on the selected items. In order to allow comparability with the other methods, these sum scores were converted to equivalent values on the IRT ability scale using the EAP approach of Thissen et al. (1995). The reliabilities of the EAP ability estimates (derived from sum scores) were then calculated. The same approach was used to estimate the reliability of the original full-length test. All model fitting and calculations relating to IRT were performed using the R package mirt (Chalmers, 2012). If we denote the estimate of each student's IRT ability estimate as $\hat{\theta}_i$ and the uncertainty around this estimate as $SE(\hat{\theta}_i)$ then the formula to estimate reliability is:

$$Reliability = \frac{Var(\hat{\theta}_i)}{Var(\hat{\theta}_i) + Mean(SE(\hat{\theta}_i)^2)}$$

Note that for the two separate IRT approaches (GRM and PCM), reliability was calculated on each model's own terms. That is, if the CAT or fixed form was derived from the GRM, then the reliability index was also calculated using this model. If the CAT or fixed form was derived using the PCM, then reliability was also estimated using this model. For this reason, the reliability indices from the different models are not directly compared.[9] For both the full-length test and the random selection of items, both reliability types of index were calculated. Note that, although calculated differently, both reliability coefficients (measured on scales from 0 to 1) can be interpreted in a similar way to more familiar indices such as Cronbach's alpha.

Results

To begin with, we examine the results relating to reliability. These are shown in Figure 1 in two panels relating to the two separate IRT models that can be used to estimate reliability. Each point on the chart represents the reliability of a full-length assessment (the x-axis) and the extent to which this reliability changes (the y-axis) under various approaches to selecting only half the items for each student. Thus, for each assessment the chart includes three points in each panel (one relating to each method) and these are positioned in a vertical line. For example, the leftmost set of points relate to an assessment with an original full-length reliability just above 0.65. Selecting half the items using a CAT-like approach based on a GRM barely reduced the reported reliability. In contrast, in this instance, a fixed form based on the GRM reduced reported reliability by about 0.03 and selecting half the items at random reduced the reliability by about 0.12.

The overall pattern of results in Figure 1 is as expected. CAT-like approaches led to lower reductions in reliability relative to the full-length test than selecting a single fixed form for all students. Although care is needed with the comparison, when judged on their own terms, the extra emphasis on selecting highly discriminating items based on the GRM (and giving more weight to them in scoring) led to smaller reductions in reliability than the CAT-like approach based on the Rasch PCM. Indeed, in one case, through giving more weight to scores on highly discriminating items, the CAT-like approach appears to lead to improved reliability relative to the original full-length test despite consisting of only a subset of the items for each

[9] Although it is possible to estimate the reliability of scores derived using one model based upon another model, it is not particularly straightforward. It is also not something I have ever seen done in practice. For these reasons it is avoided in this article.

student. Among the two fixed form approaches in each panel of Figure 1, selecting items in an optimal manner based on an IRT model led to higher reliabilities than selecting items at random.

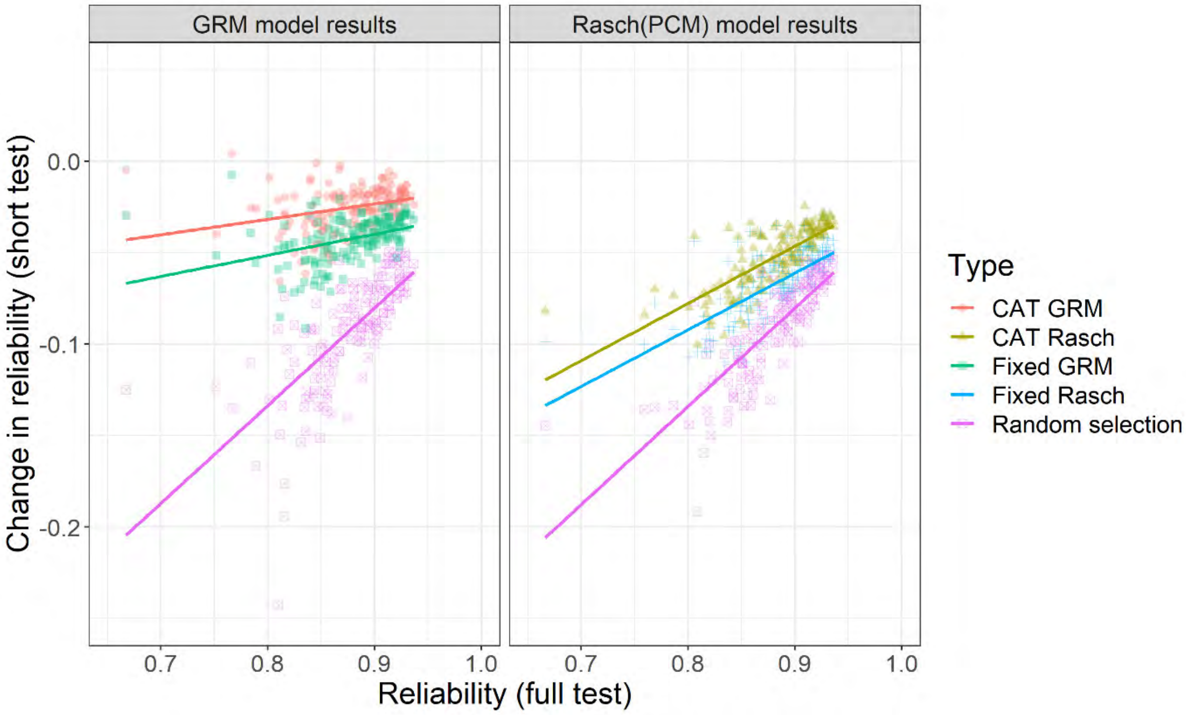

Figure 1: Original full-length test reliabilities and changes in reliability under each of the methods for selecting a subset of items for use with each student. Regression lines have been added to aid interpretation. Results are split by the IRT model used to calculate reliability.

Of most interest in this research is the extent to which the results relating to the superior reliability coefficients of CAT-like approaches in Figure 1 translate into superior predictive value. This is explored in Figure 2. Figure 2 is designed to follow the same pattern as Figure 1 but plots predictive values for the full-length test and changes in predictive values rather than reliabilities. Note that although predictive value can be directly compared across all methods (i.e., between PCM and GRM approaches), for consistency with Figure 1 the split by IRT model is retained.

As can be seen, Figure 2 creates a rather different impression to Figure 1. The advantages of the CAT-like approaches over other methods are reduced relative to the gaps shown in Figure 1. Most surprisingly, the gap between the CAT-like approach based on the Rasch PCM and fixed item selection based on the same model has vanished. On the other hand, the gaps between choosing optimal fixed form tests (using either the PCM or GRM) and selecting fixed form tests at random remain strongly evident.

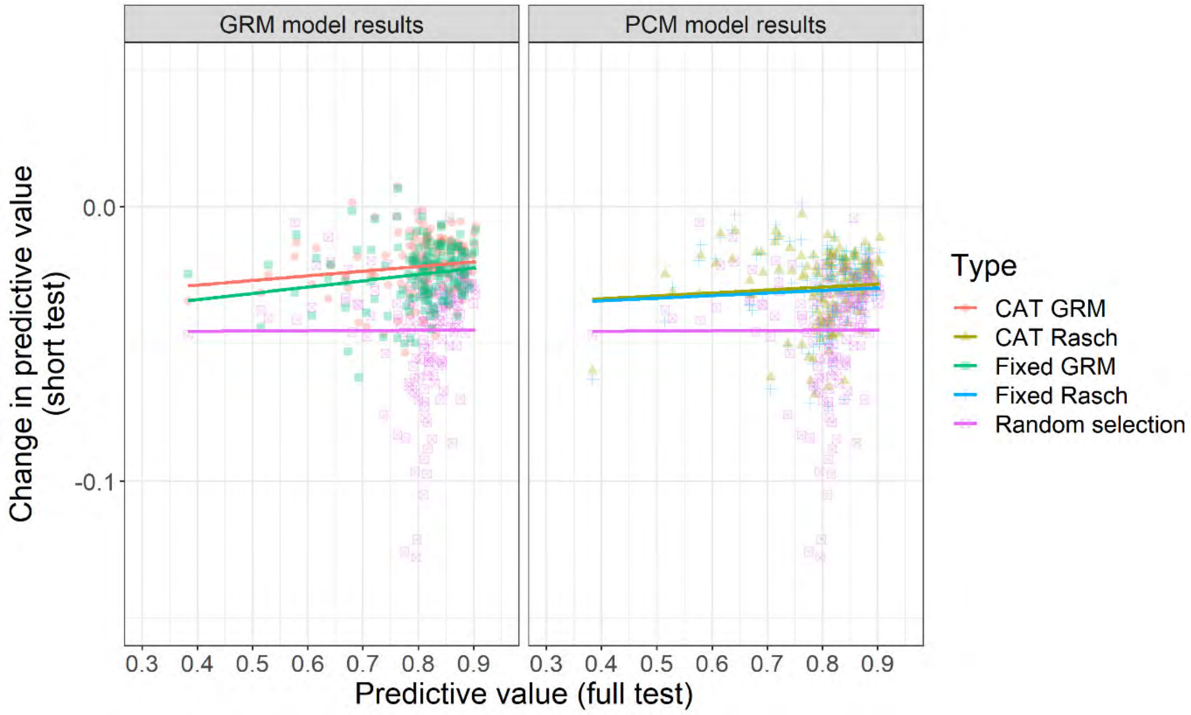

Figure 2: Original full-length test predictive values and changes in predictive value under each of four methods for selecting a subset of items for use with each student. Regression lines have been added to aid interpretation (the lines for CAT Rasch and Fixed Rasch are almost identical).

In order to interpret the two figures, it is helpful to have some idea of how much we would expect changes in reliability to impact upon predictive value. This can be calculated using the following simple formula based upon classical test theory:

$$\begin{aligned}&\mathit{Expected\ change\ in\ predictive\ value}\\&\qquad= \mathit{Original\ predictive\ value}\left(\frac{\sqrt{\mathit{New\ reliability}}}{\sqrt{\mathit{Original\ reliability}}}-1\right)\end{aligned}$$

Changes in reliability relative to the full-length test for the CAT-like GRM approach are compared to changes in predictive value in Figure 3. Each point in the chart represents an assessment. The jagged red line represents the expected change in predictive value based on the change in reliability using the formula above. The line is jagged as the change in predictive value depends not only upon the change in reliability but also upon the original values of predictive value and reliability. As can be seen, although there are exceptions, for the majority of assessments the change in predictive value is much worse than would be expected given the reported changes in reliability coefficients.

If the analysis in Figure 3 is reproduced using simulation, then changes in predictive value are far closer to the predicted values based on the above formula. In other words, the failure of changes in estimated reliabilities to be reflected in changes in predictive value must relate to some form of lack of fit in the underlying IRT model. This will be discussed more later. A particularly striking feature of Figure 3 is the weak relationship between changes in estimated reliability and changes in predictive value. A possible explanation for this is that, in reality, shortening a test (whether using a CAT or otherwise) not only alters reliability but also has some slight impact upon the construct being measured. The changes may either strengthen or weaken the relationship with external measures of achievement. This could lead to the noisy pattern we see in Figure 3.

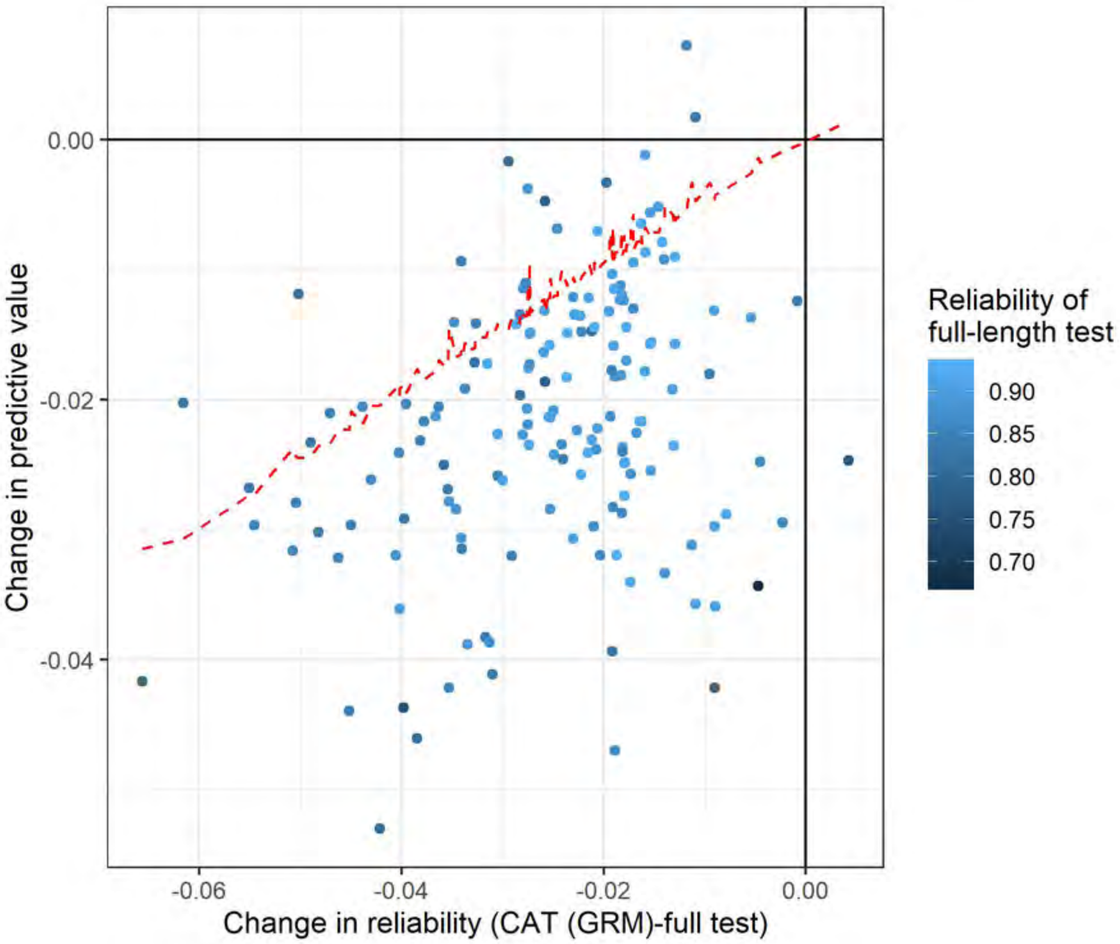

Figure 3: Changes in reliability against changes in predictive value from applying a CAT-like approach based on the GRM. The red line indicates the expected change in predictive value based on a formula from classical test theory.

Table 1 shows the mean predictive value and (relevant) reliabilities across all 159 assessments from each approach as well as the full-length assessments. Note the need to report results to three decimal places in order to properly reveal findings. The highlighting in Table 1 is used to group methods using the same model for item selection.

Table 1 repeats many of the findings described above in a different way. For example, the gap in reliability between a CAT-like and fixed test based on the PCM (of 0.015) does not translate into any meaningful difference in average predictive values (0.001). Similarly, the gap between CAT-like and fixed tests where item selection is based on the GRM is also much smaller in terms of predictive value (0.003) than in terms of reliability (0.018).

The final three columns attempt to convert the mean reliabilities and predictive values into equivalent test lengths relative to a full-length test. The columns based on reliabilities use the Spearman-Brown formula (Spearman, 1910) to convert the mean reliabilities into an equivalent test length compared to the full-length test. The use of the Spearman-Brown formula in this way effectively assumes that items are selected at random. Reassuringly, Table 1 shows that the mean reliabilities of half-length tests selected at random are indeed, according to the Spearman-Brown formula, equivalent to a randomly selected test of about half the length of the full test. The various optimal approaches to item selection for both CAT-like and fixed form tests perform better in terms of reliability. Despite only requiring half the items from the full-length test they achieve an average reliability equivalent to a randomly selected test of between 59 and 81 per cent of the length.

A similar process can be used to generate equivalent test lengths based on predictive value. First, the formula provided earlier is used to convert mean predictive values into equivalent reliabilities. These are then converted into equivalent test lengths using the Spearman-Brown formula. These test lengths are generally lower than those based on reliability coefficients—especially for the CAT-like tests. For example, while the reliability coefficients might lead us to believe that a half-length CAT (based on the GRM) was worth a randomly selected test of 81 per cent length, predictive value suggests it may only be as good as a randomly selected test of 69 per cent length. The only item selection method (besides random) where the equivalent relative length is just as high whether it is based on predictive value rather than reliability is the creation of a fixed form test based on the Rasch PCM. This may be because the rather conservative nature of this approach (essentially just picking items of about the right difficulty for the average student) has less scope for over-optimism about reliability. Also, being a fixed form test, it avoids the need to provide comparable scores for students that have taken different items and the associated additional reliance on assumptions from a given model.

Table 1: Reliabilities, predictive values and associated equivalent test lengths for various approaches to test construction.

		Mean across 159 assessments of...			**Equivalent relative random length based on mean...**		
Method	**Scoring method**	GRM reliability	PCM reliability	Predictive value	GRM reliability	PCM reliability	Predictive value
Full-length test	**Sum score**	0.878	0.881	0.806	100%	100%	100%
CAT (GRM)	IRT	0.853	-	0.785	81%	-	69%
Fixed (GRM)	Sum score	0.835	-	0.782	71%	-	66%
CAT (Rasch PCM)	IRT	-	0.829	0.777	-	65%	62%
Fixed (Rasch PCM)	Sum score	-	0.814	0.776	-	59%	60%
Random	Sum score	0.786	0.791	0.761	51%	51%	50%

Model fits

As mentioned above, repeating the entire exercise using simulated rather than real data leads to much closer agreement between changes in reliability and changes in predictive value. As such, the fact that for our CAT-like approaches higher reliabilities hardly translate into higher predictive values must in some way mean that the model assumptions are not correct. This section discusses the various ways in which real data may not conform to an IRT model and the extent to which this is practically detectable.

The first thing to note is that, in terms of the indices of model fit typically used in IRT, our data did not reveal any obvious problems. Firstly, we consider the fit of the Rasch PCM model. The fit of each item in each data set to the Rasch model was evaluated using inlier-sensitive or information-weighted fit (INFIT) and outlier-sensitive fit (OUTFIT) (Linacre, 2002). Of the 4970 items in the analysis (across all data sets) only 70 (1.4 per cent) had values for these fit indices outside of the range between 0.5 and 1.5 which, according to Linacre (2002), is required to ensure items are "productive for measurement". Only 11 items in total (0.2 per cent) had values of either INFIT or OUTFIT in excess of 2 indicating severe lack of fit. In other words, the vast majority of items displayed a level of fit with the Rasch model that would be deemed acceptable in most operational contexts. Nonetheless, even the relatively small amount of lack of fit in the data appeared to be enough so that apparent gains in reliability may not translate into improvements in predictive value.

We next consider the fit of the GRM models to the data. To check this, overall goodness of fit statistics (root mean square error of approximation RMSEA and Standardized Root Mean Square Residual SRMSR, see Maydeu-Olivares, 2013) were calculated for each of the real data sets. Using these metrics, it was determined that 152 out of 159 of the data sets had values for RMSEA below the level of 0.05 which was recommended by Maydeu-Olivares (2013) as indicating adequate fit. The very largest value of RMSEA across all data sets was only slightly above this threshold at 0.07. Similarly, for 154 of 159 data sets, the value of SRMSR (an easier to understand metric that simply calculates how far pairwise item correlations in each data set are from their predicted values based on GRM on average) was below 0.05—a "substantively negligible amount of misfit" (Maydeu-Olivares, 2013, p.84). The largest value of SRMSR was also 0.07. In other words, by any normal operational definition, the GRM had a very good fit to all of the data sets in the analysis.

Despite the relatively good fit of the data to the various IRT models described above, it is possible that even the small amounts of lack of fit were sufficient to mean that differences in reliability between different techniques did not translate into differences in predictive value. This indicates that the issues shown in the above analysis are not easily detectable simply by looking at the outputs of IRT analyses.

The above measures of model fit are internal in the sense that they look at the extent to which relationships between items within the same test adhere to expectations. However, they do not reflect all of the assumptions of the IRT model. Perhaps the most crucial assumption of IRT in our context is the definition of measurement error. Usually, and certainly in nearly all simulation studies, measurement error is thought of as entirely random and, thus, unrelated to any external variables. However, this highly simplified conception of measurement error may not reflect reality. In particular, improvements in reliability indices via optimal item selection may not simply mean the removal of purely random measurement error. Rather, they may represent a change in emphasis regarding which specific pieces of knowledge are regarded as particularly pertinent to the construct and which are not. In other words, different approaches to item selection may lead to changes to the construct being assessed. Such changes may or may not be desirable dependent upon the purpose of the assessments. However, we need to remain aware of this potentially unintended consequence of switching to a CAT format and not assume that results from reliability coefficients and simulation studies tell the full story.

Discussion

In many ways, the analysis in this article supports the "useful" nature of IRT models and, in particular, their value developing CATs. On average, test scores derived from a simulated CAT process had higher predictive value than any single fixed test across students. Similarly, there were no instances where using a CAT and the associated algorithm for producing student scores led to markedly lower predictive value than using a random selection (and in most cases it was better). Thus, the article is not a criticism of the use of CATs in themselves. What is at stake here is the rather more technical, but nonetheless important, topic of whether we are able to accurately evaluate test quality based on the output from IRT analyses alone, or whether we risk deceiving ourselves that changes are leading to improved validity when in fact they do not. That is, whether a focus on reliability indices risks overselling the advantages of CATs.

The results of analysis show that, relative to fixed form tests, expected advantages in test quality (based on reliability indices) may not always necessarily translate into verifiably higher predictive values. Having said this, the differences between expectations based on reliability and actual predictive values were often quite small in real terms.

It is worth admitting that very few people are likely to care about the levels of difference in reliability (or predictive value) described in this article. For example, how many people would really care about whether an assessment's correlation with achievement more widely is 0.78 or 0.77? However, the point is that the results here form part of a wider body of work questioning whether computer adaptive testing will necessarily result in improved test quality in every context. For example, previous research (Veldkamp, 2013) has already demonstrated how the uncertainty in the estimated parameters of items used in a CAT may mean that they are less effective than thought. More generally, the issue is that in a CAT we are highly reliant on the accuracy of an IRT model for correctly scaling the scores of students who have taken different sets of items against one another. If the underlying IRT model is not correct in every respect, this may lead to some degree of error in this process.

With these risks in mind, and in the context of high-stakes examinations covering a broad range of content such as GCSEs and A Levels, it is worth considering whether the potential benefits of computer adaptive testing are sufficient relative to the added difficulty in ensuring comparability between scores from different students. It is interesting to note that, in practice, CATs do not always lead to the level of improvement in reliability that might be hoped for. For example, ETS researcher Martha Stocking once quipped that real tests often had so many additional constraints such as ensuring content coverage and avoiding overexposure of individual items that most CATs were actually BATs (barely adaptive tests) (Chuah et al., 2006). Given the likely requirement to ensure that examinations continue to cover the majority of the taught curriculum for each student, this would be a particular risk in the context of qualifications such as GCSEs and A Levels.

In considering the value of CATs, it is worth noting that many of their benefits relate to the application of computer-based testing more generally rather than the adaptive nature of the tests. For example, van der Linden and Glas note advantages such as "the possibility for examinees to schedule tests at their convenience; tests are taken in a more comfortable setting and with fewer people around than in large scale paper-and-pencil administrations; electronic processing of test data and reporting of scores are faster; and wider ranges of questions and test content can be put to use" (van der Linden & Glas, 2010, page vi). All of these advantages are good reasons to explore the possibility of extending the use of computer-based testing in England. Chasing high reliability coefficients through CATs should very firmly stay in second place.

References

Bramley, T. (2021, March 31). Online assessment - the robustness and resilience of the exam system (part 1). *Cambridge Assessment Website blog.* https://www.cambridgeassessment.org.uk/blogs/the-robustness-and-resilience-of-the-exam-system-part-1/

Benton, T. (2017, November). *Pooling the totality of our data resources to maintain standards in the face of changing cohorts.* Paper presented at the 18th annual AEA-Europe conference, Prague, Czech Republic. http://www.cambridgeassessment.org.uk/Images/429428-pooling-the-totality-of-our-data-resources-to-maintain-standards-in-the-face-of-changing-cohorts.pdf.

Benton, T. (2018, November). *Exploring the relationship between optimal methods of item scoring and selection and predictive validity.* Paper presented at the Association for Educational Assessment – Europe conference, Arnhem/Nijmegen, The Netherlands. https://www.cambridgeassessment.org.uk/Images/525258-exploring-the-relationship-between-optimal-methods-of-item-scoring-and-selection-and-predictive-validity.pdf.

Chalmers, R. P. (2012). *mirt:* A Multidimensional Item Response Theory Package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. http://www.jstatsoft.org/v48/i06/.

Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for CAST item parameter estimation. *Applied Measurement in Education, 19*(3), 241–255. https://doi.org/10.1207/s15324818ame1903_5.

Guilford, J. P. (1941). A simple weight scoring for test items and its reliability. *Psychometrika, 6*(6), 367–374. https://doi.org/10.1007/BF02288593

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice, 8*(1), 35–41. https://doi.org/10.1111/j.1745-3992.1989.tb00313.x

Kreiter, K. D., Ferguson, K., Gruppen, L. D. (1999). Evaluating the Usefulness of Computerized Adaptive Testing for Medical In-course Assessment. *Academic Medicine, 74*(10), 1125–1128. https://doi.org/10.1097/00001888-199910000-00016.

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions, 16*(2), 878. https://www.rasch.org/rmt/rmt162f.htm.

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement, 11*, 71–101. https://doi.org/10.1080/15366367.2013.831680

Meadows, M. (2021, June 15). Speech at City of London Schools Conference 2021. GOV.UK. https://www.gov.uk/government/speeches/dr-michelle-meadows-speech-at-city-of-london-schools-conference-2021.

OECD. (2019). *PISA 2018 Technical Report*. Chapter 2 - Test Design and Test Development. https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018%20TecReport-Ch-02-Test-Design.pdf./

Revelle, W. (2020) *psych: Procedures for Personality and Psychological Research*. https://CRAN.R-project.org/package=psych. Version = 2.1.3.

Spearman, C. (1910). Correlation Calculated from Faulty Data. *British Journal of Psychology, 3*, 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Straetmans, G. J. J. M., & Eggen, T. J. H. M. (1998). Computerized Adaptive Testing: What It Is and How It Works. *Educational Technology, 38*(1), 45–52. https://www.jstor.org/stable/44428447

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. S. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*(1), 39–49. https://doi.org/10.1177%2F014662169501900105

Van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of Adaptive Testing.* Springer.

Van der Linden, W. J., & Pashley, P. J. (2010). Item Selection and Ability Estimation in Adaptive Testing. In Van der Linden, W. J., & Glas, C. A. W. (Eds.), *Elements of Adaptive Testing.* Springer.

Veldkamp, B. P. (2013). Ensuring the future of Computerized Adaptive Testing. In Eggen, T. J. H. M. & Veldkamp, B. P. (Eds.), *Psychometrics in practice at RCEC* (pp.137–150). RCEC. https://ris.utwente.nl/ws/files/253342308/Eggen2012psychometrics.pdf#page=43.

Veldkamp, B. P., & Verschoor, A. J. (2019). Robust computerized adaptive testing. In Veldkamp, B. P. & Sluijter, C. (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp.291–305). Springer. https://library.oapen.org/bitstream/handle/20.500.12657/22945/1007216.pdf?sequence=1#page=291.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40*(5), 663–705. https://doi.org/10.3102%2F00346543040005663

Weiss, D.J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. *Applied Psychological Measurement, 6*(4), 473–492. https://doi.org/10.1177%2F014662168200600408