Cambridge Assessment

# Exploring the relationship between optimal methods of item scoring and selection and predictive validity

*Conference Paper*

**Tom Benton**

Presented at the 19th annual conference of the Association for Educational Assessment in Europe,
Arnhem-Nijmegen, The Netherlands,
November 2018

## Author contact details:

Tom Benton
Assessment Research and Development,
Research Division
Cambridge Assessment
The Triangle Building
Shaftesbury Road
Cambridge
CB2 8EA
UK

benton.t@cambridgeassessment.org.uk

http://www.cambridgeassessment.org.uk

As a department of the University of Cambridge, Cambridge Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

## How to cite this publication:

Benton, T. (2018, November). *Exploring the relationship between optimal methods of item scoring and selection and predictive validity.* Paper presented at the 19th annual conference of the Association for Educational Assessment in Europe, Arnhem-Nijmegen, The Netherlands.

## Acknowledgment:

I would like to thank Stephen Cromie, Cambridge Assessment International Education, for instigating this research and providing the initial avenues for analysis.

## Introduction

One of the questions with the longest history in educational assessment is whether it is possible to increase the reliability of a test simply by altering the way in which scores on individual test items are combined to make the overall test score. Most usually, the score available on each item is communicated to the candidate within a question paper. The score they are awarded on the test as a whole is then calculated simply as the sum of these item scores. However, throughout the history of assessment, psychometricians have been tempted to try to improve on this simple and transparent process with the aim of making the resulting scores more reliable.

The simplest way in which scoring might be altered is to assign a weight to each item and then create candidate scores as a weighted sum of item scores rather than a simple sum. The aim is that highly reliable, (presumably) higher quality items are given more weight than those that appear to have less relevance to the construct being measured. This goal seems reasonable as, after all, it is unlikely that all items are of exactly equal relevance and placing greater emphasis on candidate achievement on the most reliable items should improve the quality of the final scores. In other words, it should mean that candidates are more likely to be placed in the correct rank order in terms of their true abilities in some domain. One early attempt at this type of rescoring, though by no means the first, can be found in Guilford (1941).

Over the following decades this question has been repeatedly explored within the research literature. A review of the substantial amount of research on this topic up until 1970 is provided by Wang and Stanley (1970). This paper notes that “numerous studies have been performed demonstrating that item weighting is futile” (p. 688) and concludes that “weighting of test items was shown repeatedly to be ineffective, or so slightly effective as to be impractical.” (p. 699).

Since this work, similar problems under slightly different guises have been investigated in the psychometric literature. For example, Knott and Bartholomew (1993) investigated achieving optimal test-retest reliability when something is known about the test-retest correlation of individual items.

Kane and Case (2004) investigated the optimal way to combine two scores – one a highly reliable objective test score (e.g. from a multiple choice test) and one from a more subjective scoring (e.g. an essay scoring). They note that the fact that two such scores may be added together need *not* assume that they are both measuring the same construct, although it does imply the (not unreasonable) assumption that added ability in one element may compensate for a lack of ability in another in practice[1]. They conclude that giving extra weight to the more reliable of the two observed scores will increase reliability and, *up to a point*, will increase

[1] Specifically they suggest that, in the case of combining scores for a college admissions test consisting of verbal and quantitative subtests, it is reasonable to assume that they are “at least partially compensatory in practice; in terms of success in college, strength in one set of competencies can compensate for weakness in another” (. 224).

validity in terms of how well the overall score predicts a target composite. Importantly however, they note that if the two scores are deliberately designed to measure distinct skills "it is generally *not* a good strategy to simply select the weights in the observed composite to maximize reliability" (p. 239, emphasis in original article) as this will lead to an overall decrease in validity. Similar findings are provided by Rudner (2001).

Despite the wealth of existing research on this topic, there are particular reasons for the further investigations within this paper. An initial motivation came from the observation (expanded upon later) that the corrected item-total correlation for, for example, a single 4-mark item is often less than that for four distinct 1-mark items combined into a 4-mark item. This may indicate that performance on such items has been systematically over-rewarded in exams and may be in need of correction. This also has implications for the ways in which we judge item quality for the purposes of refining existing tests or designing new ones.

A second motivation for the research comes from the observation that, although it is a fact almost universally acknowledged that item weighting is an ineffective approach to improving reliability (e.g. Wang and Stanley, 1970), the same statistics that are widely discarded when it comes to weighting items are often recommended when it comes to the process of evaluating items – perhaps for the purpose of test construction, or for the purposes of evaluating an existing test. This is a little curious as test construction can itself be seen as a form of item weighting with some items given a weight of 1 (i.e. included in the test), and some given a weight of zero (i.e. not included in the test). The difference with traditional item weighting studies is, of course, that whereas in item weighting the default choice is that all items retain a weight of 1, in item selection we must by definition exclude at least some items (i.e. set their weights to zero). Nonetheless, it was of interest to explore how well statistical methods of automatic test construction perform in practice.

Finally, since the early studies on item weighting and selection reviewed by Wang and Stanley, the use of Item Response Theory (IRT) for test scoring has become far more widespread. It was of interest to examine the performance of IRT-based approaches to test scoring and item selection compared to more traditional methods.

Throughout this paper the focus of analysis will be to evaluate the performance of different approaches against real data sets. In particular, the central question will be whether test scores derived in particular ways are better at predicting achievement more widely. Although not a pure measure of validity, which would require a bespoke and specific definition of what each test was trying to achieve along with an independent measure of whether this had been done, the focus upon predictive value does at least allow us to understand whether increases in (estimated) reliabilities actually translate into any meaningful gains when the test scores are used to attempt to predict something about students more widely. As such, it helps us to avoid the pitfall noted by Kane and Case (2004) of aiming purely to maximise reliability.

## Data

The analysis was based upon item level data from 515 Cambridge Assessment examinations taken between 2013 and 2017 across a very wide range of subjects including Mathematics,

Sciences, Modern Languages, and Humanities. The assessments were chosen so that they were each taken by at least 5,000 candidates, contained more than 10 items of which none were optional[2], and contained at least one 1-mark item and at least one item worth more than 1 mark. The various tests included in the analysis were taken by up to 65,000 candidates with the median entry size being just over 8,000. They consisted of between 11 and 78 items with the median number being 30. The maximum scores on the tests ranged between 30 and 130 (median of 60). The maximum scores of the items included in analysis ranged between 1 and 24. That said, the majority of tests consisted entirely of items worth up to 5 marks. Across all of the assessments, a total of 15,892 items were included in analysis.

Throughout analysis the predictive value of test scores (however they were produced) was evaluated via the correlation with the external ISAWG[3] for each candidate (Benton, 2017). The 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.

Given the focus in this paper on predictive value, it is worth noting that previous research comparing reliability and predictive validity has sometimes found that increasing one may not lead to an increase in the other. Specifically the "attenuation paradox" of Loevinger (1954) showed how, if item discrimination parameters become too high, it can lead to a decrease in predictive validity. Similarly, the issue of "capitalization on chance" (Veldkamp, 2013) shows how, due to the uncertainty in item parameter estimates, decisions intended to maximise reliability may not translate into real gains. However, neither of these objections are an issue in the present study. Firstly, as is easily shown by simulations of the type described in Appendix 1, the attenuation paradox only truly becomes an issue if item discriminations are *extremely* high and, as such, is unlikely to be an issue in practice. Secondly, the issue of capitalization on chance is avoided by the fact that the analysis is restricted to assessments with large numbers of candidates where there is little uncertainty in the item parameters. As such, throughout this paper, studying the empirical predictive value of test scores provides a sensible mechanism to evaluate different methods of item scoring and selection.

**Item tariff, R_rest and optimal weights**

Before evaluating different approaches to test scoring and construction, it is worth using the data to illustrate the links between item tariffs and typical measures of item quality that prompted this research in the first place.

Across the almost 16,000 items from all 515 assessments in the analysis, Figure 1 shows the relationship between item tariff (i.e., the number of marks available on each item) and the correlation between the scores on the item and the total score on the remainder of the same test (i.e., corrected item-total correlation or R_rest)[4]. The solid blue line summarises this relationship as a smooth line. As can be seen, there is some evidence that R_rest values

[2] That is, within the test being studied, all candidates were asked to attempt to answer all of the items.

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

[4] To allow us to focus on the most relevant part of the figure a handful of items with negative R_rest values have been removed from this and subsequent charts.

increase as the number of available marks on an item increases. However, the dashed red line provides a comparator based upon the Spearman-Brown formula (Brown, 1910). Specifically, the mean value of the R_rest coefficients for 1-mark items (0.38) is squared to give an estimate of the typical reliability of a 1-mark item (0.14)[5]. This value is then plugged into the Spearman-Brown prediction formula[6] to estimate the expected reliability of a test consisting of between 1 and 25 single mark items and the square root taken to convert this into an R_rest coefficient. As can be seen, by this method, the R_rest values of multi-mark items are generally lower than would be achieved by replacing them with 1-mark items up to the same number of marks. For example, by reading the value of the solid line for 6-mark items and finding the equivalent R_rest value for multiple 1-mark items, it appears that an average 6-mark item is perhaps equivalent to just three 1-mark items. In other words, multi-mark items do not appear to give as much value (in terms of reliability) per mark as 1-mark items.

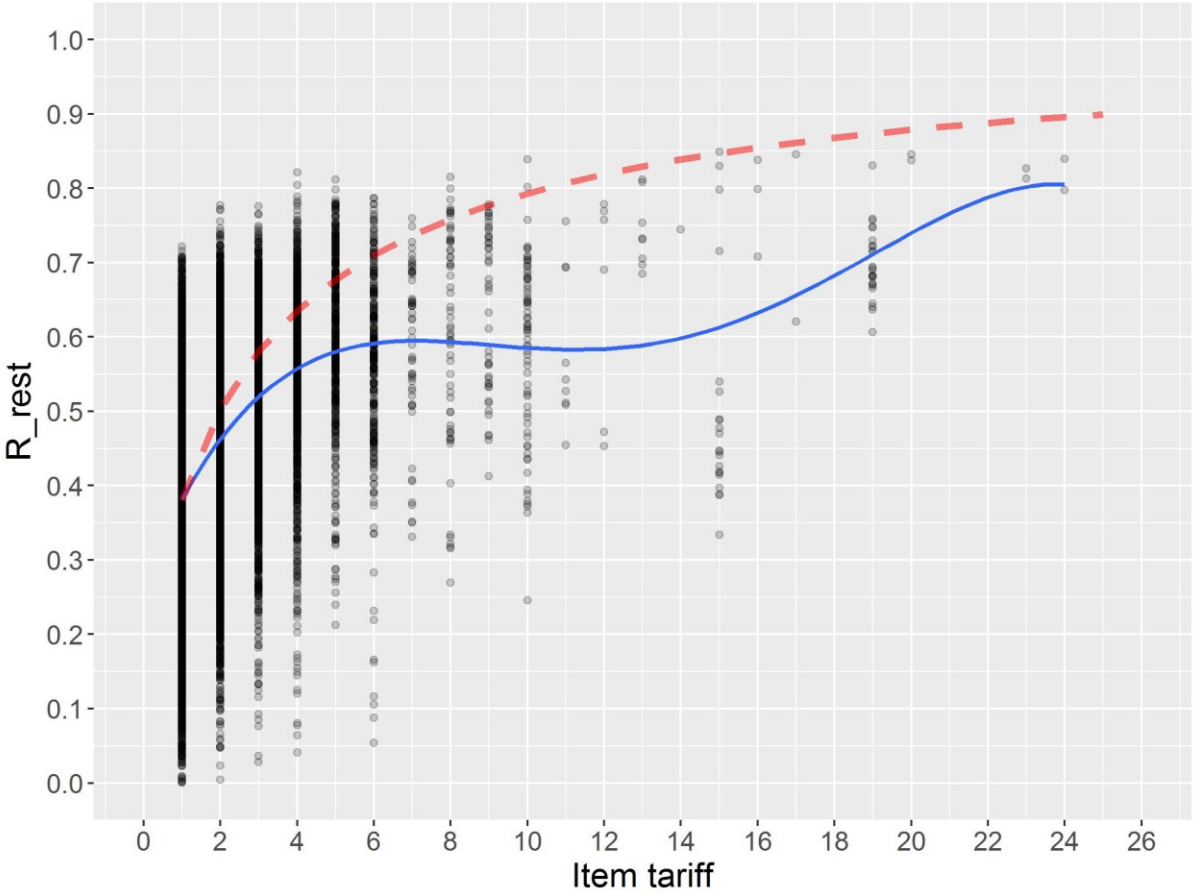


***Figure 1: The relationship between item tariffs and corrected item-total correlations (R_rest). The blue line shows the estimated relationship between the two quantities. The red dashed line shows an extrapolation from the 1-mark value based upon the Spearman-Brown formula.***

[5] Squaring the R_rest coefficient is based on an analogy with test level reliability. For a whole test, the test reliability is an estimate of the correlation with "true ability" (i.e., the score on a perfectly reliable test) squared. For these calculations we use an approximation based on imagining that the (remainder of the) full test is perfectly reliable so that the correlation with the full test score squared is an item-level reliability coefficient.

[6] https://en.wikipedia.org/wiki/Spearman%E2%80%93Brown_prediction_formula.

That said, focussing entirely on reliability (which is what R_rest does) puts us in serious danger of falling into precisely the trap that Kane and Case (2004) warn of. With this in mind, Figure 2 shows the relationship between the R_rest value for each item and their correlation with the external measure of ability (the external ISAWG). Figure 2 reveals a strong level of similarity between the two correlation coefficients. The general pattern reveals that correlations with the external ISAWG (R_crit) tend to be of similar magnitude to the corrected item-total correlations (R_rest) but just slightly smaller.

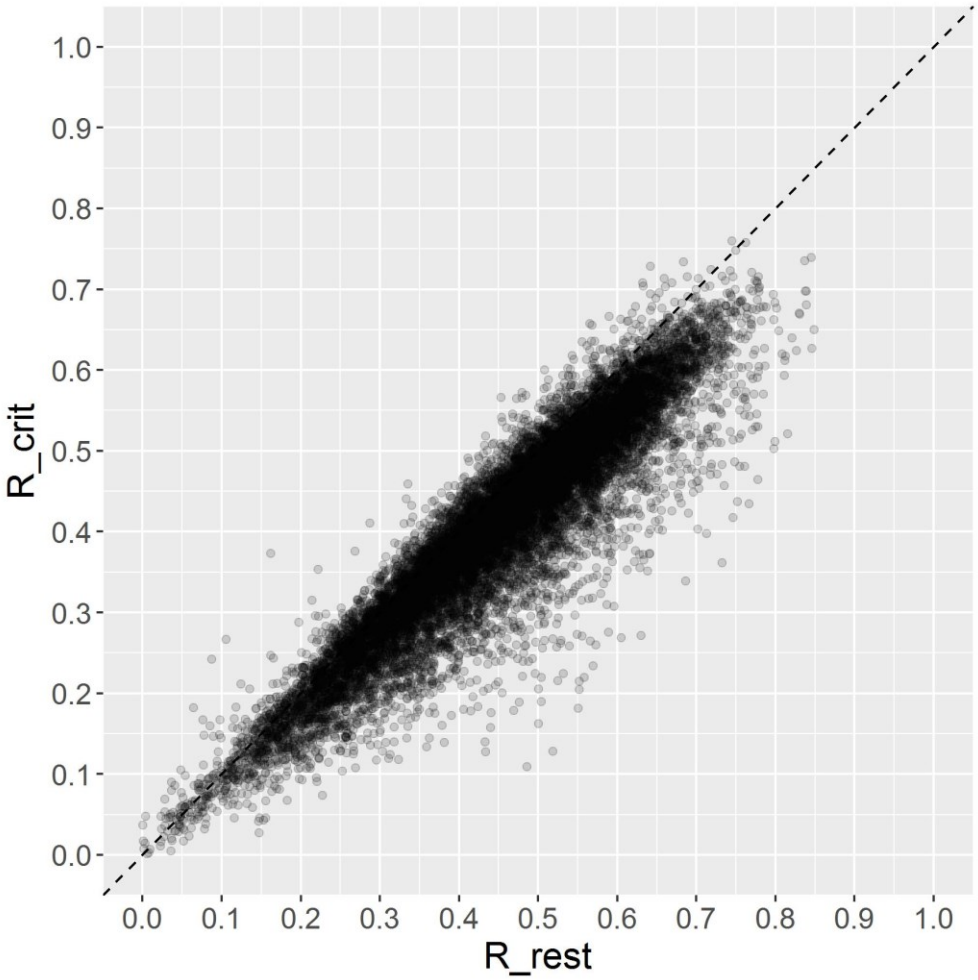


***Figure 2: The relationship between corrected item-total correlations (R_rest) and correlations between scores on each item and the external ISAWG (R_crit).***
***The dashed line highlights the location of the line of equality.***

The results in Figure 2 may appear to suggest that R_rest coefficients, that more or less entirely concern the reliability of various items, can be straightforwardly used to infer the relative weighting that should be given to items to maximise predictive validity. However, this impression is somewhat punctured by further results in Figure 3. This chart plots the R_rest values against the coefficients from regression of the external ISAWG on each set of test items (i.e. the coefficients from 515 separate regression analyses). These regressions were fitted using the non-negative least squares method using the R package *nnls* (Mullen & van Stockum, 2012). For each individual test, the regression coefficients were rescaled so that the maximum weighted score available on the test was equal to the original maximum test

score. Figure 3 shows that, despite the strong relationships seen in Figure 2, the R_rest values display almost no relationship at all with the optimal regression coefficients[7].

The reason for this difference is that a high R_rest coefficient simultaneously tells us two things. The usual interpretation is that the item is accurately measuring the same construct as the remainder of the test – in other words, it is a good item. However, the high coefficient could also be interpreted as telling us that the item is measuring something that has already been adequately measured by the remainder of the test – in other words, the item is possibly unnecessary. To take an extreme example, an R_rest value of 1 would actually mean that we could already know how candidates would perform on this item from their performance on the rest of the test, so there would be no need for them to take it and certainly no need to include the given item in a predictive regression equation. For this reason, despite their strong association with R_crit, R_rest values are insufficient to help us identify the optimal weights that are most predictive of an external ability measure.

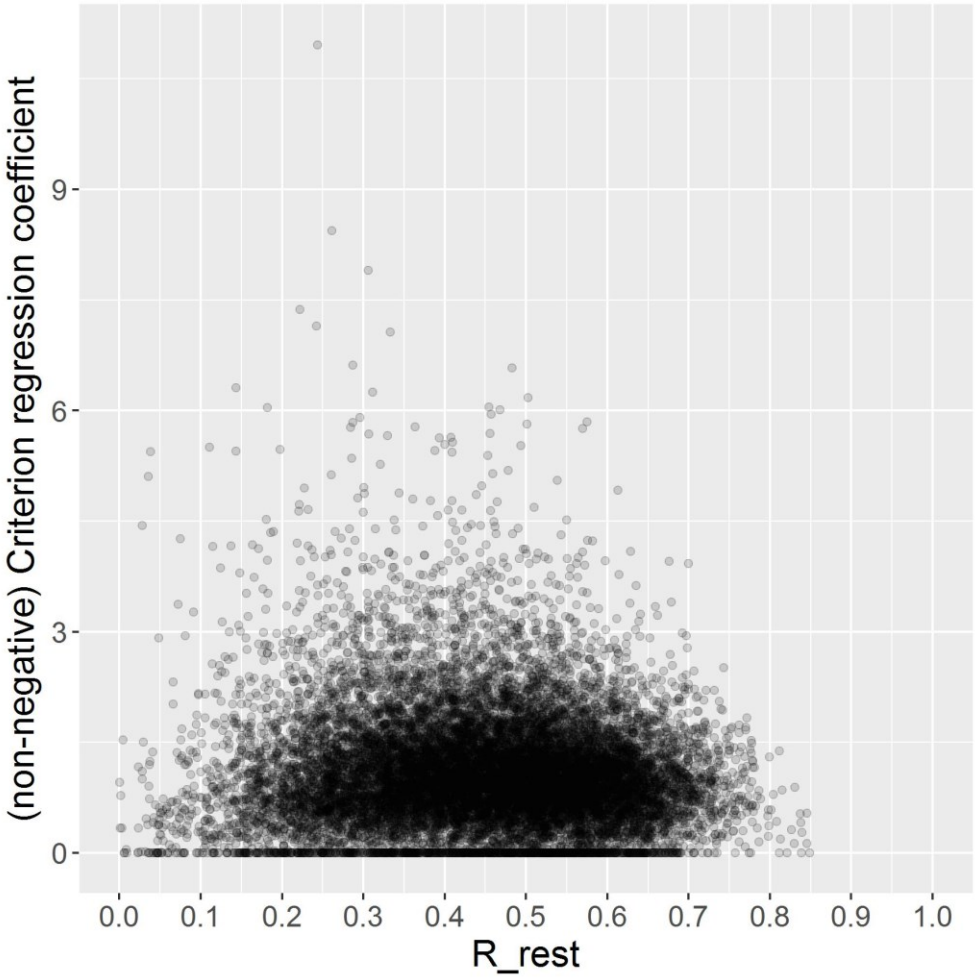


***Figure 3: The relationship between corrected item-total correlations (R_rest) and regression coefficients from non-negative regression of the ISAWG on individual item scores.***

[7] The results are essentially unchanged if negative weights are allowed but the experiment itself becomes less realistic.

To finish this section, Figure 4 takes us full circle and shows how item tariff relates to the optimal regression coefficients – that is, the optimal relative weights that should be given to scores from different items. Whilst the slight negative slope of the fitted line in the figures provides some indication that high tariff items should be given less weight, the relationship is very slight. Further details on how average regression coefficients vary with item tariff are given in Table 1.

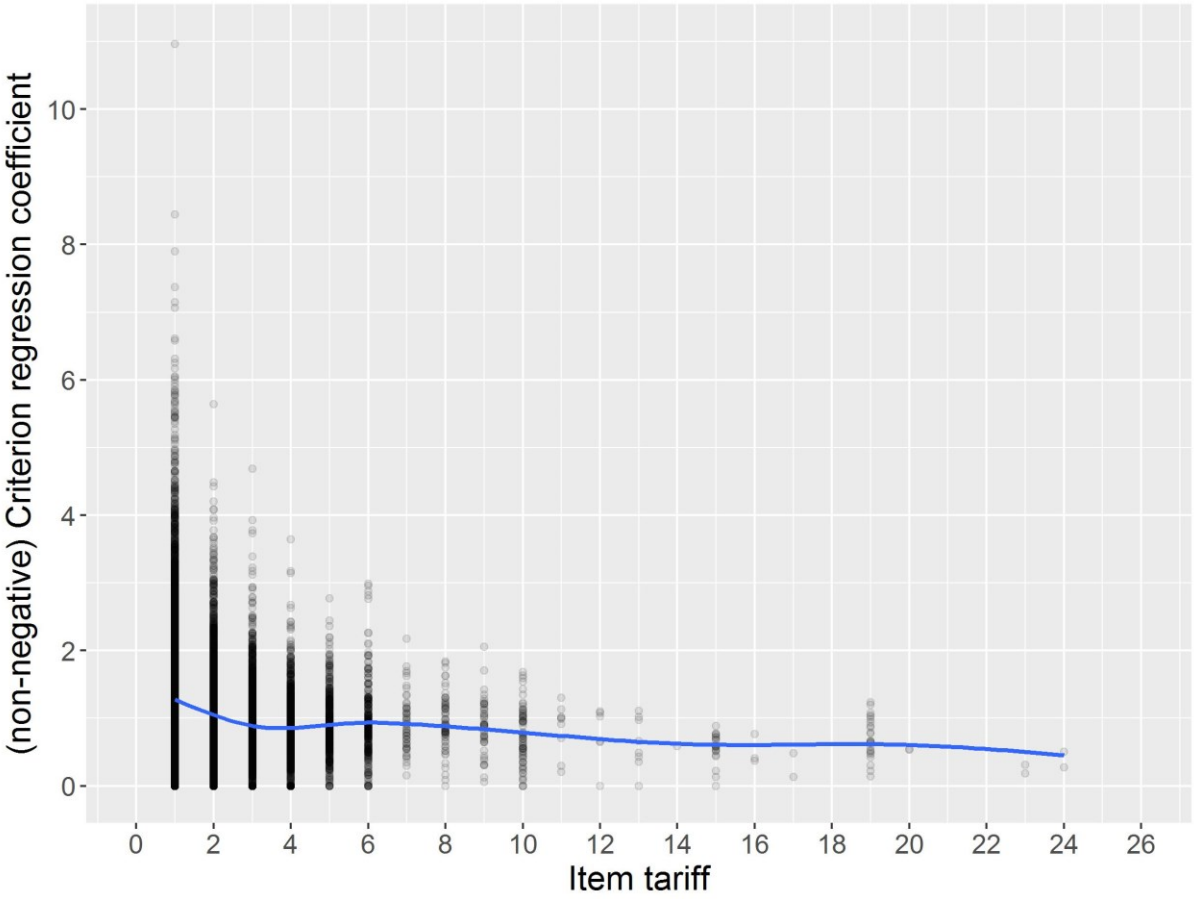


***Figure 4: The relationship between item tariffs and regression coefficients from non-negative regression of the ISAWG on individual item scores.***

***Table 1: Mean correction item-total correlations (R_rest), correlations with external ISAWG (R_crit) and coefficients from non-negative regression for items of varying tariffs.***

| Item tariff | Mean R_rest coefficient | Mean R_crit coefficient | Mean regression coefficient | N items |
|---|---|---|---|---|
| 1 | 0.38 | 0.34 | 1.29 | 7164 |
| 2 | 0.46 | 0.42 | 1.01 | 4578 |
| 3 | 0.52 | 0.48 | 0.92 | 2083 |
| 4 | 0.54 | 0.50 | 0.87 | 997 |
| 5-9 | 0.59 | 0.52 | 0.91 | 891 |
| 10-14 | 0.60 | 0.48 | 0.73 | 92 |
| 15 or more | 0.65 | 0.53 | 0.59 | 57 |

Table 1 indicates that, in general, if our aim was to optimise the correlation of individual test scores with a measure of general ability (which it isn't) then, on average, total scores from 1-mark items should be multiplied so they contribute an additional 30 per cent to the marks, scores on items with between 3 and 9 marks should be adjusted to contribute perhaps 10 per cent less, and scores on items with more marks than this should perhaps contribute 30 per cent less. It is worth noting that the size of these coefficients very closely matches those in a paper by Howard Wainer who declared "it don't make no nevermind" (Wainer, 1976). Specifically, his paper asserted that where the optimal values of some regression coefficients were up to three times the size of the optimal values of others, using these optimal values in place of unit weights for all items would not lead to noticeable improvement in predictive power. Although Laughlin (1978) noted some caveats to this claim, the fact remains that the differences in regression coefficients in Table 1 are highly unlikely to lead to substantial changes in the rank order of candidates. Combining this observation with the fact that the purpose of scoring is by no means necessarily to optimise the correlation of test scores with a general measure of ability (the ISAWG) in any case, these results do not provide any strong evidence to suspect that extended response questions are awarded too many marks.

Despite the above reservations, a brief analysis was completed to explore the effect of applying a rescoring strategy based upon the mean regression coefficients in Table 1, such that more weight was given to 1-mark items and less to those with higher tariffs. The results revealed that the correlations between original and rescored test totals were universally high with a minimum of 0.985 and a median value of 0.999. Furthermore, the impact of the rescoring strategy on predictive correlations was almost negligible with the median correlation increasing from 0.777 to 0.778. As such, although there is some evidence that extended response items may be over-rewarded at present (e.g. Figure 1), the available gains from making changes to address this do not provide a compelling case for action.

The following sections will expand upon the relationship between optimising reliability and optimising predictive value and then tackle the more interesting, but related, problem of the best way to go about selecting which items to include in a test. Importantly, these following sections will go beyond linear methods of combining scores (such as might be generated by linear regression) and incorporate an analysis of the effectiveness of IRT in improving reliability.

## Optimal test scoring

### Method

This section gives details of a more thorough study of the impact of different scoring methods on reliability and predictive value. Across each of the 515 data sets used in the analysis, the following methods of producing test scores based on all test items were tried:

1. **Raw sum scoring** (i.e. the usual default). Note that for the purposes of the analysis in this section, scoring using Rasch analysis would lead to exactly the same results.
2. Weighted scoring devised to maximise the value of **Cronbach's alpha**. This metric was chosen as it is perhaps the most widely used reliability coefficient in educational

assessment. This is actually achieved by assigning weights to items as defined by the first component in a principal components analysis (Lord, 1958).

3. Weighted scoring devised to maximise the value of **McDonald's omega**. This was used as McDonald's omega is often asserted to be a more robust measure of test reliability than alpha (Hayes and Pritchard, 2013, Revelle and Zinbarg, 2009). Details of the method used to maximise Omega are given in Armor (1973).
4. The weighted scoring scheme devised by **Guilford** (1941). In essence, this method gives more weight to items with greater item-total correlations.
5. Scoring based upon IRT using a **graded response model** (GRM). This model was chosen as it allows for the fact that the discrimination parameters may vary between items. It also allows for the possibility that the gaps in difficulty between marks within an item may be inconsistent. For example, the possibility that more additional credit should be awarded for achieving 2 marks rather than 1 on an item than should be given for achieving 1 mark rather than zero. EAP[8] estimates of ability were used (Kim and Nicewander, 1993).
6. Scoring based upon IRT using the **generalized partial credit model** (GPCM). This model allows for the fact that different items may have differing discriminations but is designed so that, within each item, each additional mark will yield the same amount of additional credit. Both the GPCM and the GRM were fitted using the R package *mirt* (Chalmers, 2012).
7. Scoring based upon direct linear **regression** of the external variable (the ISAWG) on the individual item scores within each test. This method directly optimised the outcome of interest. Although in practice such criterion variables are rarely available at the time at which test scores are being awarded, this method was used to provide a sense of the largest possible increases in predictive value that could possibly be achieved by amending the method of scoring. As described earlier, linear regression was constrained so that no negative coefficients were allowed.

For each of methods 2 to 7 the improvement in score reliability over and above using simple sum scoring (method 1) was calculated. In each case this was done on the given method's own terms. For example, for method 2 (maximise alpha), the change in Cronbach's alpha between raw sum scoring and scoring specifically devised to optimise Cronbach's alpha was calculated. Similarly for method 5 (the GRM), an estimate of the reliability of sum scores based upon the fitted GRM was compared to the reliability of the final EAP estimates of ability. In the absence of an obvious alternative, for method 4 (Guildford's method) reliability was quantified in terms of Cronbach's alpha. Method 6 (linear regression) does not have a specific reliability metric associated with it. For this reason, to allow this method to be treated consistently with all the others, the difference in Pearson correlation with the external ISAWG between the linear regression method of scoring and raw sum scoring was used as a proxy for difference in reliability.

[8] EAP stands for expected a posteriori. These ability estimates are each candidate's expected level of ability, taking into account the population distribution of abilities and their performance on individual items.

Along with the measures of reliability, for each of methods 2 to 7, the change in the Spearman correlation with the external ISAWG between raw sum scoring and each of the 'optimal' methods was calculated.

**Results**

Out of the 515 data sets used in analysis, there were six where either the GRM or GPCM failed to converge and so, for consistency, these were removed from all analyses at this stage. The results of the analysis for the remaining 509 data sets are shown in Table 2. As can be seen, raw score reliabilities were typically just below 0.9. As expected, raw score reliabilities calculated using either omega or IRT were slightly higher than those calculated using Cronbach's alpha. The median Pearson correlation of raw scores with the external ISAWG was just above 0.8.

Table 2 shows that efforts to optimise any of the various reliability measures were able to achieve slight increases on average with median increases in reliability of around 0.01. However, of more interest for this research study is the final column of Table 2. This shows that with the exception of directly optimising the correlation with the external ISAWG, none of the methods led to noticeable improvements in the predictive value of test scores. In other words, none of the methods specifically designed to improve reliability were found to have a noticeable effect on predictive value. Indeed the values in this table are shown to 4 decimal places just to allow any differences at all to be visible. Interestingly, it can be seen that using the GRM approach to creating test scores was more likely to lead to a *decrease* in predictive correlations than an increase.

***Table 2: Median values of raw score reliabilities and correlations with external ISAWG, along with median improvements based upon different scoring methods.***

| Scoring methods | Median values of… | | | |
|---|---|---|---|---|
| | Raw score reliability | Improvement in reliability using 'optimal' score | Raw score (Spearman) correlation with ISAWG | Improvement in correlation using 'optimal' score |
| **Classical** | | | | |
| Max Alpha | 0.8848 | 0.0143 | 0.8160 | 0.0008 |
| Max Omega | 0.9019 | 0.0094 | 0.8160 | 0.0006 |
| Guilford | 0.8848 | 0.0141 | 0.8160 | 0.0013 |
| **IRT** | | | | |
| GPCM | 0.8912 | 0.0092 | 0.8160 | 0.0005 |
| GRM | 0.8912 | 0.0115 | 0.8160 | -0.0002 |
| **Regression** | | | | |
| Linear regression | 0.8260 | 0.0131 | 0.8160 | 0.0116 |

One issue with the analysis in Table 2 is that, because the estimated reliabilities for most tests are already very high, and the achieved improvements in reliabilities so small, it is perhaps to be expected that there would be little gain in predictive value. To address this issue, Figure 5 shows how actual improvements in predictive value relate to estimated improvements in reliability based upon different scoring methods. Each point on this chart represents one of the tests being analysed, and the figure plots the change in estimated reliability against the change in predictive value. Any assessments where *any* of the methods

resulted in negative loadings being associated with *any* items were removed as, in practice, we would never penalise a candidate for answering an item correctly. This means that Figure 5 is restricted to 484 assessments. This chart shows that even where large improvements in reliability are predicted, these are rarely associated with concomitant increases in predictive value. This is particularly true for the IRT-based methods where the data sets apparently associated with the biggest opportunities for increases in reliability were often those where IRT-based scoring led to decreases in predictive value. The red dashed lines in Figure 5 show the expected increases in predictive correlation given the increases in estimated reliability[9]. As can be seen, even in cases where a particular scoring method (such as the Guilford approach) appears to be associated with slight improvements in predictive value, these generally fall well short of the level that would be expected[10].

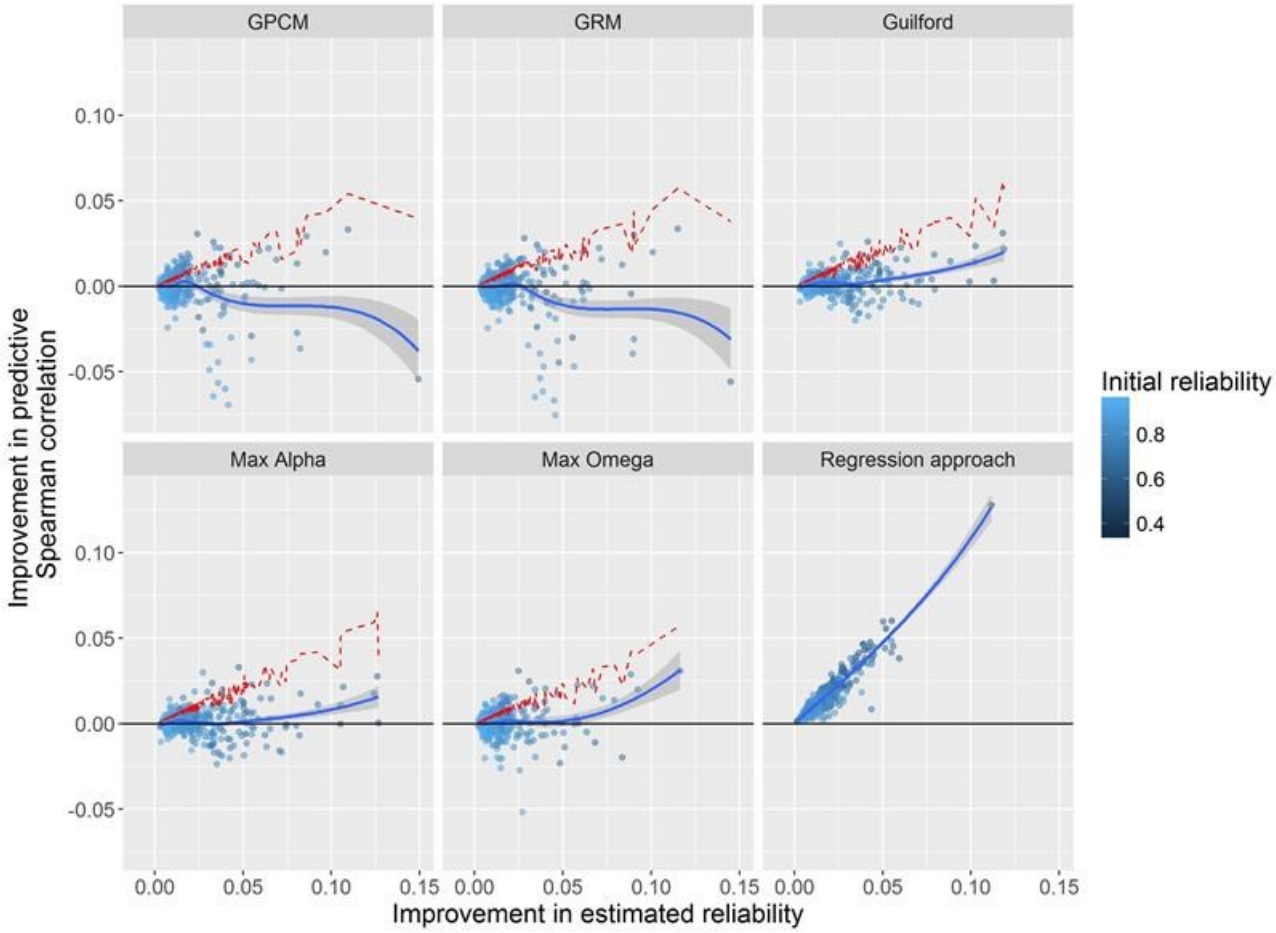


***Figure 5: Relationship between improvements in estimated reliability and improvements in predictive value.***
***The blue lines show the smoothed relationship between these two values. The red dashed line shows the theoretical increase in correlation that would be expected given the increase in reliability.***

[9] $Expected\ new\ correlation = Original\ correlation * \sqrt{\frac{New\ estimated\ reliability}{Original\ estimated\ reliability}}$.

[10] Note that for the regression approach the original 'reliability' is really just the Pearson correlation with the external ISAWG. Therefore, in this case, the chart simply shows the difference between Pearson and Spearman correlations.

**Further analyses based on split-halves**

In some ways, the general findings from the above analysis, that different methods of scoring tests are unlikely to yield genuine improvements in the predictive value of test scores, are entirely to be expected – they are exactly the findings that have repeatedly been found in decades of research on the same subject. However, the fact that scoring based on IRT-based methods may in fact reduce the validity of test scores, whilst at the same time being associated with claims of increased reliability is new. This finding is potentially important, as with the increased role of technology in educational assessment, and the widespread use of computer adaptive tests, there are numerous opportunities for traditional raw sum scoring to be replaced by an IRT-based scoring method.

With this in mind, further analysis looking specifically at the IRT-based approaches to test scoring was done. The aim of analysis was to see whether the result in the previous section (that IRT-based scoring did not generally lead to an increase in predictive value) was repeated if the reliance on the ISAWG as an external measure of ability was replaced with something else.

For this analysis, each of the 515 tests was initially split randomly into two halves so that each part contained half the number of marks of the original full test. As far as possible, this was done so that the distribution of item maxima was consistent between the two halves (i.e. both halves contained similar numbers of 1-mark items, similar numbers of 2-marks items, and so on)[11]. Using the first half of the items only, three methods of producing test scores were trialled: raw sum scores, EAP ability estimates based upon the GRM, and EAP ability estimates based upon the GPCM. The reliabilities of each of these scores were estimated along with each of their correlations with the raw sum scores on the second half of the test. As before, the aim of analysis was to see whether estimated improvements in reliability based upon IRT-based scoring actually translated into improved correlations.

Note that, even though the scores on the second half of the test were always represented as a sum score, this need not imply that the sum score on the first half will be the best predictor. According to an idealised IRT model, item scores are made up of a true element (ability) and a purely random element (the probability of getting an item correct given ability). Theoretically, any method that reduces the influence of the random element on candidate scores ought to lead to improved correlation with any external measure of ability even if it is calculated in a different way. This experiment allows to us to test the extent to which this is really the case.

The results of this analysis are shown in
Table 3. This analysis is restricted to the 500 data sets where neither of the methods produced any negative discrimination parameters. In addition, in order to investigate whether the results of analysis might be explained by multidimensionality, the method of Velicer (1976) was used to estimate the correct number of dimensions in the test. The last few rows on
Table 3 restrict analysis to the 326 tests where the true number of dimensions was estimated to be equal to one.

[11] This was achieved using functions within the R package *KernEqWPS* which is available from https://github.com/CambridgeAssessmentResearch/KernEqWPS.

As can be seen, since we are now dealing with half-tests, the starting raw score reliabilities were a little lower than in the earlier analysis. As in the earlier analysis, IRT scoring methods were associated with small increases in the estimated reliabilities of the scores. However, once again, these were not found to translate into meaningful noticeable increases in the correlation with the raw scores from the remaining half of the test. The results for apparently unidimensional tests were not appreciably different from those across all tests.

Further details are given in Figure 6. Even if we ignore the outliers on the right hand sides of these charts, we can see that increases in estimated reliability did not translate into the expected gains in predictive correlations. The results in this section are marginally more positive than those in the previous one (Figure 5). For example, instead of a median increase in predictive correlation below 0 for the GRM, in almost three-quarters of cases the predictive correlation from IRT scoring was higher than that from raw scoring. Nonetheless, the results confirm that a narrow focus on optimising reliability as defined by IRT can lead to overoptimistic expectations regarding increases in validity.

***Table 3: Median values of raw score reliabilities and correlations with other half, along with median improvements based upon different scoring methods.***

| | **Median values of…** | | | |
|---|---|---|---|---|
| **Scoring method** | **Raw score reliability** | **Improvement in reliability using 'optimal' score** | **Raw score (Spearman) correlation with half 2** | **Improvement in correlation using 'optimal' score** |
| **All 500 data sets** | | | | |
| GPCM | 0.8095 | 0.0139 | 0.8061 | 0.0036 |
| GRM | 0.8096 | 0.0170 | 0.8061 | 0.0029 |
| **326 unidimensional tests only** | | | | |
| GPCM | 0.7715 | 0.0160 | 0.7712 | 0.0039 |
| GRM | 0.7711 | 0.0201 | 0.7712 | 0.0034 |

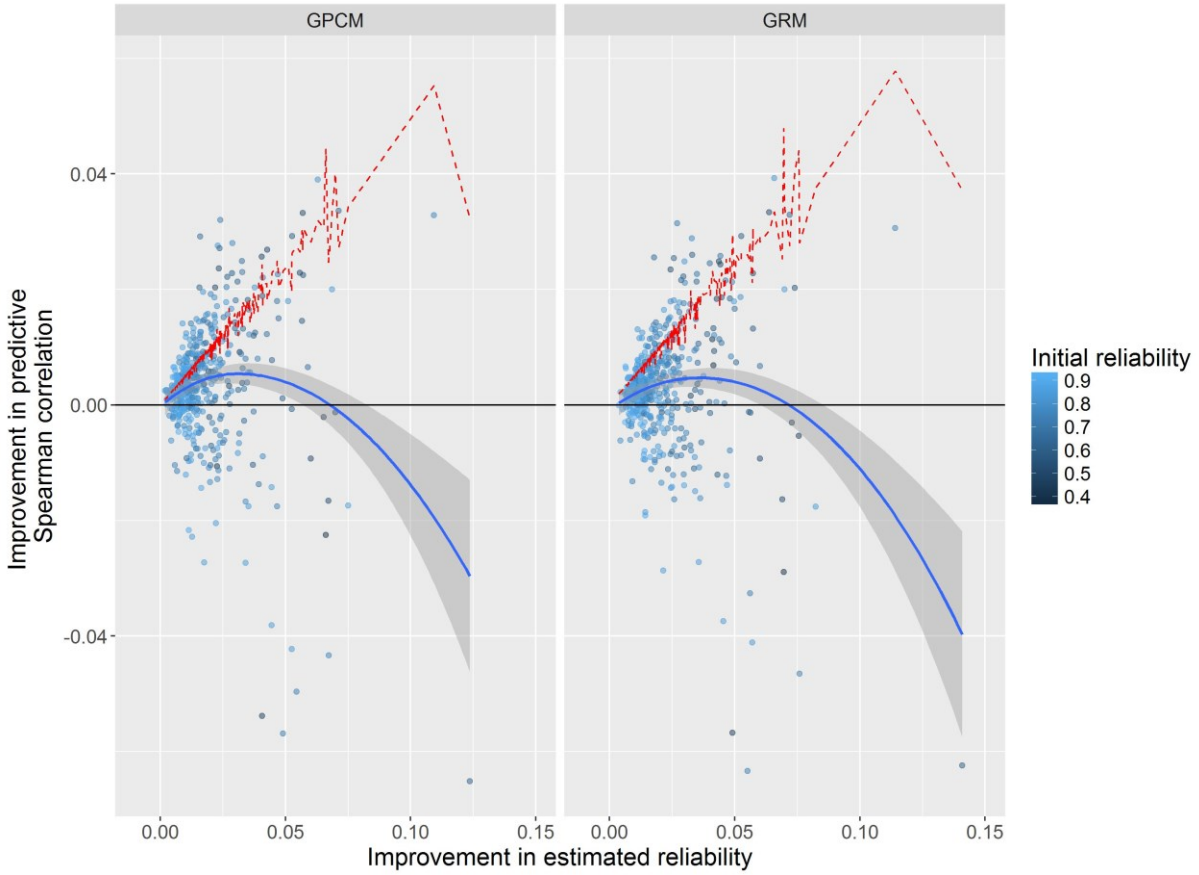


***Figure 6: Relationship between improvements in estimated reliability and improvements in predictive value for the split-half experiment.***
***The blue lines show the smoothed relationship between these two values. The red dashed line shows the theoretical increase in correlation that would be expected given the increase in reliability.***

**Summary**

This section has explored the impact of different approaches to test scoring and shown that whilst various more complex approaches to scoring (optimise alpha, IRT methods, and so on) are associated with increases in estimated reliability coefficients, these do not generally translate into the improvements in predictive power that might be hoped for. This leads to some scepticism as to whether the improvements in reliability are themselves real, or whether they are illusory.

On the one hand, these findings are entirely obvious and expected – they fit with decades of research on the same topic. On the other hand, from a practical perspective, we know that it is all too easy to focus purely on optimising reliability. Indeed, much psychometric research into the effectiveness of different IRT methods will simulate data such that random error on items is exactly that – random. If this were actually true, then any effort to reduce the impact of these 'random' measurement errors on test scores should (almost certainly[12]) lead to

[12] Some discussion of the "attenuation paradox" (Loevinger, 1954), which provides a counterexample to this idea, is provided in the discussion section. For the present, it suffices to say that it is not relevant to the results presented in this section.

improvements in the efficacy of test scores in all respects. Similarly, the vast majority of practical applications of IRT rely upon unidimensional models where any lack of fit in such models is dismissed as being too small to be worried about. In one sense this is fine as IRT is supposed to be used to tackle practical problems and need not necessarily be 'true'. However, as we use assumption laden models to describe the behaviour of tests, we must beware of being seduced by the statistics that are produced as a by-product of analysis – particularly reliability.

Having seen that theoretical gains in reliability from using IRT (or other methods) to create test scores need not translate into actual practical gains in predictive value, we now turn our attention to another possible application of psychometrics – automatic test construction.

## Optimal item selection

### Method

For any given test, the aim of analysis in this section was to trial different methods of constructing half-length tests from the full tests that were in some sense optimal. Specifically, for each of the 515 tests being analysed, various methods were trialled of selecting items from within these tests so that the total of the item maxima was half the size of the maximum number of marks available on the original test. The following methods were trialled:

1. **Random selection** of items. This is done so that, as far as possible, the proportion of items with various maximum available scores would match the original full test. For example, the proportion of items that were worth 1-mark would be as close as possible to the proportion in the original test.
2. Maximum total test information based upon the **generalised partial credit model** (GPCM). Item information functions are widely used in computer adaptive testing to help select the most appropriate item for a given candidate. In the context of a fixed test (i.e. all candidates get the same items), we can average item information across candidates for the same purpose and then note that the sum of item information functions is the total test information function. Mixed integer programming was used to find the combination of items leading to the greatest test information, such that the sum of the item maxima equalled the target level (see, for example, Diao & van der Linden, 2011).
3. Maximum test information based upon the **graded response model** (GRM). As above, but using a slightly different approach to IRT modelling.
4. Maximum test information based upon the **Rasch partial credit model**. For the previous section of this paper (test scoring), there was no need to trial Rasch analysis as the resulting ability estimates have a one-to-one correspondence to raw sum scores. However, for the item selection problem, Rasch analysis provides an approach that uses the systematic framework of IRT without the dangers associated with attempting to utilise differing discrimination parameters within the model.
5. Selection of items that **maximise Cronbach's alpha**. This was trialled as it represents a fairly common approach to item selection, particularly when researchers are interested in reducing the length of scale whilst retaining reliability (Thompson, 1990).

6. Selection of items that **maximise the sum of the item-total covariances**. Given the frequent criticisms of alpha, it was of interest to trial alternative non-IRT based techniques. This method simply quantified the quality of each item in terms of its covariance with the test total. The use of covariance (rather than correlation) meant that the method prioritised items with high standard deviations. For example, in the case of 1-mark items, this would mean prioritising items that roughly half of students answered correctly. On the other hand, it may also mean prioritising items with missing marks in the middle of the scale (e.g. all candidates awarded either 0 or 2 marks).
7. Selection of items that **maximise the sum of the R_rest correlations**. As above but with a focus on (corrected) item-total correlations rather than covariance.
8. Selection of items that **maximise the sum of the item variances**. This final method ignored issues of correlation and simply prioritised choosing items with a much variation in their scores as possible.

With the exceptions of methods 1 (random selection) and 5 (maximise alpha), all other item selection tasks were completed with the help of mixed integer programming. Mixed integer programming is a method for approaching optimisation problems where the parameters being optimised are all integers. In this special case, the parameters being optimised are in fact all binary denoting whether or not each item should be included in the particular selection. Mixed integer programming allows us to optimise the sum of some given item characteristic (e.g. item information) subject to the constraint that the sum of the maxima of the included items must be equal to half the full test maximum. These elements of analysis were completed using the R package *ompr* (Schumacher, 2017). Maximisation of Cronbach's alpha was done using a stepwise method.

For each selection of items, the reliability of the raw sum scores was calculated. In addition, the correlation of the raw sum of the selected items with the external ISAWG was calculated. This provided some idea of the impact of the different item selection algorithms on the predictive value of test scores. The impact of using alternative scoring algorithms (e.g. IRT scoring) alongside the item selection methods was also investigated. However, as found in the previous section, it was not found to have any positive impact on predictive validity and so, for brevity, the results are omitted.

**Results**

As in some earlier analyses, any tests where any of the IRT methods resulted in negative discrimination parameters were removed from analysis. This was particularly important for this analysis as negative discrimination parameters will still be associated with positive item information functions leading to potentially misleading results. The results are restricted to the remaining 487 tests.

The results of analysis are shown in Table 4. As expected, the algorithms specifically designed to optimise particular reliability indices achieved this. For example, the highest median values of Cronbach's alpha were achieved by the selections of items designed to maximise that criteria. Similarly, the GRM and GPCM methods led to the highest reliability

indices as estimated using the GRM[13]. However, the fairly major differences in reliability between these methods and random selections of items were not reflected in equally large improvements in the predictive value of test scores. In particular, whilst maximising alpha led to the highest levels of reliability (as measured in this particular way), it was also associated with one of the lowest levels of predictive value only just ahead of random selection.

Whilst both the GRM and GPCM methods were associated with noticeable increases in reliability over the Rasch method, the differences in predictive value between these methods were tiny. Likewise, whilst item selection based on non-IRT statistics (aside from alpha) were not associated with particularly high level of reliability, each of these methods was associated with similar median predictive correlations to the more complex methods.

***Table 4: Median correlations with the ISAWG and reliability indices for sum scores from items from the various item selection algorithms.***

| | Median across 487 assessments of… | | |
|---|---|---|---|
| **Item selection method** | **Correlation with ISAWG** | **Alpha** | **Sum-scale reliability (estimated using the GRM)** |
| Select at random | 0.770 | 0.799 | 0.810 |
| **IRT methods, maximising:** | | | |
| Test information (GPCM) | 0.796 | 0.854 | 0.862 |
| Test information (GRM) | 0.797 | 0.854 | 0.861 |
| Test information (Rasch) | 0.789 | 0.823 | 0.826 |
| **Classical methods, maximising:** | | | |
| Cronbach's alpha | 0.783 | 0.862 | 0.851 |
| Sum of item-total covariances | 0.795 | 0.846 | 0.845 |
| Sum of R_rest values | 0.791 | 0.848 | 0.846 |
| Sum of item variances | 0.786 | 0.817 | 0.818 |

Further details on these results are given in Figure 7 and Figure 8. Figure 7 shows the distribution of the differences in reliability and predictive value between sum scores from each of the item selection methods and original full length tests. This figure provides visual confirmation that, whilst random item selection or reliance on the Rasch model may be associated with apparently larger reductions in reliability, such differences are far less evident when the resulting scores are assessed in terms of predictive value.

[13] In most cases (360 out of 487) the GRM and GPCM methods both selected exactly the same set of items. Note that, in contrast, maximising alpha only led to the same selection of items as the GRM in 25 instances and maximising the sum of item variances *never* led to the same selection of items as maximising the GRM. Note that these item selections maximise test information rather than directly optimising reliability. This is why the GPCM method leads to a very slightly higher median reliability value than the GRM selection method.

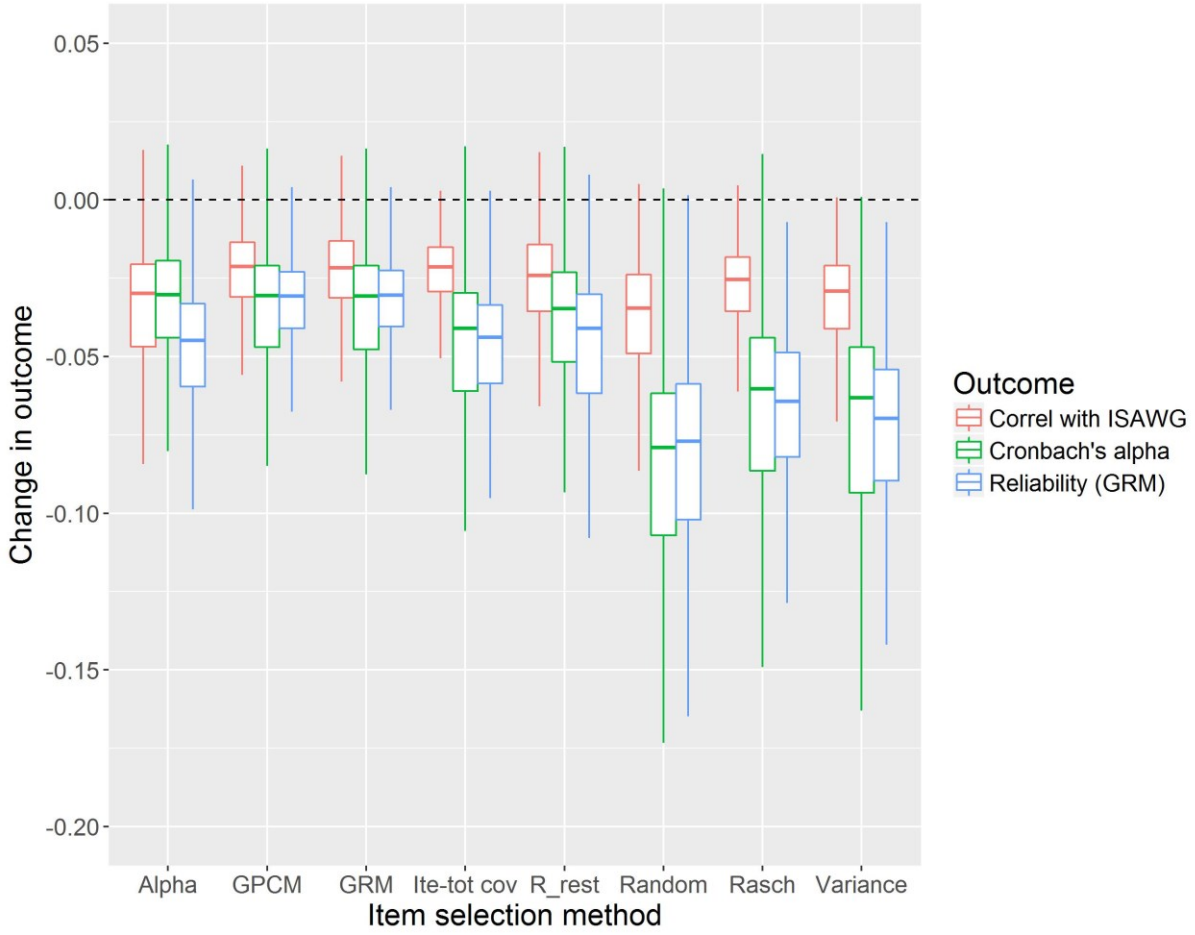


***Figure 7: Distribution of differences in reliability and predictive value of various item selection methods from the original full length test***

Figure 8 allows us to zoom in on these results more closely by presenting the mean differences in reliability and predictive value between the full length tests and the various item selection methods. Confidence intervals accounting for variation in results across the 487 tests in analysis are also included. This chart confirms that, whilst maximising alpha was related to the highest levels of (one measure of) reliability, it was one of the worst possible methods as regards ensuring that the derived tests have predictive value. The GRM and GPCM methods of item selection led to apparently important gains in terms of reliability over other methods. However, these same methods led to little benefit when evaluated in terms of the predictive value of the selected items.

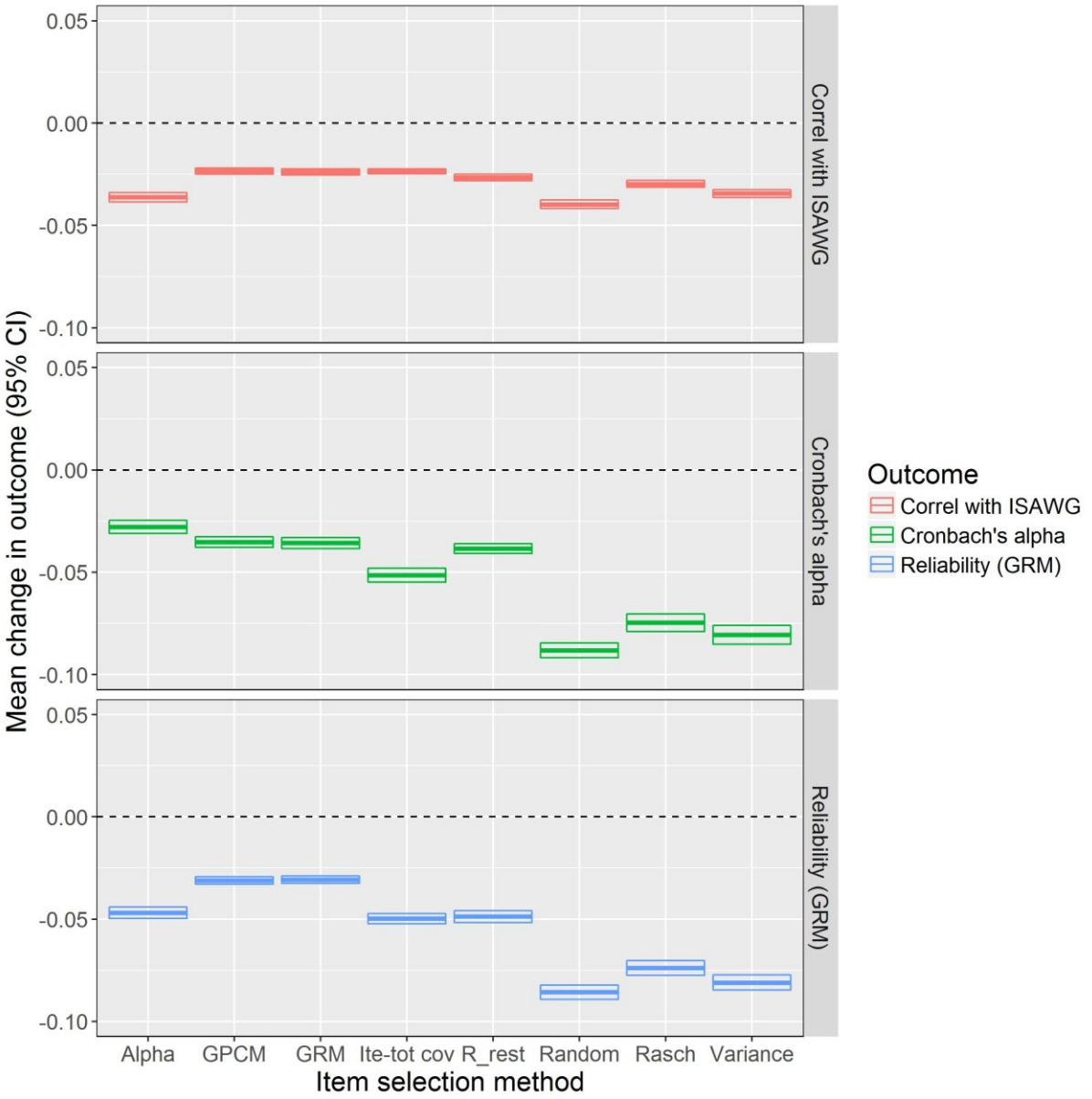


***Figure 8: Mean and confidence interval of differences in reliability and predictive value of various item selection methods from the original full length test***

**Summary**

In some ways the results of this section match those of the previous one: purported gains in reliability from using complex methods do not translate into hoped-for improvements in predictive value. That said, in contrast to the previous section, it is worth noting that the most complex IRT methods do in fact lead to the selections of items with the greatest predictive value – albeit only marginally. In this way, the use of such methods is supported by the results in this section. However, as with the previous section, the results also urge caution in drawing conclusions based purely upon reliability indices.

## Discussion

Devising methods for choosing items to include in a test, and for creating test scores based on those items, is one of the oldest areas of research in educational assessment. With the increased desire for both on-demand testing and computer adaptive testing, these topics are now of more importance than ever before. This research study has provided an empirical investigation of these issues. The research here is based upon real empirical data and summarises results across hundreds of real assessments. This is a major difference to most published research studies, which either rely purely upon simulations or restrict analysis of real data to a handful of assessments.

In many ways, the results with regard to different methods of test scoring are entirely expected and yield the same findings as decades of previous research on this topic. It is clear that reweighting items is "futile" (Wang & Stanley, 1970, page 688) in that it rarely leads to large improvements in reliability and, as often as not, fails to improve predictive value. Most importantly this research has shown that even where different scoring methods, such as IRT-based methods, do lead to apparent gains in reliability, these rarely translate into expected increases in predictive value.

This latter finding is important for the way in which reliability indices are interpreted. Usually, for the purposes of reliability, we think of test scores consisting of a 'true' element relating to candidates' underlying ability and an 'error' element, which in most discussions (and nearly all simulation studies) is assumed to be completely random. If this conceptualisation of reliability were true, then, broadly speaking, any increases in reliability indices would mean decreasing the impact of these random fluctuations. It would therefore also necessarily lead to improvements in predictive value regardless of the outcome we were interested in[14].

The result from this research that increases in reliability often do not lead to increases in predictive value reveals the falsehood of this conceptualisation of error. Increases in reliability do not necessarily mean decreasing the impact of purely random error but simply a decreased emphasis on item scores that are deemed non-discriminating by a particular psychometric model. This reminds us that, in fact, 'errors' (or rather deviations from expectations) in educational assessment are not random at all, but rather they are the responses of human beings to a given stimulus based on their own knowledge and experience. This is not to say that psychometric models or reliability indices should not be used. After all, they can be helpful in practical tasks such as test construction, equating, or in predicting the likely variability in test scores if candidates were to take another version of the test (Benton, 2013). However, it is crucial that the statistical by-products of such models, such as reliability indices, are treated with appropriate care.

This research also has implications for the use of IRT scoring methods in applications such as computer adaptive tests. Most research into this topic (e.g Wang & Vispoel, 1998; Penfield & Bergeron, 2005; Tao, Shi & Chang, 2012) uses simulation studies, as opposed to real data sets, to help identify the most effective methods to use. Such research studies are doubtless helpful in pointing practitioners towards the best methods to use for estimating

[14] The "attenuation paradox" (Loevinger, 1954) provides a counter example to this statement. However, as discussed later in this section, it is not really relevant to the findings in this particular research study.

candidates' ability and I am not suggesting that the recommendations should be ignored. However, in any practical situation where we use more complex methods of generating candidate scores, it is clear we must be wary in how we interpret the reliability indices from such methods. The real gains in the quality of test scores from using any given method may be rather less than is suggested by the associated reliability indices.

The research in this paper has also considered the effectiveness of different methods of automated item selection for fixed (as opposed to computer adaptive) tests. In some ways the results of this section reflect those from the section on test scoring. Apparent gains in reliability from using more complex methods of item selection do not generally translate into expected improvements in predictive value. In particular, selecting items to maximise Cronbach's alpha was found to be one of the least effective methods resulting in predictive correlations hardly any higher than selecting items at random. That said, it was also clear that IRT-based methods based upon the graded response model and the generalised partial credit model were among the most effective methods for this task. Thus, the research here confirms that these are good approaches to item selection. The only cautionary note is that, if such methods are used, the resulting apparent benefits in terms of reliability should be treated with care. For example, where item selection within computer adaptive tests is based upon complex IRT models, we must be cautious about making claims about potential reductions in test length based upon reliability statistics alone (examples of such claims can be found in, for example, Weiss, 1982).

The analysis of item selection methods also revealed that a number of simpler methods were nearly as effective as the more complex IRT approaches. In particular, item selection based upon the Rasch model led to predictive correlations almost as good as other IRT approaches, despite the fact that the associated reliability indices were apparently inferior. This indicates that, in automated test construction, the most important consideration is to ensure that the included items are of the appropriate difficulty for the intended population. Considerations of item discriminations should be firmly in second place.

A number of non-IRT based approaches, such as maximising the sum of the item-total covariances or the sum of the corrected item-total correlations, were also found to perform very well. This confirms that automated test construction need not necessarily be reliant on an IRT approach to item analysis.

In several places throughout this paper we have seen that improvements in estimated reliability may not lead to increased predictive correlations. As discussed earlier in this paper, similar results have been shown before in the psychometric literature through the "attenuation paradox" (Loevinger, 1954) and "capitalization on chance" (Veldkamp, 2013). However, neither of these issues explain the results in this present paper (see Appendix 1 for more detail around the attenuation paradox). To further confirm that neither the attenuation paradox nor capitalisation on chance were an issue for the present study, several analyses of methods of test scoring and item selection (particularly those using the graded response model) were rerun but replacing the original real data sets with simulated data based upon the graded response model. When this was done, the concerns identified in this paper vanished. For example, recreating Figure 5 based upon simulated data found that increases in predictive correlations were almost exactly in line with expectations given improvements in

reliability. As such, the findings in this paper cannot be dismissed as relating to the research method. Rather, they relate to the fact that, although practically useful, theoretical models of assessment data are never precisely true. It is, therefore, crucial that any claims about the benefits of particular methodologies are based upon real data rather than pure theoretical argument.

## References


Armor, D. J. (1973). Theta reliability and factor scaling. *Sociological Methodology*, *5*, 17-50.

Benton, T. (2013). Exploring equivalent forms reliability using a key stage 2 reading test. *Research Papers in Education*, *28*(1), 57-74.

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.

Brown, W. (1910). Some Experimental Results in the Correlation of Mental Abilities. *British Journal of Psychology*, *3*, 296–322.

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/.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, *35*(5), 398-409.

Guilford, J.P. (1941). A simple weight scoring for test items and its reliability. *Psychometrika*, 6(6), 367-374.

Hayes, M. & Pritchard, J. (2013). *Estimation of internal reliability*, Ofqual/13/5258. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/378056/2013-01-17-estimation-of-internal-reliability.pdf.

Kane, M., & Case, S. (2004). The Reliability and Validity of Weighted Composite Scores, *Applied Measurement in Education*, *17*(3), 221-240.

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*(4), 587-599.

Knott, M., & Bartholomew, D.J. (1993). Constructing measures with Maximum Reliability. *Psychometika*, *58*(2). 331-338.

Laughlin, J.E. (1978). Comment on "Estimating coefficients in linear Models: It don't make no nevermind". *Psychological Bulletin*, *85*, 247-253.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, *51*(5), 493.

Lord, F. M. (1955). Some perspectives on" the attenuation paradox in test theory.". *Psychological bulletin*, *52*(6), 505.

Lord, F. M. (1958). Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, *23*(4), 291-296.

Mullen, K.M., & van Stockum, I.H.M. (2012). *nnls: The Lawson-Hanson algorithm for non-negative least squares (NNLS)*. R package version 1.4. Retrieved from https://CRAN.R-project.org/package=nnls.

Penfield, R. D., & Bergeron, J. M. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Applied Psychological Measurement*, *29*(3), 218-233.

Revelle, W., & Zinbarg, R. (2009). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74* (1), 145-154.

Rudner, L. (2001). Informed test component weighting. *Educational Measurement; Issues and Practice*, *20*(1), 16–19.

Schumacher, D. (2017). *ompr: Model and Solve Mixed Integer Linear Programs*. R package version 0.7.0. Retrieved from https://CRAN.R-project.org/package=ompr.

Tao, J., Shi, N. Z., & Chang, H. H. (2012). Item-weighted likelihood method for ability estimation in tests composed of both dichotomous and polytomous items. *Journal of Educational and Behavioral Statistics*, *37*(2), 298-315.

Thompson, B. (1990). ALPHAMAX: A program that maximizes coefficient alpha by selective item deletion. *Educational and Psychological Measurement*, *50*(3), 585-589.

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). Enschede, The Netherlands: RCEC.

Wainer, H (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*, 213-217.

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, *40*(5), 663-705.

Wang, T., & Vispoel, W.P. (1998). Properties of Ability Estimation Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, *35*(2), 109-135.

Weiss, D.J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing, *Applied Psychological Measurement*, *6*(4), 473-494.

## Appendix 1: R code for investigating the "attenuation paradox"

The code below can be used to repeat the kind of analyses of the attenuation paradox used by Loevinger (1954). This concerns the fact that higher item discriminations can potentially lead to increases in classical reliability indices going alongside *decreases* in validity (i.e. the correlation between test scores and true abilities). The code takes about 10 seconds to run on my own desktop computer.

The simulation shows that, as noted by Lord (1955), the 'paradox' is far less evident once the non-linear nature of the relationship between total scores and ability is accounted for. It also shows that even in a worst case scenario (few items, all of equal difficulty) extremely high discrimination parameters are required in order for the attention paradox to occur. Finally, if IRT-based estimates of reliability are used, we can find that these also decrease if the item discriminations become too high – thus completely removing the main concern regarding the "attenuation paradox" as it relates to the present research paper.

```
library(mirt)
n=10000#number of candidates in each simulation
nite=10#number of items - worst case scenario (hardly any)
sddif=0#SD of item difficulties - worst case scenario (all items of equal difficulty)
abil=rnorm(n)#simulate some "true" candidate abilities
difs=rnorm(nite,0,sddif)#simulate some item difficulties
slopes=seq(0.8,6,0.2)#item discrimination parameters to investigate
Scorrs=rep(NA,length(slopes))#placeholder for spearman correlations
Pcorrs=rep(NA,length(slopes))#holder for spearman correlations of scores with abilities
Rels=rep(NA,length(slopes))#placeholder for reliabilities

for(i in 1:length(slopes)){
as=rep(slopes[i],nite)#set up discrimination parameters for each item
ds=difs*as#convert item difficulties to item intercepts
#now simulate item scores
ites=simdata(a=as,d=as.matrix(ds),itemtype="dich",Theta=as.matrix(abil))
#record Pearson and Spearman correlations of total scores with true abilities
Scorrs[i]=cor(rowSums(ites),abil,method="spearman")
Pcorrs[i]=cor(rowSums(ites),abil)
Rels[i]=marginal_rxx(mirt(ites,1,TOL=0.05))#quick IRT reliability index
}

plot(slopes,Scorrs,type='l',xlab="Discrimination parameters"
       ,ylab="Correl of total score with true ability")
#dotted line shows main problem is with Pearson correlations
#that fail to account for non-linearity in the relationship
lines(slopes,Pcorrs,lty=2)
#show that IRT-based estimates of reliability peak in exactly the same place
lines(slopes,sqrt(Rels),lty=3)
```