versa); or whether focusing effort on constructing parallel (equally difficult) tests would be a better use of available resource.

## References

Benton, T. (2020). How useful is comparative judgement of item difficulty for standard maintaining? *Research Matters: A Cambridge Assessment Publication, 29*, 27–35.

Bramley, T. & Benton, T. (2017). *Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests*. Paper presented at the annual conference of the Association for Educational Assessment-Europe (AEA-Europe), Prague, Czech Republic, 9–11 November, 2017.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59–88.

Brennan, R. L. (Ed.). (2006). *Educational Measurement* (4th ed.). Washington, DC: American Council on Education/Praeger.

Cizek, G. J. (1993). Reconsidering Standards and Criteria. *Journal of Educational Measurement, 30*(2), 93–106.

Cizek, G. J., & Earnest, D. S. (2015). Setting performance standards on tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp.212–237). New York: Routledge.

Dwyer, A. C. (2016). Maintaining equivalent cut scores for small sample test forms. *Journal of Educational Measurement, 53*(1), 3–22.

Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a consistent unit of scale between the responses of students and judges in standard setting. *Applied Measurement in Education, 27*(1), 1–18.

Kim, S., von Davier, A. A., & Haberman, S. (2008). Small-sample equating using a synthetic linking function. *Journal of Educational Measurement, 45*(4), 325–342.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*. (2nd ed.). New York: Springer.

Impara, J.C., & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions of the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69–81.

Livingston, S.A., & Kim, S, (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*, 330–343.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods and Perspectives* (pp.175–217). Mahwah, NJ: Lawrence Erlbaum Associates.

Lorge, I., & Kruglov, L. (1953). The improvement of estimates of test difficulty. *Educational and Psychological Measurement, 13*, 34–46.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Journal of Educational Measurement, 47*(1), 54–75.

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In P. W. Holland & D. B. Rubin (Eds.), *Test Equating* (pp. 309–317). New York: Academic Press.

# How useful is comparative judgement of item difficulty for standard maintaining?

**Tom Benton** Research Division

## Introduction

Developing a way to accurately estimate the relative difficulty of two tests before any students have taken them has long been a holy grail in test development. At one time or another, various organisations have explored how well we can discern the relative difficulties of assessments without actually trialling them with students. Recent research on this topic has been produced by Cito in the Netherlands (van Onna, Lampe, & Crompvoets, 2019), ETS in the United States (Attali, Saldivia, Jackson, Schuppan, & Wanamaker, 2014) and Cambridge Assessment in the UK (Curcin, Black, & Bramley, 2009). Item trialling is often undesirable as it places some of the burden of test development upon schools and students, and can lead to concerns over the security of items.

If accurate predictions of item difficulty were possible then, in the context of UK examinations, this would mean being able to accurately set grade boundaries for this year's GCSE exams before any students have attempted the paper. It would also provide an alternative to the current approach of "comparable outcomes" to awarding and its inherent implication that (broadly speaking) the percentage of pupils achieving high grades will not change from the previous year (Benton, 2016). Outside of the UK context, being able to accurately predict the difficulty of items might allow "lowering the sample sizes required for item pretesting, leading to lower costs and increased security of items" (Attali et al., 2014, p.7).

The previous article (Bramley, 2020) has considered the extent to which a particular form of expert judgement (the 'mean estimation' variant of the Angoff Method) might provide sufficiently accurate information on the relative difficulty of two tests. The present article explores the value of expert judgements of item difficulties derived in a different manner – by comparative judgement (CJ).

In this context, a CJ study requires expert judges to sort sets of items according to their perceived difficulty (PD). The rationale for using CJ is that previous research has indicated that judges tend to "be good at predicting the relative difficulties of items but not absolute levels" (Mislevy, Sheehan, & Wingersky, 1993, p.59). Placing items in a rank order of difficulty is conceivably a more intuitive task then estimating the proportion of minimally competent candidates who will answer them correctly, as must be done under the Angoff Method. As such, a CJ approach may be considered likely to provide better estimates of the relative difficulty of items (Attali et al., 2014).

One type of CJ exercise is a pairwise comparison study where judges are shown two items at a time, and must simply decide which of the pair is more difficult to answer correctly (see, for example, Ofqual, 2015). An alternative approach is rank ordering. As an example of this method, Curcin et al. (2009) presented judges with packs of four items which they had to place into order of difficulty. In either a pairwise comparison or a rank ordering study, each item is typically included in multiple different comparisons undertaken by different judges. The results from all the judgements of all the items by all the judges are combined into a single data set and analysed using the Bradley-Terry model (or similar) to place all of the items on a continuous scale from the easiest to the most difficult. The position of each item on this scale gives a value of PD that might then be used to allow us to infer the relative difficulty of two tests overall.

Of course, it is possible to use a rank ordering approach to judging item difficulty without the need to employ the Bradley-Terry model. For example, perhaps the earliest rank ordering study of this kind (Lorge & Kruglov, 1952) required judges to review all of the items across two tests at once and place them into a single rank order. Having done this, the average rank assigned to a given item across all of the judges provides an estimate of PD.

Existing literature suggests several ways in which estimates of PD from a CJ study might be used to infer the relative difficulty of tests. Usually these rely on linking the estimates of PD for each item to empirical difficulties as defined using item response theory (IRT) or Rasch analysis. An extreme example of this approach, taken by Holmes, Meadows and Stockford (2018) and Bramley (2010), is simply to use estimates of PD from a Bradley-Terry model directly as substitutes for empirical estimates of Rasch difficulty. There seems little justification for this approach. The former relate to the probability of a judge considering one item in a pair more difficult than another, whereas the latter relate to the probability of students answering an item correctly. These are clearly distinct concepts and there is no obvious justification for using one as a substitute for the other. In other cases, PD is used as an input to a statistical model to help infer the likely location of empirical item difficulties (e.g., Mislevy et al., 1993) or to calibrate item difficulties that have been separately estimated for two tests onto the same scale (van Onna et al., 2019).

Two immediate problems occur in attempting to relate estimates of PD to IRT difficulty parameters. Firstly, from a practical point of view, it is not obvious how items with more than one mark available should be treated. CJ studies will usually only provide a single value for the PD of each item but in order to use IRT it is necessary to estimate the difficulty of each mark within the item. Secondly, the use of IRT makes it difficult to either calculate or communicate the likely accuracy with which such methods can actually equate two tests.

The aim of this article is to simplify the evidence on the value of comparative judgements of item difficulty for estimating the overall difficulty of tests. To begin with, I will review the evidence on the strength of the relationships between estimates of item difficulty derived from CJ and actual empirical difficulties. After this, I will show how we can combine perceived item difficulties with simple (non-IRT) statistical methods to estimate the relative difficulty of two tests. Crucially, the simple approach will also allow us to assess exactly how accurate equating tests based purely on PD is likely to be in general.

## How strong is the relationship between estimates of PD from paired comparisons and empirical item difficulty?

To investigate the relationship between PD derived from a CJ study and actual item difficulties, I used data from The Office of Qualifications and Examinations Regulation (Ofqual, 2015). This study of the relative difficulty of various Mathematics exams included items from six legacy GCSEs in Mathematics offered by OCR. The estimates of PD of each item were published as part of the study (Ofqual, 2015, Appendix B, pp.140–146) and it was possible to link them to empirical data on the performance of students on the same questions and evaluate the strength of the association. Each of the assessments was taken by more than 5,000 candidates, providing ample data for empirical estimates of item difficulty.

Table 1 provides further details of the assessments included in analysis. They each contained between 20 and 40 items. Each test contained both single-mark (dichotomous) and multi-mark (polytomous) items. The empirical difficulty of each item was estimated using its facility. Item facilities are usually presented on a scale from 0 to 100, and represent the mean score on an item expressed as a percentage of the maximum score available. For the items in these six tests the mean facility was close to 50% meaning that, for a typical item, candidates achieved about half of the available marks on average. The standard deviations (SD) of the facilities across items are also shown in Table 1. These show that, although the average facility was generally close to 50% in each test, items with a wide range of difficulties were included.

The most common way to evaluate the strength of the association

**Table 1: Correlations between PD and facility for the six Mathematics GCSE papers**

| *Unit* | *Tier* | *Number of items* | *Number of marks* | *Mean Facility* | *SD of Facilities* | *Correlation of PD and Facility* | *Residual SD of Facilities* |
|---|---|---|---|---|---|---|---|
| Unit 1 | Foundation | 30 | 60 | 47.9 | 28.2 | -0.57 | 23.6 |
| Unit 1 | Higher | 27 | 60 | 44.9 | 19.8 | -0.44 | 18.1 |
| Unit 2 | Foundation | 28 | 60 | 45.6 | 21.1 | -0.50 | 18.6 |
| Unit 2 | Higher | 22 | 58[1] | 50.6 | 20.9 | -0.26 | 20.6 |
| Unit 3 | Foundation | 39 | 100 | 58.2 | 23.6 | -0.57 | 19.7 |
| Unit 3 | Higher | 35 | 100 | 68.7 | 22.2 | -0.48 | 19.8 |

[1] The original test had 23 items and 60 marks available. However, one item was omitted from Ofqual's study and so only 22 items and 58 marks are included here.

between two quantities is to calculate a correlation coefficient. The (Pearson) correlations between PD and empirical facilities are also shown in Table 1[2]. The negative sign of these correlations is expected – items that are perceived to be more difficult are answered correctly less often.

Of more interest is the size of these correlations. In Table 1 the sizes of the correlations between PD and facility range from 0.26 to 0.57. However, it is not immediately clear how to interpret these values. Clearly, there is some relationship between the perceived and actual difficulty of items. However, is the relationship strong enough to be of any value in judging the relative difficulty of, and ultimately in equating, two tests?

To investigate this, I compared the strength of the correlations in Table 1 to the correlations between the overall item facilities and facilities based on very small samples of candidates. For example, for any of the above tests, we might select just one candidate. Then, for each item, the facility (based on this one candidate) is zero if they get the item completely wrong, 100 if they get it completely right, and something in between if they achieve some but not all of the marks. The correlation between the item facilities based on this one candidate and item facilities based on the full population can then be calculated. The procedure was repeated 100 times[3] for each of the six assessments to get an idea of what correlation between a facility from one candidate and the overall facility we might expect. The same method was then repeated to estimate the predictive value of data from small samples of two, three, four, five, six, seven, eight, nine and ten candidates.

The results are shown in Figure 1. The dotted lines represent the size of the correlations between PD and facility for each test (see Table 1). The boxplots show the distribution of correlations between facilities from small samples and overall facilities for each sample size across the replications. As can be seen, in most cases, the correlation between PD (based on a CJ study) and facility is very similar to what we might expect to achieve on average by using a sample of *just one candidate*. With a sample of five we can virtually guarantee that the data from even so few candidates will be more predictive of actual item difficulty than a

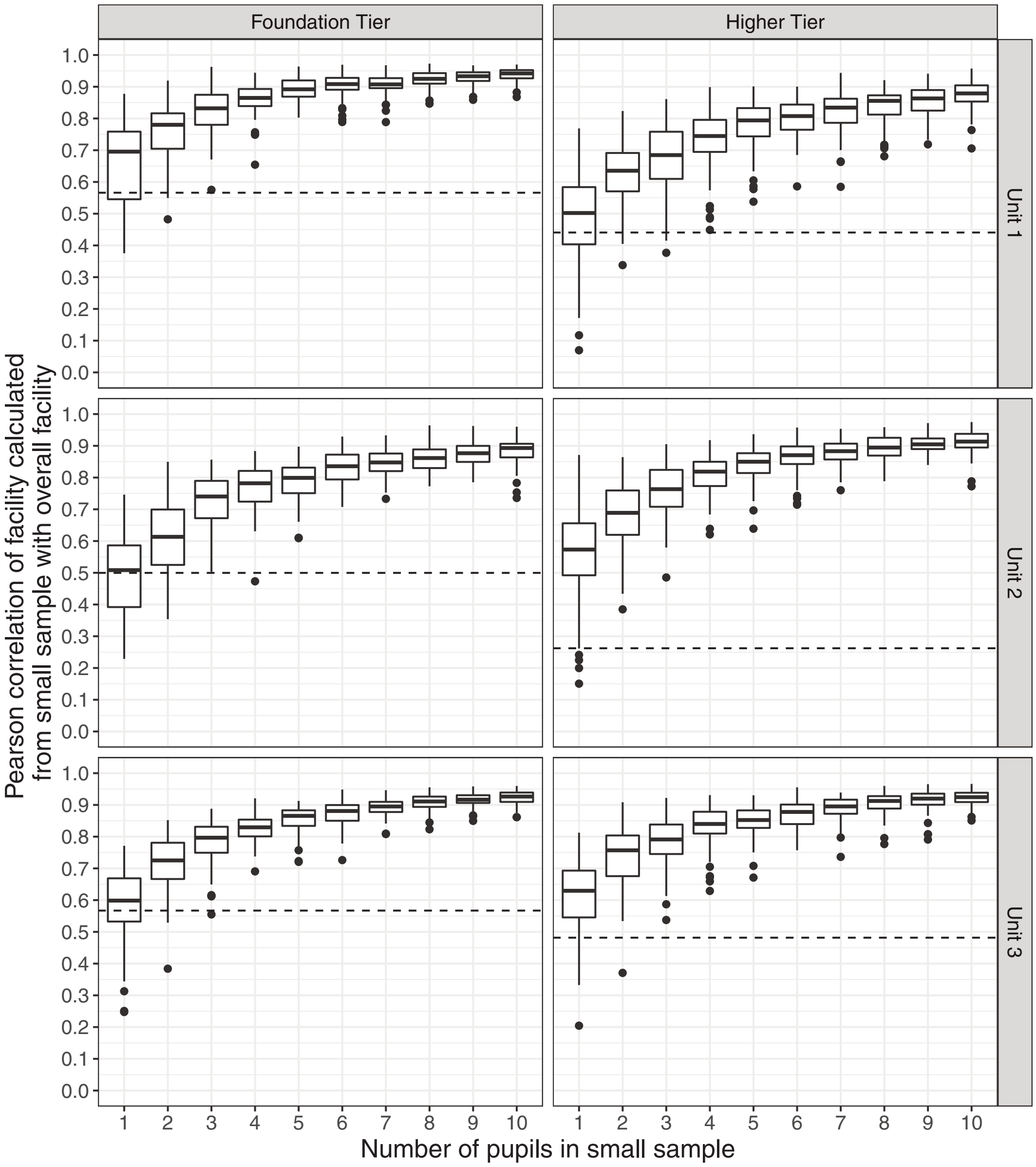

**Figure 1: Correlations between facilities calculated from small subsamples and full sample facilities for six GCSE Mathematics assessments. The dotted lines show the size of the Pearson correlation between PD and facility in each case.**

[2] Spearman correlations were of very similar magnitude and (for brevity) are not shown.

[3] That is, with 100 separate candidates.

CJ study. This immediately suggests that CJ exercises to estimate PD are a very weak source of evidence about the difficulty of items.

## An alternative to correlations – the actual accuracy of predictions

The above analysis suggests that PD cannot be seen as a strong form of evidence. However, it does not necessarily mean such information is useless. Were it possible to get even a single typical student with the correct level of exam preparation and motivation to trial a paper before it goes live, and this could be done without any security concerns, this would likely be considered a very useful resource to test developers. However, in reality this is tricky. For this reason, despite the above results, it is still of interest to explore the accuracy with which PD can predict actual item difficulty in more detail.

Like the analysis above, most studies evaluating the predictive value of PD focus upon correlation coefficients. However, looking at correlations alone can provide a misleading picture. The reason for this is that correlations will tend to increase along with the spread of values included in the study. For example, imagine that some basic Mathematics questions asking students to add two single digit numbers had been added to the above exams. Such questions would have been self-evidently easier than any other items in the exams and the vast majority of students would have answered them correctly. Thus, including such questions would make it easier for experts to correctly distinguish the relative empirical difficulty of at least some of the items, and the correlations between PD and facility would increase.

This same effect exists (perhaps in a less extreme form) whenever we use correlations to assess the strength of associations. In the current context, the greater the spread of actual item difficulties within a test, the easier it will be for experts to discern this, and the higher the correlation between PD and facility will be. Ultimately, a more useful way to understand the value of PD is to actually calculate how accurately we can predict item facilities. That is, if we were to use PD to predict the likely facility of a new item (for the same population of students), how close would that prediction be to the actual facility?

Because it is helpful for the calculations that follow in the next section, rather than evaluating the average size of differences between predicted and actual values (the mean absolute differences), we will actually use the square root of the mean squared differences (the residual standard deviation). Residual standard deviations are higher than mean absolute differences but, very broadly speaking, can be interpreted in the same way in that both give an idea of the typical difference between predicted and actual facilities.

Residual standard deviations of facilities given PD are provided in the final column of Table 1. They can actually be calculated from the overall standard deviation of facilities in each test, the correlation between facility and PD, and the number of items in the test using the formula below:

$$Residual\ SD\ of\ Facility = SD(Facility)\sqrt{\frac{(1-r^2)(\mathrm{n}-1)}{(\mathrm{n}-2)}} \qquad (1)$$

where $r$ is the correlation between PD and facility, and n is the number of items in the test. The terms relating to the number of items in the above formula adjust for the fact that a regression line will be fitted to the existing data on actual facilities. As such, without this correction, we would probably overestimate the likely accuracy of future predictions in brand new data sets.

The values in Table 1 indicate that PD allows empirical facilities to be predicted to within a little under 20% on average[4]. Crucially, by comparing these values to the overall standard deviation of facilities in each test (as Table 1 shows, these are around 22), we can see that this level of accuracy is only marginally better than could be achieved by simply guessing that all items would have a facility near the average for that test. That is, having a value for the PD of each item only marginally improves the accuracy with which we can predict empirical difficulty. We can also see that the tests displaying the highest correlation between PD and facility are not associated with predictions of facility actually being more accurate. For example, the Unit 1 Foundation tier test displayed one of the highest correlations between PD and facility but also had the highest residual standard deviations (i.e., the worst predictive accuracy).

Figure 2 allows a visual exploration of the same idea. The charts show the associations between PD and facility with a regression line shown in blue in each case. We can see that, within each assessment there is a clear relationship between PD and facility. However, there is nearly the same spread in empirical facilities for any fixed value of PD as there is overall. Around a third of the empirical facilities in Figure 2 are more than 20 percentage points away from what would be predicted based on PD.

The above results illustrate the problem of relying on correlations alone to assess the value of PD. To illustrate this further, analysis was also undertaken evaluating the association between PD of items and the percentage of candidates that answered each item fully correctly (i.e., achieved all of the available marks). This analysis showed that the correlations increased and now ranged between -0.39 and -0.73. This could be taken as indicating that PD was more predictive of the percentage of candidates answering items fully correctly than of facility. However, further analysis revealed that the actual accuracies of predictions were, in fact, worse than the accuracies of predictions of item facility (shown in Table 1). Specifically, even as the correlations increased, the residual standard deviations also increased. In other words, reporting correlations alone may not give the most appropriate picture of the value of PD.

## The actual accuracy of predictions based on PD for previous studies

The results above give a potentially disappointing picture of the accuracy with which PD can predict empirical difficulty. This is in contrast to some academic literature on this subject which presents a more positive picture of the potential of PD. To investigate this discrepancy further, results from a number of previous studies of this type are shown in Table 2. These results should not be taken as representing a systematic review of all of the articles on this issue. They simply reflect a number of articles that I am familiar with, presenting a range of views on the value of PD.

All of the studies listed in Table 2 except one (Humphry, Heldsinger, & Andrich, 2014) made use of a CJ approach to eliciting item difficulties. This exception was included simply to represent the fact that studies using methods other than CJ also occasionally claim that judges can successfully estimate the relative difficulty of items and to ensure that at least one such non-CJ study was subjected to some scrutiny.

For each study in Table 2, I have identified the reported correlation between PD and facility. In several cases, the authors reported Spearman

[4] The mean absolute difference between predicted and actual facilities is 16%..

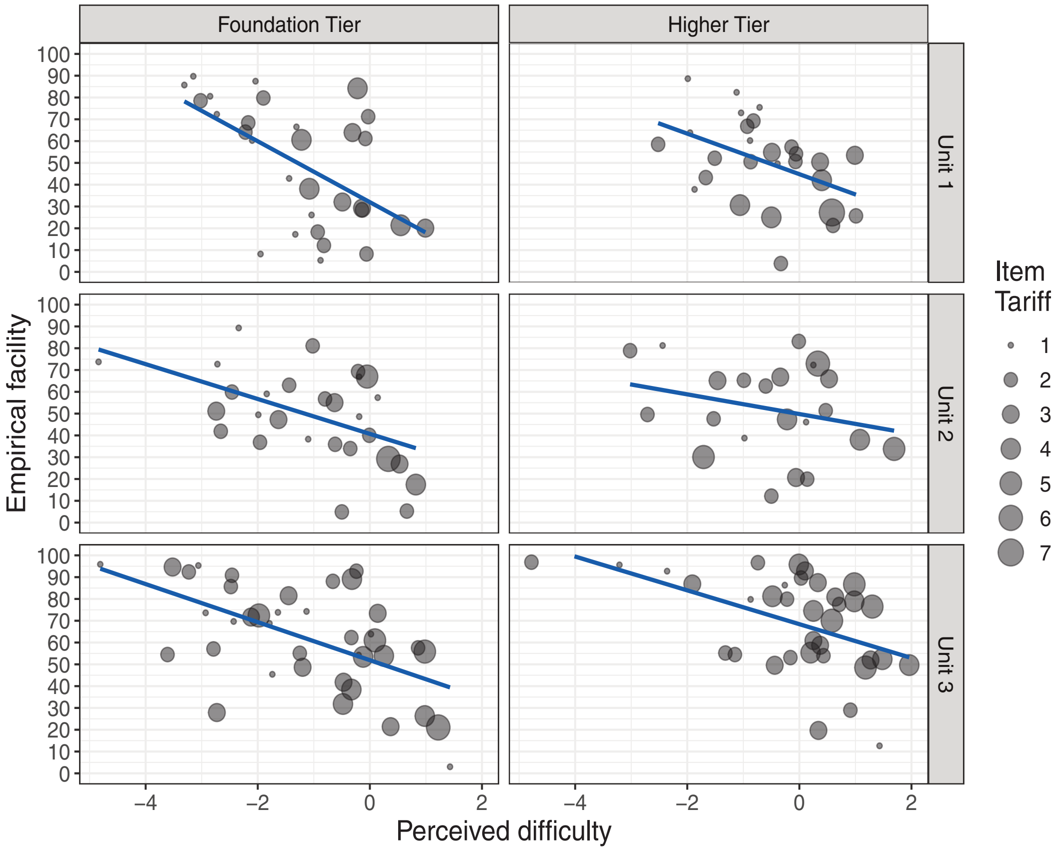

**Figure 2: Relationship between PD and item facility for each Mathematics GCSE unit.**

(i.e., rank correlations) rather than Pearson correlations, however, this should not make a major difference to results. Indeed, in several cases, by analysing data within scatterplots I was able to verify that the Pearson correlations would give similar values. The article by Humphry et al. (2014) did not present the correlation at all and it was necessary to estimate its value based on data contained within a figure in the article. As can be seen, the values of the correlations vary considerably between studies. For example, on the one hand, the study by Attali et al. (2014) found a correlation between PD and empirical difficulty close to 0.8, leading the authors to conclude that "contrary to previous investigations, judges are able to discriminate quite well between easier and harder items when they are given a comparative judgment task" (p.6). At the other end of the spectrum, Curcin et al. (2009) found a correlation of just 0.18 between PD and facility for a particular multiple choice test and were left to "speculate why rank ordering failed to elicit consistently valid judgements about question difficulty" (p.6). Thus, focussing purely on correlations gives the impression of CJ sometimes being a highly effective means to estimate item difficulty and sometimes being almost entirely ineffective.

Table 2 also provides the standard deviations of the facilities of items included in each study and, based on these and the formula provided above, the estimated residual standard deviation of facilities. Note that

**Table 2: Correlations and residual SD of facilities for studies in existing literature**

| *Author* | *Assessments studied* | *Method* | *Number of items* | *SD of Facilities* | *Correlation of PD and Facility* | *Residual SD of Facilities* |
|---|---|---|---|---|---|---|
| Humphry et al. (2014) | Multiple choice Year 7 reading test | Angoff procedure | 35 | 25.3 | -0.80 | 15.4 |
| Lorge and Kruglov (1952) | High school admissions tests in arithmetic | All items ranked by four judges. Average the item rankings. | 86 | 25.1 | -0.84 | 13.7 |
| Attali et al. (2014) | SAT Mathematics | Ranking of seven items by one judge | 7 | 28.2 | -0.79[5] | 18.9 |
| Ofqual (2018) | AS Level Mathematics | Formal pairwise comparisons | 554 | 19.2 | -0.49 | 16.8 |
| Ofqual (2017) | GCSE Biology | Formal pairwise comparisons | 351 | 23.4 | -0.50 | 20.3 |
| Curcin et al. (2009) | Multiple choice test in Administration | Formal rank ordering | 30 | 22.4 | -0.34 | 21.4 |
| Curcin et al. (2009) | Multiple choice test in Road Haulage and Passenger Transport | Formal rank ordering | 30 | 16.9 | -0.18 | 16.9 |

[5] Median value for correlation across 825 packs of seven items each of which were assessed by a single judge.

in only one of the studies (Lorge & Kruglov, 1952) was the standard deviation of item facilities reported in the original work. In the other cases, it was necessary to either calculate the standard deviation by reading the empirical data for individual items from plots, or to estimate it by careful reading of the information that was provided. For this reason, although Curcin et al. (2009) studied eight separate tests, only two indicative examples are included in Table 2. Similarly, although Ofqual (2017) analysed six separate Science assessments (two Biology, two Chemistry and two Physics tests), with correlations between PD and facility ranging from -0.50 to -0.36 (see Ofqual, 2017, Appendix C), it was only possible to include a single Biology test in the analysis here.

Studying the residual standard deviation of facilities given PD leads to a very different set of conclusions to looking at correlations alone. In particular, we can see that, despite the range of correlations in Table 2, there is far less difference in the actual accuracy with which PD could predict facility in each study. In particular, we can see that the high correlation reported by Attali et al. (2014) was associated with a residual standard deviation of 18.9, whereas for the study with the low correlation reported by Curcin et al. (2009) the residual standard deviation was 16.9. That is, contrary to what we might expect given the tone of the conclusions, the latter study was able to produce more accurate predictions of item facilities than the former.

The point here is not that PD is always equally predictive of actual difficulty. It is perfectly plausible that it is easier for judges to assess the relative difficulty of arithmetic questions aimed at pre-high school children (Lorge & Kruglov, 1952) than it is to perform the same task for Biology questions aimed at older teenagers (Ofqual, 2017). However, it is clear that previous attempts to use PD cannot be classified as successes or failures based on the correlation between PD and facility alone. More importantly, the actual accuracy with which PD can predict empirical difficulty is similar enough across both previous studies (Table 2) and the data sets we are analysing in this article (Table 1), that the following sections regarding the accuracy with which we can equate tests based on PD should generalise reasonably well across contexts beyond GCSE Mathematics.

## How accurately can we equate tests using PD?

We have seen above that PD is not a particularly good predictor of empirical difficulty and that the level of accuracy is fairly consistent across different studies. However, it might be hoped that, once PDs are aggregated across items to whole tests, the various errors will cancel out, leading to a good way of judging the relative difficulty of assessments. For example, Holmes et al. (2018) conclude that “providing there are no systematic biases in judging expected difficulty of items from different exam boards, the median and spread of predicted item difficulty for a paper will represent the actual difficulty of that paper reasonably well” (p.386).

This section considers this question in more detail. That is, given the accuracy with which we can predict item facilities, how well can we predict the difficulty of a whole test? For simplicity, we will imagine that the difficulty of a test is adequately represented by the mean score that would be achieved on it by a given population of students. For example, imagine we knew that the mean score on last year’s test was 50 out of 100. Then, using the PD of items and a predictive model devised using the previous year’s empirical data, we predicted that for the same set of students, the mean score of this year’s test would be 55 out of 100. We might reasonably conclude that this year’s test was five marks easier than last year’s. Thus, we might expect that grade boundaries should be about five marks higher this year than last year. Such reasoning is the basis for a type of statistical equating (i.e., calibrating tests against one another), known as mean equating. For grade boundaries that are reasonably close to the mean this approach is fairly easy to justify. If grade boundaries are a long way from the mean then this method is less justifiable. However, it may still provide a useful starting source of evidence. For example, it might be used as an input to a more sophisticated small-sample approach, such as circle-arc equating (Livingston & Kim, 2009). Either way, exploring the accuracy with which we expect to be able to predict the mean test scores for a fixed population provides a simple mechanism to help us explore the value of PD for equating tests.

On the basis of the above argument, we approximate how accurately we can equate tests using PD by the accuracy with which we can predict test means. Specifically, we want to calculate the standard error of a predicted test mean based upon the PDs of all the items in the test – that is, the expected root mean square error. To begin with, we note that the mean score on a whole test is simply the sum of the mean scores on the individual items. That is,

$$\textit{Predicted test mean} = \sum \textit{Predicted item mean} \qquad (2)$$

Next, if we assume a best case scenario that errors in predicted item means are independent (as opposed to consistently systematically biased), then the squared error of the predicted test mean (technically referred to as the “variance”) will be equal to the sum of the squared error in the predicted item means. Mathematically, we write these concepts as:

$$SE(\textit{predicted test mean}) = \sqrt{Variance(\textit{test mean} \mid PDs)} = \sqrt{\sum Variance(\textit{item mean} \mid PD)} \qquad (3)$$

Next, we note that the mean score on an item is just its facility (divided by 100) multiplied by the number of available marks. As such, the squared error in the item mean will be the square of the error in the facility multiplied by the maximum number of marks. From Tables 1 and 2 we can see that the residual standard deviation of item facilities given PDs tends to be about 20 (20% of the item maxima). Thus, the expected squared error of the mean score on an item given PD will be equal to $(0.2 * \textit{item max})^2$. Thus, continuing the mathematical formula above, the standard error with which we can predict the mean score on a test using item PDs is approximated by:

$$SE(\textit{predicted test mean}) \approx \sqrt{\sum (0.2 * \textit{item max})^2} = 0.2\sqrt{\sum \textit{item max}^2} = \frac{RSSIM}{5} \qquad (4)$$

The final term in the above equating (the RSSIM) was introduced in Benton (2019) and is just the square root of the sum of the squared item maxima. Its occurrence in the above formula relates to the fact that we would expect to be able to predict mean scores from item PDs more accurately for a test consisting of many low tariff items than that for one consisting of a few items worth many marks. This makes sense as if we have a greater number of items there is more chance for errors in predicted item means to cancel out.

The above formula suggests that the accuracy with which we can predict test means based on item PDs is approximated by the RSSIM divided by five. Table 3 illustrates the implications of this formula for the six GCSE Mathematics tests in analysis. For each of the tests we have calculated the RSSIM. For example, the Unit 1 Foundation tier test consisted of twelve 1-mark items, ten 2-mark items, four 3-mark items and four 4-mark items. Thus, the RSSIM was equal to:

$$\sqrt{(12 * 1^2 + 10 * 2^2 + 4 * 3^2 + 4 * 4^2} = \sqrt{152} = 12.3.$$

Using the above formula, we can estimate the expected standard error of a predicted test mean as a fifth of this value. For example, for the Foundation tier Unit 1, we estimate that the standard error of predicted test mean is 12.3/5=2.5 marks. For this same test, using common statistical practice, we can infer that a 95% confidence interval for the predicted mean would cover a range of plus or minus double this standard error, that is plus or minus roughly five marks. Thus, by assuming that the accuracy with which we can predict the mean gives a good approximation to how accurately we can position grade boundaries; we infer that on the basis of PD alone, the grade boundaries for this test might reasonably be set anywhere within a range of ten marks. Given that this whole test has a maximum of just 60 marks, this is only the vaguest sense of where grade boundaries should be placed.

**Table 3: Expected accuracies with PD**

| *Unit* | *Tier* | *Number of items* | *Number of marks* | *RSSIM* | *Standard errors on predicted boundaries* | *Width of 95% confidence interval for boundaries* |
|---|---|---|---|---|---|---|
| Unit 1 | Foundation | 30 | 60 | 12.3 | 2.5 | 9.7 |
| Unit 1 | Higher | 27 | 60 | 13.4 | 2.7 | 10.5 |
| Unit 2 | Foundation | 28 | 60 | 13.0 | 2.6 | 10.2 |
| Unit 2 | Higher | 22 | 58 | 13.9 | 2.8 | 10.9 |
| Unit 3 | Foundation | 39 | 100 | 18.2 | 3.6 | 14.2 |
| Unit 3 | Higher | 35 | 100 | 18.3 | 3.7 | 14.4 |

## Further illustration using split halves

This section provides an empirical illustration of the extent to which PDs of items might allow an accurate assessment of the relative difficulty of tests. To explore this, I used data from the Foundation tier Unit 3 paper. This paper was chosen as it included the largest number of items and also displayed the largest correlation between perceived and empirical item difficulties (see Table 1).

The items within this paper were randomly split into two half-papers, such that each half-paper consisted of 50 marks. The empirical mean score of each half-test was calculated to indicate the actual relative difficulty of the two tests. The weighted mean PD was also calculated for each test with more weight given to items worth a larger number of marks[6]. Giving more weight to PDs of items with more marks reflects the way we would use PD to predict mean test scores. The weighted mean PD provided an indication of the overall PD of each half-test. This process (split items into half-tests; calculate mean scores; calculate mean weighted PDs) was then repeated 100 times.

Across the 100 replications, Figure 3 compares the difference in PD of half-tests to the difference in means (i.e., actual difficulty). As can be seen, there is clearly some relationship between perceived and actual difficulty but it could hardly be described as providing an accurate basis for equating. For example, it is clear that where half-tests are of equal PD, there are instances of one half-test being around five marks

[6] Using the weighted median PD was also trialled but was found to make no substantial difference to the final results.

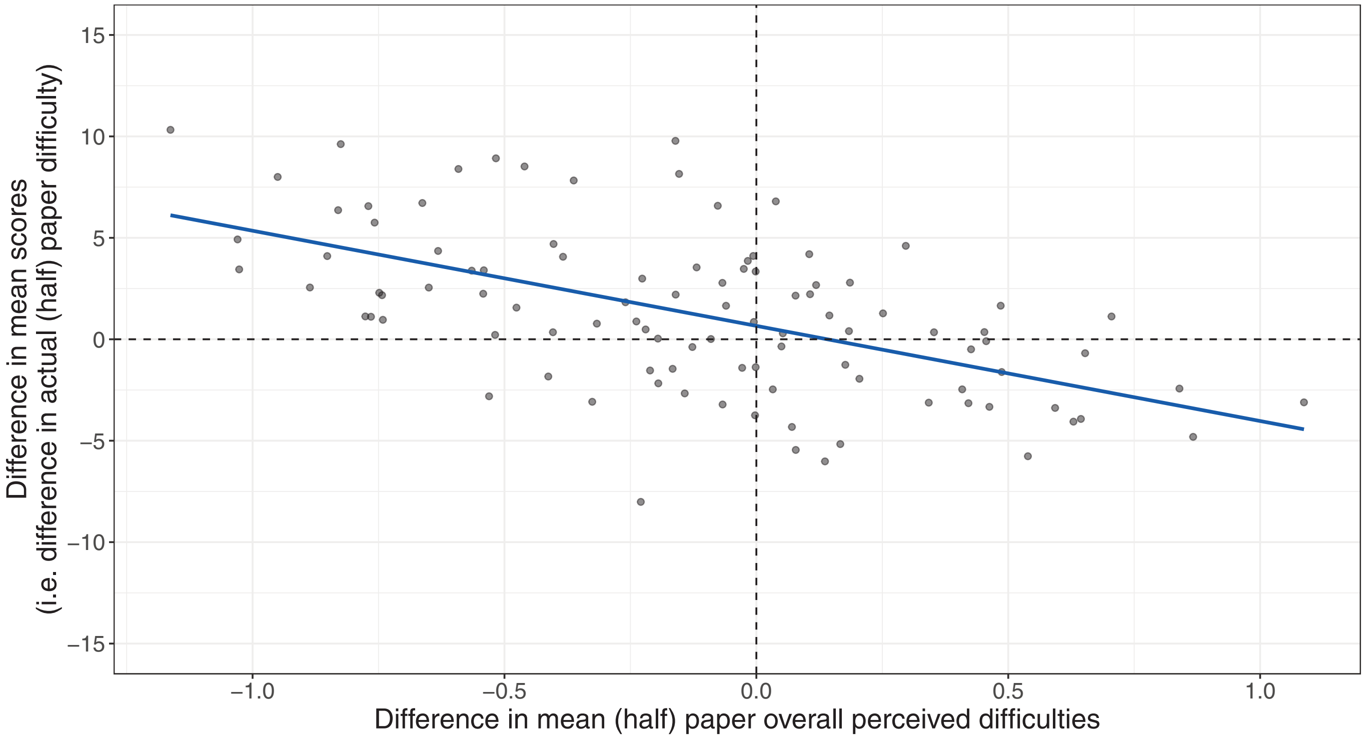

**Figure 3: The relationship between differences in mean PD of two half-papers and differences in empirical mean scores. A fitted regression line is shown in blue.**

(out of 50) easier than the other. Furthermore, PD does not even provide a reliable idea of the direction of difference in difficulty. Specifically, in 28 out of the 100 cases analysed, the direction of difference in PD was inconsistent with the direction of difference in means (i.e., a half-paper was perceived as more difficult when it was actually easier).

## Conclusion

This article has shown that PD is not a particularly good predictor of actual item difficulties. Specifically:

- On average, the predictive value of PD derived from a CJ exercise is roughly equivalent to the value of empirical data from just one student.
- For a fixed level of PD, item facilities have a standard deviation of about 20 percentage points. In other words, items that are perceived as equally difficult can have substantially different empirical difficulties.
- Broadly speaking, this level of predictive accuracy holds even for existing studies that have reported high correlations between perceived and empirical difficulty.

The analysis comparing the relative value of small samples of students and expert judgement suggests the intriguing possibility that we may generate better evidence if the subject experts involved in studies of item difficulty were replaced with the same number of students taking part in item trials. It may be argued that, because students would be far less well prepared and motivated in such an exercise than in a high-stakes exam, this would not provide an accurate idea of the actual relative difficulty of items. However, this view is not necessarily supported by the evidence. For example, in developing tests for use in primary school schools in England, the Standards and Testing Agency (STA) routinely trials all items with around one thousand pupils before they are used in live settings (STA, 2018). Estimates of item difficulty from these low-stakes trials are very highly correlated with difficulties estimated using live exam data. For example, for the Key Stage 2 Mathematics test taken in 2017 the correlation between IRT difficulty thresholds estimated from the two different sources of data was 0.976[7].

This article has also shown that it is not correct to assume that, when aggregated to the level of whole test papers, PD will provide an accurate means to judge the relative difficulty of two assessments. Equation 4 shows how to estimate the likely reliability of setting grade boundaries based on PD. For the GCSE Mathematics tests explored in this article the formula suggested that, given item PDs, a given grade boundary could reasonably be positioned at any score across a range covering roughly one sixth of the maximum available. This estimated level of accuracy is consistent with the simulations reported in Bramley (2020) for the situation where Angoff judgements of difficulty and actual item difficulties have a correlation of 0.6. This level of precision cannot be described as an accurate idea of where grade boundaries should be positioned. It should be noted that even achieving this level of accuracy requires an assumption that the items in the tests being compared come from the same "population" in some sense, and that there is no systematic reason for the items in one test being harder (relative to PD) than those in another. In practice, such systematic biases can occur. For example, Ofqual (2017, Appendix A) provides an example where, given equal PD, items produced by one assessment agency were systematically harder than those produced by another. As such, the formula provided in this article should be seen as giving a best-case view of the accuracy of the approach.

Finally, it is worth noting that, within each of the GCSE Mathematics tests that were analysed, variations in item facilities were barely any lower for fixed levels of PD than they were overall. That is, simply guessing that each item would display an average level of difficulty for the given test provides nearly as accurate a prediction of individual item difficulties as a full CJ exercise to elicit PDs. This, in turn, implies that knowing the PDs of new items, and the relationship between PD and empirical difficulty in the past, is hardly any more informative than just knowing the average difficulty of items within a particular paper historically.

Thinking about this from a practical perspective, the issue we are trying to solve is that we do not know how difficult the questions in current exam papers are. However, we do know how difficult they tended to be in the past. The results in this article indicate that PD does not add much new useful information to this. Therefore, we must conclude that the accuracy of using PDs to set grade boundaries is hardly any different from simply assuming that tests made to the same design specifications will always be equally difficult over time and, thus, that grade boundaries should remained fixed. The idea of using fixed grade boundaries over time has been suggested before (Bramley, 2012; Bramley, 2018). Whilst I am not necessarily recommending such an approach, it has to be conceded that it is hard to be in favour of the use of PDs for setting grade boundaries whilst objecting to the use of fixed grade boundaries.

In fact, for GCSEs and A Levels in England, PD already plays some role in the creation of test papers. Specifically, item writers are already required to identify the target level of each item. Usually this is expressed in terms of the grades available on the test and indicates the expected level of skill required to answer the question. As items are assembled into tests, a specification grid is used to ensure that the proportion of items at each level (as well as the balance of different topics) is kept consistent from year to year. Thus, a mechanism already exists to ensure that PD should remain reasonably constant over time, and, as such, we might expect the grade boundaries to remain constant. Given this, setting grade boundaries using CJ of the perceived difficulties of items could be seen as an expensive way of deciding that we ought to keep grade boundaries fixed over time. Whether this is a good idea is a wider question for further research.

## References

Attali, Y., Saldivia, L., Jackson, C., Schuppan, F., & Wanamaker, W. (2014). Estimating Item Difficulty With Comparative Judgments. *ETS Research Report No. RR-14-39*. Princeton, NJ: Educational Testing Service. Retrieved from: http://dx.doi.org/10.1002/ets2.12042.

Benton, T. (2016). Comparable Outcomes: Scourge or Scapegoat? Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Benton, T. (2019). Which is better: one experienced marker or many inexperienced markers? *Research Matters: A Cambridge Assessment publication, 28*, 2–10.

[7] Correlation provided by the Standards and Testing Agency under the Freedom of Information Act 2000.

Bramley, T. (2010). *Locating objects on a latent trait using Rasch analysis of experts' judgments*. Paper presented at the conference "Probabilistic models for measurement in education, psychology, social science and health", Copenhagen, Denmark, June 2010.

Bramley, T. (2012). *What if the grade boundaries on all A level examinations were set at a fixed proportion of the total mark?* Paper presented at the Maintaining Examination Standards seminar, London.

Bramley, T. (2018). When can a case be made for using fixed pass marks? *Research Matters: A Cambridge Assessment publication, 25*, 8–13.

Bramley, T. (2020). Comparing small-sample equating with Angoff judgment for linking cut-scores on two tests. *Research Matters: A Cambridge Assessment publication, 29*, 23–27.

Curcin, M., Black, B. & Bramley, T. (2009). *Standard maintaining by expert judgment on multiple-choice tests: a new use for the rank-ordering method*. Paper presented at the British Educational Research Association Annual Conference, University of Manchester, 2–5 September 2009.

Holmes, S. D., Meadows, M., Stockford, I., & He, Q. (2018). Investigating the Comparability of Examination Difficulty Using Comparative Judgement and Rasch Modelling. *International Journal of Testing, 18*(4), 366–391.

Humphry, S., Heldsinger, S., & Andrich, D. (2014). Requiring a Consistent Unit of Scale Between the Responses of Students and Judges in Standard Setting, *Applied Measurement in Education, 27*(1), 1–18.

Livingston, S. A., & Kim, S. (2009). The circle-arc method for equating in small samples. *Journal of Educational Measurement, 46*(3), 330–343.

Lorge, I., & Kruglov, L. (1952). A suggested technique for the improvement of difficulty prediction of test items. *Educational and Psychological Measurement, 12*(4), 554–561.

Mislevy, R.J., Sheehan, K.M., & Wingersky, M. (1993). How to equate tests with little or no data, *Journal of Educational Measurement, 30*(1), 55–78.

Ofqual (2015). *A comparison of expected difficulty, actual difficulty, and assessment of problem solving across GCSE Maths sample assessment materials*. Ofqual/15/5679. Coventry: Ofqual. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/429117/2015-05-21-gcse-maths-research-on-sample-assessment-materials.pdf.

Ofqual (2017). *GCSE science: An evaluation of the expected difficulty of items*. Ofqual/17/6163. Coventry: Ofqual. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/592709/gcse-science-an-evaluation-of-the-expected-difficulty-of-items.pdf.

Ofqual (2018). *A level and AS mathematics: An evaluation of the expected item difficulty*. Ofqual/18/6344. Coventry: Ofqual. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/676730/A_level_and_AS_mathematics_An_evaluation_of_the_expected_item_difficulty.pdf.

STA (2018). National curriculum test handbook: 2018 Key stages 1 and 2. STA/19/8312/e. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/765749/2018_NCT_Handbook_PDFA.pdf.

van Onna, M., Lampe, T., & Crompvoets, E. (2019). *Equating by pairwise comparison*. Presentation at the 20th annual AEA-Europe conference, Lisbon, Portugal.