
Article:
Homer, MS orcid.org/0000-0002-1161-5938 and Darling, JC orcid.org/0000-0001-7176-5957 (2016) Setting standards in knowledge assessments: comparing Ebel and Cohen via Rasch. Medical Teacher, 38 (12). pp. 1267-1277. ISSN 0142-159X

https://doi.org/10.1080/0142159X.2016.1230184

Setting standards in knowledge assessments: comparing Ebel and Cohen via Rasch

Short title

Setting standards in knowledge assessments

Authors

Matt Homer and Jonathan C Darling

Leeds Institute of Medical Education

School of Medicine

University of Leeds

LS2 9JT

UK

Corresponding author

Matt Homer, m.s.homer@leeds.ac.uk , +44(0) 113 343 4654

Abstract

Introduction

It is known that test-centred methods for setting standards in knowledge tests (e.g. Angoff or Ebel) are problematic, with expert judges not able to consistently predict the difficulty of individual items. A different approach is the Cohen method, which benchmarks the difficulty of the test based on the performance of the top candidates.

Methods

This paper investigates the extent to which Ebel (and also Cohen) produces a consistent standard in a knowledge test when comparing between adjacent cohorts. The two tests are linked using common anchor items and Rasch analysis to put all items and all candidates on the same scale.

Results

The two tests are of a similar standard but the two cohorts are different in their average abilities. The Ebel method is entirely consistent across the two years, but the Cohen method looks less so, whilst the Rasch equating itself has complications – for example, with evidence of overall misfit to the Rasch model and change in difficulty for some anchor items.

Conclusion

Based on our findings, we advocate a pluralistic and pragmatic approach to standard setting in such contexts, and recommend the use of multiple sources of information to inform the decision about the correct standard.

Practice points

- Standard setting in high-stakes knowledge tests is always a challenge.
- Equating tests through common anchors and Rasch analysis can give insight into the consistency of standards.
- Ebel standards appear consistent year-on-year but those set by the Cohen method appear less so.
- Rasch equating is itself challenging, with evidence of overall misfit and anchor shifts in difficulty.
- There is no ‘gold standard’ for standard setting in knowledge tests, and all sources of evidence are useful.

Notes on contributors

Matt Homer is an Associate Professor, working in both the Schools of Medicine and Education. His medical education research focuses on psychometrics, particularly related to OSCEs and knowledge tests.

Jonathan Darling is Senior Lecturer in Paediatrics and Child Health and Associate Director of Student Support for the MBChB programme. He has a particular interest in item quality and standard setting, and use of OSCEs for assessment of paediatric skills.

Glossary terms

Rasch analysis

Rasch analysis is a robust psychometric approach to assessing the quality of items and tests based on the assumption that the test is measuring a single underlying latent trait (e.g. clinical knowledge). The Rasch analysis produces an estimate for each person of their ability, and for each item its difficulty. These estimates are all on the same interval level scale.

Introduction

Setting the standard ‘correctly’ remains a fraught issue in many assessment contexts, and this is particularly true in high-stakes fields such as medicine (Cusimano, 1996; Cizek and Bunch, 2007; Downing et al., 2003; Norcini, 2003). In performance tests such as OSCEs, there are problems of assessor bias that might impact on decision making (Fuller et al., 2011; Homer et al., 2015; Pell et al., 2010; Pell et al., 2015). In automatically-marked written knowledge tests (e.g. single best answer test items) there are no such problems, but nevertheless the task of ensuring that pass/fail decisions are robust and consistent across cohorts remains a significant challenge. As a consequence of these difficulties, there is a very wide range of literature on different methods of standard setting, comparing their advantages and disadvantages across a range of assessment formats and settings (Jalili et al., 2011; Livingston and Zieky, 1982; Maccann, 2009; Schneid et al., 2014; Taube, 1997). In medical education, recent years have generally seen a move towards a more defensible, criterion-based, approach to standard setting, where judgements are made as to what is an acceptable level of performance for the

minimally competent candidate (Ricketts, 2009). This absolute approach contrasts with arguably less defensible, norm-referenced approaches where a pre-determined proportion of the cohort is set to pass/fail a priori.

In test-centred standard setting, items are scrutinised by experts and judgements are made as to their difficulty, often again in terms of the hypothetical 'borderline' or minimally competent candidate (Cizek and Bunch, 2007, page 48; Jaeger, 1991). One such method is the (modified) Angoff, but the main method investigated in this paper is that of Ebel – where experts judge test items not only on a single dimension of difficulty (e.g. Easy, Medium, Hard) but also on curricula/objective relevance (e.g. Essential, Important, Acceptable) (Ebel, 1972; Skakun and Kling, 1980). Judges also then determine the expected proportion of minimally competent candidates that would correctly answer each of the nine combinations of the difficulty and relevance dimensions. More details of the precise nature of the Ebel method under investigation in this paper will be given later.

There is compelling evidence that test-centred approaches to setting standards on medical education knowledge assessments are far from perfect (Clauser et al., 2008; Clauser et al., 2009; Clauser et al., 2014; Margolis and Clauser, 2014; Mee et al., 2013; Margolis et al., 2016; Shulruf et al., 2016). Experts are generally found to be poor at judging the difficulty of items, even with normative data. For example, in an experimental study (Clauser et al., 2009), judges were not able to make consistent predictions about item difficulty, and when given normative data (i.e. item performance statistics) revised their judgments to more closely match this data, even when this data was randomly generated. More recent work (Margolis et al., 2016)

shows that judges have higher expectations of performance on items that they can correctly answer, even having corrected for item difficulty and judge stringency. Even when Angoff judges agree, this does not automatically imply that the standard is actually appropriate .

Whilst there are a range of papers that provide evidence of the problematic nature of Angoff standard-setting, there is much less published research on the Ebel method (Downing et al., 2003; Homer et al., 2012), but since Ebel involves estimates of item difficulty for the minimally competent candidate very like Angoff, it is likely to have similar problems. Work by Homer et al (2012) demonstrates that this is indeed the case – examiners tend to rate easy questions harder and harder questions easier than the empirical data suggests they should.

More recently, the 'Cohen' method of standard setting has been advocated (Cohen-Schotanus and van der Vleuten, 2010; Taylor, 2011). This is seen as a pragmatic method to be employed 'in-house' (i.e. in individual medical schools) where the top performing candidates (e.g. the 90^{th} or 95^{th} percentile) in an examination are used to estimate or benchmark the difficulty of the test, and then the standard is set at a proportion of this benchmark, possibly with an additional correction for guessing depending on the local context (Taylor, 2011). Hence, Cohen is perhaps best described as a modified norm-reference approach with in-built adjustment for test difficulty. The key assumption of the Cohen approach is that the top-performers are consistent (i.e. of equal ability) over time – so that the standard set, based on the performance of this sub-group, is itself consistent. This assumption is thought to be more valid when the 90^{th} or 95^{th} percentile (and above) is used as the benchmark,

rather than just the single top-performing candidate, since use of group data is less vulnerable to outlier bias (Cohen-Schotanus and van der Vleuten, 2010). Taylor (2011) has confirmed this to be the case in a particular UK context.

The context and purpose of this paper

At our institution, we have been aware for some time of the many challenges of robust standard setting, and of the potential problems with the Ebel method that we have been employing as our sole standard setting method over many years. We have evidence that for many items, assessor judgements of their performance do not always match that seen in practice (Homer et al., 2012). As a consequence, in parallel with Ebel, we have been comparing standards year-on-year through linking tests with common anchor items using Rasch analysis (Bhakta et al., 2005; Bond and Fox, 2007; Downing, 2003; Tavakol, 2013). In essence, through the use of common items and the Rasch measurement model, all students and items in successive years are put on the same student ability/item difficulty scale and this allows the equivalence or otherwise of the Ebel standards in successive years to be compared.

Using the same examination data, we have recently also begun experimental work with the Cohen method as another approach to standard setting (Cohen-Schotanus and van der Vleuten, 2010; Taylor, 2011). Again, through the Rasch linking of tests, we are able to compare high-performing students year-on-year and to determine the extent to which this sub-group are of constant ability, and thereby gain insight into the validity of the Cohen method.

In this paper, we present analysis of two successive annual knowledge test assessments (i.e. student responses and standard setting metrics), in order to answer these research questions:

- How similar are the Ebel and Cohen set standards across the two years, and how well are these standards maintained year-on-year?
- How easy would it be to implement a Rasch test-equating methodology to maintain appropriate standards?

We also discuss a range of practical issues that this research has generated, and make some general suggestions as to the way forward with standard setting in such contexts.

Methods

Overview of examinations and linking

The tests we consider are multi-specialty summative knowledge assessments taken in the summers of 2014 and 2015 respectively by successive cohorts - fourth year students on a five-year undergraduate medical degree programme. We chose this particular year group to investigate because the multi-speciality nature of the summative exam is a particular challenge to standard setting using methods such as Ebel since experts have to judge items across a range of medical specialities (Homer et al., 2012). There were 293 candidates in 2014 and 274 in 2015.

Each test originally consisted of 200 single best answer items across two papers (Paper 1: 75 Extended matching questions and 75 multiple choice questions; Paper

2: 50 multiple choice questions each based on a different clinical image) (Case and Swanson, 2001). All items were scored 0 or 1. These papers form the first part of a sequential examination in two parts (Pell et al., 2013). If students perform sufficiently well on the first part they do not have to sit the second part of the examination. The work in this paper focuses entirely on the first part of the sequence that all students sit, and we refer to this part as the 'test' throughout.

Following a post-test item-screening analysis, two items were removed from the 2014 test because of poor psychometrics (e.g. negative item-corrected total correlation) (Case and Swanson, 2001). For similar reasons, nine poorly performing items were removed from the 2015 test (this includes one item that had a facility of 100% - such an 'extreme' item cannot be included in the Rasch analysis and so has been removed entirely from the rest of the analysis). Hence, the two tests as discussed in the remainder of this paper consist of 198 and 191 items in 2014 and 2015 respectively. Fifty items ('anchors') are common to both tests – these are spread across papers 1 and 2 and had been selected because they covered all five course modules being examined (Psychiatry; Paediatrics and Child Health; Gynaecology, Obstetrics and Sexual Health; Emergency and Critical Care; and Cancer and Continuing Care), and demonstrated good psychometric properties when used in earlier assessments.

Ebel standard setting

We use Ebel as our main standard setting method. Each item is tagged by 'difficulty' and 'relevance' by a group of judges – and for each combination of these two ratings an expected performance for the 'minimally competent' is set by these judges. This

performance grid and the item allocations for 2014 and 2015 are given in Tables A1-A3 in the Appendix. The Ebel-set pass mark for each test is the sum of the item allocations (Tables A2 and A3 respectively) weighted by the values in Table A1 (Cizek and Bunch, 2007, Chapter 5).

These values are adjusted upwards through the addition of two standard errors of measurement (SEM) (Hays et al., 2008; McManus, 2012; Pell et al., 2013) to avoid, or at least minimize, the number of false positives in the first part of the sequential test – that is, students 'passing' this part through measurement error working in their favour. Typically the SEM is of the order 5 marks (i.e. 2.5%; for the full sequence the adjustment is modified to the addition of only a single SEM (Pell et al., 2013)). In the remainder of this paper we shall refer to the Ebel + 2 SEM score as the 'Ebel passing score'. We have used an Ebel process since 2009, and over time have made adjustments, for example to the Ebel grid and the general process, informed by our work with the Rasch method (Homer et al., 2012). Consequently, for the actual standard setting process in 2014 we used a slightly different approach to that presented in this paper, but for consistency and ease of presentation we have applied an identical method across both years in the work in this paper.

Cohen standard setting

Very recently, we have begun exploring an alternative approach to Ebel – the Cohen method (Taylor, 2011; Cohen-Schotanus and van der Vleuten, 2010). Under Cohen, we calculate the 90^{th} (or 95^{th}) percentiles of the student total score in each of our two tests. This allows us to compare the difficulty of each test based on the performance

of the top candidates and the assumption that this performance is consistent year-on-year.

Rasch analysis – model fit and anchor consistency

All items and students from both years are combined in a single data set to carry out a ‘concurrent calibration’ in the terminology of Kolen and Brennan (Kolen and Brennan, 2014, Chapter 6). The Rasch analysis (Tennant and Conaghan, 2007) is carried out using RUMM2020 software (Andrich et al., 2002). The common anchor items are used to ‘equate’ the two assessments – i.e. to put all students and items on the same ability/difficulty scale. Note that Rasch item ‘difficulty’ estimates should be distinguished from Ebel item difficulty categorisations. The overall mean Rasch item difficulty across all items is set by default to zero in RUMM2020, and student ability is then calculated relative to this (arbitrary) benchmark.

In a Rasch analysis, the usual statistical methodology of adjusting the model to fit the data is turned on its head in order to produce proper ‘objective’ measurement (Panayides et al., 2009; Bond and Fox, 2007). In other words, the item response patterns have to fit the (Rasch) model. If not, then misfitting items might need to be removed and all estimates re-calculated in order to improve the quality of the measurement. We carry out the standard measures of Rasch model fit including tests of item fit, person (i.e. student) fit, overall model fit, and uni-dimensionality (Tennant and Conaghan, 2007).

Table 1 shows how many items were included our main analysis.

Item type	No. of unique items	Total no. of items
Common anchors	50	50
2014	148	198
2015	141	191
Total	339	339

Table 1: Number of items in equating analysis

The Rasch analysis to link the two tests is based on the underlying assumption that the anchor items perform in exactly the same way across the two assessments (i.e. have the same difficulty for students of the same ability). It is of course possible that this is not the case. Item performance could change over a year for a number of reasons – for example, the focus of teaching in certain topics might change, teachers themselves, or the pedagogy employed, could be different, medical guidance could change, and item position in the test could have an effect (Albano, 2013). The linking of the two tests is vulnerable to any such shifts in the anchor items' psychometric properties. Hence, an important aspect of Rasch test equating is a differential item analysis (DIF) for the anchors across tests (Bond and Fox, 2007). We therefore also consider in this paper whether there is evidence that at a given level of the latent trait (e.g. student ability in terms of clinical knowledge), there is evidence of differential performance on the anchor items when comparing the two year groups.

Comparisons of standards, tests and cohorts

Rasch student ability and item difficulty estimates were exported to a spreadsheet program and to SPSS (version 20) for additional analysis – to plot graphs, and to compare anchor and non-anchor item difficulty, test difficulty, and cohort ability. We use simple inferential statistics to do this - error bars, independent-sample t-tests,

ANOVA, and Pearson correlation coefficient r as the effect size measure (Cumming, 2011). Through the Rasch equating we are then able to judge the efficacy (or otherwise) of the Ebel and Cohen methods of setting and maintaining standards in these two knowledge tests.

Results

Ebel standard setting

The Ebel 'passing scores' for the first part of sequential test are given in Table 2:

Year	No. of items=Maximum available score	Ebel passing score		No. of students brought back for second part of sequence	
		Raw	Percentage	Raw	Percentage
2014	198	110	55.6	22	8.0
2015	191	105	55.0	25	8.5

Table 2: Ebel standards as set in 2014 and 2015

Hence, according to Ebel, the two tests are of very similar difficulty, with 2015 being slightly harder (i.e. having a slightly lower passing score – 55.0% compared to 55.6% in 2014). The proportion of the cohort being brought back for the full sequence is very similar (8.0% and 8.5% respectively).

Cohen standard setting

The 90th and 95th percentiles for the two cohorts based on the total scores in the respective tests are shown in Table 3.

Year	No. of items=Maximum available score	90th percentile of total score		95th percentile of total score	
		Raw	Percentage	Raw	Percentage
2014	198	147.0	74.2	152.3	76.9
2015	191	152.0	79.6	156.0	81.7
Difference in percentage scores (2015 – 2014)			5.3		4.8
Indicative difference in Cohen standard (2015 – 2014) (see text for explanation)			0.75×5.3 = 4.0		0.72×4.8 = 3.5

Table 3: 90th and 95th percentiles in 2014 and 2015

Note the actual passing score under Cohen is set as a fraction of the 90th or 95th percentile, possibly with a correction for guessing (Taylor, 2011). Using the 90th percentile to replicate our Ebel standard in 2014 (see Table 2), this fraction is 0.75 (=55.6/74.2). The equivalent for the 95th percentile is 0.72 (=55.0/76.9). Hence, the final row in Table 3 presents the relevant proportion of the difference in the total score percentile as a percentage (2015 – 2014). This gives an indication of the difference in passing standards Cohen would produce in the two tests – 4.0% using the 90th percentile, and 3.5% if using the 95th percentile.

If we assume, as Cohen requires (Cohen-Schotanus and van der Vleuten, 2010; Taylor, 2011), that the high-performing group is stable year-on-year, the data in Table 3 shows that the two tests vary in difficulty, with 2014 being harder (as the 90th and 95th percentile are scoring approximately 5% lower in 2014, which translates to

Cohen standards differing by the order of 4%). This clearly contradicts the evidence from the Ebel standard setting presented above which indicates that the tests had very similar levels of difficulty.

We move on now to the Rasch analysis which will enable us to disentangle test and cohort effects with the aim of showing of these standard setting approaches is 'working' best.

Rasch model fit

There are problems with overall Rasch model fit – the item-trait interaction (i.e. overall misfit to the model) is significant (Chi-square=3,446, df=2,712, p<0.001). Formally, this indicates that the ordering of the difficulty of the items is different for students of different ability, whereas one of the assumptions of the Rasch model is that this ordering should be consistent across the ability range (Tennant and Conaghan, 2007). However, at the individual item level only two items out of the 339 (0.6%) show misfit to the Rasch model based on Bonferroni corrected p-values.

The overall internal consistency reliability - the Person separation index (Tennant and Conaghan, 2007) - is good at 0.89, and there is little evidence of violation of local item independence: 0.2% of item-pairs have a residual correlation above r=0.2. This is a measure of how strongly item responses correlate, having controlled for the main trait being measured, and one of the assumptions of the Rasch model is that items are locally independent – i.e. do not correlate once the main dimension has been accounted for (Tennant and Conaghan, 2007). In simple terms, this is good evidence that the items are measuring a single trait.

To keep the presentation of the results relatively concise, all 339 items (as per Table 1) have been kept in the analysis, based on the view that a few problematic items/items pairs will make little difference to the overall quality of the measurement given the large number of items involved in total.

Only the anchors are taken by all candidates, and so the blocked nature of the equating design means there is 'missing' data for the non-anchors. This design is called 'common items, non-equivalent groups' in the terminology of Kolen and Brennan (2014, Chapter 4). In this design, it is not possible to assess the uni-dimensionality of the construct during the joint Rasch analysis. However, separate exploratory factor analyses indicate that the individual tests themselves were uni-dimensional in the sense that all items loaded on to a single dominant factor in each case.

Finally, the person-item location distribution (i.e. the joint graph of item difficulty and student ability) is shown in Figure 1, and indicates that both tests have good 'targeting' – the range of item difficulties is wider than the range of the student ability estimates (Tennant and Conaghan, 2007).

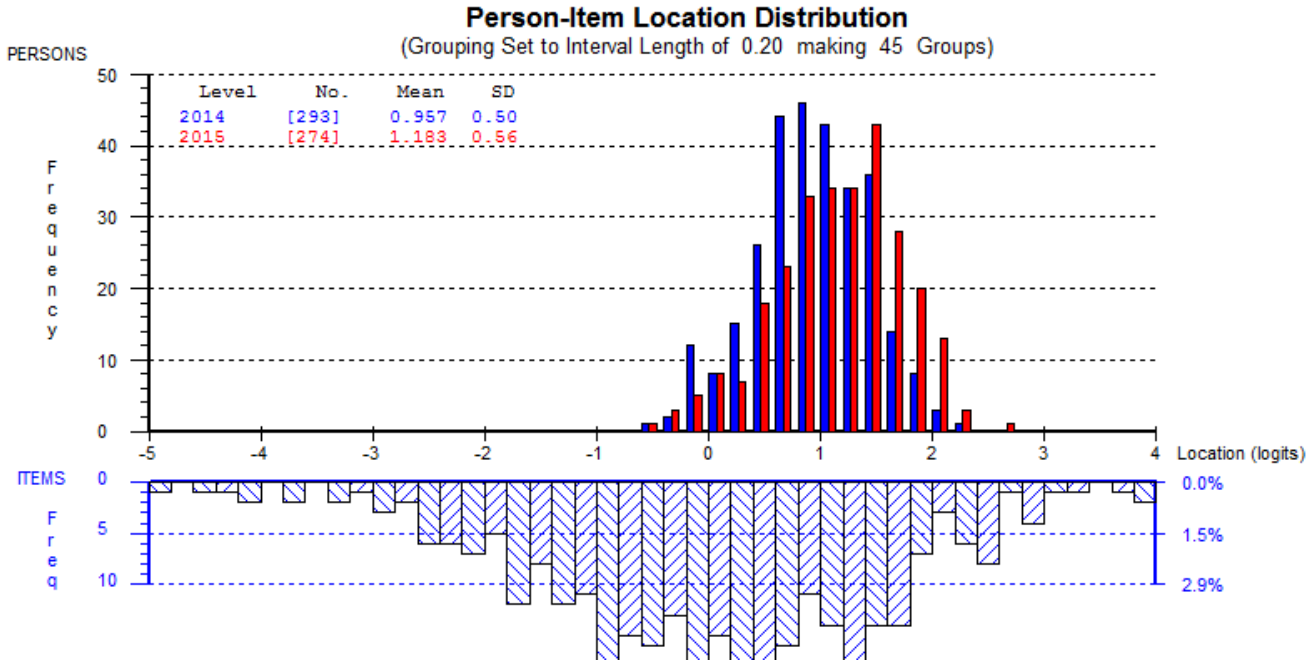

Figure 1: Student-item location distribution by year

This is important since good targeting is a pre-condition for successful Rasch measurement that produces robust item and student estimates.

We comment in detail later on the apparent differences in ability across the two year groups shown in Figure 1.

In summary, the overall fit to the Rasch model is satisfactory in most aspects, bar the overall test of model fit. This could be seen as a problem in terms of confidence in the 'specific objectivity' of the measurement the tests provide (i.e. that the measurement estimates are independent of the sample items and persons (Tennant and Conaghan, 2007). However, in terms our main concern, robust test equating, we argue this is not a substantial problem, providing the anchors are functioning well. We shall return to the overall misfit issue in the Discussion but first investigate the performance of the anchors in more detail.

The consistency of the anchors across the two tests

The key issue with regard to the anchors and the robustness of the equating is the extent to which these 50 items behave the same for students of the same ability across both groups of students. Figure 2 includes all 50 anchor items and compares their facilities in 2014 (x) and 2015 (y). With perfect consistency in performance and assuming equal ability across the two cohorts, we would see all items lying on the line y=x shown in Figure 2. Later we shall see that this latter assumption is not quite correct, and that the cohorts differ in average ability. Hence the comparison of facilities shown in Figure 2 should be regarded as illustrative of the differential item functioning (DIF) rather than completely precise. The actual determination as to whether or not an item suffers from DIF is based on significance tests as part of the Rasch analysis in the software, and does take account of any differences in ability between the cohorts (Tennant and Conaghan, 2007).

We find evidence of DIF in 11 anchor items out of the 50 (22%) when comparing across the two years. This is a change in difficulty of these items for students of the same ability across the two years – in essence, those far from the line y=x in Figure 2 are those with the strongest DIF.

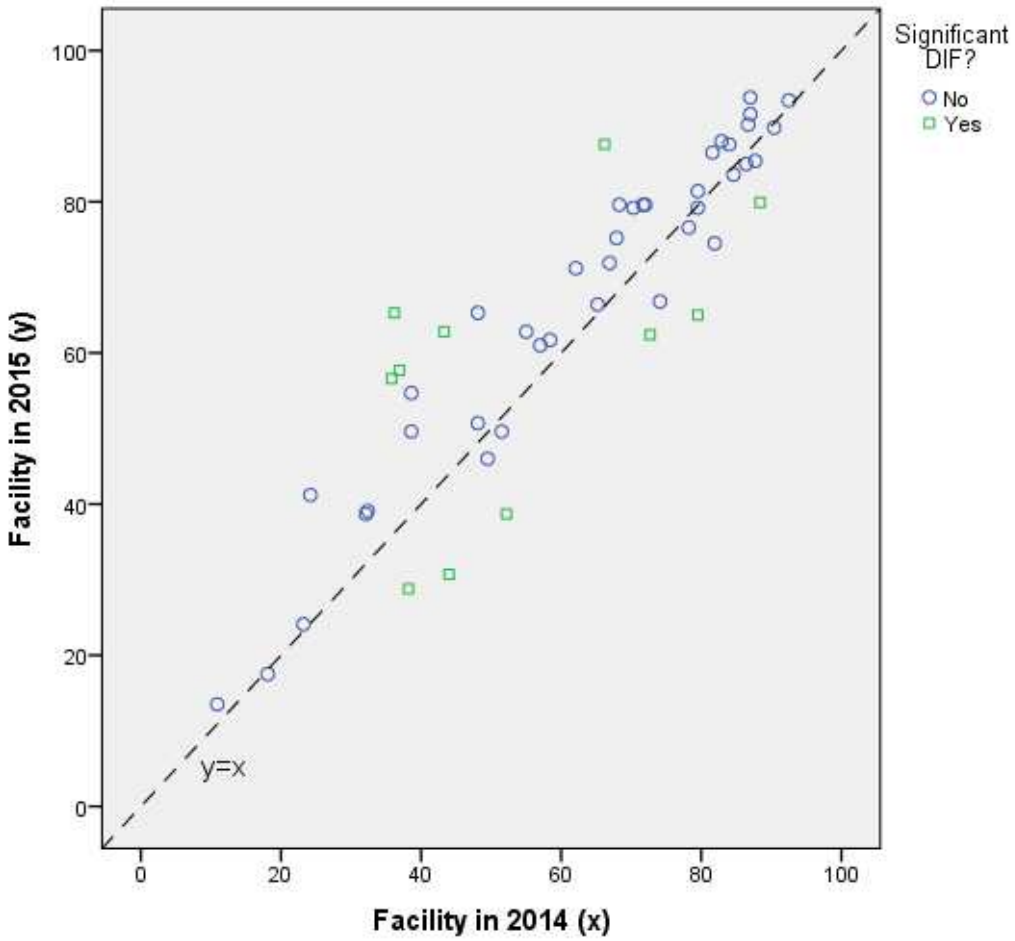

Figure 2: Anchor facilities, with those with DIF by year identified

We see in Figure 2 that these differences are in both directions and the overall effect is not particularly strong – the mean change in facility across the group of 11 anchors with significant DIF is 3.8% (higher in 2015). We have checked the impact of removing these anchors and re-running all subsequent equating analyses. All substantive findings that follow are unaffected by this, and so we have made the parsimonious decision to keep these anchors in the analysis, in part to keep the presentation as simple as possible. We are also of the view that removal of any items from the analysis degrades the sampling of the domain being tested and takes the analysis further away from that based on the actual tests sat (Case and Swanson, 2001, page 8).

Comparison of test difficulty

Having checked the Rasch model fit and anchor consistency, we now in a position to compare the tests using the 50 common anchor items to estimate all item difficulties/student abilities on a common scale through Rasch analysis of all students and items.

Figure 3 shows an error bar (i.e. mean and 95% confidence intervals for this mean) comparing average Rasch difficulty estimates of all items in the two tests – 2014 non-anchors, 2015 non-anchors and common anchors:

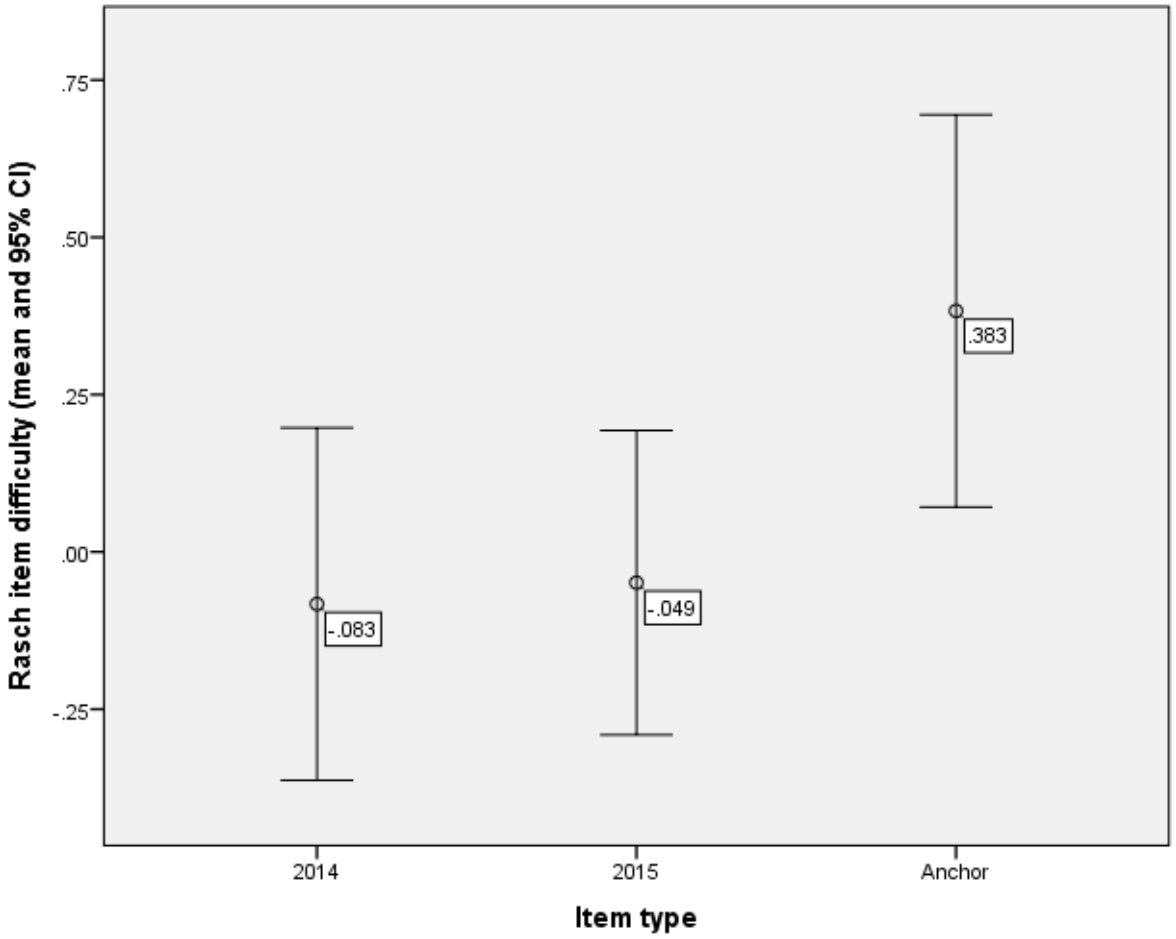

Figure 3: Mean Rasch item estimates – 2014 non-anchors, 2015 non-anchors and anchors

The vertical scale is the usual Rasch logit scale (Tennant and Conaghan, 2007) where higher values correspond to more difficult items. Figure 3 indicates that the 2015 test is very slightly more difficult on average than that in 2014, and that the anchors are more difficult than the non-anchor items in both tests – in terms of facilities, this latter difference is of the order of 5%. An overall ANOVA test indicates that the differences between these three mean logit scores are not statistically significant ($F(2,366)=1.85$, $p=0.16$, r-squared=0.011).

That the two tests are very close in difficulty is confirmed in Figure 4 which shows the expected percentage score on the two tests (y) against (Rasch) candidate ability (x).

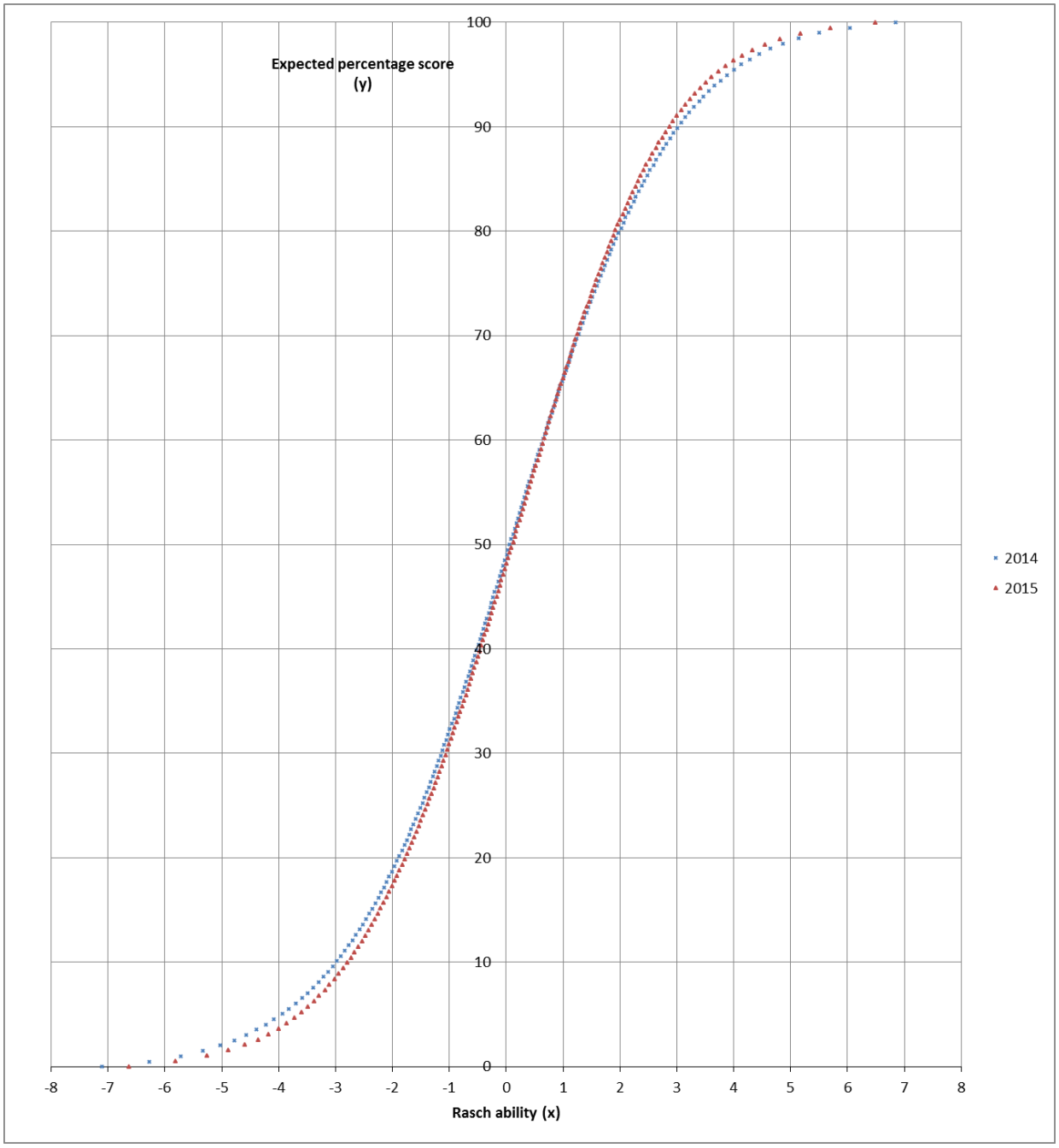

Figure 4: Comparison of test difficulty in 2014 and 2015

From Figure 1 earlier we see that in terms of ability both cohorts are centred at approximately 1 on the x-axis, and we note that the graphs in Figure 4 are very close at this value. Typically, a student of this ability in each cohort would score approximately 66% on either test.

We have also confirmed that the tests are of a similar standard of difficulty using an alternative method, Tucker equating (Kolen and Brennan, 2014, Chapter 4), a classical test theory approach to test equating through common items. We find, for example, that a student scoring 55.6% in 2014 (i.e. the Ebel passing score in Table 2) would score 54.6% in 2015 according to the Tucker equating (the Rasch equating gives 55.0% for this). This alternative approach to equating therefore confirms that the test in 2015 is slightly more difficult than that in 2014 and gives very similar results to the Rasch equating.

Comparison of Ebel pass marks

Figure 5 shows the test characteristics curve (Tennant and Conaghan, 2007) for the 2014 test. This shows how students with higher ability (i.e. to the left on the x-axis) score more highly on the test (as one would expect). The Ebel passing score in 2014 (110 marks, 55.6% - see Table 2) has been added so that the equivalent logit location (ability score) can be read off on the horizontal Rasch scale (0.381).

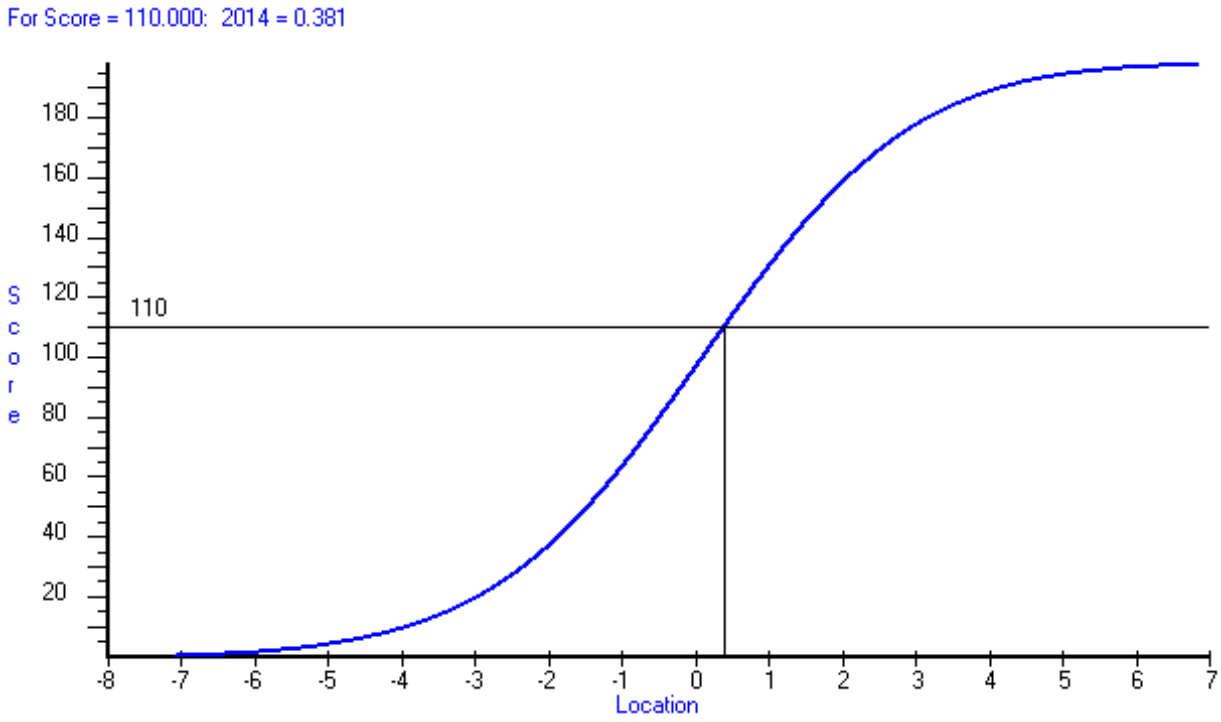

Figure 5: Test characteristic curve and Ebel standard 2014

In other words, the Ebel pass mark in 2014 is equivalent to 0.381 logits on the common 'Rasch' scale. This is the expected Rasch ability of someone at this total raw score, and can be thought of as the Ebel standard for 2014 in Rasch logits.

The equivalent analysis for 2015 indicates that the passing score (105 marks, 55.0% - Table 2) is equivalent to 0.375 logits in 2015 – this is the corresponding Ebel standard in 2015 on the same scale as that for 2014.

We now see clearly that the difference between the Ebel standards across the two years is essentially zero on the common Rasch scale (0.381-0.375=0.006 logits, equivalent to less than 2% of a single mark on either test). Hence, this analysis provides good evidence that the Ebel standard setting has maintained an equivalent standard year-on-year.

Comparison of student ability

Returning to the student ability distribution shown in the upper half of Figure 1, we can compare the two cohorts in this regard. We find that the mean Rasch estimates of ability in the two groups are different:

Mean logit ability score in 2014 = 0.957

Mean logit ability score in 2015 = 1.183

The mean difference between the two cohorts based on all items is therefore 0.226 (=1.183-0.957), and this is statistically significant ($t=5.04$, $df=565$, $p<0.001$, $r=0.21$). Converting this difference from logits to raw scores on the 2014 test shows that the 2015 cohort would score approximately 7 marks (3.5%) more highly on average in this test compared to the 2014 cohort.

Despite the clear evidence of differences in the mean ability, the proportion being brought back for the second sequential test based on the Ebel standard is similar in the two groups – see the two columns to the right of Table 1. Given that these standards have been shown to be equivalent, this tells us that whilst the ability distributions in 2014 and 2015 have different average locations, the proportions in the key pass/fail region are in fact similar.

With relevance to the Cohen standard setting, we also find that the Rasch estimates of the 90^{th} and 95^{th} percentiles of student performance compare almost exactly with the figures presented above in Table 3 based on total score (i.e. a classical)

analysis. This is additional confirmation that the ability level of the highest performing students (e.g. the 90th or 95th percentiles) is **not** stable year-on-year.

Summary of results

For clarity, we summarise our key findings as follows:

- The tests in 2014 and 2015 are of a very similar level of difficulty.
- The Ebel set standards year-on-year are very similar.
- The two cohorts are of a different average ability, with 2015 being a more able group.
- The 90th and 95th percentiles of the two cohorts are also quite different across years, again with the 2015 group of higher ability.

Methodologically, we have seen generally good fit to the Rasch model at the item-level, but there are problems with the overall Rasch model fit, and differential item functioning for a proportion of anchors. However, based on the additional Tucker equating, we have good evidence these problems do not impact on our substantive findings when equating the two tests.

Discussion

This study compares standard setting approaches across two successive years of tests, linked by common anchors, and our main findings are clear. We have found good evidence that the Ebel standards are equivalent year-on-year, but that the underlying assumption of the Cohen approach is undermined. The highest

performing students are **not** of similar abilities in the two year groups meaning that such students cannot automatically be used to benchmark the difficulty of a test as Cohen suggests (Cohen-Schotanus and van der Vleuten, 2010; Taylor, 2011). Returning to the Ebel method, it seems that at the item-level there have been problems in getting 'it right' (Homer et al., 2012; Clauser et al., 2008; Clauser et al., 2009; Clauser et al., 2014; Mee et al., 2013; Margolis et al., 2016), but through our work using Rasch alongside Ebel, we are now more confident that examiner judgements are appropriate, certainly across the tests as a whole.

The Rasch approach to measurement and to test equating is undoubtedly attractive and, hypothetically, very robust (Tavakol, 2013; Panayides et al., 2009; Bond and Fox, 2007). However, in practice the exercise is to an extent problematic. There are some problems with overall fit to the Rasch model, although at item-level the fit is good. There is also evidence of shift in difficulty for a number of common anchors although the overall impact on the equating is minimal. However, other classical equating approaches, for example Tucker (Kolen and Brennan, 2014, Chapter 4), do not provide any opportunity to look at shifts in anchor items, so Rasch is certainly advantageous in this respect.

Ideally, future research should investigate potential reasons for the differential item functioning of the anchors, for example, considering if there have been obvious changes in teaching, or important changes in medical practice or some other effect. This would give additional insight into the robustness or otherwise of the Rasch approach, but we are confident that despite these problems the key findings across the research are secure, certainly in terms of the main equating results.

Technically speaking, Rasch analysis requires statistical or psychometric (and software) expertise that is not always available in individual medical schools (Andrich et al., 2002; Tennant and Conaghan, 2007). The analysis, necessarily post-hoc, is also time-consuming and obviously has to take place before any exam decisions can be made. We emphasise that the overall misfit to the model has not been resolved in the work we have presented, and that with the number of items in these types of assessments, serious attempts to deal comprehensively with this misfit would prove a very drawn out process. This in itself is a useful 'null' finding worthy of publishing; modern measurement approaches such as Rasch analysis are not a complete panacea in these contexts, and our work suggests that overall misfit might well have to be tolerated to a degree. The alternative might be to spend a lot of time, for example, throwing away 'bad' items, which then has its own impact on domain sampling of the assessment, and hence its validity (Cook, 2014; Kane, 2013; Kane, 2001).

By way of contrast, most of the work related to 'Ebeling' can be done well before the examinations take place – even though the process of producing the item-level judgements themselves is time-consuming for the team of academic staff involved. A major benefit of the Cohen approach is that it requires only simple analysis of total scores on tests and can be done quickly once test scores become available (Cohen-Schotanus and van der Vleuten, 2010; Taylor, 2011).

In terms of standard setting policies, our evidence suggests that multiple approaches to this process might be sensible. The pragmatic use of all available evidence, and

comparing the information from multiple methods, gives deeper insights and avoids falling into the perhaps simplistic trap of attempting to discover the single ‘best’ standard setting method and sticking with it. In other words, we believe that all practical sources of evidence as to the validity of any standard should be employed, as is the case in the wider justification of the validity of an assessment and/or its outcomes (Kane, 2001; Kane, 2013; Cook, 2014).

Although standard-setting methods such as Angoff and Ebel can produce apparently very precise pass marks, this apparent precision is both misleading and misplaced. We propose a more nuanced ‘composite’ approach, bringing together the strengths of several methods to construct a pass mark ‘window’ within which the examiners as a team make a final judgment, based on the strength of the different sources of information available. This approach is in tune both with Livingstone and Zieky’s original call for ‘reality checks’ when standard setting through assessor judgements (Livingston and Zieky, 1982), and also with the move towards accepting a place for subjectivity and judgement in assessment (Hodges, 2013; Schuwirth and Vleuten, 2006).

Finally, we comment on one finding that we didn’t really expect to see - the large difference between the two student groups in terms of their average ability. One might have expected that year-on-year a student body of size at least 250 wouldn’t collectively vary that much in ability. We find that this is not the case, and again this underlines the importance of not relying heavily on normative methods for setting standards. Our curriculum for years 1-3 changed in 2010, and a proportion of

students in 2014 would have been the first year of that curriculum. It is possible that the differences we see between cohorts are to an extent an artefact of this change (e.g. in part due to a bedding down process in the new curriculum). However, this is completely speculative as we do not have data available to confirm or deny these potential effects. Both groups did the same Year 4 curriculum, which is that being being tested in the assessments we are investigating. It seems, however, very likely that the Cohen method is picking up the difference in abilities of the two groups rather than the difficulty of the test (as it is intended to). We will take care to monitor the Cohen method and the relative ability of student cohorts over the next few administrations of these assessments to gain a better understanding of the degree of variation between cohorts and stability or otherwise of the standard setting approaches we employ.

Conclusion

Our key overall message from this work is that there is no single 'gold-standard' method for setting and/or maintaining standards in these contexts. We have shown that both of the two methods investigated in this paper (Ebel and Cohen) have advantages, but also that they bring problems too. Our work indicates that Ebel is 'better' than Cohen in producing a more consistent standard, in part due to the fact that the assumption of Cohen that the highest performing 5^{th} or 10^{th} percentile of students is of the same ability year-on-year in not always true. However, we would need more data over several cohorts to fully confirm that, in our context at least, Ebel is indeed more consistent.

We also find that relatively 'heavy duty' equating approaches (Rasch) are practically challenging – both from a technical point of view, and when producing standards in time-pressured contexts. These findings are important in the UK context, where a national licensing examination is being developed (General Medical Council, 2015) partly informed by studies comparing current standards across medical schools. Even with a single medical school, our work indicates how difficult it is to compare cohorts and standards with confidence and rigour.

In light of the work presented in this paper, we therefore advocate a pluralistic and pragmatic approach to standard setting in single medical schools, and perhaps more widely. We recommend the use of multiple sources of information to inform the decision about the correct standard. However, the precise system for making these decisions needs further consideration, and we hope in the near future to develop a practical 'how to guide' for setting standards using diverse ranges of evidence.

We conclude by encapsulating our main message in this (mis-) quote from the English statistician George Box[1]:

> Essentially, all standard setting approaches are wrong, but some are useful.

References

Albano, A.D. 2013. Multilevel Modeling of Item Position Effects. Journal of Educational Measurement. **50**(4),pp.408–426.

[1] https://en.wikiquote.org/wiki/George_E._P._Box

Andrich, D., Sheridan, B. and Luo, G. 2002. RUMM2020: Rasch unidimensional models for measurement. Perth, Western Australia: RUMM Laboratory.

Bhakta, B., Tennant, A., Horton, M., Lawton, G. and Andrich, D. 2005. Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. BMC Medical Education. **5**(1),p.9.

Bond, T.G. and Fox, C.M. 2007. Applying the Rasch Model: Fundamental Measurement in the Human Sciences 2nd ed. Psychology Press.

Case, S. and Swanson, D. 2001. Constructing Written Test Questions for the Basic and Clinical Sciences [Online]. Philadelphia: NBME. [Accessed 3 July 2013]. Available from: http://www.heacademy.ac.uk/resources/detail/resource_database/SNAS/Constructing_Written_Test_Questions_for_the_Basic_and_Clinical_Sciences.

Cizek, G.J. and Bunch, M.B. 2007. Standard setting a guide to establishing and evaluating performance standards on tests [Online]. Thousand Oaks, Calif.: Sage Publications. [Accessed 21 August 2013]. Available from: http://SRMO.sagepub.com/view/standard-setting/SAGE.xml.

Clauser, B.E., Harik, P., Margolis, M.J., McManus, I.C., Mollon, J., Chis, L. and Williams, S. 2008. An Empirical Examination of the Impact of Group Discussion and Examinee Performance Information on Judgments Made in the Angoff Standard-Setting Procedure. Applied Measurement in Education. **22**(1),pp.1–21.

Clauser, B.E., Mee, J., Baldwin, S.G., Margolis, M.J. and Dillon, G.F. 2009. Judges' Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. Journal of Educational Measurement. **46**(4),pp.390–407.

Clauser, J.C., Margolis, M.J. and Clauser, B.E. 2014. An Examination of the Replicability of Angoff Standard Setting Results Within a Generalizability Theory Framework. Journal of Educational Measurement. **51**(2),pp.127–140.

Cohen-Schotanus, J. and van der Vleuten, C.P.M. 2010. A standard setting method with the best performing students as point of reference: practical and affordable. Medical Teacher. **32**(2),pp.154–160.

Cook, D.A. 2014. When I say… validity. Medical Education. **48**(10),pp.948–949.

Cumming, G. 2011. Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. New York: Routledge.

Cusimano, M.D. 1996. Standard setting in medical education. Academic Medicine: Journal of the Association of American Medical Colleges. **71**(10 Suppl),pp.S112–120.

Downing, S.M. 2003. Item response theory: applications of modern test theory in medical education. Medical education. **37**(8),pp.739–745.

Downing, S.M., Lieska, N.G. and Raible, M.D. 2003. Establishing passing standards for classroom achievement tests in medical education: a comparative study of four methods. Academic Medicine: Journal of the Association of American Medical Colleges. **78**(10 Suppl),pp.S85–87.

Ebel, R.L. 1972. Essentials of educational measurement [Online] xii, 370 p. ; Englewood Cliff, N.J., US: Prentice-Hall, Inc. [Accessed 15 January 2010]. Available from: http://openlibrary.org/b/OL17728612M/Essentials_of_educational_measurement.

Fuller, R., Homer, M. and Pell, G. 2011. What a difference an examiner makes! __Detection, impact and resolution of 'rogue' examiner behaviour in high stakes OSCE assessments.

General Medical Council 2015. GMC Council approves development of UK medical licensing assessment. [Accessed 18 July 2016]. Available from: http://www.gmc-uk.org/news/26549.asp.

Hays, R., Gupta, T.S. and Veitch, J. 2008. The practical value of the standard error of measurement in borderline pass/fail decisions. Medical Education. **42**(8),pp.810–815.

Hodges, B. 2013. Assessment in the post-psychometric era: learning to love the subjective and collective. Medical teacher. **35**(7),pp.564–568.

Homer, M., Darling, J. and Pell, G. 2012. Psychometric characteristics of integrated multi-specialty examinations: Ebel ratings and unidimensionality. Assessment & Evaluation in Higher Education. **37**(7),pp.787–804.

Homer, M., Pell, G., Fuller, R. and Patterson, J. 2015. Quantifying error in OSCE standard setting for varying cohort sizes: A resampling approach to measuring assessment quality. Medical Teacher.,pp.1–8.

Jaeger, R.M. 1991. Selection of Judges for Standard-Setting. Educational Measurement: Issues and Practice. **10**(2),pp.3–14.

Jalili, M., Hejri, S.M. and Norcini, J.J. 2011. Comparison of two methods of standard setting: the performance of the three-level Angoff method. Medical Education. **45**(12),pp.1199–1208.

Kane, M. 2001. So Much Remains the Same: Conception and Status of Validation in Setting Standards In: G. J. Cizek, ed. Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 53–58.

Kane, M.T. 2013. Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement. **50**(1),pp.1–73.

Kolen, M.J. and Brennan, R.L. 2014. Test Equating, Scaling, and Linking: Methods and Practices. Springer Science & Business Media.

Livingston, S.A. and Zieky, M.J. 1982. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. [Accessed 19 March 2015]. Available from: http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED227113.

Maccann, R.G. 2009. Standard setting with dichotomous and constructed response items: some rasch model approaches. Journal of Applied Measurement. **10**(4),pp.438–454.

Margolis, M.J. and Clauser, B.E. 2014. The Impact of Examinee Performance Information on Judges' Cut Scores in Modified Angoff Standard-Setting Exercises. Educational Measurement: Issues and Practice. **33**(1),pp.15–22.

Margolis, M.J., Mee, J., Clauser, B.E., Winward, M. and Clauser, J.C. 2016. Effect of Content Knowledge on Angoff-Style Standard Setting Judgments. Educational Measurement: Issues and Practice. [Online]. [Accessed 15 February 2016]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/emip.12104/abstract.

McManus, I.C. 2012. The misinterpretation of the standard error of measurement in medical education: a primer on the problems, pitfalls and peculiarities of the three different standard errors of measurement. Medical teacher. **34**(7),pp.569–576.

Mee, J., Clauser, B.E. and Margolis, M.J. 2013. The Impact of Process Instructions on Judges' Use of Examinee Performance Data in Angoff Standard Setting Exercises. Educational Measurement: Issues and Practice. **32**(3),pp.27–35.

Norcini, J.J. 2003. Setting standards on educational tests. Medical Education. **37**(5),pp.464–469.

Panayides, P., Robinson, C. and Tymms, P. 2009. The assessment revolution that has passed England by: Rasch measurement. British Educational Research Journal. [Online]. [Accessed 18 November 2009]. Available from: http://www.informaworld.com/10.1080/01411920903018182.

Pell, G., Fuller, R., Homer, M. and Roberts, T. 2013. Advancing the objective structured clinical examination: sequential testing in theory and practice. Medical Education. **47**(6),pp.569–577.

Pell, G., Fuller, R., Homer, M. and Roberts, T. 2010. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. Medical Teacher. **32**(10),pp.802–811.

Pell, G., Homer, M. and Fuller, R. 2015. Investigating disparity between global grades and checklist scores in OSCEs. Medical Teacher.,pp.1–8.

Ricketts, C. 2009. A plea for the proper use of criterion-referenced tests in medical assessment. Medical Education. **43**(12),pp.1141–1146.

Schneid, S.D., Armour, C., Park, Y.S., Yudkowsky, R. and Bordage, G. 2014. Reducing the number of options on multiple-choice questions: response time,

psychometrics and standard setting. Medical Education. **48**(10),pp.1020–1027.

Schuwirth, L.W.T. and Vleuten, C.P.M. 2006. A plea for new psychometric models in educational assessment. Medical Education. **40**(4),pp.296–300.

Shulruf, B., Wilkinson, T., Weller, J., Jones, P. and Poole, P. 2016. Insights into the Angoff method: results from a simulation study. BMC medical education. **16**(1),p.134.

Skakun, E.N. and Kling, S. 1980. Comparability of Methods for Setting Standards. Journal of Educational Measurement. **17**(3),pp.229–235.

Taube, K.T. 1997. The incorporation of empirical item difficulty data into the Angoff standard-setting procedure. Evaluation & the Health Professions. **20**(4),pp.479–498.

Tavakol 2013. Psychometric evaluation of a knowledge based examination using Rasch analysis : An illustrative guide: AMEE Guide No. 72. MEDICAL TEACHER. **35**(1),pp.74–84.

Taylor, C.A. 2011. Development of a modified Cohen method of standard setting. Medical Teacher. **33**(12),pp.e678–e682.

Tennant, A. and Conaghan, P.G. 2007. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis and Rheumatism. **57**(8),pp.1358–1362.

Keywords

Psychometrics

Standard setting

Written assessment

Funding

This work is not tied to any external funding

Ethics

The University of Leeds gave permission for this anonymised data to be used for research. The co-chairs of the University of Leeds School of Medicine ethics committee confirmed to the authors that formal ethics approval for this study was not required as it involved the use of routinely collected student assessment data which were fully anonymised prior to analysis.

Acknowledgments

We thank Godfrey Pell and Jason Ward for their helpful comments and conversations relating to the early development of this work.

Declaration of interests

The authors have no declarations on interest to report.

Appendix 1 – Ebel judgements

2014 and 2015		Relevance		
		Essential	Important	Acceptable
Difficulty	Easy	0.75	0.55	0.15
	Medium	0.60	0.45	0.12
	Hard	0.40	0.20	0.07

Table A1: Proportion of minimally competent students who will get the correct answer

2014		Relevance			Total
		Essential	Important	Acceptable	
Difficulty	Easy	31	21	0	52
	Medium	45	70	5	120
	Hard	0	20	6	26
Total		76	111	11	198

Table A2: Item allocations in 2014

2015		Relevance			Total
		Essential	Important	Acceptable	
Difficulty	Easy	42	22	2	66
	Medium	32	54	9	95
	Hard	1	20	9	30
Total		76	96	20	191

Table A3: Item allocations in 2015