Keeping up Appearances: maintaining standards in Hong Kong's Language Proficiency Assessment for Teachers of English (LPATE)

Dr. Neil DRAVE
Hong Kong Examinations and Assessment Authority, Hong Kong
ndrave@hkeaa.edu.hk

Abstract
The Language Proficiency Assessment for Teachers of English (LPATE) assesses the English proficiency of those who wish to teach English in Hong Kong schools. The LPATE is a standards-referenced assessment administered by the Hong Kong Examinations and Assessment Authority (HKEAA) and is a high-stakes test in the sense that candidates who do not reach the minimum standard (the 'benchmark') are not allowed to teach English. Since the test outcome is so crucial, both for the individual candidate and the teaching profession, it is important that the appropriate standards are implemented and maintained. Without consistency in this regard, the various parties which make use of the test results would be unable to use them to make employment-related decisions. The prevailing standard must be applied equally to all candidates and the standard of performance expected must be very similar from year to year. This paper explores the issue of standards within the context of teacher education and assessment in Hong Kong. It discusses how the standards were originally set for the LPATE and what procedures and practices are in place to maintain them.

Keywords: language proficiency, standards, expert judgement

Introduction
Anyone wishing to teach English in a Hong Kong public school must demonstrate that they are proficient in the English language. They can do this by attaining a prescribed minimum score on the Language Proficiency Assessment for Teachers of English (the LPATE)[1].

The LPATE is a high-stakes assessment because people who do not meet its minimum standard are not allowed to teach English in a school. Since its introduction in 2001, it has been a high-profile, and often controversial, part of the Hong Kong Government's strategy for making teaching more professional (Qian, 2008; Drave 2006), a worldwide trend which has given rise to many similar certification tests (e.g. the LPTT for language teachers in Australia; Elder, 2001).

Hong Kong's education community has also come to accept the need for greater teacher professionalism and certification, and the LPATE is now a widely trusted test. Hong Kong universities require their education students to take the LPATE, and many schools view it as a desirable qualification even for teachers with degrees in English. Recent years have seen increased academic debate about, and research on, what language teachers should know and be able to do (e.g. teachers' language awareness; Andrews, 2007), and this trend may be partly attributable to the implementation of the LPATE.

As Hong Kong's first major standards-referenced assessment, the LPATE has also been at the forefront of the movement to make assessments more 'meaningful' by reporting results as proficiency levels, a practice which will be adopted for the new Hong Kong Diploma of Secondary Education when it is implemented in 2012.

All of these factors – the gate-keeping function, high profile within education and pioneering reporting method – make the LPATE an important and influential assessment. It is therefore crucial that the standards which its candidates must meet are well thought out, clearly articulated and reliably implemented from year to year.

In the remainder of this paper, I address the issue of standards. I describe the current LPATE standard setting and maintenance mechanisms, focusing on the setting of cut scores, and

[1] There is also a Putonghua equivalent, the LPAT-P. Those with relevant degrees are exempted from the LPATE.

evaluate these in light of current research. I conclude with a list of issues which need to be considered when setting standards.

Structure of the LPATE

The LPATE has five papers: Paper 1 Reading, Paper 2 Writing (Part 1: Composition, Part 2: Error correction and explanation), Paper 3 Listening, Paper 4 Speaking, and Paper 5 Classroom Language Assessment (administered by the Education Bureau of the Hong Kong Government and for serving teachers only). Based on their performance in a paper, a candidate is awarded a proficiency level indicated by a number between 1 and 5, with a Level 3 corresponding to the level of language proficiency required to 'pass' the assessment, i.e. to be able to take up a post as an English teacher in a school.

Paper 2 Part 1 (Composition), Paper 4 (Speaking) and Paper 5 (CLA) are direct-performance assessments which employ explicitly defined scales and descriptors. The method of setting the standard in these 'productive' skills is quite straightforward: samples of performance from pervious years (e.g. candidate compositions, videos of Speaking tests etc.) are reviewed by senior examiners, who decide what level they demonstrate, and these are then shown to new markers in a standardisation and training session. The aim is to ensure that the observed performance is matched with the descriptors at each scale level and that the required standard does not change from year to year.

The remaining LPATE papers (Reading, Writing [Error correction and explanation] and Listening) are indirect, paper-and-pencil tests which give rise to numerical scores. The original proficiency levels for the papers were defined by a panel of experts using data from Language Benchmark Pilot tests, administered in 1999 (Coniam and Falvey, 1999). There were revisions to the format of the LPATE in 2007 and these necessitated a re-setting of the standards, which was undertaken using the same procedures as are currently used for the LPATE (see below). It is therefore the equivalents of these levels which need to be identified for the papers in each administration to ensure that the prevailing standards match those which are publicly available.

LPATE employs a 'conjunctive' rather than a 'compensatory' scoring method (Zieky and Perie, 2006, 10; Hambleton and Pitoniak, 2006, 450), meaning that a candidate has to pass all scales in all papers to be benchmarked: there is no provision for failure on one scale to be compensated for by a pass on another. Conjunctive methods are considered appropriate for tests of separate language skills (Zieky and Perie, 2006, 10) since it is fair to expect a candidate to be strong in all areas. However, the LPATE employs within-paper conjunctive scoring, a practice which may lead to lower pass rates than compensatory methods (Coniam and Falvey, 2001) and which has to be taken into consideration when setting standards.

Standard setting methods

The task of the 'standards-maintaining' exercise (Bramley and Black, 2008, 3) for the paper-and-pencil tests aims to identify points on a score scale that divide the observed test score into classifications on a five point scale. Decisions have to be made as to what score ranges are considered equivalent to LPATE proficiency levels 1, 2, 3, 4 and 5, i.e. what the cut scores should be (Kane, 2001, 55). There is no foolproof way of making these classification decisions and all methods have their weaknesses. For example, any score is subject to measurement error, and this holds true for cut scores as well. There are also political and other pressures on test administrators, meaning that the process of determining cut scores may be as much political as it is educational (Zieky, 2001, 46; Kane, 2001, 58; Cizek et al., 2004, 32). Nevertheless – or perhaps because of this – setting cut scores is a crucial task.

It is common to categorise standard setting methods into those based on judgements about test questions and of test takers' work (Hambleton and Pitoniak, 2006; Zieky and Perie, 2006; Cizek et al., 2004). In the former category are the Nedelsky method for MC items, the Bookmark method, in which items are ordered according to difficulty, and the Ebel method, which requires judgements about item relevance/importance as well as difficulty. Testee work is monitored in the Borderline and Contrasting Groups methods, and in the Body of Work method, in which

holistic ratings are given to 'response booklets' containing a number of test-taker performances. LPATE draws upon both categories of methods to maintain its standards.

LPATE practice

The two methods of determining cut scores in the LPATE are Expert Judgement and Rasch analysis. The Expert Judgement method (which might more properly be termed a 'procedure' since it can be used with a number of methods) is a variation of the Extended (or Modified) Angoff Method (Hambleton and Pitoniak, 2006; Zieky and Perie, 2006), used for many purposes, including linking language tests to the CEFR (Tannenbaum and Wylie, 2009), the certification of doctors in the US (Morrison et al., 2009; Clauser et al., 2009), and of medical translators in Japan (Kozaki, 2004). In this method, a panel of expert judges is asked to estimate the proportion of target examinees (in our case, minimally competent teachers of English in Hong Kong schools) that would probably be able to answer correctly each of the test items. The proportion assigned to each item may also be conceptualized as the probability that a single target examinee can answer that item correctly. The judges' estimations are expressed in the form of a numerical value between zero and one: '0.65' would mean a 65% probability of a correct response, for example. A cut score is then obtained by summing these values across all items. The Angoff Method originally required simple 'Yes' or 'No' responses and was tailored to MC items, but this modification has proven to be very suitable for constructed-response items of the kind found in the LPATE assessment (Cizek, 1993, 1996; Berk, 1986, 1995).

We supplement Angoff in two ways: first by providing descriptive test statistics and other relevant data (e.g. normative or impact figures); and second, by giving judges an opportunity to directly review live candidate scripts (cf. Hambleton and Pitoniak, 2006, 447).

Before the judges make their final decision on what cut scores to recommend, they are given additional data in the form of cut scores suggested by a one-parameter IRT (Rasch) analysis. Rasch has many uses in the social sciences (Bond and Fox, 2002) and is widely used for setting educational standards (e.g. in pharmaceutical examinations, Jackson et al., 2002). Our data come from common person equating of pre-tests and anchor tests, the results of which are extrapolated to the live versions of the tests. Rasch can be used to provide data which might help judges decide on their probabilities, such as item difficulty figures (Bond, 2003, 191; MacCann and Stanley, 2006; Tiratira, 2009), but in LPATE it is used only for test equating.

There are many things to consider when planning and running a standard setting exercise (Hambleton and Pitoniak, 2006; Cizek, 2001). The current LPATE procedure may be summarised as follows:

1. Pre-testing: we administer an anchor test for each LPATE paper, which is published and for which the cut scores have already been set, and a trial version of the live test (with more items) to pre-test candidates.
2. Test equating: after marking the pre-test, and the deletion of unsuitable items, common person test equating is undertaken. Using the RUMM computer program, the pre-test candidates' scores on both tests are converted to logits (Bond and Fox, 2001; Luo et al., 2001; Yu and Osborn Popp, 2005, describe how to use the WINSTEPS program to do this).
3. Extrapolation. The logit values which correspond to the published (anchor) Level 2, 3, 4 and 5 cut scores are extrapolated to the 'live' tests. First, the logit values which correspond to the appropriate cut scores on the anchor test are identified. Second, the same logit values on the 'live' test are identified. Third, the 'live' test scores which correspond to these logit levels are read off and recorded for later use.
4. Statistics. When the live test has been administered and marked, test statistics are calculated. The most useful are the measures of central tendency, the IF index (item facility, i.e. percentage correct) and the ID index (item discrimination, how well each item was answered by the best and worst candidates).

An Expert Judgement exercise (in the form of a round-table meeting) is then carried out to:

1. Gather information from a source other than the Rasch analysis to help determine the cut scores.

2. Ensure that the determination of the cut scores for the three papers is transparent to, and monitored by, members from various stakeholder groups. This is important because it ensures that the process has face validity.

The expert group normally has between 10 and 15 participants, which is an optimal number (Zieky and Perie, 2006). It includes primary and secondary school principals and teachers, lecturers from tertiary institutions, representatives from committees which designed and moderated the test papers and representatives from the Hong Kong Government's Education Bureau, which originally commissioned the LPATE.

Judges are sent an information pack containing the test papers and instructions, and asked to do the papers as candidates before they attend the standard setting meeting. Each panel member then estimates the probability of a just-qualified teacher answering each item correctly, as described previously. The estimates are entered into an Excel spreadsheet, outliers are removed and the probabilities summed to give preliminary cut scores.

Standard setting meeting

The expert panel begin by setting the cut score for Level 3 on one of the papers (normally the Reading), as this is the crucial 'benchmark' level.

The panel is first presented with the probability scores of all members and the preliminary cut score for Level 3. The (Level 3) paper cut scores of individual panel members are compared with the cut scores suggested by the Rasch analysis. If the results of the panel are reasonably coherent and deemed close enough to the Rasch scores, the process ends and the panel makes a firm recommendation to the approving body. If the results of the panel are reasonably coherent but not close to the Rasch cut score, there are two options:

1. The panel provides an explanation to support its recommendations; or
2. The panel may revise their item probabilities.

If the second of these options is chosen, additional information is provided to the panel to help them decide whether to change their probabilities (e.g. test statistics from the live administration, impact data on how different cut scores would affect the passing percentage). A new cut score is calculated, after removing outliers, and presented to the panel for their consideration. If there is still a difference between this figure and the one suggested by the Rasch, or if there is little difference but the panel so wishes, the panel will analyse samples of performance by viewing the scripts of candidates who have been awarded different raw scores. Judges may be directed to consider candidate performances on items at certain levels of difficulty (as defined by different IF scores), which makes this part of the process similar to the Bookmark method. A cut score is then decided upon, taking into account a one or two mark margin of error (depending on the paper), which is incorporated to account for the fact that a cut score has a standard error (Jaeger and Mills, 2001, 329). This procedure is then repeated for all the levels on all the papers.

Certain features of the procedure used in LPATE (discussion among judges, the opportunity to re-rate, consideration of item statistics and impact data) are known to lead to greater inter-expert consistency, which may be considered a positive feature. Giving judges access to test data leads to more realistic judgements as then they know how the candidates (not just minimally competent ones) have actually performed (Zieky and Perie, 2006, 10). This is a legitimate approach and should not be seen as 'contaminating' the standard setting process (Zieky and Perie, 2006; Zieky, 2001, 37). It also ensures that there is less of a mismatch between expert judgements of what candidates can do (which may be partly prescriptive – see below) and what they actually can do. However, in some contexts such features may give rise to standards which are too conservative (Cisek, 2001, 11).

Strengths of the current method

The main strengths of the current LPATE procedures are:

1. The method has been used for many years and all those involved know the process thoroughly. It therefore requires minimal (re)training to use.

2. The expert group is relatively stable, with a low turnover of members from year to year, and has representatives from the major stakeholder groups.
3. The group members have relevant subject knowledge and professional expertise, so they are able to make informed judgements (Cizek et al., 2004, 34).
4. The standard setting process has objective (i.e. statistical) and subjective components, which provide a check on each other.
5. The process involves consideration of pre-test and live test information, as well as impact data, which gives the judges all they need to make principled and realistic decisions about cut scores.
6. The process – at least at the meeting stage – is efficient because much of the time-consuming work (e.g. assigning probabilities, test equating) is done before the actual standard setting process gets underway.
7. The Extended Angoff method leads to a single, clear cut score for each level.
8. The entire process conforms to (most of) the commonly accepted guidelines for good standard setting practice (see Hambleton and Pitoniak, 2006 for a list of these).

Issues

While the current process is generally satisfactory, certain aspects could be improved upon. The main issues are as follows:

1. Difficulty of judging probabilities.
    a. The Angoff method is not easy to implement as it relies on experts' ability to internalise and apply standards from year to year as they operationalise the concepts of 'benchmark level' and 'just-qualified candidate' (Zieky, 2001, 35-37; Jaeger and Mills, 2001, 335; Bramley and Black, 2008, 9). Judges' estimations of difficulty may therefore be inaccurate. There is some evidence, for example, that easier items are systematically judged to be more difficult than they really were, and vice versa (Hambleton and Pitoniak, 2006). In LPATE, we attempt to control for this by providing actual test data. One possible enhancement to the current procedure would be to undertake IRT analysis of the live test performance and supply the resulting data to judges.
    b. Judges may not be clear about their task as they are not normally asked to make probability judgements. A useful indication of whether the judges understand what they have to do is to look at the probabilities assigned to MC items, which for a 4-option item should be $\geq$0.25 i.e. 1 divided by the number of choices (Zieky and Perie, 2006, 14).
2. Gate keeping. Since the LPATE is a test for teachers, with teachers and teacher trainers on the expert panel, some members may be prescriptive rather than descriptive when making judgements, i.e. they may assign probabilities according to what they think candidates should be able to do rather than what they are probably capable of. Prescriptivism can be used to set cut scores, as in the Jaeger method (Hambleton and Pitoniak, 2006, 442), but it leads to high standards, as the experts exercise their perceived gate keeping function.
3. Groupthink. Like any group, an expert judgement group is subject to 'groupthink' (Janis, 1972) which manifests itself as behaviour consistent with a belief that any decision made by a group is inherently and necessarily better than that made in another way. This may lead to a rejection of the Rasch evidence, for example, or to the stifling of dissenting voices in the group discussion phase.
4. Pre-testing and Ras(c)h decisions. The cut scores provided by Rasch analysis are useful only in so far as the pre-test and test equating data are reliable. There are always likely to be question marks over this, however, and care must taken to ensure that
    a. the pre-test candidate population is large and heterogeneous enough to enable accurate calibration of items (Hambleton and Jones, 1993);
    b. testees are motivated to answer to the best of their ability;
    c. each test item discriminates equally (Henning et al., 1985);
    d. items are pitched at an appropriate level of difficulty for the pre-testees, so that item difficulty can be reliably estimated (Bond and Fox 2001, 58);
    e. the marking of the anchor test is consistent from year to year;
    f. the live test is marked in the same way as the pre-test.

Also, any lack of understanding of the Rasch model could lead to inappropriate interpretation of its results or to judges simply ignoring them.

5. Judging candidate proficiency. The final stage of the expert judgement process is (almost always) the viewing of candidate work, particularly for setting cut scores for levels other than Level 3. There are various matters to take into account when engaging in this practice.
    a. Because of the relatively small candidature, there are only a small number of scripts at particular mark levels, particularly at the top and bottom of the ability range. It may therefore be difficult to get a sense of the quality of candidate performance at each level.
    b. LPATE is not a criterion-referenced test and therefore it is difficult to give a definite point of mastery/non-mastery which would correspond to a 'benchmark' level.
    c. Descriptors are widely considered to be a useful tool for guiding the viewing of candidate scripts/performances and making judgements about them (Cizek at al., 2004, 34; Alderson, 1995, 76; Zieky and Perie, 2006; Kane, 2001, 56). LPATE uses these for the papers which assess productive proficiency (Writing Part 1, Speaking, CLA) but not for other papers because it is felt that any set of descriptors would be too general to be useful for analysing a candidate's performance on an item-based test. Care must be taken, therefore, to ensure that judgements are made by comparing candidate performance to the 'putative standard' which has emerged from previous discussion, rather than simply to other candidates, which would contravene the spirit of the standards-referenced assessment method.
    d. Due to time constraints, there may be a tendency to look only at certain key items, which may then become the de facto criterial items for benchmarking. This tendency may be exacerbated by the fact that scripts are presented to the meeting on a computer screen rather than on paper as this means that judges' attention at any one time is focused on specific part of a script.
    e. Judges must be familiar with the criteria used to judge candidate answers so that they do not make judgements based on irrelevant criteria (e.g. handwriting).
6. Consistency across papers. Angoff sums the scores on items to arrive at a cut score, and therefore produces a standard which is compensatory (Hambleton and Pitoniak, 2006, 450); candidates can be weak in one of the three passages used to test Reading, for example, but as long as they are strong enough in the other passages they can pass, as there is just one cut score. However, the direct tests of writing and speaking are not compensatory (or are so in a very limited sense). We therefore have different papers being treated in different ways. There is a need for more research on the consequences of this and on how we can ensure that the standard of proficiency required by different papers is similar.
7. Leader integrity. The person who leads the standard setting process, provides the data and leads the discussion is likely to be in a very powerful position and can influence the ultimate outcome of the process. It is therefore a matter of the utmost importance that this person is reliable, that they are independent and do not have a vested interest in seeing certain proportion of candidates pass/fail, and that they keep a record of the process which can be scrutinised by all concerned. Many of the issues mentioned above can be managed if the standard setting process is handled with integrity and professionalism.

Conclusion

In this paper I have focused on the process of maintaining standards for the three LPATE papers which use cut scores. Given the high-stakes nature of the assessment, it is important that the method adopted is undertaken with great transparency and rigour. There is no perfect procedure, of course, and the one currently used is not without its issues. The HKEAA is confident, however, that the current practice is theoretically defensible and effectively monitored, and that it gives rise to standards which are quite stable from year to year.

References

Alderson, J. Charles. (1995). Bands and scores. In Alderson, C. and B. North (Eds.), Language Testing in the 1990s: the Communicative Legacy. Hemel Hempstead: Phoenix ELT, 71-86.

Andrews, S. (2007). Teacher Language Awareness. Cambridge: CUP.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Berk, R. A. (1995). Something old, something new, something borrowed, a lot to do! Applied Measurement in Education, 8(1), 99-109.

Bond, T. (2003). Validity and assessment: a Rasch measurement perspective. Metodologica de las Ciencias del Comportamiento, 5(2), 179-194.

Bond, T. and C. Fox. (2001). Applying the Rasch Model. Mahwah, NJ: Lawrence Erlbaum.

Bramley, T. and B. Black. (2008). Maintaining performance standards: aligning raw score scales on different tests via a latent trait created by rank-ordering examinees' work. Paper presented at the 3rd International Rasch Measurement conference, University of Western Australia, Perth, January 2008.

Cizek, G. (1993). Reconsidering standards and criteria. Journal of Educational Measurement, 30, 93-106.

Cizek, G. (1996). Standard-setting guidelines. Educational Measurement: Issues and Practice, 15, 13-21.

Cizek, G. (Ed.) (2001). Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum.

Cizek, G. (2001). Conjectures on the rise and call of standard setting: introduction to context and practice. In Cizek, G. (Ed.), Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum, 3-17.

Cizek, G., M. Bunch and H. Koons. (2004). Setting performance standards: contemporary methods. NCME Instructional Module. Available online: http://www.ncme.org/pubs/items/Setting%20Performance%20Standards%20ITEMS%20Module.pdf

Clauser, B., J. Mee, S. Baldwin, M. Margolis and G. Dillon. (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), 390-407.

Coniam, D. and P. Falvey. (1999). Setting standards for teachers of English in Hong Kong – the teachers' perspective. Curriculum Forum, 8(2), 1-23.

Coniam, D. and P. Falvey. (2001). Awarding passes in the Language Proficiency Assessment for Teachers of English: different methods, varying outcomes. Chinese University of HK Education Journal, 29(2), 23-35.

Drave, N. (2006). The Language Proficiency Assessment for Teachers of English (LPATE) as an instrument of educational change. Proceedings of the 9th Academic Forum on English Language Testing in Asia, 18-40.

Elder, C. (2001). Assessing the language proficiency of teachers: are there any border controls? Language Testing, 18(2), 149-170.

Hambleton, R. and M. Pitoniak. (2006). Setting performance standards. In Brennan, R. (Ed.), Educational Measurement. Westport, CT: Praeger, 433-470.

Hambleton, R. and R. Jones. (1993). Comparison of Classical Test Theory and Item Response Theory and their applications to test development. National Council on Measurement in Education, Instructional topics and educational measurement series, 253-262. Available online: http://www.ncme.org/pubs/items/24.pdf

Henning, G., T. Hudson and J. Turner. (1985). Item response theory and the assumption of unidimensionality for language tests. Language Testing, 2(2), 141-154.

Hurtz, G. and M. Auerbach. (2003). A meta-analysis of the effects of modifications to the Angoff Method on cutoff scores and judgment consensus. Educational and Psychological Measurement, 63(4), 584-601.

Jackson, T., J. Draugalis, M. Slack, W. Zachry and J. D'Agostino. (2002). Validation of authentic performance assessment: a process suited for Rasch modeling. American Journal of Pharmaceutical Education, 66, 233-243.

Jaeger, R. and C. Mills. (2001). An integrated judgement procedure for setting standards on complex, larger-scale assessments. In Cizek, G. (Ed.), Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum, 313-338.

Janis, Irving L. (1972). Victims of Groupthink. Boston: Houghton Mifflin Company.

Kane, M. (2001). So much remains the same: conception and status of validation in setting standards. In Cizek, G. (Ed.), Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum, 53-88.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. Language Testing, 21(1), 1-27.

Luo, G. A. Seow and L. C. Chew. (2001). Linking and anchoring techniques in test equating using the Rasch model. Available online: http://www.caaconference.com/pastConferences/2001/proceedings/h1.pdf

MacCann, R. and G. Stanley. (2006). The use of Rasch modeling to improve standard setting. Practical Assessment Research and Evaluation, 11(2). Available online: http://pareonline.net/getvn.asp?v=11&n=2

Morrison, H., H. McNally, C. Wylie, P. McFaul and W. Thompson. (2009). The passing score in the Objective Structured Clinical Examination. Medical Education, 30(5), 345-348.

Qian, D. (2008). English language assessment in Hong Kong: a survey of practices, developments and issues. Language Testing, 25(1), 85-110.

Tannenbaum, R. and E. Caroline Wylie. (2009). Using standard-setting methodology for linking assessment scores to proficiency scales: TOEFL iBT and TOEIC assessment exemplars. ETS Research Spotlight, 2, 10-14.

Tiratira, N. (2009). Cutoff scores: the basic Angoff method and the Item Response Theory method. The International Journal of Educational and Psychological Assessment, 1(1), 39-47.

Yu, C. H. and S. Osborn Popp (2005). Test equating by common items and common subjects: concepts and applications. Practical Assessment Research and Evaluation, 10(4). Available online: http://pareonline.net/getvn.asp?v=10&n=4

Zieky, M. (2001). So much has changed: how the setting of cutscores has evolved since the 1980s. In Cizek, G., Setting Performance Standards: Concepts, Methods, and Perspectives. Mahwah, NJ: Lawrence Erlbaum, 19-51.

Zieky, M. and M. Perie. (2006). A Primer on Setting Cut Scores on Tests of Educational Achievement. Princeton, NJ: Educational Testing Service.