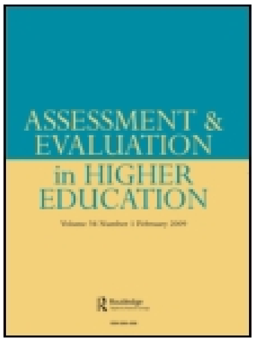
ASSESSMENT & EVALUATION in HIGHER EDUCATION

Assessment & Evaluation in Higher Education

Routledge
Taylor & Francis Group

ISSN: 0260-2938 (Print) 1469-297X (Online) Journal homepage: https://www.tandfonline.com/loi/caeh20

Criterion-referenced and norm-referenced assessments: compatibility and complementarity

Beatrice Lok, Carmel McNaught & Kenneth Young

To cite this article: Beatrice Lok, Carmel McNaught & Kenneth Young (2016) Criterion-referenced and norm-referenced assessments: compatibility and complementarity, Assessment & Evaluation in Higher Education, 41:3, 450-465, DOI: 10.1080/02602938.2015.1022136

To link to this article: https://doi.org/10.1080/02602938.2015.1022136

Published online: 27 Mar 2015.

Assessment & Evaluation in Higher Education, 2016
Vol. 41, No. 3, 450–465, http://dx.doi.org/10.1080/02602938.2015.1022136

Criterion-referenced and norm-referenced assessments: compatibility and complementarity

Beatrice Lok[a], Carmel McNaught[b] and Kenneth Young[b]*

[a]Centre for the Advancement of Learning and Teaching (CALT), University College London, London, UK; [b]The Chinese University of Hong Kong, Hong Kong, China

The tension between criterion-referenced and norm-referenced assessment is examined in the context of curriculum planning and assessment in outcomes-based approaches to higher education. This paper argues the importance of a criterion-referenced assessment approach once an outcomes-based approach has been adopted. It further discusses the implementation of criterion-referenced assessment, considering to what extent the criteria and standards adopted are implicitly norm referenced. It introduces a compatible interpretation of criterion-referenced and norm-referenced assessments in higher education, and illustrates how their combined use can avoid grade inflation and also provide useful information to educators, employers and learners. Instead of seeing criterion referencing and norm referencing as a dichotomy, assessment in higher education benefits from their synthesis through a feedback loop that emphasises alignment between learning and assessment; such feedback and alignment are essential features of quality assurance and enhancement.

Keywords: criterion referencing; norm referencing; outcomes-based approaches; grade inflation

Introduction

Assessment of student learning is an essential element in higher education. Raw scores in assessment are seldom reported as such, but are typically coarse-grained into broad bands or grades, for example, letter grades. In criterion referencing, each student is judged against predetermined absolute standards or criteria, without regard to other students; thus, it is possible for a majority to obtain the top grade, say ‘A’, or, conversely, for none to do so. In norm referencing, a predetermined percentage of students (usually with some margin of flexibility) would obtain a certain grade; if the entire class is inadequate, there would still be the same number of ‘A’s, and, conversely, if the entire class is outstanding, the same number of ‘D’s must still be awarded.

It is clear that neither approach, in its pure form, is appropriate in extreme scenarios. Most teachers, if pressed, would confess to a pragmatic hybrid, expecting some level of absolute performance and at the same time respecting the convention on grade distribution. This approach is coming under increasing stress for a variety of reasons. The hybrid lacks conceptual clarity, for example, in situations when the two considerations come into conflict. At a micro-level, such lack of clarity invites arbitrary and inconsistent practice. At a macro-level, it sits ill with the modern

*Corresponding author. Email: kyoung@cuhk.edu.hk concept of quality assurance, which demands conscious decisions based on unambiguous policy, monitoring the consequences, and then reflecting on the policy and practice for continuous improvements. Clear quality-assurance policy is usually predicated on an outcomes-based approach to curriculum development, in which clear assessment criteria are necessary. For these reasons, there has been continuing debate about criterion-referenced and norm-referenced assessment, so much so that they are often seen as mutually exclusive options in a binary choice. This paper examines the debate in a broad context, argues that criterion-referenced and norm-referenced approaches to assessment should be viewed as compatible and complementary, and puts forward a model that transcends the dichotomy, so that the two approaches work together consistently in a synergistic feedback loop that is consonant with modern conceptions of quality assurance.

Purposes of assessment

It is impossible to discuss the relative merits of criterion referencing vs. norm referencing without considering the purposes of assessment.

Formative assessment

Formative assessment aims to provide feedback to learners and focuses on improvements facilitated by information on what has been mastered and where weaknesses lie. For this purpose, criterion referencing appropriately reflects the level of achievement of an individual, whether in self- or diagnostic assessments. The depth of feedback can be varied according to the purpose and design of assessment. In the simplest case, the feedback may be expressed as a single 'cut-off point' (i.e. pass or fail) with no differentiation about degrees of attainments. More commonly, further diagnostic details are given. As examples, the computer-based Test of English as a Foreign Language is reported with a score out of 300 and with indicators of proficiency level on specific skills; while an essay can be marked with detailed written comments against criteria specified in a rubric, with or without an overall grade. Feedback in formative assessment assists learners to understand the depth of the criteria and, in forcing the marker to be conscious of different levels of attainment with respect to the criteria, helps to sharpen the scoring.

Summative assessment

Summative assessment provides scoring information that can be read and easily understood by a range of stakeholders, including external stakeholders; for this reason, and in contrast to the feedback function in formative assessment, the term 'feedout' is sometimes used. Most public examinations at the end of secondary school are summative in nature.

Because criteria (and their interpretation) are likely to be unfamiliar to external stakeholders, there is often the need to express summative assessment in norm-referenced terms. Most employers would find a statement such as 'in the top quartile of University X' useful, if they have some familiarity with the university. Many graduate schools ask referees to state the percentile position of the applicant (e.g. 'truly outstanding, top 5%'), incidentally betraying a possible lack of confidence in grades and grade distributions.

Learners themselves also want to know where they stand amongst peers because the answer indicates their competitiveness amongst the group. School students particularly need this type of information when competing for limited university places. In these circumstances, summative assessment with a norm-referenced element is essential because there is truth in the statement that 'the real world is norm referenced and not criterion referenced': if there are 20 desirable jobs for 100 applicants, the top 20% (no matter what their absolute standards may be) will get them; the same is true for all quota-limited contexts: undergraduate places (especially in highly sought-after institutions and programmes), graduate school admission, awards of scholarships, etc.

Increasing tension

Whether the assessment is for formative or summative purposes, the tension between the criterion-referenced and norm-referenced methodologies has been accentuated by two developments: the trend towards outcomes-based approaches in higher education and the phenomenon of grade inflation, leading to the need for instruments to combat it.

Outcomes-based approaches

In an outcomes-based approach, a clear articulation of desired learning outcomes takes centre stage in the curriculum, and drives the design of the learning environment and process. The choice of content, the design of learning activities and the nature of assessment should be aligned with the stated learning outcomes.

The concept of outcomes-based approaches was influenced by the reform of education in the USA in the second half of the twentieth century. The complex political (e.g. the cold war) and ideological (e.g. race relations) reasons that drove the reform are beyond the scope of this paper. While the main emphasis was on school education, a flow-on effect into higher education took place – in terms of both funding and policy.

A number of education problems were tackled in this reform era (Spady 1994):

(1) 'Clarity of focus': educators must clearly illustrate the intended learning outcomes with students and prioritise them in planning, teaching and assessment;
(2) 'Designing down': intermediate outcomes are valuable in scaffolding or supporting achievement of the broader outcomes;
(3) 'High expectations': the level of acceptable performance should be raised; this entails abandoning a norm-referenced assessment approach and opening access to all students for high-level learning;
(4) 'Expanded opportunity': while the learning outcomes might be fixed, the time to achieve could vary.

Carroll (1963) also argued that uniform minimum standards should be clearly stated and understood, and that students should be given the time to achieve these standards. Bloom (1968) followed Carroll's thinking and developed the notion of mastery learning, which has become a core principle of outcomes-based approaches through all levels of education. Parallel processes to those in the USA also took place in most of the English-speaking world outside Asia.

Much of the early emphasis on specification of reliably measurable outcomes can be traced to the behaviourist tradition in education with its roots in Skinnerian psychology (e.g. Tyler 1949). However, particularly since the 1980s, the challenge of the articulation and measurement of complex and multifaceted educational outcomes has received increasing attention (Ewell 2009).

Of course, conversations and policy at national and institutional levels often take time to filter down to departments and individual teachers. Outcomes-based approaches adopted by a university for quality-assurance (and perhaps even pedagogical) reasons may not have actually made much difference to the learning experience of students; the decoupling of policy and rhetoric from practice is not unknown in other aspects of higher education. The adoption of outcomes-based approaches is not an on/off switch, and should not be portrayed as easily achievable.

For example, outcomes which are clearly defined and easily measurable can be quite low level in cognitive demand. However, the essential skills and capacities that should be fostered in university involve a variety of intellectual, interpersonal and personal capabilities, such as critical thinking, creative thinking, self-managed learning, adaptability, problem-solving, communication skills, interpersonal skills and group work, and computer literacy. Articulating these desired outcomes in terms of concrete criteria is challenging and possibly self-defeating; moreover, teachers need to be able to couch discipline-domain outcomes within these broader capabilities. One solution to this dilemma is to use a checklist/grid/taxonomy to monitor the balance of stated outcomes at both programme and course level. Two main taxonomies are that of Bloom (1956) and its revision (Anderson and Krathwohl 2001; Krathwohl 2002); and the Structure of the Observed Learning Outcomes taxonomy (Biggs and Collis 1982) and its simplification (e.g. McNaught, Cheng, and Lam 2006).

Outcomes-based approaches require clearly stated assessment criteria pegged to the declared outcomes, and therefore argue for – indeed mandate – the use of criterion referencing in assessment. The strong synergy between outcomes-based approaches and criterion referencing means that curriculum developers should address assessment processes explicitly and should not leave the implementation of outcome statements to the whim of individual teachers. To abandon or minimise explicit criteria in assessment would undermine the entire curriculum design. Achieving good curriculum alignment is an iterative process. Thus, outcomes-based approaches should focus not only on the reliability of educational measurement, but also on the principle of curriculum alignment (Biggs 1999; Biggs and Tang 2011) in curriculum development (Shay 2008). Outcomes-based approaches, and the quality framework of which they are a part, refer not only to assessments, but to the entire curriculum, with a focus on alignment between clearly articulated outcomes, learning and pedagogy targeted towards those outcomes, and assessments against the same outcomes.

Grade inflation

Grade inflation erodes public trust: the public may be inclined to believe that an ‘A’ grade equates to the top 10 or 20% of the class; in practice, if only criterion-referenced assessment is adopted, considerably more can be awarded an ‘A’. A recent large-scale study (Rojstaczer and Healy 2012) has documented massive grade inflation in American colleges and universities, with ‘A’ grade now accounting for 43%, up from 28% in the 1960s. Grade inflation is exacerbated by certain circumstances. If failure or multiple attempts are costly for the student – in terms of money and effort, or the need to repeat a year – there may be pressure to award passing grades. Indeed, the most rapid grade inflation in the USA occurred during the period of the Vietnam War (Rojstaczer and Healy 2012), when poor grades could jeopardise exemption from military service (for the male students). A culture of students as consumers with attendant rights and a litigious atmosphere in some societies add to the pressure on teachers; moreover, stringent grading impacts negatively on student evaluation of teaching, which increasingly is seen to have an effect on the career of the teacher.

Moderation by a third party can alleviate any pressure on the teacher, overt or otherwise, but probably the only effective instrument to combat grade inflation is a robust grade-distribution guideline.

Dilemma

One is therefore caught in an apparent dilemma: the modern quality framework pushes institutions towards outcomes-based approaches and therefore criterion-referenced assessment, but the abandonment of norm-referenced assessment and especially grade-distribution guidelines can allow rampant grade inflation. It is therefore imperative to re-examine the true natures of both criterion referencing and norm referencing in order to find a way out of the dilemma.

What do criterion referencing and norm referencing really mean?

The precise term ‘criterion referenced measurement’ was proposed by Glaser and Klaus (1962) in education and educational psychology to indicate measures of proficiency on an individual achievement basis. Glaser (1963, 519), in the original application in a psychological arena, suggested the term ‘criterion-referenced measures’ and highlighted the advantages of measuring achievement based on standards rather than on norms. He highlighted the notion of a criterion-referenced measure as ‘a continuum of knowledge acquisition’ that focused on the development of an individual within a particular period of time, leading to a very different score interpretation from a norm-referenced one. Bond (1996) and Huitt (1996) compared norm-referenced and criterion-referenced tests in various dimensions: (1) intended purposes; (2) selection of test content; (3) test interpretation; and (4) item characteristics. Their papers focused on the differences between the two approaches, especially factors that are important in choosing between the two; however, there was no discussion about their combined use.

The ‘criterion referencing only’ camp is typified by Gipps (1994) and Nightingale et al. (1996), who discussed the disadvantages of norm referencing and supported the use of criterion referencing for assessment. For instance, Gipps (1994) reviewed different types of assessments, both norm-referenced and criterion-referenced. She compared their purposes and effects on learning, and argued that criterion-referenced assessment – through the conscious incorporation of criteria that relate to higher-order cognition skills – ensures that these abilities are given adequate attention. Nightingale et al. (1996) highlighted three assessment principles which support the use of criterion referencing: (1) a variety of clearly stated learning outcomes introduces flexibility in making reasonable judgements about students' achievement levels across a number of different but related aspects; (2) explicit goals motivate learning; and (3) a holistic judgement across complex criteria is more aligned with graduate capabilities in higher education. Nevertheless, teachers are often reluctant to fully embrace criterion referencing, if only because writing clear criteria for different levels of performance is difficult, involving some knowledge of educational theory not always familiar to teachers appointed as subject experts.

Neil, Wadley, and Phinn (1999) provided a brief review of both approaches and proposed a generic framework that acknowledged the strengths and weaknesses of criterion referencing for assessing undergraduate written work. They stressed that criterion referencing can emphasise higher-order cognitive skills, such as critical thinking, and clear and incisive writing – though one must caution that such criteria are difficult to interpret without a context, a point to which we shall return. Wiliam (1996) also reviewed the respective drawbacks of criterion referencing and norm referencing, and argued that neither approach was able to offer an adequate and authentic assessment of performance in higher education.

The differences and similarities between criterion-referenced and norm-referenced assessments are summarised in Table 1. These distinctions are mainly conceptual, but blurred in practice.

Table 1. Comparison between criterion-referenced and norm-referenced assessments.

	Criterion-referenced assessment	Norm-referenced assessment
Purpose	• Reflect the progress of development of individual students	• Produce an order of student ranking relative to a group
Design of assessment tasks	• Align with content and expected outcomes	• Discriminates high and low achievers
Unit of score interpretation	• Individual	• Group
Score presentation	• Grades linked to criteria	• Grades, derived from raw scores, usually presented in a bell curve and often coarse grained into letter grades
Advantages	• Able to reflect actual performance and progress of each student • Allows explicit incorporation of higher-order cognitive skills, so that assessment can better reflect these abilities	• Able to compare individuals within the group • Can be statistically adjusted to have a prescribed width (standard deviation) • Able to avoid grade inflation
Disadvantages	• No control on grade distribution, which can lead to grade inflation • Less able to compare individual performance against the cohort	• Group performance can influence individual grades, which can be unfair • May fail to show the range of actual differences amongst individual students

This present paper does not intend to review the literature on comparing and contrasting the two approaches in detail; rather, it aims to explore the operational compatibility of criterion referencing and norm referencing in higher education assessment.

Many assessment strategies are compatible with both criterion-referenced and norm-referenced grading methods, and the boundary between the two is fuzzy. Nightingale et al. (1996, 9) noted that it was difficult to distinguish some assessment strategies as ‘entirely norm-referenced or criteria-referenced’. This corresponded to the argument of Rowntree (1987) that criterion referencing and norm referencing share more commonalities than are generally perceived and any rigid adherence to either is problematic. In contrast, Frisbie (2005) argued that the different score interpretations led to the perception of a dichotomy that requires a binary choice between criterion referencing and norm referencing, diverting attention away from the possibility of their synergistic use in one single assessment.

Compatibility and complementarity

This paper rejects the fallacious ‘either-or’ view of criterion referencing and norm referencing, and argues that they are compatible and complementary. We examine this from three perspectives: dual reporting; a one-way transformation; and a feedback loop.

Dual reporting

There are many examples in which both criterion referencing and norm referencing are present with both dimensions reported in some form. For example, dual reporting (e.g. Johnny is able to do X, Y, Z (criteria) and ranks in the top 15% (norm information) of the class in this subject) is commonly used in many schools in Hong Kong, with the report card providing for each subject both a grade or numerical mark (e.g. 60 marks) as well as a position-in-class indicator (e.g. rank 35 out of 60) for the subject. In this highly competitive education system, the position-in-class indicator assumes high importance.

Dual reporting can also be found in the California Critical Thinking Skills Test (CCTST), a popular standardised critical-thinking test in higher education. CCTST aims to measure an individual’s critical-thinking capability in various skill areas, and its report provides an overall score and scale scores together with a norm-referenced percentile ranking. In addition, it offers descriptions of the strength of both scores (Facione and Facione 2009).

The SAT (2015), (formerly Scholastic Aptitude Test, Scholastic Assessment Test and the SAT Reasoning Test) is generally seen as norm referenced (scores are interpreted as percentiles), but one can argue that there are definite criteria associated with the assessment (Angoff 1974). The national College English Test (CET) in China provides ‘the dual function’ using criteria related to norms (Zheng and Cheng 2008, 416), with benchmarks adopted for the two levels (i.e. CET-4 and CET-6), but with scores also equated with predetermined norms (Yang and Jin 2001).

Often in dual reporting, the normative information emerges a posteriori and as a fact (e.g. a score of 650 happens to equate to the top decile), whereas, strictly speaking, norm referencing imposes grade distributions a priori and as a policy (e.g. no more than 10% of the candidates shall be awarded ‘A’s). The terminology

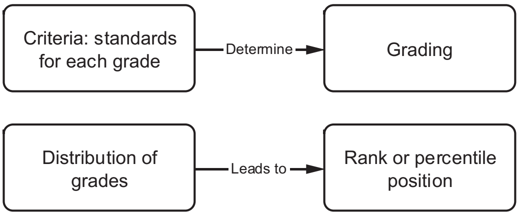
Criteria: standards for each grade
Determine
Grading
Distribution of grades
Leads to
Rank or percentile position

Figure 1. Dual reporting.

(e.g. deciles vs. letter grades) reflects this distinction. However, for a large and stable candidature, it no longer matters in practice which comes first. Imagine that one starts with a criterion-referenced approach, so that scores are tied to absolute performance. It is found, empirically, that over a number of years the median score in, say, the SAT is about 500 (National Center for Education Statistics 2013). Then, because the ability of the candidature is stable, in the next year, the logic can be reversed: one defines 500 to be the median, which is a norm-referenced prescription. Betebenner (2009) also suggested the combined use of norm referencing and criterion referencing as a way to profile student growth and development. This state of affairs can be represented by two free-floating boxes – free floating in the sense that they are not linked in design, only a posteriori (Figure 1).

One-way transformation from norm referencing to criterion referencing

As discussed, the criteria in ostensibly criterion-referenced schemes are often implicitly based on norms derived from a cohort or group, typically in the recent past. This can be clearly illustrated by an extreme (if somewhat facetious) example: an appropriate criterion for primary school mathematics might be ‘ability to add’, while that for mathematics majors in university might be ‘ability to solve differential equations’. This is so obvious (and the opposite so ludicrous) that one sometimes forgets that the only justification is empirical – that *most* primary school pupils can add but cannot solve differential equations, and *all* mathematics majors in university would have no trouble adding. About these seemingly trivial statements, two remarks can be added. First, the words ‘most’ and ‘all’ betray the norm referencing to a group. Second, these statements are empirical and not a priori: one can imagine that six-year-old Martians might be able to solve differential equations. In other words, to determine whether the criteria are appropriate requires that one look empirically at the ability and performance of the group. Once the need for such scrutiny is acknowledged, then one must also admit the possibility of a mismatch being discovered over time (of course not, in reality, as extreme as the facetious example might suggest), and the consequent need to deal with the mismatch. This is a key message of this paper, to which we shall return.

Not only is the *definition* of the criteria often norm referenced, but their *interpretation* must likewise be made in the context of a group. It has been argued that criteria can never be clear and explicit enough. The meanings of criteria can only be operationalised through the active engagement of students and teachers in co-constructing and interpreting their own understandings (O’Donovan, Price, and Rust 2004; Rust, O’Donovan, and Price 2005; Shay 2008). In this sense, criterion

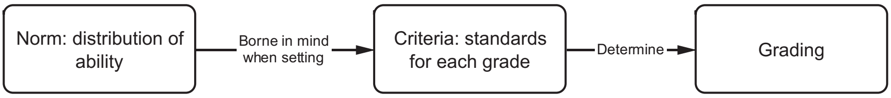
Norm: distribution of ability
Borne in mind when setting
Criteria: standards for each grade
Determine
Grading

Figure 2. One-way (norm-referenced→criterion-referenced) transformation.

referencing needs to be anchored to a specific context (Sadler 2005). If 'cogent argument' is adopted as a criterion for grading, the interpretation must be different when applied to a final-year thesis in a prestigious institution, and when applied to a freshman term-paper in an institution with a non-competitive intake. If this criterion is part of the rubric associated with an 'A' grade, surely it must mean 'cogent argument at a level that is unusual *for such a cohort*' – again betraying an implicit norm reference. Making the norm reference explicit would render it difficult to defend a large percentage of 'A's on the excuse that criterion referencing should not be bound by percentage distributions. This consideration is especially important in universities, where higher-order skills such as 'ability to synthesise a cogent argument' or 'critical thinking' are increasingly emphasised, and these are especially prone to wide interpretation, in contrast to skill-based criteria such as 'typing at 60 words per minute' or 'the ability to add'.

There are other reasons why criterion referencing may require the monitoring of norm-based distributions, for example, when consistency across multiple assessors in interpreting a set of criteria has to be maintained. In this regard, Orr (2007) stressed that assessment is co-constructed in communities of practice, implicitly based on past (normative) knowledge. In large public examinations in which scripts are randomly assigned to markers, even if there are descriptors attached to each grade, it is common to ensure that the distributions of scores are broadly similar across markers. Norm referencing, as a result, becomes a strategy for checking on decisions made in a criterion-referenced fashion.

Thus, behind the criterion-referenced rubric (that the students see) is an assumption about norms (that the teachers and assessors should be aware of). This state of affairs can be represented as in Figure 2.

Feedback loop

The third model, extending the one just described, involves three elements: norm referencing, criterion referencing and actual grading, all in a feedback loop, which can be iterated many times (Figure 3). First, assumptions about the ability or performance of the cohort (norm referencing) are used to derive a set of criteria, which are then incorporated into assessment rubrics (criterion referencing). Second, these are applied in assessment to arrive at a set of grades. The markers should simply follow the rubrics, i.e. act in a criterion-referenced fashion. Third, after the grading is done, the actual student performance is monitored. Theoretically, the distribution may be very different from that assumed at the beginning; but with experienced teachers, good moderation and/or other adjustments, after a few cycles discrepancies would be small and would appear only slowly over time (as student abilities change), and the criteria should thus be suitably revised for future assessments. For example, if a high percentage of students are scoring 'A's according to the rubric, this new empirical evidence should inform the normative assumptions, leading to a revised set of

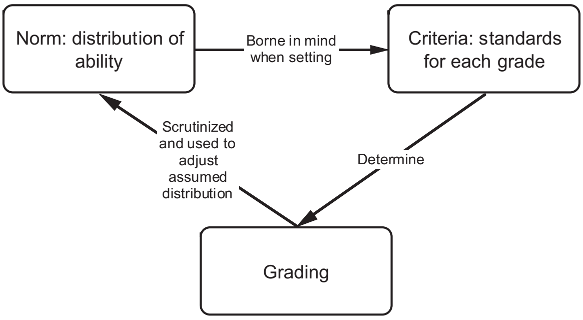
Norm: distribution of ability
Borne in mind when setting
Criteria: standards for each grade
Scrutinized and used to adjust assumed distribution
Determine
Grading

Figure 3. Feedback loop.

criteria, thus ensuring that the rubrics for the top grade remain sufficiently challenging and the educational experience sufficiently rewarding; the opposite adjustment should be made if the original criteria were discovered to be too demanding. One might say this is changing the goal posts, but that is not an appropriate metaphor, since the procedure is *not* to change the goal posts during the game, but only for the *next* game, and in a way that is announced before the next game starts. The three steps together constitute a feedback loop, which is consonant with the modern concept of fitness-for-purpose, or the quality loop sometimes abbreviated as ADRI (approach–deploy–review–improve) (Woodhouse 2003).

Thinking in terms of such a feedback loop serves to underscore a number of features:

- There is no need to choose between norm referencing and criterion referencing. They are both present.
- Not only are they both present, but with the caveat about minor adjustments from year to year, they are consistent. Thus, it is possible both to define rubrics (criterion referencing) *and* to prescribe grade-distribution guidelines (norm referencing), provided the latter contains a degree of flexibility.
- The presence of norm referencing and criterion referencing in a loop enables the generation of both useful feedback to learners and useful summative information to external stakeholders.
- The use of criteria allows meaningful reference to higher-order learning outcomes. While these are inevitably ambiguous and even unknown to external stakeholders, the simultaneous use of norm referencing allows the interpretation of these criteria to be supported by norm comparisons, and to guard against grade inflation.
- Since these steps are all in a loop, there is no need to argue which one comes first.
- The entire approach is coherent with modern quality-assurance and fitness-for-purpose concepts.

Taken together, these features constitute a theoretical basis for clarifying the implicit hybrid model cited at the beginning of this paper.

It may be added that this nuanced model is unnecessary in contexts that require only certification of skills – in which case results are typically scored as pass/fail, multiple attempts are usually allowed, and there is no worry about grade inflation or an overwhelming passing rate. The example of driving tests comes to mind, and also other skill-based assessments including those in the use of information technology, some of which are in fact likened to driving tests, for example, the International Computer Driving Licence.

Other issues

Norm referencing to which population

There is one further problem: in norm referencing, what is the reference population for deriving the norm? The obvious answer seems to be the actual class being examined but, very often, reference is made (implicitly) to a larger population. If several instructors teach and examine different sections of the same course, there is usually some procedure to homogenise or moderate the assessment. Then it is legitimate to have a very high percentage of 'A's in one section if enrolment patterns for the better students are for some reason skewed. In schools, the grading is often referenced to the estimated distribution in the analogous public examination, which is useful feedback in preparing for the latter. In higher education, external examiners (a practice now diluted in many nations) allow reference to a group of peer institutions, or the students therein. The ideal is that a first-class honours from any university means the same thing – a statement that few would now take seriously. Thus, the basis of comparison in norm referencing can be quite complex. In short, there is often an implicit norm reference to a larger (external) population, whether this is channelled formally through an external examiner or, more often, informally by the teacher benchmarking against standards elsewhere (often relying on her/his own experience in graduate school, as a tutor or teaching assistant).

Anchoring norm referencing to a larger population enhances the reliability of an assessment; however, the intended learning outcomes central to an outcomes-based curriculum can only be assessed with any validity by a criterion-referenced approach. The feedback loop offers flexibility for educators to make iterative adjustments so that claims about both reliability and validity are possible.

Assessments based on progression

It should be mentioned, if only in passing, that some assessments fall into neither a criterion-referenced nor a norm-referenced framework, but focus more on progress, and in this sense are largely formative in nature.

Ipsative referencing (Hughes 2011) is a comparison of an individual's current performance with a previous performance. Such referencing tracks the progress of individual learners over time, rather than in comparison with other learners or against preset criteria. For example, the pre-course and post-course assessments of a module measure growth and added value. However, such feedback may not be transferable beyond the module, as the learning outcomes may vary.

Some assessments focus less on outcomes and more on process; in these cases, diplomas or certificates only attest that the student (or participant) has gone through a process, with focus on participation and course completion. One example relates to physical education classes that are required in many universities, for example, in North America. Because the range of physical ability amongst students is very broad, a passing grade can usually be achieved by just showing up and trying conscientiously. In such cases, ipsative referencing can be introduced to measure the learning gain of individuals throughout the process for different purposes. In continuing professional development, very often only attendance is required.

Conclusion

There are two main principles of assessment: (1) to grade students with differing levels of knowledge and skills; and (2) to reflect the growth of individual knowledge and skills. Teachers have to balance the assessment focus between these principles and develop an appropriate assessment strategy. Considering the competitiveness of the current job market for graduates, an assessment profile which can reflect not only absolute competence and learning development, but also the relative position within a cohort is useful. In addition, students' confidence about the validity of assessment measures needs to be maintained. In terms of norm referencing and criterion referencing, there are theoretically four possibilities:

(1) *Choose only one of norm referencing or criterion referencing*: we have argued that this is a false dichotomy and a trap.
(2) *Use both, but without relationship*: a dual reporting as shown in Figure 1, which is possible and rather common in schools.
(3) *Use both, in a synergistic way*: as suggested in Figure 3 and the discussion around it.
(4) *Use neither*: which is the worst of all possible worlds. In other words, there is no clear articulation of rubrics (thus, not criterion referenced) and no reasonably robust percentage grade guidelines (thus, not norm referenced). The teacher just grades somewhat crudely 'by experience'; and the lax practice is sometimes defended in the name of teacher autonomy or even academic freedom. A close approximation to this disaster is to abandon norm referencing for a very soft criterion-referenced approach, with criteria (a) so vague that they can mean anything, and (b) not reviewed so that the majority can satisfy the rubrics for the top grade. This is how grade inflation happens; and the danger still lurks in many universities.

In summary, the debate on norm-referenced and criterion-referenced approaches to assessment should not be seen as a binary choice fraught with tension; rather, the two are compatible and complementary. Through the perspective of a feedback loop, compatibility is ensured, in a dynamic sense, and the complementary use of criterion referencing and norm referencing is embedded in the design. These concepts and devices for internal consistency should also be supplemented by external anchors: to larger peer groups in setting normative expectations, and to desired outcomes in setting the assessment criteria. The model of a complementary criterion-referenced and norm-referenced loop provides a useful conceptual framework for current discussions of learning assessments in higher education, for which a possible implementation plan would contain the four essential domains: institution-wide normative guidelines, subject-specific criteria, a grading and feedback loop, and quality-assurance checks.

Normative guidelines

First, institution-wide grade-distribution guidelines need to be adopted and publicised for the information of external stakeholders. There must be mechanisms to ensure that these guidelines are adhered to and rare exceptions are documented. Such practice should be uniform across an institution (ideally even over an entire sector), at least for each level of learning, and need make no reference to subject content.

Subject-specific criteria

Second, for each subject (either an undergraduate degree or a component module, e.g. one course over one term), the intended learning outcomes need to be stated as part of the curriculum design, and these should be further articulated into grading rubrics or criteria. The criteria should span different levels of sophistication, including some that (in a formative sense) stretch the more able students, or (in a summative sense) are likely to be met only by a minority who are then to be awarded the top grades. The range of sophistication promotes depth of learning. The formulation of these criteria would, of course, be subject-specific, and would rely on assumptions about student ability presumably drawn from past cohorts and/or lateral benchmarking to an external reference.

Grading and feedback loop

Third, students should theoretically take note only of the criteria and the rubrics, as targets for their learning. Likewise, graders should only take note of the criteria and the rubrics in assessing student attainment. Moderation comes at this stage in checking that there is consistency (both internally across sections and also externally in comparison to selected peer groups by an external examiner, for example) in interpreting the criteria. Most importantly, where the adoption of a set of criteria consistently results in conflicts with the institution-wide normative grade-distribution guidelines, there has to be a discussion on possible revision of the criteria or of their interpretation in the next offering. Such conflicts, if managed in time, should never become serious.

Quality-assurance process

Educational processes are not well served by either hard rules ('accountability') or the total absence of rules ('autonomy'). A healthy balance can only be struck if there is a quality-assurance process which checks that broad normative guidelines exist and are well understood, that there is broad (but not necessarily literal) adherence to these guidelines, that exceptions are documented and, most importantly, that conflicts are dealt with in the feedback loop. The quality-assurance process is needed at multiple levels: within a department monitoring its different courses, within the institution to monitor the different teaching units and even across the sector by a quality agency advising institutions under its purview – in all cases, not just ticking check boxes but helping the units to develop robust processes.

It must be said that criterion referencing and norm referencing can be brought into consistency much more simply, by mandating shifts in the grade boundaries whenever grade distributions fall outside guidelines – a practice that is not uncommon. But such a mechanical adjustment would engender a sense of passive compliance with apparently bureaucratic rules, and will have no effect on learning and teaching in subsequent years. On the other hand, the feedback loop provides the opportunity for alignment, leading to changes not only in assessment, but through the revision of intended learning outcomes, also in learning and teaching. Only when all components of the curriculum are considered together can there be the alignment that lies at the heart of outcomes-based approaches.

Acknowledgements

The use of a synergistic loop to align criterion-referenced and norm-referenced assessment arose out of the development of an assessment policy at The Chinese University of Hong Kong in 2010, in which the second and third authors were involved. Some of the ideas in a rudimentary form are contained in internal documents, for example, *Assessment of Student Learning in Taught Programmes,* adopted in 2010 and found in http://www.cuhk.edu.hk/policy/assessment/AssessmentofStudentLearning.pdf; Appendix 2 therein contains the loop corresponding to Figure 3 in this paper. This assessment policy is also referenced in the University's overall *Quality Manual*, http://www.cuhk.edu.hk/clear/qm/index.html, adopted in 2014. We thank colleagues who engaged in debate and contributed to the development of the ideas.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Beatrice Lok is a visiting lecturer of the Centre for the Advancement of Teaching and Learning (CALT) at University College London. She was a postdoctoral fellow of the Centre for Learning Enhancement And Research at The Chinese University of Hong Kong. She is interested in the areas of quality enhancement in higher education, second-language learning motivation and second-language development of non-native speaking students.

Carmel McNaught is Emeritus Professor of Learning Enhancement and former director of the Centre for Learning Enhancement And Research (CLEAR) at The Chinese University of Hong Kong. Since the early 1970s, Carmel has worked in higher education in Australasia, southern Africa and the UK in the fields of chemistry, science education, second-language learning, eLearning, and higher education curriculum and policy matters. She is actively involved in several professional organizations and is a fellow of the Association for the Advancement of Computers in Education; is a university quality-assurance consultant and auditor in Australia and Hong Kong; has served on the editorial boards of 18 international journals; and is a prolific author with well over 300 academic publications. Her recent publications and activities can be viewed at http://www.cuhk.edu.hk/clear/people/Carmel.html. She is currently a consultant, working mostly in Australia, Hong Kong, New Zealand, Singapore, the UAE and the UK.

Kenneth Young is master of CW Chu College and also professor of Physics at The Chinese University of Hong Kong. For a number of years until 2011, he was Pro-Vice-Chancellor of the University with responsibility for education. Currently, he is a member of the Quality Assurance Council under the Hong Kong University Grants Committee. He also chairs the Curriculum Development Council in Hong Kong, and is a member of the Education Commission. He is a fellow of the American Physical Society.

References

Anderson, L. W., and D. R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Boston, MA: Allyn & Bacon.

Angoff, W. H. 1974. *Criterion-referencing, Norm-referencing and the SAT.* Research Memorandum. Princeton, NJ: Educational Testing Service.

Betebenner, D. W. 2009. "Norm- and Criterion-referenced Student Growth." *Educational Measurement: Issues and Practice* 28 (4): 42–51.

Biggs, J. 1999. "What the Student Does: Teaching for Enhanced Learning." *Higher Education Research & Development* 18 (1): 57–75.

Biggs, J. B., and K. F. Collis. 1982. *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. New York: Academic Press.

Biggs, J., and C. Tang. 2011. *Teaching for Quality Learning at University.* 4th ed. Maidenhead: Open University Press.

Bloom, B. S., ed. 1956. *Taxonomy of Educational Objectives: The Classification of Educational Goals: Handbook I, Cognitive Domain*. New York: Longman.

Bloom, B. 1968. "Learning for Mastery." *Evaluation Comment* 1 (2): 1–5.

Bond, L. A. 1996. "Norm- and Criterion-referenced Testing." *Practical Assessment, Research & Evaluation* 5 (2). http://PAREonline.net/getvn.asp?v=5&n=2.

Carroll, J. 1963. "A Model of School Learning." *Teachers College Record* 64: 723–733.

Ewell, P. T. 2009. "Assessment, Accountability, and Improvement: Revisiting the Tension." NILOA Occasional Paper No. 1. Urbana: University of Illinois and Indiana University, National Institute for Learning Outcomes Assessment.

Facione, N. C., and P. A. Facione. 2009. *California Critical Thinking Skills Test: Test Manual*. Millbrae: California Academic Press.

Frisbie, D. A. 2005. "Measurement 101: Some Fundamentals Revisited." *Educational Measurement: Issues and Practice* 24 (3): 21–28.

Gipps, C. 1994. *Beyond Testing: Towards a Theory of Educational Assessment*. London: Falmer Press.

Glaser, R. 1963. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." *American Psychologist* 18 (8): 519–521.

Glaser, R., and D. J. Klaus. 1962. "Proficiency Measurement: Assessing Human Performance." In *Psychological Principles in System Development,* edited by R. M. Gagne, 419–474. New York, NY: Holt, Rinehart & Winston.

Hughes, G. 2011. "Towards a Personal Best: A Case for Introducing Ipsative Assessment in Higher Education." *Studies in Higher Education* 36 (3): 353–367.

Huitt, W. 1996. "Measurement and Evaluation: Criterion- versus Norm-referenced Testing." *Educational Psychology Interactive*, Valdosta, GA: Valdosta State University. http://www.edpsycinteractive.org/topics/measeval/crnmref.html.

Krathwohl, D. R. 2002. "A Revision of Bloom's Taxonomy: An Overview." *Theory into Practice* 41 (4): 212–218.

McNaught, C., K. F. Cheng, and P. Lam. 2006. "Developing Evidence-based Criteria for the Design and Use of Online Forums in Higher Education in Hong Kong." In *User-centered Design of Online Learning Communities*, edited by N. Lambropoulos and P. Zaphiris, 161–184. Hershey, PA: Information Science Publishing.

National Center for Education Statistics. 2013. *Digest of Education Statistics, 2012* (NCES 2014–015), Table 171. U.S. Department of Education. http://nces.ed.gov/programs/digest/d12/tables/dt12_171.asp.

Neil, D. T., D. A. Wadley, and S. R. Phinn. 1999. "A Generic Framework for Criterion-referenced Assessment of Undergraduate Essays." *Journal of Geography in Higher Education* 23: 303–325.

Nightingale, P., I. T. Te Wiata, S. Toohey, G. Ryan, C. Hughes, and D. Magin. 1996. *Assessing Learning in Universities*. Sydney: Professional Development Centre, University of New South Wales.

O'Donovan, B., M. Price, and C. Rust. 2004. "Know What I Mean? Enhancing Student Understanding of Assessment Standards and Criteria." *Teaching in Higher Education* 9: 325–335.

Orr, S. 2007. "Assessment Moderation: Constructing the Marks and Constructing the Students." *Assessment & Evaluation in Higher Education* 32 (6): 645–656.

Rojstaczer, S., and C. Healy. 2012. "Where A is Ordinary: The Evolution of American College and University Grading, 1940–2009." *Teachers College Record* 114 (7): 1–23.

Rowntree, D. 1987. *Assessing Students: How Shall We Know Them?* Revised ed. London: Kogan Page.

Rust, C., B. O'Donovan, and M. Price. 2005. "A Social Constructivist Assessment Process Model: How the Research Literature Shows Us This Could Be Best Practice." *Assessment and Evaluation in Higher Education* 30 (3): 231–240.

Sadler, D. R. 2005. "Interpretations of Criteria-based Assessment and Grading in Higher Education." *Assessment & Evaluation in Higher Education* 30 (2): 175–194.

SAT. 2015. *Understanding Your Scores*. Accessed January 20. http://sat.collegeboard.org/scores/understanding-sat-scores

Shay, S. 2008. "Beyond Social Constructivist Perspectives on Assessment: The Centering of Knowledge." *Teaching in Higher Education* 13 (5): 595–605.

Spady, W. 1994. *Outcome-based Education: Critical Issues and Answers*. Arlington, VA: American Association of School Administrators.

Tyler, R. W. 1949. *Basic Principles of Curriculum and Instruction*. Chicago: The University of Chicago Press.

Wiliam, D. 1996. "Standards in Examinations: A Matter of Trust?" *Curriculum Journal* 7 (3): 293–306.

Woodhouse, D. 2003. "Quality Improvement through Quality Audit." *Quality in Higher Education* 9 (2): 133–139.

Yang, H., and Y. Jin. 2001. "Score Interpretation of CET." *Foreign Language World* 81: 62–68.

Zheng, Y., and L. Cheng. 2008. "Test Review: College English Test in China." *Language Testing* 25: 408–417.