Higher Education Academy

TRANSFORMING TEACHING
INSPIRING LEARNING

Assessment and feedback in higher education

A review of literature for the Higher Education Academy

Brad Jackel, Jacob Pearce, Ali Radloff and Daniel Edwards

In partnership with: Australian Council for Educational Research

ACER

Contents

Acknowledgements

The report authors acknowledge the support and assistance of the Higher Education Academy in the development of this document. Particular thanks go to Sarah Cutforth, Abbi Flint, Samuel Elkington and Sam Burgum. We would also like to thank our ACER colleagues who supported us in this work, particularly Juliette Mendelovits and Patricia Freeman.

Executive summary

This literature review has been undertaken for the Higher Education Academy (HEA). It explores recent scholarly contributions in the area of assessment and feedback in higher education (HE).

As outlined in the HEA's terms of reference for this work, the contents of this review are designed to "help practitioners, policy makers and researchers to focus more effectively on relevant questions, issues or sources of evidence to inform their own research or practice." The review is intended to provide an insight into contemporary scholarly work in this area, and the approaches to searching the literature and compiling this document reflect this aim.

The review explores two key questions, and is structured around this inquiry.

The first question examined is:

> What key elements are highlighted in the literature as fundamental to the development of assessment in higher education that is high quality and improves student learning?

In exploring this issue, the review identifies a number of principles and a number of processes considered to be fundamental to quality assessment that achieves these aims. These fundamentals have been shaped by the scholarly discussion and evidence provided in the current literature and by the work of the HEA itself, specifically through the 'tenets' of the HEA's (2016a) Transforming Assessment Framework.

The fundamentals discussed included:

Assessment for learning, which is almost universally accepted in the literature as an approach to learning in higher education. Consolidating this practice is now a key area for higher education practitioners to direct focus.

Aligned and fit-for-purpose assessment: the literature shows that in a constructively aligned system, the intended learning outcomes are explained to students, well-defined and transparent performance standards exist, and assessment tasks are relevant and measure students' progress against the learning outcomes using the performance standards (Biggs 1996, 2003). Students in courses that are more constructively aligned are "more likely to adopt deep learning approaches and less likely to use surface learning approaches" in their studies (Wang *et al.* 2013).

Collaborative construction of standards, whereby assessment standards are developed in a collaborative fashion involving staff, students and other stakeholders: "given that assessment standards are socially constructed, staff and students need to engage in dialogue about standards to understand what is required from, and entailed in, the assessment process" (HEA 2016a, p. 3).

Integrating assessment literacy with learning is another aspect strongly advocated through the HEA's framework and again builds on the engagement between teachers and students in assessment, ensuring that assessment discussions are integrated into the discourse of academic staff in their communication with students.

Defensibility of professional judgements: as highlighted by Wyatt-Smith and Klenowski (2013), while standards play an important role in the consistency, transparency and fairness of assessment, standards and criteria are not in themselves sufficient to improve assessment processes: "Assessment is dependent upon professional judgement; confidence in which depends on forums for developing and sharing standards within and between academic, disciplinary and professional communities" (HEA 2016a, p. 3). While academic judgments are valuable and integral to the operation of higher education, they should be continuously reviewed through critical reflection in a collaborative setting to ensure that they are reliable, consistent, transparent, fair and defensible (Bloxham, den-Outer, Hudson and Price 2016).

The limits of assessment: "Limits to the extent that standards can be articulated explicitly must be recognised since ever more detailed specificity and striving for reliability, all too frequently, diminish the learning experience and threaten its validity"(HEA 2012, p. 20). This 'fundamental' also highlights that learning is essentially a non-linear process and as such there are obvious limitations to the extent to which assessment can and should be used.

The second question explored in the review is:

> How are the fundamentals of assessment manifested in recent literature on assessment in higher education?

Feedback/feed-forward is an area of significant scholarly interest over the past five years. The literature shows a robust and defensible model for what good feedback should achieve and practitioners as taking defensible measures to reach those goals. However, efforts to empirically demonstrate the impact of these theoretical models need to be more robust than the current norm.

Peer-assessment and self-assessment: the literature on peer and self-assessment is extensive. Some ambiguity around the precise definitions of these terms leads to ambiguity around exactly what constitutes successful practice and how to measure that success. Peer and self-assessment are used for both formative and summative assessment; however, while its formative use appears to hold the most promise the majority of empirical evidence presented in the research is dependent on calibration of summative grades.

Summative assessment has been highlighted in this section of the review for the opposite reason; that is, the scarcity of literature on summative assessment in comparison to formative assessment is problematic. It is important to note this because summative assessment remains necessary for certification and is therefore highly influential in shaping the behaviour of students' choices and priorities regarding their learning (Yorke 2011). Summative assessment continues to be a significant part of assessment practice in higher education and its importance should be reflected in the literature.

Technology and assessment: the use of technology within assessment is a common subject in the literature. The wide reach and diversity of technology in assessment practice is highlighted, as is the way in which this space is continuously evolving. A key argument posed for use of technology is the way in which it can be a conduit for 'authentic assessment' (Craft and Ainscough 2015). However, there are words of caution on this front elsewhere (Jackel 2014).

Psychometrics are identified as playing an increasing role in assessment in higher education in the recent literature. Articles highlighted the powerful use of such tools when applied in the right context (Hardy and Aruguete 2014), but also the need for perspective when utilising the tools in the assessment of folios, or presentations (Aryadoust 2015). The work in this area is growing and psychometrics are increasingly used for validating forms of assessment that students are undertaking and as a tool in researching the use of or development of assessments.

Other areas, briefly discussed as relevant to the contemporary literature in this area included continuous assessment, inquiry-based learning, 'open' versus 'closed' book examinations, 'meta-reflection', institutional culture towards assessment, differentiated assessment, issues with combining scores and empirical errors in evaluative assessment.

A third area of the review offers some reflections from the researchers. Key gaps identified in the literature include exploring the *summative/formative dichotomy* in approaches to discussion of assessment and the '*aggregation issue*'. Emerging themes in relation to *programmatic* assessment and *assessment of learning outcomes* are also examined in this reflection section.

The literature review concludes with a summary of findings from three perspectives.

For *institutional policy*, the literature review concludes that there should be leadership in setting principles for improving the assessment literacy of both students and academics; for embedding the knowledge and application of the ‘fundamentals’ listed above; and for facilitating the development of more programmatic approaches to design and development of assessment.

For *educational practice* the review findings suggest that the areas of feedback/feed-forward and peer/self-assessment are theoretically well based, but continue to lack a strong empirical basis for best practice. Continuing the themes highlighted in earlier work by the HEA (2012), further development of evaluative designs to inform innovation appears to be an area for future focus. Collaboration between discipline experts and assessment experts could be one way of promoting this and improving assessment literacy at the same time.

The review also revealed significant emphasis on ‘formative’ assessment, and a notable lack of discussion on ‘summative’ assessment. While the distinction between assessment *of* learning as opposed to *for* learning is conceptually useful and defensible, it can obscure the fact that even the most summative of summative assessment has an enormous impact on where students invest their time. Rather than continuing to emphasise and reiterate the formative/summative dichotomy, it would perhaps be useful to start from the observation that all assessment is formative in some sense, while only some assessment is both formative and summative.

For the *higher education sector*, further leadership relating to the areas highlighted here such as scholarship, collaboration and literacy of assessment would be beneficial. The current policy climate in the UK, with the implementation of the Teaching Excellence Framework (TEF), suggests the beginning of a new era of assessment and quality assurance. In such contexts, it is important that assessment is seen as important for improving learning, is fit-for-purpose, is defensible and recognises its limits.

Introduction

This literature review has been undertaken for the Higher Education Academy (HEA). It explores recent scholarly contributions in the area of assessment and feedback in higher education (HE).

The report is structured as follows. This introduction provides some background and a context for the approach taken to developing this review and the 'lens' through which the review was undertaken. The review then comprises three broad sections:

- 'The fundamentals' – a section that explores the significant principles underpinning best practice in assessment. Drawing on the HEA's Transforming Assessment Framework (2016a), this section expands the reach of these ideas based on areas identified through the literature. Essentially, this section explores the following research question: 'What key elements are highlighted in the literature as fundamental to the development of assessment in higher education that is high quality and improves student learning?'
- 'Practice and innovations' – an exploration of recent literature that demonstrates the kinds of assessment practice currently being undertaken in higher education and links this application to the concepts discussed in the previous section of the review. This section is guided by the research question: 'How are the fundamentals of assessment manifested in recent literature on assessment in higher education?'
  In addressing this question, a number of key areas of literature are explored. The most significant in terms of recent attention in the literature are feedback and peer-review/self-review, and these are described in detail. Topics such as technology, summative assessment, and the use of psychometrics are also explored. The section finishes with a brief outline of other emerging trends in practice identified in the literature;
- 'Reflections and setting directions' – this section summarises the gaps in current literature identified by the authors, particularly in relation to gaps previously identified in work of the HEA through the *A Marked Improvement* (HEA 2012) report. The discussion begins with the identified gaps, and then includes reflections on trends in the literature relating to emerging principles and future directions.

The report concludes with some suggestions for relevant stakeholders based on the findings of the review.

Background and approach

The core stakeholders of the HEA are the intended audience of this review; practitioners and researchers with interest in the scholarship of assessment in higher education. As outlined in the HEA's terms of reference for this work, the contents of this review are designed to "help practitioners, policy makers and researchers to focus more effectively on relevant questions, issues or sources of evidence to inform their own research or practice." The review is intended to provide insights into contemporary scholarly work in this area, and the approaches to researching and compiling this document reflect this aim.

Based on guidance from and parameters set by the HEA in scoping this project, the research team undertook a broad scan of the international literature covering the area of assessment in higher education. The section below outlines the approach taken in researching, collating and developing this document.

The timing of the previous work by the HEA helped to define some parameters for this work. Given the publication of the HEA's *A Marked Improvement* in 2012, this review focuses on work published following the development of that report (i.e. 2011 onwards), although the researchers have included notable earlier works where helpful.

A systematic process was followed in searching within and across key databases of bibliographic information. The referencing/bibliography tool Zotero was used throughout to store searched materials and to search within identified documents. Initial search terms were kept purposefully broad (e.g. assessment, higher education, universities, etc.) in order to capture the most comprehensive range of work possible and to avoid missing new areas of assessment research.

In addition, the research team undertook focused journal-specific searching. This process involved sourcing every article from the journal *Assessment and Evaluation in Higher Education* going back to January 2011, and examining in detail the following journals for relevant articles on assessment published since January 2011:

- *Studies in Higher Education;*
- *Innovations in Education and Teaching International;*
- *Quality Assurance in Education;*
- *Issues in Educational Research;*
- *International Journal of Higher Education;*
- *Higher Education Research and Development;*
- *Education and Training;*

- *Teaching in Higher Education;*
- *Perspectives: Policy and Practice in Higher Education;*
- Higher Education;
- Quality in Higher Education;
- Research in Higher Education;
- Journal of Higher Education Policy and Management.

Google Scholar was then used to find articles concerned with higher education and assessment that had been published in 2011 or more recently. As Google Scholar ranks search results by citation count, this enabled us to identify with ease those high-impact articles published within the specified date range that had not already been sourced. As a final step the team obtained a number of recently published books on assessment in higher education that had been cited in many articles sourced from the initial searches.

The process resulted in sourcing approximately 850 papers, books, and chapters of potential relevance to the literature review. As noted above, these were collated and stored during the search process in a shared online Zotero library. Once collated, the next step involved reading all the abstracts and removing articles that, despite keywords and titles, were not relevant to the research questions being explored. Approximately 200 articles were removed from consideration following this initial scanning process.

During this process, the researchers also undertook a process of 'tagging' articles and chapters with keywords relevant to the literature review. The collation platform being used by the team for storing the searched articles facilitated this process well.

Manual tagging in this way allowed the research team to develop methodically an overview of research priorities emerging within the field of literature and to quickly apply filters to the library to identify all sourced articles that were relevant to a particular topic within assessment in higher education. For example, we could filter to examine 'feedback' and then filter again by other terms such as 'UK-centric'. Once complete this allowed the team to efficiently remove articles that, while not entirely irrelevant (they were important to know about), were tangential to the focus of this literature review. Custom keywords, developed by the researchers, also offered a tool for dealing with regional differences in jargon. For example, 'assessment' in a US context is nearly always concerned with assessment *of* institutions, and less so assessment *within* institutions.

This approach helped identify trends in the recent literature, key areas of focus, and apparent gaps in the literature. The searching and tagging/categorisation closely informed the way in which the review has been structured.

At a number of points during the development of this work, the research team consulted with HEA researchers. This included consulting the HEA Assessment Lead to further define processes, parameters and approaches, and ensure consistency with the expectations of HEA and relevant stakeholders.

A number of key decisions relating to the content and focus of the review were made at various points during the development of this work. Following the guidance of HEA, and keeping in mind the intended audience, the researchers concentrated on assessment practice and implementation that is practitioner-focused and emphasises aspects of learning that can be influenced/enacted at the coal-face of a university — that is, with and for students. Assessment trends and practice relating to higher-level aspects such as professional accreditation, government quality assurance processes, and 'big picture' learning outcomes assessment programmes have been largely (although not entirely) left out of the scope of this particular work. In addition, the abundant literature exploring student evaluations of teaching and learning (and sometimes by implication, assessment) has not been covered extensively here.

The fundamentals

This section summarises the elements highlighted in the research literature as essential to the conceptualisation, development and implementation of high quality assessment in higher education. It is framed by the research question:

> What key elements are highlighted in the literature as fundamental to the development of assessment in higher education that is high quality and improves student learning?

The key elements discussed here are seen as independent of both content and context. In other words, these elements are important to the development of high quality assessment regardless of the assessment content and where the assessment is taking place.

The HEA has developed a framework for transforming assessment in higher education (HEA 2016a). This framework outlines the tenets the HEA has identified that inform and underpin the transformation of assessment practice. These tenets were described in detail in *A Marked Improvement* (HEA 2012) and originated in the 2007 and 2009 Assessment Standards Knowledge exchange (Oxford Brookes University 2014). The HEA Framework also describes the processes by which assessment can be transformed in higher education. The framework has helped to inform the discussion of fundamental principles in this section. In addition to the framework, the discussion has been complemented and augmented by trends, evidence and emphases in the contemporary literature relating to assessment in higher education.

This section explores the following 'fundamental principles':

- assessment for learning;
- aligned and fit-for-purpose assessment;
- collaborative construction of standards;
- integrating assessment literacy with learning;
- defensibility of professional judgements;
- the limits of assessment.

Fundamental principles

Assessment for learning

The HEA framework asserts that assessment is an activity that should be used to inform and improve student learning: "Learning and assessment should be integrated and fully aligned" (HEA 2016a, p. 19). Rather than being used solely to ascertain whether learning outcomes have been achieved, assessment should provide feedback on performance in a way that contributes to students' learning (Sadler 1998).

Assessment *for* learning – often referred to as formative assessment – occurs when assessment and learning are integrated. While assessment *of* learning – or summative assessment – remains an important part of evaluating student learning at the completion of a programme of study, assessment for learning can be used to provide rich, meaningful and timely feedback to students on their learning and progress throughout a programme of study. Assessment for learning can take many forms, and may be either formal or informal (Yorke 2003).

Assessment for learning can help students understand the expected standards in a programme of study (Yorke 2003). It can also help motivate students to study and allows them to reflect on their learning (Weurlander *at al.* 2012). Consistent and timely feedback provided on formative assessment can enable students to learn more efficiently and effectively (see e.g. Richardson and Healy 2013). Detailed data from formative assessment can also be used by teachers, tutors and other practitioners to identify gaps in students' learning and help them target teaching (Bennett 2011).

Over the past 20 years, the principle of assessment for learning has been increasingly embraced globally. As Rhodes (2016) explains:

> Assessment is more efficient and effective for institutions because it now can provide students with feedback on defined strengths and weaknesses in formative as well as summative detail not available previously. In this more robust feedback rubric environment, faculty are discovering the power of working and talking with colleagues about student learning improvement and how assessment results can help them improve assignments and their own sense of efficacy.
> *(Rhodes 2016, p. 112)*

While assessment of learning and assessment for learning are often viewed as distinct types of assessment, in reality the differences are often more subtle (Bennett 2011; Yorke 2003). Some assessment may be both summative and formative. For example, an end of module assignment or examination will have a strong formative component in driving what and how students learn during the module (Bennett 2011). Students may also receive feedback on summative assessment that they can feed-forward into future learning. Ideally, summative assessment should measure students' knowledge and understanding, but should also support and feed-forward into future learning.

Closely related to the notion of assessment for learning is the idea of 'educational rationale'. This argues that each piece of assessment in a programme should have a clear educational value and an articulated rationale. This means that assessment should be a primary consideration in course design and should be focused on student learning (Carless 2007, 2015a, 2015b; Clements and Cord 2013). The design of assessments is therefore a crucial component of curriculum design.

In discussing directions for change in assessment practice, Medland (2016) notes that:

> There is now an extensive body of literature that can inform stakeholder understanding of how a shift in culture might be brought about in higher education, from the current dominant discourse of the testing culture, towards an assessment for/as learning culture … It calls for assessment to be a central aspect of curriculum design and development. *(Medland 2016, p. 91)*

One fundamental aspect of assessment that is detailed in the HEA framework (2016a) is the process of developing assessment infrastructure. This process supports "change, which includes institutional assessment regulations, and the use of technologies to enhance assessment practice, improve feedback and streamline assessment management (e.g. e-submission, e-feedback)" (HEA 2016a, p. 3). This is closely linked to all aspects of assessment quality. Enhancing the processes and technology used to administer assessment and provide feedback supports assessment for learning.

Aligned and fit-for-purpose assessment

Building on the notion of assessment *for* learning are the inter-related concepts of constructive alignment and fit-for-purpose assessment. Constructive alignment takes place when learning outcomes, curriculum, teaching and learning strategies and assessment tasks and methods are all aligned (Biggs and Tang 2011; Boud and Falchikov 2006). Assessment that is fit-for-purpose occurs when assessment methods and tasks are valid, relevant and authentic and focus on "the demonstrable achievement of intended programme outcomes" (HEA 2016a, p. 3). Essentially, constructive alignment takes place at the programme level, while fitness-for-purpose usually focuses on individual assessment tasks. Each of these concepts is discussed below in the context of recent literature.

Tam (2014) elaborates on the concept of constructive alignment by arguing that:

> outcomes, teaching and learning, and assessment need to be aligned to achieve consistency and coherence in the design process, resulting in instruction and assessment that are designed to address the intended learning outcomes. *(Tam 2014, p. 166)*

In a constructively aligned system, the intended learning outcomes are explained to students, well-defined and transparent performance standards exist, and assessment tasks are relevant and measure students' progress against learning outcomes using the outlined performance standards (Biggs 1996, 2003). This alignment assures that content in the curriculum is linked to the stated learning outcomes of a programme, and is mapped to assessment in a fair way.

Although constructive alignment is broadly accepted as having a positive impact on the quality of assessment and student learning outcomes (Biggs 2003), there has been, to date, limited research critically evaluating the idea of constructive alignment. One recent paper found that students in courses that are more constructively aligned are "more likely to adopt deep learning approaches and less likely to use surface learning approaches" in their studies (Wang *et al*. 2013).

Curriculum review and development is one of the fundamental processes for assessment outlined in the HEA framework (2016a). This process is an essential part of constructive alignment. It involves reviewing curricula and developing programmes of assessment that are fully aligned with the intended learning outcomes. The HEA framework posits, "that enhanced assessment practice can be effectively integrated into institutional processes" (HEA 2016a, p. 3).

Assessment frameworks are another way by which constructive alignment can be achieved. Assessment frameworks are distinct from curriculum frameworks. While curriculum frameworks often articulate the desired learning outcomes of a programme of study, an assessment framework provides more detail of *what* the desired learning outcomes are and *how* they are to be assessed (Jago 2009):

> Assessment frameworks provide a structured conceptual map of the learning outcomes of a programme of study. Where curriculum frameworks detail what is to be taught, assessment frameworks detail what is to be assessed … Further, an assessment framework details how an assessment is to be operationalised. It combines theory and practice, and explains both the 'what' and the 'how'. (*Pearce et al. 2015, p. 111)*

Assessment frameworks allow the process of constructive alignment between curriculum and assessment to be thoroughly mapped, and for the rationale of the assessment tasks to be articulated.

Fitness-for-purpose relates to individual assessment tasks that exist within an overall assessment framework. A fit-for-purpose assessment is one that provides a "valid assessment of the achievement of intended programme outcomes" (HEA 2012, p. 19). An assessment is fit-for-purpose when the methods used to assess a particular learning outcome are relevant, appropriate and focused on measuring the achievement of intended learning outcomes (van der Vleuten *et al*. 2012). Developing a fit-for-purpose assessment requires consideration of the content, context, level, and student cohort (Brown and Race 2012).

The notion of 'fitness-for-purpose' is inherently linked to assessment quality and could be considered a functional definition of assessment quality (Schuwirthand Van der Vleuten 2011). Richardson (2015) notes that "Ideally, the form of assessment on a particular module should be determined by the design of the module and in particular by the module's intended learning outcomes" (p. 440). That is, assessment is fit-for-purpose when it is likely to assess what it was intended to assess, as defined in a curriculum or assessment framework. This is analogous to older conceptions of validity, that is: Does the instrument measure what it was intended to measure?

It is a widely held view that assessment should always be fit-for-purpose (Brown and Race 2012). It is difficult to disagree with such a claim, since if the purpose of an assessment can be articulated, then the assessment should conform to this purpose. Essentially, the greatest consideration when designing an assessment should be given to whether the selected assessment task can adequately assess the intended content in a way that is relevant and appropriate. When considering the fitness-for-purpose of assessments, relevance becomes paramount.

Collaborative construction of standards

Students typically receive a score or grade on completed assessment that reflects their performance on the assessment against some pre-determined performance standards. The process of setting these pre-defined standards is an important part of the assessment cycle, and requires dedicated time and resources. Determining a performance standard or an achievement scale for an assessment or a single piece of assessment should be a constructive and collaborative process.

Although they are accepted as a necessary component of assessment, standards are inherently difficult to define (Bloxham 2012). Sadler (2013) provides a particularly useful definition of a standard as a "definite degree of academic achievement established by authority, custom, or consensus and used as a fixed reference point for reporting a student's level of attainment" (p. 15).

Standards are often poorly articulated, unclear and difficult to use (Sadler 2014; Weisler 2016). They are often used to compare student performance relative to that of other students in the cohort rather than making comparisons against a fixed reference point (Sadler 2014). However, when they are clearly articulated, and when students engage with them, performance standards help improve transparency of assessment and student learning (see e.g. Hendry, Armstrong and Bromberger 2012).

The HEA framework (2016a, p. 3) explains that "given that assessment standards are socially constructed, staff and students need to engage in dialogue about standards to understand what is required from, and entailed in, the assessment process." This, as well as the link between students' understanding of standards and their learning, underscores the importance of constructing standards in a collaborative fashion, in a way that engages both staff and students.

The establishment of standards cannot be a 'one-off' event. Expectations for performance can evolve over time. The application of standards in assessment (through the use of evaluative criteria in rubrics or the practice of making professional judgments regarding the quality of student work) is an inexact process and requires constant review. Sadler (2014, p. 274) notes that, "consensus on the wording of the outcome statements does not necessarily result in consensus on underlying achievement standards." Thus, critical reflection of standards should occur on an ongoing basis, with dedicated times and formalised processes built into an assessment cycle.

> In many higher education contexts, decisions on assessment and grading … are devolved to individual academics, small teams or program directors. A growing practice has been to develop explicit descriptions of expected standards so they can be used by students (as producers) and academic appraisers … If all relevant parties work to the same set of specifications, the belief is that appropriate levels of consistency and comparability will result. *(Sadler 2014, pp. 273–74)*

Students also need to be engaged in the standard-setting process and given clarity regarding expectations. Student satisfaction in relation to assessment is hindered when they simply do not know what is expected of them. “It is when learners share an understanding of academic and professional standards in an atmosphere of mutual trust that learning works best” (HEA 2012, p. 20).

An oft-cited study by McDowell and Sambell (1999), that evaluates the role of students as stakeholders in the assessment process, is relevant here. Their analysis identified some major concerns about including students in judging the fitness for purpose of assessment. Importantly, while there is some concern that students are “insufficiently knowledgeable and informed about assessment”:

> the data show that many students are able to think about assessment in quite sophisticated ways; this suggests that they may be at least as well-informed as some other stakeholders if not better informed … this is not to suggest that students should be the sole judges of assessment quality since their knowledge and experience is clearly limited, only that their views should be considered alongside others. *(Sambell 1999, p.121)*

This emphasises that while students should be included in the development of assessment standards, they are not the only stakeholders. While students provide useful input into assessment, as they are not assessment experts, it is important that other stakeholders are involved in the process.

Integrating assessment literacy with learning

Assessment literacy can be defined as students’ ability to understand the purpose and processes of assessment, and accurately judge their own work (Smith, Worsfold, Davies, Fisher, and McPhail 2013). The HEA framework (2016a) advocates integrating assessment literacy into course design in order to help students understand and engage with the standards. However, it is not just the literacy of students that is important in this regard, as noted by Price *et al*. (2011), developing assessment literacy among academics is critical to enabling informed debates on approaches that are implemented.

In *A Marked Improvement* (HEA 2012), this is articulated in the following way:

> Active engagement with assessment standards needs to be an integral and seamless part of course design and the learning process in order to allow students to develop their own, internalised conceptions of standards and to monitor and supervise their own learning. *(HEA 2012, p. 21)*

One of the fundamental processes articulated in the HEA framework (2016a) is that of dialogue and building a mutual understanding of the importance and role of assessment between staff and students. This fundamental process is closely linked with assessment literacy. Through collaborative construction of standards and by actively engaging students with assessment standards and processes, assessment literacy among students can be raised.

Rhodes (2016) notes that:

> Assessment is being integrated into the daily activities of faculty and other educational professionals as they teach, construct, and deliver programs, design projects and high-impact practices, and challenge students to become more adept and practiced at exercising higher-order abilities such as analysis, problem solving, and evaluation. *(Rhodes 2016, p. 112)*

Integrating assessment literacy into learning encourages students to become autonomous learners with the capacity to reflect on and review their own learning progress, as well as the progress of their peers. Being an active agent in this process can help students to develop a sophisticated understanding of assessment.

Attempts to engage students with assessment standards and processes have ranged from developing assessment criteria and performance standards and making them available to students to actively engaging students with these standards (O'Donovan, Price, and Rust 2008). Simply presenting written descriptions of performance standards to students in the form of rubrics or scoring guides does not appear to adequately improve students' assessment literacy or learning (O'Donovan, Price, and Rust 2004; Reddy and Andrade 2010). Engaging students with standards through self-assessment and peer-assessment tasks appears to encourage critical reflection by students (Boud, Cohen and Sampson 1999; Dochy, Segers, and Sluijsmans 1999). Discussing and applying assessment standards to exemplars appears to have a small but positive impact on assessment literacy and student learning (Rust, Price, and O'Donovan 2003; Smith *et al*. 2013).

The principle of integrating assessment literacy with learning is closely aligned with the tenet that assessments should be implemented using transparent and fair processes. Transparency around assessment appears to increase students' confidence with assessment processes (Bamber 2015). Further, there may be a negative relationship between students' perceptions of the transparency of assessment goals, standards and grading processes and fear of failure and academic pressure (Suhre, Jansen and Torenbeek 2013).

The past decades have seen a movement towards increasing transparency in assessment (Bloxham *et al*. 2016), yet there are still concerns about the rigour and reliability of assessment processes (Sadler 2013). In a recent New Zealand study (Fletcher *et al.* 2012), the common disparity between faculty perceptions and student perceptions of the role and importance of assessment are highlighted. Transparency can be facilitated using some of the same approaches used to increase students' assessment literacy, such as sharing and discussing performance standards, scoring guides, rubrics or exemplars with students (Bell, Mladenovic and Price 2013; Hendry *et al.* 2012; Jonsson 2014). Involving students more actively in assessment, for example, by incorporating self-assessment or peer-assessment activities, can also increase transparency (Falchikov and Goldfinch 2000; Nicol, Thomson and Breslin 2014).

Fairness relates to the degree to which an assessment provides an equivalent measure of student ability independent of the non-relevant personal characteristics of the students. Assessments should not disadvantage any particular group of students – such as those from a particular socio-economic or cultural background, or those of a particular gender – in a systematic way. At an individual item level, this means that assessments need to address issues of potential bias. At the process level, fairness relates to the consistency of student experience and evaluation in any assessment process. Inconsistency in assessment experience is contentious, open to external criticism and even litigation. Students are stakeholders who want fairness and consistency in assessment. Clarity in expectations (through integrating assessment literacy in course design) goes some way to achieving this.

Defensibility of professional judgments

Despite continuing efforts to ensure transparency and fairness, some form of variability is inevitable when setting and scoring assessment, and assessors can often be inconsistent when scoring assessment (Bloxham *et al*. 2016). While standards play an important role in the consistency, transparency and fairness of assessment, standards and criteria are not in themselves sufficient to improve assessment processes (Wyatt-Smith and Klenowski 2013). This is particularly the case with more complex assessment tasks that have no single “correct” solution (Wyatt-Smith and Klenowski 2013).

“Assessment is dependent upon professional judgement; confidence in which depends on forums for developing and sharing standards within and between academic, disciplinary and professional communities” (HEA 2016a, p. 3). In the assessment of highly complex learning, professionals (including content experts, lecturers, teachers and tutors) often make holistic judgements regarding student learning rather than overly mechanistic ones. These judgments are valuable, but should be continuously reviewed through critical reflection in a collaborative setting to ensure that judgements are reliable, consistent, transparent, fair to students, and ultimately defensible (Bloxham *et al*. 2016)

Bloxham, Hughes and Adie (2016) note that “designing quality assessment that generates credible evidence of student achievement, and that can be consistently and reliably judged by those assessing the work, is a contentious and complex task.” (p. 638). The quality of judgements is often demonstrated by their adherence to criteria, the credibility of evidence, the existence of shared standards and the consistency of judgments (Bloxham, Hughes and Adie 2016). Activities that demonstrate moderation of judgements include peer scrutiny, double and collaborative marking processes, random checking, and consensus moderation discussions (Bloxham, Hughes and Adie 2016). The process of “moderation with a focus on community building has been shown to add to assessors’ assessment literacy as well as knowledge of standards” (p. 648).

The limits of assessment

Assessment can only measure so much. The HEA framework (2016a) recognises that assessment lacks precision. In *A marked improvement*, the point is made in the following way:

> Limits to the extent that standards can be articulated explicitly must be recognised since ever more detailed specificity and striving for reliability, all too frequently, diminish the learning experience and threaten its validity. *(HEA 2012, p. 20)*

Since the 1990s, there has been a proliferation of standards and learning outcomes documents in the higher education realm (Hudson, Bloxham, den Outer and Price 2015). Some researchers have criticised such documents, arguing that they “are based on a misconception of standards as explicit and absolute, when in fact standards have tacit and contextual qualities that make it impossible to codify them fully” (see e.g. Hudson *et al*. 2015; Sadler 1985). This principle holds that “there are important benefits of higher education which are not amenable either to the precise specification of standards or to objective assessment" (HEA 2012, p. 20).

This principle also aims to acknowledge that learning is a complex and non-linear process, which occurs across multiple different educational contexts. Attempting to precisely define all the meaningful competencies or learning outcomes of a programme is an inherently reductive process, which somewhat artificially constrains the transformative experience of learning. In the words of the HEA, "There are some aspects of learning that cannot be reasonably assessed", and "the effects of assessment extend over and above the intended outcomes' (HEA 2012, p. 20).

Practice and innovations

Introduction

The 'Practice and innovations' section contains a number of sub-sections detailing the way in which current literature relating to assessment in higher education applies the fundamentals discussed in the section above.

Framed by the research question below, this section reviews the literature based on the most significant issues being highlighted in scholarly research in assessment.

> How are the fundamentals of assessment manifested in recent literature on assessment in higher education?

The two most substantial issues identified in the literature and discussed in this section are feedback/feed-forward, and peer-assessment/self-assessment. In addition to discussion of these areas, exploration of recent literature in relation to the use of technology in assessment, the treatment of summative assessment and the role of psychometrics in current assessment literature are also examined in this section. The section finishes by looking at a number of other emerging practices in the recent literature that were considered noteworthy.

Feedback/feed-forward

A substantial proportion of the literature on assessment focuses on feedback and/or feed-forward. Given an otherwise positive book review recently complained the reader was left "pondering the precise definition" (Bamkin 2013, p. 128) of feed-forward, it is worth stating that "feed-forward refers to 'timely and constructive feedback' that feeds into the next assignment" (Wimshurst and Manning 2013, p. 451), that feed-forward is the aspect of feedback that encourages and allows improvement and, as such, should be a "necessary characteristic of feedback" (Boud and Molloy 2013, p. 702).

Many authors note the problem of modular university subjects and the fact that end of module feedback is difficult or impossible to feed-forward into future assessment, given that the module has finished (Higgins, Hartley, and Skelton 2002; Hughes, 2011). This has led to a focus on feed-forward prior to final assessment using a number of strategies such as exemplars (Hendry *et al.* 2012; Scoles, Huxham and McArthur 2013), continuous assessment marking (Hernández 2012), peer assessment, and feedback on drafts.

In general, the field demonstrates a shift away from feedback as justification or explanation of a mark to feedback being primarily about its ability to feed forward into future learning. Building on current theories about formative assessment as well as research showing feedback to be a source of student dissatisfaction (Boud and Molloy 2013), there is broad consensus in the literature about what good feedback should be and should do. Orsmond *et al.* (2013) present a useful summary of the received model of 'new feedback', which is shown in Table 1.

1. Model of 'New Feedback'	
New feedback delivery	**Standard feedback delivery**
Encourages dialogue between giver and receiver of feedback	Monologue often tutor directed one way feedback
Involves peers	Does not involve peers
Explicitly encourages self-assessment/regulation	Does not explicitly encourage self-assessment/regulation
Feedback on assignment process	Feedback on assignment product
Students encouraged to be *proactive* in working with feedback	Students encouraged to be *reactive* in working with feedback

(Orsmond et al. 2013, p. 244)

While this model demonstrates a well-developed, coherent and conceptually persuasive approach to a new type of feedback delivery, it also illustrates some of the weaknesses of the literature, one of which is a tendency to dichotomise. While the distinction between, say, assessment *of* learning as opposed to *for* learning is conceptually useful and defensible; it can obscure the fact that even the most summative of summative assessment has an enormous impact on where students invest their time. In this sense summative assessment of learning has, and has always had, a strong formative component in shaping what students make an effort to learn. Too much focus on this dichotomy can lead to a situation where innovation is located within formative assessment alone. The possible impacts on student learning of innovation in summative assessment are left under-researched.

However, the biggest problem with the field is that while the pedagogical theories behind the current consensus on feedback are sound, solid evidence for effectiveness is correspondingly thin. This leads to an abundance of assertions that various strategies 'may', 'might', 'could', 'can' or 'appear to' improve learning. Given current attempts to reform feedback were initially driven by student complaints that feedback was "one of the least satisfactory aspects of their university experience" (Carless 2015a, p. 189), it is perhaps unsurprising that the literature shows an over-reliance on student and teacher perceptions of feedback as research data.

Student perceptions are vital, and closing the gap between student and teacher perceptions of feedback is important (Adcroft and Willis 2013), as is (for example) understanding the degree to which higher education students' concept of feedback is mediated by their secondary experience (Beaumont, O'Doherty and Shannon 2011). However, "student perceptions of feedback and its educational effectiveness are not necessarily the same" (Denton and Rowe 2015, p. 1101) and observations such as "the exemplar facility was highly valued by students, although there were no quantitative effects such as an increase in students' assignment marks" (Handley and Williams 2011, p. 95) should cause concern. Moreover, studies based on student perceptions often make observations such as, "much of what the students had to say about their experiences with assessment feedback problems agreed with research documented elsewhere" (Ferguson 2011, p. 59). The agreement Ferguson identified in 2011 raises the question of why student perception of feedback remains the dominant approach today.

Where the literature is lacking is obvious:

> while theoretical justification for the apparent benefits of feed-forward is strong, the empirical support for such confidence remains slight *(Wimshurst and Manning 2013, p. 451).*
>
> There have been very few published reports of investigations in higher education where the feedback loop has been completed, and I have been unable to find any previous report where the loop has been completed by the teacher who originally provided the feedback. *(Donovan 2014, p. 1019)*

> Although there are many suggestions within the feedback literature for strategies ... there are few empirical studies which test the effectiveness of these interventions. Evidence of the effectiveness of such interventions is greatly needed. *(Ali, Rose and Ahmed 2015, p. 582)*

Boud and Molloy state that "if there is no discernable effect, then feedback has not occurred" (2013, p. 1019), a point which Boud has been making for some time: "feedback … is only worthy of the name if … teachers can detect in the work of students that the information they have provided has made a difference" (2007, p. 18). In reviewing recent literature on this subject it is difficult to avoid concluding that future research need to focus on identifying the impact of feedback in the work of students. The number of researchers beginning to express frustration with this gap is encouraging. It suggests that future research should begin to focus on valid and reliable measurements of the impact of feedback.

Recently, there have been some notable attempts to measure the impact of feedback, although often these studies have quite limited sample sizes. Scoles *et al*. (2013) used feed-forward exemplars to help students prepare for summative assessment and consequently showed an improvement in students' results. Vardi (2013) conducted a similar study with two student cohorts, and managed to identify why improvement was shown in one cohort but not in the other, which would not have been possible without empirical evidence. Walker and Hobson (2014) pinpointed a lack of alignment between staff and student perceptions and intervened with feed-forward towards summative assessment. As a result of this intervention, they showed that students had an improved pass rate.

The benefits of empirical evidence are particularly evident in the sophisticated analysis presented by Nicol, Thomson and Breslin (2014). Here the authors use surveys and focus groups to investigate the cognitive processes that led to empirical improvements in performance measured elsewhere (Cho and Cho 2011) following the student provision of feedback to peers. Nicol *et al*. build on Cho and Cho's empirical results, using students' perceptions of feedback to argue that, "feedback production is recognised as just as valuable for learning as feedback receipt" (Nicol *et al.* 2014, p. 120).

In essence, the field has a robust and defensible model for what feedback should achieve and is taking defensible measures to reach those goals. However, the definition of success in reaching those goals needs to move towards more objective measures. Measuring improvements in performance following pedagogical intervention is difficult, but this does not mean it should be avoided. It is necessary to measure empirical improvement in order to move the field past the continued reproduction of perception surveys about case-study implementations of a consensus theoretical position.

Peer-assessment and self-assessment

While peer and self-assessment have been extensively researched, there is often real ambiguity as to exactly what these terms mean within the context of individual papers. Is 'peer assessment' the assignation of a grade, the production of purely formative feedback, a combination of the two, or something else altogether? The distinction between peer evaluation and peer feedback was made explicit some time ago (Falchikov and Blythman 2001), yet despite this many papers require close reading to discern exactly what the authors mean by 'peer assessment'.

This is not merely a question of semantics. In an update of Topping's inventory (1998) of peer assessment, Gielan, Dochy and Onghena (2011) urge using the inventory to guide the description of peer assessment practices in future studies, as "when this happens, performing replication studies, studying interaction effects between certain variables and comparing and synthesising findings on peer assessment should all become more straightforward" (2011, p. 150). Later that year, in a review of the literature, and after teasing out the various meanings of 'peer assessment' Gielen *et al.* (2011) noted that:

> The formulation of new goals for peer assessment, beyond being an assessment tool, has introduced new expectations … The problem is not so much that different goals, and thus different definitions of quality, exist: the problem is that a clear view on the relationship between these goals and definitions is getting lost … The risk here is that one takes a certain conceptualisation of quality for granted, without questioning the alignment with the implicit goal assumptions. *(Gielen et al. 2011, pp. 720–1)*

The authors go on to express the hope that "this overview should help researchers and practitioners to be more explicit about their goals of using peer assessment, and should clarify the relationship with appropriate quality criteria" (Gielen *et al*. 2011, p. 732). A similar plea could be made regarding the work of Taras (2010) in categorising the various definitions of 'self-assessment'.

The problem of a lack of alignment between peer assessment's new "implicit goal assumptions" and "certain conceptualisations of quality" are real. The literature on peer assessment is not short of empirical data. It is not difficult to take peer or self-assigned grades, compare them either with each other or with tutor grades, then publish a correlation. However, while there is much data, whether it is the right data is another question. If we are, for example, increasingly interested in the degree to which the *production* of peer assessment has an impact on a student's ability to self-assess, then correlations between peer and tutor grades are not highly relevant.

Persuasive arguments have been made over the past decade that "the dominance of peer assessment processes using grades can undermine the potential of peer feedback for improving student learning" (Liu and Carless 2006, p. 279), yet the relative ease of publishing studies that correlate grades guarantees their continued production even when they do not answer the questions we now need to answer. When a 2016 meta-analysis of peer versus instructor ratings reports results that do not "depart much from what was reported by Falchikov and Goldfinch (2000)" (Li *et al*. 2016, p. 254) there is a clear implication that the field is in danger of treading water. Reinholz makes this point directly, arguing that most studies on peer assessment focus "on calibration of instructor and peer grades, rather than learning opportunities" (Reinholz 2016, p. 301). He goes on to provide a "framework for understanding the learning that takes place through peer assessment by connecting peer and self-assessment", offered as a "guide for future research studies to test and validate the model" (Reinholz 2016, pp. 312–13).

The biggest question about research into peer assessment is whether we use it as a formative or summative pedagogical tool. There are cases where using it summatively is clearly useful, such as where peer assessment is used to allow teachers to "design assessment that unravels group and individual contribution" (Tucker 2013, p. 74) when assigning individual marks for group projects. A number of papers address this (see e.g. Ko 2014; Neus 2011; Spatar, Penna, Mills, Kutija and Cooke 2015).

However, it is sometimes unclear what the continued calibration of peer and tutor grades hopes to achieve. The implicit claim of that calibration is that peer grades could potentially replace tutor grades. What seems to be overlooked is that in a controlled setting where peers know their grades are being compared with a tutor's, there is no motive to game the system. But when you change the context of an assessment you can change the variable it measures: if peer evaluation became the basis on which people were allowed to prescribe medication or argue a case in court, we would have a compelling motive to collude. This is not to indulge in cynicism, simply to note the proportion of the literature focused on integrity issues such as plagiarism (see e.g. Bruton and Childers 2016; Chew, Ding and Rowell 2015; Newton 2016; Sutherland-Smith 2014) and to observe that this research was produced for a reason. Regardless of how well we calibrate peer and instructor grades, many papers detail formidable problems with the summative use of peer assessment. "Perceptions of unfairness" (Carvalho 2013, p. 503) are common, students are "highly critical of it as a summative practice" (Patton 2012, p. 719) and when "asked whether their peers should have a greater say in their overall grade the majority disagreed" (McGarr and Clifford 2013, p. 677). Furthermore, evidence "that students would need to complete six assessments each with groups of at least three assessors per assignment in order to achieve acceptable accuracy (Jeffery *et al*. 2016, p. 139) suggests that inasmuch as part of the drive towards peer assessment is to get "students to do the work of … overworked academic and sessional staff" (Wilson, Diao and Huang 2015, p. 28) the solution could create more work than the problem.

Peer assessment as a formative practice is far more promising. Students are positive after having done it and prefer it to summative peer assessment: "the students' proposed model of peer assessment addressed their concerns … by increasing formative feedback and limiting its potentially damaging impact as a summative exercise" (Patton 2012, p. 729).

The relationship between peer-assessment and self-assessment needs to be addressed at this point. By 'self-assess' we mean the meta-cognitive ability to critically judge the value of one's own work, not self-assess in the summative sense, the utility of which is limited by remarkably consistent findings that average students over-estimate their abilities while more able students are either more accurate or too pessimistic (Jewell 2013; Boud, Lawson and Thompson 2015; Jewell 2013; Kun 2016; Ünaldı 2016). Students in the middle of the scale often cannot identify what they don't understand. Students at the top of the scale usually can. This is not a problem if the self-assessment we are talking about is formative – we can help the less able and overly optimistic students to see problems in their own work that they could not identify unaided. But plainly it is a problem if the self-assessment is contributing to a grade. Additionally, given the increasing internationalising of student cohorts as policy positions tertiary degrees as an exportable commodity, the degree to which culture and national background impacts self-assessment (see e.g. Humburg and van der Velden 2015) should counsel caution for its summative use; also an issue with summative peer assessment (Li *et al*. 2016).

Research is consistently suggesting that the biggest benefit of peer assessment is its impact on the ability of students to self-assess in the meta-cognitive sense. A study of 800 students found that:

> 72% of the students agreed or strongly agreed that the useful part of the peer assessment was giving feedback, rather than what was received. It thus seems [it was] the development of critical higher order outcomes that students found most useful in improving their work. In this context, reliability of assessments becomes of secondary concern as it is the process of giving formative peer assessment that holds the greatest potential *(Mostert and Snowball 2013, p. 684).*

That essential point is echoed repeatedly throughout the field (see e.g. Chambers, Whannell, and Whannell 2014; Yucel, Bird, Young, and Blanksby 2014). Mulder *et al*. argue that:

> helping students to become skilled peer-assessors enables them not only to develop important high-order (generic) skills such as critical evaluation and communication, but also to develop important skills of self-assessment – arguably one of the most important goals of a higher education. *(Mulder, Baik, Naylor and Pearce 2014, p. 672)*

Their research found that student perceptions of the benefits of peer assessment are "mirrored in genuine improvement in the quality of work submitted" (Mulder *et al*. 2014, p. 671).

While most studies reiterate the importance of criteria (El-Mowafy 2014) and training (Liu and Li 2014) for accurate peer assessment, Jones and Alcock (2014) investigated the accuracy of peer assessment without any criteria at all, instead "presenting assessors with pairs of student scripts and asking them to judge which of the two students has performed 'better'" (p. 1775). They found high validity and inter-rater reliability, suggesting that, "human beings are better at comparing one object against another than they are at comparing an object against specified criteria" (p. 1785). If this is true, then the reasons why student self-assessment might be improved by the *process* of peer assessment more than the receipt of its feedback could be that the process provides them with many 'objects' with which to compare, judge and improve the quality of their own work: "*all* students still rated the marking sheet as less useful than exemplars for guiding them" (Hendry *et al.* 2012, p. 158).

The consensus position on the formative value of peer assessment is compelling and coherent, as is the emerging conviction that it is the provision more than the receipt of feedback that has the most potential. There are relatively obvious ways of beginning to investigate the validity of that conviction, such as comparing post-draft improvements between students who produced feedback with those who received it, one of a number of possible approaches that could begin to generate the kind of data the field needs. Given the importance the literature assigns to encouraging students to move beyond the passive reception of feedback towards actively soliciting and interrogating it (Boud and Molloy 2013), it should not be difficult to measure the degree to which the production of feedback achieves this goal.

As is the case with the literature on feedback (before and after), student surveys that reveal changing attitudes are useful but are not of themselves evidence of pedagogical improvement. However, the literature on peer assessment is most deficient in specifying exactly what peer assessment aims to achieve and then aligning that with empirical investigations of quality that *match those goals*. Given the importance of 'constructive alignment' in the literature on assessment in general, this should be both achievable and non-controversial.

Summative assessment

Literature on summative assessment in and of itself is, certainly when compared to the literature on formative assessment, rather scarce. Even within the context of the important question of whether it is better for lecturers to set their own exams or for the process of teaching to be decoupled from summative assessment through the use of external examinations, a 2011 literature review noted a striking lack of empirical evidence one way or the other, concluding that "in the absence of empirical evidence, the discussion about coupling or segregating teacher roles remains a highly speculative endeavour" (Sasanguie *et al.* 2011, p. 908). Within the literature on assessment, summative assessment is usually mentioned to provide a contrast with formative assessment in order that formative assessment can be explained and its benefits emphasised. This strategy continues to propagate "the unhelpful binary division between summative and formative assessment" (Boud and Soler 2016, p. 402).

Occasional efforts to expand that dichotomy have been made. Crisp (2012), for example, argues for adding "integrative assessment" and "diagnostic assessment" to the formative/summative dichotomy to form a four-way matrix. The purpose of integrative assessment "is to influence students' approaches to future learning" (Crisp 2012, p. 41) and while this is a worthy enough goal, it is not clear how success in doing so could be measured, or how this differs in any substantial way from the goals of "sustainable assessment", where "every act of assessment needs in some identifiable way to build students' capacity to manage and judge their own learning and thus equip themselves for the more challenging learning environments they will confront post-graduation" (Boud and Soler 2016, p. 410). As an aside, Crisp does make the interesting observation that "diagnostic assessment is probably the most underutilised of the current assessment formats in higher education" (Crisp 2012, p. 39). Yet even attempted expansions of this nature still propagate the division between formative and summative assessment.

The impact that summative assessment has on shaping student behaviour is well known and has been described as the “backwash” effect. Typically, the backwash effect is presented in negative terms, as a “significant” disadvantage to summative assessment (Williams 2014, p. 566): the term “backwash” itself is inherently negative. However, there is no intrinsic problem with summative assessment influencing what strategic learners choose to do. It is only an issue if the summative assessment is influencing them in ways that are not in alignment with the pedagogical goals of a given course. If backwash is a problem, it highlights a defect with a given example of summative assessment, not an essential flaw in the practise of summative assessment in general. This is a non-controversial observation. Yet the recent literature shows few attempts to harness the latent power of summative assessment in order to nudge learners in a desired or beneficial direction.

Hunt, Koenders and Gynnlid (2012) reported that when the assessment of laboratory skills moved from written reports to a demonstration of practical skills, “four out of five students reported that an increased awareness of the importance of practical laboratory skills stimulated them to greater efforts to achieve” (p. 861). Turner, Roberts, Heal and Wright (2013) attempt to harness the power of summative assessment in order to “‘engage students in productive and worthwhile learning of an appropriate” (p. 663). They note that it was “important for us, having chosen this ‘alternative’ form of assessing our students, to carry out an evaluation of its efficacy in particular from the student perspective as this is currently under-researched” (p. 671) and conclude that they were “hopeful that the alternative form of assessment we have reviewed here has a good backwash effect on students” (p. 671). It is possible to move beyond hoping for a beneficial effect. It is not difficult to imagine research into how an example of summative assessment shapes student behaviour being followed by research into how that behaviour changed after the introduction of a summative innovation.

Despite occasional examples like that of Turner *et al.* (2013) or that of Iannone and Simpson – who argue that “oral assessment may be a viable assessment method for straddling the ‘assessment for’ and ‘assessment of’ learning divide in higher education” (Iannone and Simpson 2015) – it is difficult to avoid concluding that summative assessment is, at present, under-researched, given its importance in shaping the behaviour of students’ choices and priorities regarding their learning: “Developments in [summative] assessment have lagged behind developments in curricula” (Yorke 2011, p. 117).

Technology and assessment

The number of articles dealing with information and communications technology (ICT) and assessment necessitates some attempt at summary. This is difficult for two reasons. Firstly, the literature is invariably discussing technology and something else, for example, technology and feedback (Jones, Georghiades and Gunson 2012; Knauf 2016; McCarthy 2015). Secondly, the technology itself is highly diverse, not to mention rapidly evolving.

A few examples will give some sense of the diversity. Sharrock (2015) gives an overview of the current debates around massive open online courses (MOOCs), including the assessment of student learning in that environment. Snowball (2014), in the process of detailing how the use of ICT might assist in coping with the diversity of student cohorts, provides a succinct overview of the evolving ways in which technology has recently impacted higher education. Butz *et al*. (2015) look at the challenges of attempting to ensure equivalence in hybrid learning environments, comparing “performance in relation to attendance mode” (p. 636). Williams (2014) suggests an interesting, if slightly utopian, approach to using learning analytics gleaned from student behaviour within online formative assessment tasks as a means of summative assessment. Caple and Bogle (2013) show how the use of wikis in group assessment allows for the possibilities of tracking exactly who contributed what to the final product. Carruthers *et al*. (2015) investigate audio feedback; Craft and Ainscough (2015) attempt to develop clinical role play in an online environment; Holmes (2015) considers continuous e-assessment; Jackling *et al*. (2015) use blogs to encourage reflective learning; and Lai (2012) argues for assessing not only participation but the quality of online peer engagement. These papers have almost nothing in common except that they all deal with some kind of intersection between technology and assessment.

The characteristic strengths and weaknesses of given areas of research into assessment (such as an over reliance on student surveys as research data) are mirrored in the literature dealing with the use of technology within those fields. Additionally, the findings about a given technology in a given field typically mirror those reported in the field in general: in a paper looking at blended learning environments and collaborative feedback tools, Purvis *et al*. reported that the “students realised the benefits not only of the receiving of feedback but also of the learning process of giving feedback to others” (Purvis *et al.* 2011, p.

99), a finding familiar in the general literature on both feedback and peer assessment. In this sense the challenges and opportunities of a given technology in the context of a given form of assessment are better considered within that context, as any findings are specific to both that given context and a given form of technology.

That said, some general observations can be made. The literature on technology and assessment is “extensive and broadly optimistic” (Jackel 2014, p. 71), an enthusiasm that can be problematic in a number of ways, such as clear instances of confirmation bias as well as research into technologies that are largely obsolete by the time the research is published.

More substantively, the general level of enthusiasm within the literature can lead to a situation where the tail begins to wag the dog: “such projects are often run with the technical project goals achieved but the educational goals unrealised” (Stoddart 2015, p. 586). Trenholm *et al.* (2016) imply that that any formal “educational goals” might not have even been formulated before technological innovation, when they find no “systematic link between assessment schemes and instructors’ approaches to teaching” (p. 150) in a large survey of assessors’ experiences with assessment in online mathematics courses. A number of researchers make similar points. Stödberg – in a meta-review of the literature on the use of technology in higher education assessment between 2005 and 2009 – argues that it is important that e-assessment is “driven by pedagogy instead of technology” (Stödberg 2012, p. 602), and in 2015 Henderson *et al.* (2015) observed that:

> ongoing discussions about digital technology and higher education might better balance enthusiasms for the ‘state of the art’ (i.e. what we know might be achieved through technology-enabled learning) with an acknowledgement of the ‘state of the actual’ (i.e. the realities of technology use within contemporary university contexts). *(Henderson et al. 2015, p. 308)*

While it is difficult to disagree with the observation that ongoing discussions about ICT and assessment need to better balance enthusiasm for what might be possible with what is possible, the nexus between technology and HE assessment is not going away and will only become more important. At a pragmatic level there is now, in many cases, no option but to deliver assessment through technology: it is difficult to imagine how a course delivered online could assess its students in any other way.

Assessment delivered online or via technology is not simply a replacement for assessment delivered offline. Snowball (2014) presented data showing that "more interactive learning opportunities (such as self-marking online multiple choice and graphic exercises) have a bigger impact on student performance than passive activities" (p. 836) in an experiment where one weekly face-to-face lecture was replaced with various online activities, with those activities being compared for impact. This implies that e-assessment replicates the kind of learning that goes on with face-to-face *teaching* as much as it replicates offline assessment. Finally, organisations such as Assessment and Teaching of 21st Century Skills (ATC21S) make the case that we must begin to assess variables such as "collaboration and digital literacy" (ATC21S n.d.), and it can be safely assumed that we will need to use technology in order to assess students' abilities with technology.

Indeed, the need for authentic assessment is often used to justify the implementation of computer-based assessment (Craft and Ainscough 2015): clearly it serves little purpose to test HE students' composition skills with pen and paper when they will be composing with a keyboard for the rest of their lives. In other cases it is not so clear. When looking at the difference between performance on paper and performance on screen at an item-by-item level for a tertiary entrance test, Jackel (2014) found a pattern that revealed items based on diagrammatic stimulus were harder on screen than paper. He suggested that this was due to the students' inability to scribble freely on the diagrams when they were presented on a screen, concluding that:

> we need to carefully consider what we are doing when we deny candidates the ability to annotate, the ability to use their own diagrammatic reasoning when solving certain kinds of problems.
> *(Jackel 2014, p. 76)*

Here, the 'authentic' arguments for computer-based tests become problematic, as by changing the mode of delivery we are changing how the student can solve a problem, and in doing so will begin to change the kinds of problem-solving students will attempt to learn. When we consider that Google itself advises that the "easiest and fastest way to get started" when designing software "is to sketch out your screens by hand using paper and pencils" ("Putting it All Together: Wireframing the Example App," n.d.), the idea that assessment must be delivered via technology because technology is what the students will be using in the real world is not quite as simple as it first seems. The technology itself was typically conceptualised and designed in the first instance with pen and paper: despite the potential of harnessing technology in assessment, assessment that does not harness technology is not necessarily ‘out of date’ simply because it does not.

Psychometrics

By ‘psychometrics’ we are referring to a suite of statistical and probabilistic methods such as Item-Response Theory, Rasch measurement, and factor analysis that can (in theory) be applied to almost any form of assessment to validate the degree to which the components of that assessment conform to a given model. That modelling can be, and often is, applied to psychological profiles to ensure a degree of internal consistency. However, it can be applied just as easily to a multiple-choice physics test to help weed out individual items that are measuring a variable significantly different to that measured by the rest of the test.

Properly applied, this kind of analysis is very powerful. Psychometric analysis can determine whether an instrument is producing false-positive or false-negative results. For example, psychometric analysis allows assessors to identify a multiple choice question that was so hard for a given group of students that they essentially all gave up and simply guessed: that question can then be removed from scoring.

Its utility in practice is heavily dependent on having sufficient sample sizes and on the format of the assessment. It is difficult to see how psychometrics could be applied usefully to directly validate an oral presentation, creative writing assignment or reflective folio, although it can be applied to how such forms of assessment are graded, particularly when more than one assessor is involved in grading (Aryadoust 2015). However, given adequate numbers it is relatively straightforward to use psychometrics directly in developing surveys, multiple choice questions and short answer response questions. In particular, computer-based testing in those formats opens up the possibility of assessors being provided with psychometric analysis of the test the moment the students finish taking it, giving assessors powerful insight into the degree to which a given question is measuring what it was supposed to measure. Malau-Aduli and colleagues (2014) report on exactly that process.

Hardy and Arguete (2014) used psychometric analysis in developing a diagnostic instrument to identify when intervention might be needed to prevent students dropping out from university. Cumming *et al.* (2015) used psychometric analysis to validate a group-work skills questionnaire, and Homer, Darling and Pell (2012) applied Rasch analysis to summative medical exams. Jeffrey and colleagues (2016) used psychometrics to investigate the reliability of student assessments;

Lethbridge and colleagues (2013) used them in the development of a reflective thinking questionnaire in the context of nursing; and Xie and colleagues (2014) used Rasch modelling to calibrate individual items in an item bank. A final example of psychometric analysis in assessment is Jones and Alcock's (2014) significant study into the utility of peer assessment without assessment criteria.

In general, psychometric analysis is beginning to be used in two broad contexts. Firstly, as applied to the development or validation of forms of assessment that students themselves are undertaking. Secondly, in research investigating the use of or development of assessment tools. While still in its infancy within university assessment (although already standard practice in university admissions tests), the degree to which psychometrics is beginning to impact assessment in higher education is significant; its potential is even more so.

Other trends and areas of note

Outside of the main areas of research focus there are a number of interesting studies that, while sharing many of the field's general assumptions about assessment, cannot be neatly categorised as a clear example of their application. Because these articles are highly diverse, there is no general field to summarise. However, it is necessary to give some sense of the research being conducted outside of the main areas, and it is reasonable to suppose that some of these more singular papers will eventually flower into larger sub-fields of interest within the literature on assessment in general:

- Frost, de Pont and Brailsford (2012) introduced continuous assessment in the form of ten short assignments, during the semester, worth an accumulative 20% of the final mark. Although they admit that the "ideal 'before and after' scenario for assessing the impact of the assessment innovation was not met" (2012, pp. 300–301) they conclude by asserting that continuous assessment in this (relatively typical) manner "advanced teaching objectives and contributed to appropriate learning outcomes" (p. 301). However, in a provocative counterpoint to this line of reasoning, Harland *et al.* (2015) recently produced an article detailing what happened when continuous assessment was implemented across an institution and the unintended consequences of the assessment 'arms race' this created:

> Lecturers rewarded student work with grades and controlled study behaviour with assessment. In some situations it was possible to experience hundreds of graded assessments in an academic year. Students were single-minded when it came to grades and would not work without them. These conditions contributed to competition for student attention and a grading arms race between academics and subjects. In this context, the spaces for achieving certain educational objectives, such as fostering self-motivated learners, were marginalised. Both students and lecturers were unsatisfied with this situation, but neither group could envisage radical change *(Millar and Sim 2015, p. 528);*

Harlend *et al.*'s (2015) observations imply the following: the advantages of a given innovation are typically theorised within a set of unexamined assumptions about the advantages of that innovation *in the context of a given subject or module.* Likewise, the initial research into a given innovation is typically conducted by discrete practitioners, also within the context of a given subject or module. Little thought goes into the unintended consequences of that innovation when applied system wide. Those unintended consequences in a system-wide setting can be significant and may create bigger problems than the innovation itself solved;

- Spronken-Smith *et al.* (2012) produced a student perceptions survey on student response to inquiry based learning (IBL), producing not atypical data to show that "courses using discovery-oriented IBL were more highly rated", and that "the most highly rated course design was open" (p. 57). Interesting here is their admission that, "to determine whether all types of IBL courses are promoting enhanced [intended learning outcomes] compared to more traditionally taught courses, a comparative study should be undertaken" (2012, p. 57). This points to a characteristic weakness of much research based on student responses. When comparative studies are undertaken, the results can be rather confronting: "in contrast to the message from the general assessment literature, we found that mathematics students … prefer traditional assessment methods … they perceive them to be fairer than innovative methods and they perceive traditional methods to be the best discriminators of mathematical ability" (Iannone and Simpson 2015, p. 1046). Iannone and Simpson go on to argue that the students did not "base those preferences on an unreflecting conservatism" but rather had "thoughtful reasons for the preferences", underpinned by "a concern for fairness, the abilities being assessed and the proportion of a given method in a varied assessment diet" (pp. 1062–3). They conclude, somewhat dryly, that "our findings strike a note of caution for practitioners interpreting research findings in their subject and institutional contexts" (p. 1064);

- Reimann (2011) presented students with a choice between taking an unseen exam question on content that had been taught, and being given a seen exam question on content that had not been taught. The rationale behind providing students with a seen assessment on content that requires them to actively discover and research a solution is clearly a veritable hit-list of current pedagogical convictions (assessment *for* learning, authentic assessment, inquiry based learning, the student as active learners who construct their own knowledge and so on) not to mention a large body of student perception surveys of these strategies that suggests students appreciate them. Given the option, however, Reimann's (2011, p. 277) students perceived choosing the seen question "as a risk rather than an opportunity" and take-up was low: "the option of being able to answer a seen exam question did not appear to contribute to students' readiness for risk taking and dealing with uncertainty." The sample size in question was also low, but the qualitative feedback is handled well and the degree to which these findings run counter to prevailing wisdom warrants repeating the research with larger numbers;
- Hagström and Scheja (2014, p. 244) introduce an element of meta-reflection "'to prompt students to think critically and analytically around salient features of the subject in hand" (p. 244) to a take-home examination and present data showing that with "no changes to the course other than the introduction of an element of meta-reflection" (p. 250) throughput was increased by 70-80%, concluding that "the introduction of an element of meta-reflection about how to tackle examinations both stimulated a deeper approach to learning the course contents, and significantly enhanced the aggregated student performance" (p. 251);
- the title of a recent paper by Fuller, Henderson, and Bustamante (2015) – 'Assessment leaders' perspectives of institutional cultures of assessment' – gives an adequate sense of its contents. The study offers a perspective on assessment that researcher's "could apply to their campus contexts or to future research" and is aimed at "individuals interested in probing the philosophical foundations of educational assessment and culture, and especially higher education assessment and culture" (p. 349);
- Varsavsky and Rayner (2013) present an interesting article, the contents of which are, again, neatly summarised by its title: 'Strategies that challenge: exploring the use of differentiated assessment to challenge high-achieving

students in large enrolment undergraduate cohorts'. This is not quite the 'adaptive testing' that some computer-based scaling tests use to alleviate the ceiling effect, but does give some sense of how the advantages of adaptive assessment might begin to be applied within the context of HE modules;

- Shulruf, Poole, Jones and Wilkinson (2015) present a method of using probability theories to define the exact point at which we delineate pass from fail. This is one of the very few papers that attempt to grapple with the mechanics of the way assessment is converted into numbers and some of the implications of what that means. The final sentence of the following will be considered again later, in discussing of the aggregation fallacy:

  > Our findings lead us to conclude that a method such as the [objective borderline method], that eliminates much (yet not all) of the subjective judgement required around the pass/fail cut-score, may be as sound as approaches used in some of the more established standard setting methods. In other words, when determining cut-scores is concerned, a focus on quality of the examination may be just as important as a focus on the quality of the examinees *(Shulruf et al. 2015, p. 434);*

- finally, Vaughan, Lalonde and Jenkins-Guarnieri (2014) present a dense article that proposes some methods for getting around some of the characteristic flaws of much educational research, using a "hierarchical propensity score matching method that can be utilized in contexts where randomization is not feasible and dependence between subjects is a concern" (p. 564). If the field as a whole wishes to head towards experimental designs that can empirically "show causal relationships" (2014, p. 565), then it will be necessary to begin to develop the scholarship of assessment per se to a level where researchers can productively engage with arguments such as that proposed by Vaughan *et al.* (2014).

Reflections: setting directions, identifying gaps

Based on the review of recent literature and with reference to work already undertaken by the HEA in this area, this section explores some of the key issues and themes identified in the current scholarly output on assessment in higher education. It examines gaps in the current literature and emerging principles.

Identifying gaps

The summative vs. formative dichotomy

It has already been argued that too much focus on the dichotomy of assessment *of* versus assessment *for* learning can lead to a situation where innovation in summative assessment is left under-researched. This is a problem as summative assessment has a huge bearing on where students focus their efforts. The dichotomy is also problematic because it is, in a sense, false, or at least misleading. To the degree to which summative assessment shapes the efforts of students, it is not just summative, but deeply formative. Students, understandably, focus on what is going to get them the best mark:

> For many students it is a game of tactical positioning and strategic manoeuvring to adapt and reduce these educational experiences to achievable goals to meet their ultimate goal – a good grade. This study has highlighted the influence of … prevailing assessment cultures on students' engagement in peer learning which somewhat tempers the claims made by advocates of such pedagogical approaches. *(McGarr and Clifford 2013, p. 690)*

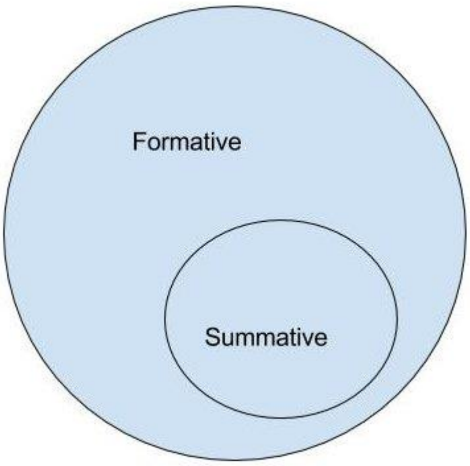
Yet, despite our awareness of the impact of summative assessment on the shaping of student learning, the literature essentially treats formative and summative assessment as headings for oppositional and mutually exclusive lists of attributes and practices:

Formative assessment (assessment *for* learning)	Summative assessment (assessment *of* learning)
A B C	-A -B -C

There is some truth in this dichotomy and it has had some real benefits. It has helped guide the field towards an understanding of how assessment can help students learn rather than merely record what they have learnt.

Yet the dichotomy creates its own problems. Formative assessment is new. Summative assessment is old. Innovation and research effort become heavily skewed towards formative assessment. We begin to overlook the fact that summative assessment never simply measured what students *had* learnt, it largely shaped what they bothered *to* learn in the first place.

Summative assessment is not going away. It will remain necessary for certification, for advancement to higher degrees, not to mention the job prospects and futures of the student. As such it will continue to have a huge impact on where students choose to invest their time and effort. Rather than continuing to emphasise and reiterate the formative/summative dichotomy, it would perhaps be useful to start from the observation that all assessment is formative in some sense, while only some assessment is both formative and summative:



That is, summative assessment is simply a subset of formative assessment. This could help foreground the formative aspects of summative assessment in the way it shapes student learning, and encourage a renewed focus on innovating summative assessment: “students’ motivation is often driven by assessment and therefore” it is important to develop a method of assessment “that both assesses and encourages” their development (Matheson, Wilkinson, and Gilhooly 2012, p. 257).

Innovation in summative assessment can be very difficult as it has a much higher risk factor than innovation in purely formative assessment. Yet, given the degree to which “students’ study habits are strongly geared to meeting the assessment requirements” (Hardy *et al*. 2016, p. 18), the potential rewards are also high. What really underpins assessment for learning is not its opposition to assessment of learning, but “the principle that all assessment, within the overall package, should contribute to helping students to learn and to succeed" (Sambell, McDowell and Montgomery 2013, p. 3).

The issue of 'aggregation'

A marked improvement alluded to an example of the failure to think enough about summative assessment and the mechanics of how we convert assessment into data in referring to the problem (or perhaps fallacy) of aggregation, where grades are arrived at by:

> artificially combining marks from different sources, which do not have equal weightings, meaning or validity. Thereby, this over-reliance on numbers can obscure learning and achievement in the search for credit equivalence and the end result is divorced from the aims of the original curriculum design. A review of assessment would reduce the proportion of work that is subject to this false maths; grading only that which can be safely graded and giving more credibility to students' results. *(HEA 2012, p. 13)*

The 'false maths' are not difficult to understand. Suppose we give three different students marks out of five for three different subjects. Our results could look like this:

	Subject A	Subject B	Subject C	Aggregate score
Student X	0	5	5	10
Student Y	5	5	0	10
Student Z	3	3	4	10

Our aggregate score tells us that these three students are all the same, but in reality the differences between them are considerable. Given only the aggregate we have no valid or defensible way of differentiating between these students. The aggregate effectively destroys the data it contains.

This is a perennial issue with the aggregation of subject scores to arrive at values such as grade point average (GPA). It is a perennial issue with admissions data where heterogeneous secondary subject scores are aggregated or where a generic admissions test which explicitly measures two or more independent variables has its section score data ignored in favour of the overall aggregate.

In terms of in-subject assessment within a course, aggregation of assessments to arrive at a final mark has been typically unproblematic. The reason is simple: when two or more assessments are essentially measuring the same variable, then adding them together is, in principle, valid. This would be the case in, for example, a Physics course where the total mark is comprised of two tests worth a total of 50% and a final exam worth 50%. However, suppose our Physics lecturer decides to innovate with the first two tests. One test is replaced with a peer-assessed exercise worth 20% of the total mark. The second test is replaced with an oral presentation worth 30% where the student has to deliver a presentation on the philosophical issues arising from the apparent non-falsifiability of string theory. Finally, students are given an exam worth 50%, which requires them to show how they can creatively apply acquired knowledge to solve a set of problems.

Considered on their own merits, each of these three assessment strategies is valid and defensible. However, the variable being measured when assigning a mark to the oral presentation is a very different variable to that being measured when completing the exam. These variables are not mutually exclusive, just independent. As such, aggregating them for an overall mark is simply not defensible.

At the top of a given scale, the 'false maths' of heterogeneous aggregation is not critical: these students did well at everything. At the bottom of the scale we know we are looking at students who did poorly at everything. However, in the middle of the scale this kind of aggregation leaves us unable to compare students in a valid or defensible manner, yet the middle of the scale is where all the hard decisions must be made, such as whether someone passes or fails or goes on to further study. The middle of the scale is where we most need the data we are using to be of high quality, and it is in the middle of the scale where a failure to think about the fallacy of aggregation often means it is of very low quality: in this case it is very difficult to defend a valid standard of pass or fail as that boundary is located within the section of the scale that cannot be relied on.

Shuruf *et al.* (2015) noted that, when determining the precise pass/fail point for a given examination, "a focus on quality of the examination may be just as important as a focus on the quality of the examinees" (Shulruf *et al*. 2015, p. 434). They were considering discrete examinations but their observation applies just as well to the total 'examination' comprised of the aggregation of a student's individual assessment scores within a subject. Unless those individual assessments are measuring a very similar variable then the overall quality of the final grade will be poor, especially in the middle of the scale.

It is quite possible that students intuitively appreciate the problem here, as suggested by the concern Iannone and Simpson (2015) found regarding the mix of assessment "methods they encounter during their degree" (p. 1046). However, regardless of whether students are aware of the problem, their assessors certainly should be. There have been some attempts to begin to address "the analytical need to ensure that data can be meaningfully aggregated" (Coates and Seifert 2011, p. 186), see, for example: Giada, Giovanni, and Vincenza (2014); Hathcoat *et al.* 2016). However, despite heterogeneous aggregation being described in language as strong as 'false maths' in *A marked improvement*, little appears to have changed in terms of practice. Examples in which this problem has not been considered at all when introducing innovations in assessment are nearly as extensive as the bibliography to this document.

Emerging ideas

The following emerging ideas were identified in the literature as supporting high-quality assessment that improves student learning. However, these ideas are not yet adopted by the wider academic community.

Programmatic assessment

An emerging principle of assessment, which has synergies with constructive alignment and assessment for learning, is the notion of programmatic assessment. Most of the work on programmatic assessment is being conducted in medical education. However, this type of assessment and aggregation in assessment could potentially be used more broadly.

This approach treats assessment as a whole programme, rather than as isolated or ad hoc assessment instruments (Schuwirth and Van der Vleuten 2011; van der Vleuten *et al*. 2012; Prideaux 2016). The strength of an assessment programme is thus judged by the programme as a whole rather than by the properties of single instruments. A variety of well-chosen (fit-for-purpose) tools can be used where appropriate, which allows for adequate sampling across the breadth and depth of the curriculum. The programmatic approach thus recognises the strengths and weaknesses of different approaches to assessment, and seeks to ensure a balance across a whole programme.

A key principle in this approach is that individual data points can be utilised for targeted feedback of learning (assessment for learning), while high-stakes decisions can be made based on the aggregation of multiple data points (van der Vleuten *et al*. 2012). For instance, the pass/fail decision for a course, or even an entire programme, can be made using rich information taken from the across the curriculum and in different assessment contexts. One cornerstone of the programmatic approach to assessment is the notion of "information richness” (Van Der Vleuten *et al.* 2015). The quality, quantity, and accuracy of the information provided by assessment are important.

With the programmatic approach, progressive assessment information is collected about students until sufficient information is available to be able to make informed judgements about their achievements. This principle acknowledges that individuals learn at different rates, and the measurement of learning should occur at multiple times throughout that process. The result of this approach is that there are multiple opportunities for gaps in learning to be identified, and remedial work to be instigated. That is, there should be no ‘surprise’ if a student fails one high-stakes instrument at the completion of a course, as there are progressive measures of learning throughout the programme. Dijkstsa *et al*. (2012) have developed fit-for-purpose guidelines for implementing programmatic assessment in different settings.

Assessment of learning outcomes

Another area of assessment that is not a focus of this document, but which is appearing more regularly in scholarly literature and policy debate in recent years, is the area of assessment of learning outcomes. This is an important area of higher education policy and one that is increasingly being married with policies of “quality assurance”, as discussed by Hathcoat and colleagues (2016):

> Widespread criticism about the value of higher education, coupled with increasing demands for accountability and transparency, reinforce the need for institutions to assess student learning and development. Assessing student learning outcomes in higher education however, remains riddled with challenges. *(Hathcoat and colleagues 2016, p. 16)*

This assessment of the learning outcomes concept is also tied closely to the almost universally adopted articulation of graduate attributes within higher education. As graduate attribute or graduate capability statements have become the norm, the issue of ‘proving’ that, or ‘measuring’ whether, these attributes have been achieved has become more prominent in discussion. On this issue, de la Harpe and David (2012, p. 507) articulate the problems in progressing these ideas at the university level:

> We have suggested that if universities are to progress the graduate attribute agenda, academic staff confidence and willingness, as well as the role of beliefs for some graduate attributes, must be acknowledged and addressed. Continuing to rely solely on teachers and their ability to integrate attributes into the formal curriculum will not deliver university graduates equipped for the rapidly changing world of work. *(de la Harpe and David 2012, p. 507)*

The conceptualisation and partial realisation of large-scale, multi-institutional, cross-national assessment of learning outcomes projects have also emerged in recent years. Projects such as the OECD's feasibility study *Assessment of Higher Education Learning Outcomes* (Richardson and Coates 2014) and the European Commission's project on measuring and comparing achievements of learning outcomes in higher education (CALOHEE 2016) are two multi-national projects instigated by inter-governmental organisations within the past five years. Further to this, within the UK context at the very least, recent policies highlighted in the Teaching Excellence Framework (TEF) and in the Government's *Success as a knowledge economy* paper pave the way for quality assurance-linked assessments of learning outcomes as a measure of teaching quality (BIS 2016). The trajectories of these initiatives will continue to attract attention and debate. While these broad initiatives appear at present somewhat distant from the coal-face of higher education and out of the control of practitioners, the almost inevitable growth in this area and its consequences for the way in which assessment is undertaken will require further attention in the future.

Conclusion and future directions

The concluding section of this review explores the broad themes identified in this paper and suggests some future directions that have emerged from the examination of the literature. These conclusions highlight findings and recommendations for three domains: institutional policy; educational practitioners; and the higher education sector as a whole. The topics raised in each of these sections overlap with one another and each issue is arguably relevant for all of these domains. Nonetheless, the structure of the conclusion helps to situate the findings of the literature in different contexts.

Institutional policy

The literature in the field could be used extensively at the institution level to frame policy, professional development and overall directions for teaching. In tandem with the HEA's Framework for Transforming Assessment, the six fundamental principles identified in the literature should be helping to frame the design and review of teaching in higher education. These are discussed extensively early in the report and include:

- assessment for learning;
- aligned and fit-for-purpose assessment;
- collaborative construction of standards;
- integrating assessment literacy with learning;
- defensibility of professional judgements;
- the limits of assessment.

Guided by this set of broad principles, and drawing on other themes from the literature, institutional policy could move towards models of programmatic assessment that are described in the literature as 'whole-of-programme' approaches towards assessment, rather than each assessment task being applied in an isolated or ad hoc manner. There are definite synergies between this approach and the almost universal embedding of graduate capability/attribute statements within higher education institutional policies. Embedding the statements of attributes into a holistic notion of assessment for each qualification arguably offers institutions not only a guide to how they want their students to develop, but guides assessment that promotes these attributes.

Institutional policy is the most useful starting point for setting principles to improve the assessment literacy of both students and academics. As discussed in the review, notable gaps exist in this area, with some examples being the continued (mis)use of aggregated marks and its pitfalls, especially in the middle-range of achievement. The HEA's *A marked improvement* noted this in 2012, but the literature does not suggest substantial change or recognition of this issue in the interim. There could be a key role for institution-level leadership to articulate and address this issue.

Educational practice

Feedback, peer-assessment/self-assessment, and the increasing use of technology in assessment are some of the areas emerging from the literature. Conversely, scholarship in the area of summative assessment and literature written by assessment experts for practitioners was found to be relatively uncommon.

The burgeoning literature relating to the area of feedback (moving towards the notion of feed-forward) as a vehicle for enabling educators to inform students and enhance their future learning is substantial. Practitioners have a robust and defensible model for what feedback should achieve and is taking defensible measures to reach those goals. The definition of success and the development of ways to measure this success are seen as important in improving the evidence base to support and enhance the use of feedback. From a practice perspective, this appears to be the next step necessary for practitioners to take in ensuring that the perceived benefits of this practice are demonstrable.

Similar issues relating to another large area of interest in assessment – peer-assessment and self-assessment – also persist as a challenge to educators. Improving the growing empirical base in this area appears to be an important step in embedding the most effective elements of this practice. Technological innovation is a related issue for educational practitioners. With the possibilities of using technology in assessment becoming broader by the semester, the caution of Henderson *et al.* (2015) is important to reiterate:

> ongoing discussions about digital technology and higher education might better balance enthusiasms for the "state of the art" (i.e. what we know might be achieved through technology-enabled learning) with an acknowledgement of the "state of the actual" *(Henderson et al. 2015, p. 308).*

The literature review highlights the significant emphasis on 'formative' assessment, and the notable lack of discussion of 'summative' assessment. While the distinction between assessment of learning as opposed to assessment for learning is conceptually useful and defensible, it can obscure the fact that even the most summative of summative assessment has an enormous impact on where students invest their time. In this sense, summative assessment of learning has a strong formative component in shaping what students make an effort to learn. Too much focus on this dichotomy could mean that innovation is located within formative assessment alone. The possible impacts of innovation in summative assessment on student learning are left under-researched.

Summative assessment is not going away. It will remain necessary for certification, for advancement to higher degrees, not to mention the job prospects and futures of the student. As such it will continue to have a huge impact on where students choose to invest their time and effort. Rather than continuing to emphasise and reiterate the formative/summative dichotomy, it would perhaps be useful to start from the observation that all assessment is formative in some sense, while only some assessment is both formative and summative. As assessment drives learning, ensuring that summative assessments are fit-for-purpose and well aligned to the desired curriculum outcomes positively effects student learning.

The final part of this report concludes with a focus on some key elements at the heart of the HEA's approach to informing and equipping higher education practitioners – the need for more "evidence-based policy or practice that has had a demonstrable impact on student outcomes including student performance, progression, engagement, satisfaction, skill acquisition and/or self-confidence" (HEA 2016b).

Proving and improving

There are opportunities for improving the 'proving' of assessment approaches in higher education that demonstrably assist learning outcomes. This was a clear point made by the HEA in *A marked improvement* (HEA 2012), and based on the review undertaken for this research, this message still needs to be emphasised in 2016. While institutions, and the higher education sector as a whole, have a role in this, the group best positioned to move this forward is the educational practitioners themselves.

As highlighted in several parts of the literature review, there are substantial issues with the dearth of solid, empirical evidence used to evaluate the effectiveness of current practice. Issues relating to empirical support for theoretical justifications are highlighted in numerous articles cited in this report (see e.g. Ali *et al*. 2015; Bloxham et al. 2016; Donovan 2014; Sasanguie *et al*. 2011; Wimshurst and Manning 2013). Much of the published evidence relating to assessment practice relies on small samples or single cohorts and produce results specific to a field of study that cannot necessarily add evidence to bolster the theoretical foundations for the approach. These limitations are well acknowledged by the authors of most studies, but in some instances the generalisation of findings is misleading. The reaction by Sasanguie *et al.* (2011) at the 'speculative' nature of much of the work in summative assessment practice following their broad literature review is telling.

The HEA's previous work in this area (2012) highlighted a related issue identified in this current review – that is, much of the scholarship in the area of assessment in higher education is developed by content experts rather than assessment experts. This would not be a problem in itself, were there a substantial overlap in these skills among individuals, but a strong grasp of assessment practice is not commonly found in faculty across the world (Price, Carroll, O'Donovan and Rust 2011; M Yorke, Bridges and Woolf 2000). The content expertise is pivotal to the development and reporting of assessment, but the reporting of the findings of research into assessment in a specific field or with a specific group of students in a generalised way may not necessarily contribute to best practice.

Drawing on ideas from Price *et al.* (2011), developing literacy relating to assessment among academics is critical to enabling informed debates on approaches that are implemented at the coal-face. Price and colleagues emphasise that "assessment standards reside in academic/professional communities" (2011, p. 484). As such, more collaboration between assessment experts and content experts on the development, design, implementation and then reporting of assessment practice would provide significant benefit in producing more empirically sound literature and in developing high quality assessment practice. This collaboration would benefit both types of practitioners in their professional development.

The higher education sector

Sectoral leadership should guide scholarship, collaboration and the improvement of assessment literacy. The 'fundamentals' identified in this literature review are elements that should be embraced at the sector level (as well as on the institution and practitioner fronts). Providing support to encourage collaborative development, implementation and evaluation of new and innovative assessment processes would benefit the sector greatly. In the UK, as a new era of assessment and quality assurance potentially begins to emerge in the implementation of new policies – particularly the Teaching Excellence Framework – the critical importance of ensuring assessment adheres to these fundamental principles has never been so great.

References

Ali, N., Rose, S. and Ahmed, L. (2015) Psychology students' perception of and engagement with feedback as a function of year of study. *Assessment and Evaluation in Higher Education*, **40** (4) 574–86. http://doi.org/10.1080/02602938.2014.936355

Aryadoust, V. (2015) Self- and Peer Assessments of oral presentations by first-year university students. *Educational Assessment*, **20** (3) 199–225. http://doi.org/10.1080/10627197.2015.1061989.

ATC21S (n.d.) Assessment and Teaching of 21st Century Skills [Internet]. Available from: http://www.atc21s.org/ [Accessed 8 July 2016].

Bamber, M. (2015) The impact on stakeholder confidence of increased transparency in the examination assessment process. *Assessment and Evaluation in Higher Education*, **40** (4) 471–87. http://doi.org/10.1080/02602938.2014.921662.

Bamkin, S. (2013) Giving students effective written feedback. *Assessment and Evaluation in Higher Education*, **38** (1) 125–9. http://doi.org/10.1080/02602938.2011.596273.

Bell, A., Mladenovic, R. and Price, M. (2013) Students' perceptions of the usefulness of marking guides, grade descriptors and annotated exemplars. *Assessment and Evaluation in Higher Education,* **38** (7) 769–88. http://doi.org/10.1080/02602938.2012.714738.

Bennett, R.E. (2011) Formative assessment: A critical review. *Assessment in Education: Principles, Policy and Practice*, **18** (1) 5–25. http://doi.org/10.1080/0969594X.2010.513678.

Biggs, J. (1996) Enhancing Teaching through Constructive Alignment. *Higher Education*, **32** (3) 347–64.

Biggs, J. (2003) *Aligning teaching for constructing learning* [Internet]. York: Higher Education Academy. Available from: https://www.researchgate.net/profile/John_Biggs3/publication/255583992_Aligning_Teaching_for_Constructing_Learning/links/5406ffe70cf2bba34c1e8153.pdf [Accessed 8 July 2016].

Biggs, J. and Tang, C. (2011) *Teaching for quality learning at university: What the student does.* (4th Ed.). Buckingham: Open University Press/McGraw Hill.

Bloxham, S. (2012) "You can see the quality in front of your eyes": Grounding academic standards between rationality and interpretation. *Quality in Higher Education*, **18** (2) 185–204. http://doi.org/10.1080/13538322.2012.711071

Bloxham, S., den-Outer, B., Hudson, J. and Price, M. (2016) Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment and Evaluation in Higher Education*, **41** (3) 466–81. http://doi.org/10.1080/02602938.2015.1024607

Bloxham, S., Hughes, C. and Adie, L. (2016) What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment and Evaluation in Higher Education*, **41** (4) 638–53. http://doi.org/10.1080/02602938.2015.1039932.

Boud, D. (2007) Reframing assessment as if learning were important. In Boud,D and Falchikov,N (ed), *Rethinking assessment in higher education*, London: Taylor and Francis, pp.14-26.

Boud, D., Cohen, R. and Sampson, J. (1999) Peer learning and assessment. *Assessment and Evaluation in Higher Education,* **24** (4) 413–26. http://doi.org/10.1080/0260293990240405.

Boud, D. and Falchikov, N. (2006) Aligning assessment with long-term learning. *Assessment and Evaluation in Higher Education,* **31** (4) 399–413. http://doi.org/10.1080/02602930600679050.

Boud, D., Lawson, R. and Thompson, D.G. (2015) The calibration of student judgement through self-assessment: Disruptive effects of assessment patterns. *Higher Education Research and Development*, **34** (1) 45–59. http://doi.org/10.1080/07294360.2014.934328.

Boud, D. and Molloy, E. (2013) Rethinking models of feedback for learning: The challenge of design. *Assessment and Evaluation in Higher Education*, **38** (6) 698–712. http://doi.org/10.1080/02602938.2012.691462.

Brown, S. and Race, P. (2012) Using effective assessment to promote learning. In Hunt, L. and Chalmers, D. (eds.), *University Teaching in Focus: A Learning-Centred Approach*. London: Routledge, Taylor and Francis Group, pp. 74–91.

Bruton, S. and Childers, D. (2016) The ethics and politics of policing plagiarism: A qualitative study of faculty views on student plagiarism and Turnitin®. *Assessment and Evaluation in Higher Education*, **41** (2) 316–30. http://doi.org/10.1080/02602938.2015.1008981.

Butz, N.T. and Askim-Lovseth, M.K. (2015) Oral communication skills assessment in a synchronous hybrid MBA programme: Does attending face-to-face matter for US and international students? *Assessment and Evaluation in Higher Education*, **40** (4) 624–39. http://doi.org/10.1080/02602938.2014.940577.

CALOHEE (2016, January 1). CALOHEE – Measuring and Comparing Achievements of Learning Outcomes in Higher Education in Europe [Internet]. Available from http://www.eurashe.eu/projects/calohee/ [Accessed 13 July 2016].

Caple, H. and Bogle, M. (2013) Making group assessment transparent: What wikis can contribute to collaborative projects. *Assessment and Evaluation in Higher Education*, **38** (2) 198–210. http://doi.org/10.1080/02602938.2011.618879.

Carless, D. (2007) Learning-oriented assessment: Conceptual bases and practical implications. *Innovations in Education and Teaching International*, **44** (1) 57–66. http://doi.org/10.1080/14703290601081332.

Carless, D. (2015a) *Excellence in University Assessment: Learning from award-winning practice*. London: Taylor and Francis.

Carless, D. (2015b) Exploring learning-oriented assessment processes. *Higher Education*, **69** (6) 963–76. http://doi.org/10.1007/s10734-014-9816-z.

Carruthers, C., McCarron, B., Bolan, P., Devine, A., McMahon-Beattie, U. and Burns, A. (2015) "I like the sound of that" – an evaluation of providing audio feedback via the virtual learning environment for summative assessment. *Assessment and Evaluation in Higher Education*, **40** (3) 352–70. http://doi.org/10.1080/02602938.2014.917145.

Carvalho, A. (2013). Students' perceptions of fairness in peer assessment: evidence from a problem-based learning course. *Teaching in Higher Education,* **18** (5) 491–505. http://doi.org/10.1080/13562517.2012.753051.

Chambers, K., Whannell, R. and Whannell, P. (2014) The use of peer assessment in a regional Australian university tertiary bridging course. *Australian Journal of Adult Learning*, **54** (1) 69.

Chew, E., Ding, S.L. and Rowell, G. (2015) Changing attitudes in learning and assessment: Cast-off "plagiarism detection" and cast-on self-service assessment for learning. *Innovations in Education and Teaching International*, **52** (5) 454–63. http://doi.org/10.1080/14703297.2013.832633.

Cho, Y.H. and Cho, K. (2011) Peer reviewers learn from giving comments. *Instructional Science*, **39** (5) 629–43.

Clements, M.D. and Cord, B.A. (2013) Assessment guiding learning: Developing graduate qualities in an experiential learning programme. *Assessment and Evaluation in Higher Education*, **38** (1) 114–24. http://doi.org/10.1080/02602938.2011.609314.

Coates, H. and Seifert, T. (2011) Linking assessment for learning, improvement and accountability. *Quality in Higher Education*, **17** (2) 179–94. http://doi.org/10.1080/13538322.2011.554308.

Craft, J. and Ainscough, L. (2015) Development of an electronic role-play assessment initiative in bioscience for nursing students. *Innovations in Education and Teaching International*, **52** (2) 172–84. http://doi.org/10.1080/14703297.2014.931241.

Cumming, J., Woodcock, C., Cooley, S.J., Holland, M.J.G. and Burns, V.E. (2015) Development and validation of the groupwork skills questionnaire (GSQ) for higher education. *Assessment and Evaluation in Higher Education*, **40** (7) 988–1001. http://doi.org/10.1080/02602938.2014.957642.

de la Harpe, B. and David, C. (2012) Major influences on the teaching and assessment of graduate attributes. *Higher Education Research and Development*, **31** (4) 493–510. http://doi.org/10.1080/07294360.2011.629361.

Denton, P. and Rowe, P. (2015) Using statement banks to return online feedback: Limitations of the transmission approach in a credit-bearing assessment. *Assessment and Evaluation in Higher Education*, **40** (8) 1095–1103. http://doi.org/10.1080/02602938.2014.970124.

BIS(2016). *Success as a knowledge economy: Teaching excellence, social mobility and student choice*. Department for Business, Innovation and Skills. [Internet] Available from: https://www.timeshighereducation.com/higher-education-white-paper-success-knowledge-economy [Accessed 8 July 2016]

Dijkstra, J., Galbraith, R., Hodges, B.D., McAvoy, P.A., McCrorie, P., Southgate, L.J., Van der Vleuten, C.P. and Schuwirth, L.W. (2012) Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Medical Education*, **12** (1) 20.

Dochy, F., Segers, M. and Sluijsmans, D. (1999) The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education,* **24** (3) 331–50. http://doi.org/10.1080/03075079912331379935.

Donovan, P. (2014) Closing the feedback loop: Physics undergraduates' use of feedback comments on laboratory coursework. *Assessment and Evaluation in Higher Education*, **39** (8) 1017–29. http://doi.org/10.1080/02602938.2014.881979.

El-Mowafy, A. (2014) Using peer assessment of fieldwork to enhance students' practical training. *Assessment and Evaluation in Higher Education*, **39** (2) 223–41. http://doi.org/10.1080/02602938.2013.820823.

Falchikov, N. and Blythman, M. (2001) *Learning together: Peer tutoring in higher education*. London: Routledge/Falmer.

Falchikov, N. and Goldfinch, J. (2000) Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, **70** (3) 287–322. http://doi.org/10.3102/00346543070003287.

Ferguson, P. (2011) Student perceptions of quality feedback in teacher education. *Assessment and Evaluation in Higher Education*, **36** (1) 51–62. http://doi.org/10.1080/02602930903197883.

Fletcher, R.B., Meyer, L.H., Anderson, H., Johnston, P. and Rees, M. (2012) Faculty and students conceptions of assessment in higher education. *Higher Education*, **64** (1) 119–33. http://doi.org/10.1007/s10734-011-9484-1.

Frost, J., de Pont, G. and Brailsford, I. (2012) Expanding assessment methods and moments in history. *Assessment and Evaluation in Higher Education*, **37** (3) 293–304. http://doi.org/10.1080/02602938.2010.531247.

Fuller, M., Henderson, S. and Bustamante, R. (2015) Assessment leaders' perspectives of institutional cultures of assessment: A Delphi study. *Assessment and Evaluation in Higher Education*, **40** (3) 331–51. http://doi.org/10.1080/02602938.2014.917369.

Giada, A., Giovanni, B. and Vincenza, C. (2014) A new indicator for higher education student performance. *Higher Education*, **68** (5) 653–68. http://doi.org/10.1007/s10734-014-9737-x.

Gielen, S., Dochy, F. and Onghena, P. (2011) An inventory of peer assessment diversity. *Assessment and Evaluation in Higher Education*, **36** (2) 137–55. http://doi.org/10.1080/02602930903221444.

Gielen, S., Dochy, F., Onghena, P., Struyven, K. and Smeets, S. (2011) Goals of peer assessment and their associated quality concepts. *Studies in Higher Education,* **36** (6) 719–35. http://doi.org/10.1080/03075071003759037.

Hagström, L. and Scheja, M. (2014) Using meta-reflection to improve learning and throughput: Redesigning assessment procedures in a political science course on power. *Assessment and Evaluation in Higher Education*, **39** (2) 242–52. http://doi.org/10.1080/02602938.2013.820822.

Handley, K. and Williams, L. (2011) From copying to learning: Using exemplars to engage students with assessment criteria and feedback. *Assessment and Evaluation in Higher Education*, **36** (1) 95–108. http://doi.org/10.1080/02602930903201669.

Hardy, J., Galloway, R., Rhind, S., McBride, K., Hughes, K. and Donnelly, R. (2016) Ask, answer, assess. York: Higher Education Academy.Available from: https://www.heacademy.ac.uk/sites/default/files/ask_answer_assess.pdf [Accessed 8 July 2016].

Hardy, P. and Aruguete, M. (2014) Needs assessment in STEM disciplines: Reliability, validity and factor structure of the Student Support Needs Scale (SSNS). *Assessment and Evaluation in Higher Education*, **39** (5) 553–62. http://doi.org/10.1080/02602938.2013.857642

Harland, T., McLean, A., Wass, R., Miller, E. and Sim, K.N. (2015) An assessment arms race and its fallout: High-stakes grading and the case for slow scholarship. *Assessment and Evaluation in Higher Education,* **40** (4) 528–41. http://doi.org/10.1080/02602938.2014.931927.

Hathcoat, J.D., Penn, J.D., Barnes, L.L.B. and Comer, J.C. (2016) A second dystopia in education: Validity issues in authentic assessment practices. *Research in Higher Education*, **57** (7) 892-912. http://doi.org/10.1007/s11162-016-9407-1.

HEA (2012) *A Marked Improvement: Transforming Assessment in Higher Education* [Internet]. York: Higher Education Academy. Available from: https://www.heacademy.ac.uk/sites/default/files/a_marked_improvement.pdf. [Accessed 8 July 2016].

HEA (2016a) Framework for transforming assessment in higher education. York: Higher Education Academy. Available from: https://www.heacademy.ac.uk/sites/default/files/downloads/transforming-assessment-in-he.pdf. [Accessed 8 July 2016].

HEA (2016b) *Invitation to tender - review of research literature*. York: Higher Education Academy.

Henderson, M., Selwyn, N., Finger, G. and Aston, R. (2015) Students' everyday engagement with digital technology in university: Exploring patterns of use and "usefulness". *Journal of Higher Education Policy and Management*, **37** (3) 308–19. http://doi.org/10.1080/1360080X.2015.1034424.

Hendry, G.D., Armstrong, S. and Bromberger, N. (2012) Implementing standards-based assessment effectively: Incorporating discussion of exemplars into classroom teaching. *Assessment and Evaluation in Higher Education*. **37** (2) 149–161. http://doi.org/10.1080/02602938.2010.515014.

Hernández, R. (2012) Does continuous assessment in higher education support student learning? *Higher Education*, **64** (4) 489–502. http://doi.org/10.1007/s10734-012-9506-7.

Higgins, R., Hartley, P. and Skelton, A. (2002) The conscientious consumer: Reconsidering the role of assessment feedback in student learning. *Studies in Higher Education*, **27** (1) 53–64. http://doi.org/10.1080/03075070120099368.

Holmes, N. (2015) Student perceptions of their learning and engagement in response to the use of a continuous e-assessment in an undergraduate module. *Assessment and Evaluation in Higher Education,* **40** (1) 1–14. http://doi.org/10.1080/02602938.2014.881978.

Homer, M., Darling, J. and Pell, G. (2012) Psychometric characteristics of integrated multi-specialty examinations: Ebel ratings and unidimensionality. *Assessment and Evaluation in Higher Education*, **37** (7) 787–804. http://doi.org/10.1080/02602938.2011.573843.

Hudson, J., Bloxham, S., den Outer, B. and Price, M. (2015) Conceptual acrobatics: Talking about assessment standards in the transparency era. *Studies in Higher Education*, 1–15. http://doi.org/10.1080/03075079.2015.1092130.

Hughes, G. (2011) Towards a personal best: a case for introducing ipsative assessment in higher education. *Studies in Higher Education*, **36** (3) 353–67. http://doi.org/10.1080/03075079.2010.486859.

Humburg, M. and van der Velden, R. (2015) Self-assessments or tests? Comparing cross-national differences in patterns and outcomes of graduates' skills based on international large-scale surveys. *Studies in Higher Education*, **40** (3) 482–504. http://doi.org/10.1080/03075079.2015.1004237.

Hunt, L., Koenders, A. and Gynnild, V. (2012) Assessing practical laboratory skills in undergraduate molecular biology courses. *Assessment and Evaluation in Higher Education*, 37 (7) 861–74. http://doi.org/10.1080/02602938.2011.576313.

Iannone, P. and Simpson, A. (2015) Students' preferences in undergraduate mathematics assessment. *Studies in Higher Education*, **40** (6) 1046–67. http://doi.org/10.1080/03075079.2013.858683.

Jackel, B. (2014) Item differential in computer based and paper based versions of a high stakes tertiary entrance test: diagrams and the problem of annotation [Internet]. In Dwyer, T. Purchase, H. and Delaney, A. (eds.) *Diagrammatic representation and inference* (Vol. 8578). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 71–77. Available from: http://link.springer.com/10.1007/978-3-662-44043-8 [Accessed 8 July 2016].

Jackling, B., Natoli, R., Siddique, S. and Sciulli, N. (2015) Student attitudes to blogs: A case study of reflective and collaborative learning. *Assessment and Evaluation in Higher Education*, **40** (4) 542–56. http://doi.org/10.1080/02602938.2014.931926.

Jago, C. (2009) A history of NAEP assessment frameworks. National Assessment Governing Board.

Jeffery, D., Yankulov, K., Crerar, A. and Ritchie, K. (2016) How to achieve accurate peer assessment for high value written assignments in a senior undergraduate course. *Assessment and Evaluation in Higher Education*, **41** (1) 127–40. http://doi.org/10.1080/02602938.2014.987721.

Jewell, S. (2013) The impact of assessment and feedback processes on student engagement in a research methods module. In Clouder, L. Broughan, C. Steventon, G. and Jewell, S. (eds.) *Improving student engagement and development through assessment: Theory and practice in higher education*. London: Taylor and Francis.

Jones, I. and Alcock, L. (2014) Peer assessment without assessment criteria. *Studies in Higher Education*, **39** (10) 1774–87. http://doi.org/10.1080/03075079.2013.821974.

Jones, N., Georghiades, P. and Gunson, J. (2012) Student feedback via screen capture digital video: Stimulating student's modified action. *Higher Education,* **64** (5) 593–607. http://doi.org/10.1007/s10734-012-9514-7.

Jonsson, A. (2014) Rubrics as a way of providing transparency in assessment. *Assessment and Evaluation in Higher Education*, **39** (7) 840–52. http://doi.org/10.1080/02602938.2013.875117.

Knauf, H. (2016) Reading, listening and feeling: audio feedback as a component of an inclusive learning culture at universities. *Assessment and Evaluation in Higher Education*, **41** (3) 442–49. http://doi.org/10.1080/02602938.2015.1021664.

Ko, S-S. (2014). Peer assessment in group projects accounting for assessor reliability by an iterative method. *Teaching in Higher Education*, **19** (3) 301–14. http://doi.org/10.1080/13562517.2013.860110.

Kun, A.I. (2016) A comparison of self versus tutor assessment among Hungarian undergraduate business students. *Assessment and Evaluation in Higher Education*, **41** (3) 350–67. http://doi.org/10.1080/02602938.2015.1011602.

Lai, K. (2012) Assessing participation skills: Online discussions with peers. *Assessment and Evaluation in Higher Education*, **37** (8) 933–47. http://doi.org/10.1080/02602938.2011.590878.

Lethbridge, K., Andrusyszyn, M-A., Iwasiw, C., Laschinger, H.K.S. and Fernando, R. (2013) Assessing the psychometric properties of Kember and Leung's reflection questionnaire. *Assessment and Evaluation in Higher Education*, **38** (3) 303–25. http://doi.org/10.1080/02602938.2011.630977.

Li, H., Xiong, Y., Zang, X., L. Kornhaber, M., Lyu, Y., Chung, K.S. and K. Suen, H. (2016) Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment and Evaluation in Higher Education*, **41** (2) 245–64. http://doi.org/10.1080/02602938.2014.999746.

Liu, N-F. and Carless, D. (2006) Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, **11** (3) 279–90. http://doi.org/10.1080/13562510600680582.

Liu, X. and Li, L. (2014) Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment. *Assessment and Evaluation in Higher Education*, **39** (3),275–92. http://doi.org/10.1080/02602938.2013.823540.

Malau-Aduli, B.S., Assenheimer, D., Choi-Lundberg, D. and Zimitat, C. (2014) Using computer-based technology to improve feedback to staff and students on MCQ assessments. *Innovations in Education and Teaching International*, **51** (5) 510–22. http://doi.org/10.1080/14703297.2013.796711.

McCarthy, J. (2015) Evaluating written, audio and video feedback in higher education summative assessment tasks. *Issues in Educational Research,* **25** (2) 153–69.

Mcdowell, L. and Sambell, K. (1999) Fitness for purpose in the assessment of learning: Students as stakeholders. *Quality in Higher Education*, 5 (2) 107–23. http://doi.org/10.1080/1353832990050202.

McGarr, O. and Clifford, A.M. (2013) "Just enough to make you take it seriously": Exploring students' attitudes towards peer assessment. *Higher Education*, **65** (6) 677–93. http://doi.org/10.1007/s10734-012-9570-z.

Medland, E. (2016) Assessment in higher education: Drivers, barriers and directions for change in the UK. *Assessment and Evaluation in Higher Education,* **41** (1) 81–96. http://doi.org/10.1080/02602938.2014.982072.

Mostert, M. and Snowball, J.D. (2013) Where angels fear to tread: Online peer-assessment in a large first-year class. *Assessment and Evaluation in Higher Education*, **38** (6) 674–86. http://doi.org/10.1080/02602938.2012.683770.

Mulder, R., Baik, C., Naylor, R. and Pearce, J. (2014) How does student peer review influence perceptions, engagement and academic outcomes? A case study. *Assessment and Evaluation in Higher Education*, **39** (6) 657–77. http://doi.org/10.1080/02602938.2013.860421.

Neus, J.L. (2011) Peer assessment accounting for student agreement. *Assessment and Evaluation in Higher Education*, **36** (3) 301–14. http://doi.org/10.1080/02602930903342315.

Newton, P. (2016) Academic integrity: a quantitative study of confidence and understanding in students at the start of their higher education. *Assessment and Evaluation in Higher Education*, **41** (3) 482–97. http://doi.org/10.1080/02602938.2015.1024199.

Nicol, D., Thomson, A. and Breslin, C. (2014) Rethinking feedback practices in higher education: A peer review perspective. *Assessment and Evaluation in Higher Education*, **39** (1) 102–22. http://doi.org/10.1080/02602938.2013.795518.

O'Donovan, B., Price, M. and Rust, C. (2004) Know what I mean? Enhancing student understanding of assessment standards and criteria. *Teaching in Higher Education,* **9** (3) 325–35.

O'Donovan, B., Price, M. and Rust, C. (2008). Developing student understanding of assessment standards: A nested hierarchy of approaches. *Teaching in Higher Education*, **13** (2) 205–17. http://doi.org/10.1080/13562510801923344.

Orsmond, P., Maw, S.J., Park, J.R., Gomez, S. and Crook, A.C. (2013) Moving feedback forward: Theory to practice. *Assessment and Evaluation in Higher Education,* **38** (2, 240–52. http://doi.org/10.1080/02602938.2011.625472.

Oxford Brookes University (2014) ASKe - Centre for Excellence in Teaching and Learning [Internet]. Available from: https://www.brookes.ac.uk/aske/ [Accessed 4 July 2016].

Patton, C. (2012) "Some kind of weird, evil experiment": Student perceptions of peer assessment. *Assessment and Evaluation in Higher Education,* **37** (6) 719–31. http://doi.org/10.1080/02602938.2011.563281.

Pearce, J., Edwards, D., Fraillon, J., Coates, H., Canny, B.J. and Wilkinson, D. (2015) The rationale for and use of assessment frameworks: Improving assessment and reporting quality in medical education. *Perspectives on Medical Education*, **4** (3) 110–18. http://doi.org/10.1007/s40037-015-0182-z.

Price, M., Carroll, J., O'Donovan, B. and Rust, C. (2011) If I was going there I wouldn't start from here: A critical commentary on current assessment practice. *Assessment and Evaluation in Higher Education*, **36** (4) 479–92. http://doi.org/10.1080/02602930903512883.

Prideaux, D. (2016) The emperor's new wardrobe: The whole and the sum of the parts in curriculum design. *Medical Education,* **50** (1) 10–12.

Purvis, A.J., Aspden, L.J., Bannister, P.W. and Helm, P.A. (2011) Assessment strategies to support higher level learning in blended delivery. *Innovations in Education and Teaching International*, **48** (1) 91–100. http://doi.org/10.1080/14703297.2010.543767.

Putting it All Together: Wireframing the Example App | Android Developers (n.d.). Available from: https://developer.android.com/training/design-navigation/wireframing.html#wireframe-digital [Accessed 6 July 2016].

Reddy, Y.M. and Andrade, H. (2010) A review of rubric use in higher education. *Assessment and Evaluation in Higher Education*, **35** (4) 435–48. http://doi.org/10.1080/02602930902862859.

Reimann, N. (2011) To risk or not to risk it: Student (non-)engagement with seen examination questions. *Assessment and Evaluation in Higher Education*, **36** (3) 263–79. http://doi.org/10.1080/02602930903311716.

Reinholz, D. (2016) The assessment cycle: A model for learning through peer assessment. *Assessment and Evaluation in Higher Education*, **41** (2) 301–15. http://doi.org/10.1080/02602938.2015.1008982.

Rhodes, T.L. (2016) Assessment: Growing up is a many-splendored thing. *Journal of Assessment and Institutional Effectiveness*, **5** (2) 101. http://doi.org/10.5325/jasseinsteffe.5.2.0101.

Richardson, M. and Healy, M. (2013) Beneath the patchwork quilt: Unravelling assessment. *Assessment and Evaluation in Higher Education*, **38** (7) 847–56. http://doi.org/10.1080/02602938.2012.731036.

Richardson, S. and Coates, H. (2014) Essential foundations for establishing equivalence in cross-national higher education assessment. *Higher Education*, **68** (6) 825–36. http://doi.org/10.1007/s10734-014-9746-9.

Rust, C., Price, M. and O'Donovan, B. (2003) Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment and Evaluation in Higher Education*, **28** (2) 147–64. http://doi.org/10.1080/02602930301671.

Sadler, D.R. (1985). The Origins and Functions of Evaluative Criteria. *Educational Theory*, **35** (3) 285–97. http://doi.org/10.1111/j.1741-5446.1985.00285.x.

Sadler, D.R. (1998) Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy and Practice*, **5** (1) 77–84. http://doi.org/10.1080/0969595980050104.

Sadler, D.R. (2013) Making competent judgments of competence. In Blomeke, S., Zlatkin-Troischanskaia, O., Kuhn, C. and Fege, J. (eds.) *Modeling and measuring competencies in higher education: Tasks and challenges*. Rotterdam: Sense Publishers, pp. 13–27.

Sadler, D.R. (2014) The futility of attempting to codify academic achievement standards. *Higher Education,* **67** (3) 273–88. http://doi.org/10.1007/s10734-013-9649-1.

Sambell, K., McDowell, L. and Montgomery, C. (2013) *Assessment for learning in higher education*. New York: Routledge.

Sasanguie, D., Elen, J., Clarebout, G., Noortgate, W.V. den, Vandenabeele, J. and De Fraine, B. (2011) Disentangling instructional roles: The case of teaching and summative assessment. *Studies in Higher Education*, **36** (8) 897–910. http://doi.org/10.1080/03075079.2010.482206.

Schuwirth, L.W.T. and Van der Vleuten, C.P.M. (2011) Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher*, **33** (6) 478–85. http://doi.org/10.3109/0142159X.2011.565828.

Scoles, J., Huxham, M. and McArthur, J. (2013) No longer exempt from good practice: Using exemplars to close the feedback gap for exams. *Assessment and Evaluation in Higher Education*, **38** (6) 631–45. http://doi.org/10.1080/02602938.2012.674485.

Sharrock, G. (2015) Making sense of the MOOCs debate. *Journal of Higher Education Policy and Management*, **37** (5) 597–609. http://doi.org/10.1080/1360080X.2015.1079399.

Shulruf, B., Poole, P., Jones, P. and Wilkinson, T. (2015) The objective borderline method: A probabilistic method for standard setting. *Assessment and Evaluation in Higher Education*, **40** (3) 420–38. http://doi.org/10.1080/02602938.2014.918088.

Smith, C.D., Worsfold, K., Davies, L., Fisher, R. and McPhail, R. (2013). Assessment literacy and student learning: The case for explicitly developing students "assessment literacy". *Assessment and Evaluation in Higher Education*, **38** (1) 44–60. http://doi.org/10.1080/02602938.2011.598636.

Snowball, J.D. (2014) Using interactive content and online activities to accommodate diversity in a large first year class. *Higher Education*, **67** (6) 823–38. http://doi.org/10.1007/s10734-013-9708-7.

Spatar, C., Penna, N., Mills, H., Kutija, V. and Cooke, M. (2015) A robust approach for mapping group marks to individual marks using peer assessment. *Assessment and Evaluation in Higher Education*, **40** (3) 371–89. http://doi.org/10.1080/02602938.2014.917270.

Spronken-Smith, R., Walker, R., Batchelor, J., O'Steen, B. and Angelo, T. (2012) Evaluating student perceptions of learning processes and intended learning outcomes under inquiry approaches. *Assessment and Evaluation in Higher Education*, **37** (1) 57–72. http://doi.org/10.1080/02602938.2010.496531.

Stödberg, U. (2012) A research review of e-assessment. *Assessment and Evaluation in Higher Education*, **37** (5) 591–604. http://doi.org/10.1080/02602938.2011.557496.

Stoddart, P. (2015) Using educational technology as an institutional teaching and learning improvement strategy? *Journal of Higher Education Policy and Management*, **37** (5) 586–96. http://doi.org/10.1080/1360080X.2015.1079401.

Suhre, C.J.M., Jansen, E.P.W.A. and Torenbeek, M. (2013) Determinants of timely completion: The impact of bachelor's degree programme characteristics and student motivation on study progress. *Higher Education Research* *and Development*, **32** (3) 479–92. http://doi.org/10.1080/07294360.2012.684374.

Sutherland-Smith, W. (2014) Legality, quality assurance and learning: Competing discourses of plagiarism management in higher education. *Journal of Higher Education Policy and Management*, **36** (1) 29–42. http://doi.org/10.1080/1360080X.2013.844666.

Trenholm, S., Alcock, L. and Robinson, C. (2016) The instructor experience of fully online tertiary mathematics: A challenge and an opportunity. *Journal for Research in Mathematics Education,* **47** (2) 147. http://doi.org/10.5951/jresematheduc.47.2.0147.

Tam, M. (2014) Outcomes-based approach to quality assessment and curriculum improvement in higher education. *Quality Assurance in Education*, **22** (2) 158–68. http://doi.org/10.1108/QAE-09-2011-0059.

Taras, M. (2010) Student self-assessment: Processes and consequences. *Teaching in Higher Education*, **15** (2) 199–209. http://doi.org/10.1080/13562511003620027.

Topping, K. (1998) Peer assessment between students in colleges and universities. *Review of Educational Research*, **68** (3) 249–76.

Tucker, R. (2013) The architecture of peer assessment: Do academically successful students make good teammates in design assignments? *Assessment and Evaluation in Higher Education*, **38** (1) 74–84. http://doi.org/10.1080/02602938.2011.604122.

Turner, K., Roberts, L., Heal, C. and Wright, L. (2013) Oral presentation as a form of summative assessment in a master's level PGCE module: The student perspective. *Assessment and Evaluation in Higher Education*, **38** (6) 662–73. http://doi.org/10.1080/02602938.2012.680016.

Ünaldı, I. (2016) Self and teacher assessment as predictors of proficiency levels of Turkish EFL learners. *Assessment and Evaluation in Higher Education*, 41 (1) 67–80. http://doi.org/10.1080/02602938.2014.980223.

Van Der Vleuten, C.P.M., Schuwirth, L.W.T., Driessen, E.W., Dijkstra, J., Tigelaar, D., Baartman, L.K.J. and van Tartwijk, J. (2012) A model for programmatic assessment fit for purpose. *Medical Teacher*, 34 (3) 205–14. http://doi.org/10.3109/0142159X.2012.652239.

Van Der Vleuten, C.P.M., Schuwirth, L.W.T., Driessen, E.W., Govaerts, M.J.B. and Heeneman, S. (2015) Twelve tips for programmatic assessment. *Medical Teacher*, **37** (7) 641–46. http://doi.org/10.3109/0142159X.2014.973388.

Vardi, I. (2013) Effectively feeding forward from one written assessment task to the next. *Assessment and Evaluation in Higher Education,* 38 (5) 599–610. http://doi.org/10.1080/02602938.2012.670197.

Varsavsky, C. and Rayner, G. (2013) Strategies that challenge: Exploring the use of differentiated assessment to challenge high-achieving students in large enrolment undergraduate cohorts. *Assessment and Evaluation in Higher Education*, **38** (7) 789–802. http://doi.org/10.1080/02602938.2012.714739.

Vaughan, A.L., Lalonde, T.L. and Jenkins-Guarnieri, M.A. (2014) Assessing student achievement in large-scale educational programs using hierarchical propensity scores. *Research in Higher Education*, **55** (6) 564–80. http://doi.org/10.1007/s11162-014-9329-8.

Walker, S. and Hobson, J. (2014) Interventions in teaching first-year law: Feeding forward to improve learning outcomes. *Assessment and Evaluation in Higher Education*, **39** (3) 326–38. http://doi.org/10.1080/02602938.2013.832728.

Wang, X., Su, Y., Cheung, S., Wong, E. and Kwong, T. (2013) An exploration of Biggs' constructive alignment in course design and its impact on students' learning approaches. *Assessment and Evaluation in Higher Education*, **38** (4) 477–91. http://doi.org/10.1080/02602938.2012.658018.

Weisler, S. (2016) Some perspectives on assessment of student learning. *Journal of Assessment and Institutional Effectiveness*, **5** (2) 117. http://doi.org/10.5325/jasseinsteffe.5.2.0117.

Weurlander, M., Söderberg, M., Scheja, M., Hult, H. and Wernerson, A. (2012) Exploring formative assessment as a tool for learning: Students' experiences of different methods of formative assessment. *Assessment and Evaluation in Higher Education*, **37** (6) 747–60. http://doi.org/10.1080/02602938.2011.572153.

Williams, P. (2014) Squaring the circle: A new alternative to alternative-assessment. *Teaching in Higher Education*, **19** (5) 565–77. http://doi.org/10.1080/13562517.2014.882894.

Wilson, M. J., Diao, M.M. and Huang, L. (2015) "I"m not here to learn how to mark someone else's stuff': An investigation of an online peer-to-peer review workshop tool. *Assessment and Evaluation in Higher Education*, **40** (1) 15–32. http://doi.org/10.1080/02602938.2014.881980.

Wimshurst, K. and Manning, M. (2013) Feed-forward assessment, exemplars and peer marking: Evidence of efficacy. *Assessment and Evaluation in Higher Education*, **38** (4) 451–65. http://doi.org/10.1080/02602938.2011.646236.

Wyatt-Smith, C. and Klenowski, V. (2013) Explicit, latent and meta-criteria: Types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy and Practice*, **20** (1) 35–52. http://doi.org/10.1080/0969594X.2012.725030.

Xie, Q., Zhong, X., Wang, W-C. and Lim, C.P. (2014) Development of an item bank for assessing generic competences in a higher-education institute: A Rasch modelling approach. *Higher Education Research and Development*, **33** (4) 821–35. http://doi.org/10.1080/07294360.2013.863847.

Yorke, M. (2003) Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education*, **45** (4) 477–501.

Yorke, M. (2011) Summative assessment: dealing with the "measurement fallacy". *Studies in Higher Education*, **36** (3) 251–73. http://doi.org/10.1080/03075070903545082.

Yorke, M., Bridges, P. and Woolf, H. (2000) Mark distributions and marking practices in UK higher education. *Active Learning in Higher Education*, **1** (1) 7–27.

Yucel, R., Bird, F.L., Young, J. and Blanksby, T. (2014) The road to self-assessment: Exemplar marking before peer review develops first-year students' capacity to judge the quality of a scientific report. *Assessment and Evaluation in Higher Education*, **39** (8) 971–86. http://doi.org/10.1080/02602938.2014.880400.

Contact us

+44 (0)1904 717500   enquiries@heacademy.ac.uk
Innovation Way, York Science Park, Heslington, York, YO10 5BR
Twitter: @HEAcademy   www.heacademy.ac.uk

The Higher Education Academy (HEA) is the national body for learning and teaching in higher education. We work with universities and other higher education providers to bring about change in learning and teaching. We do this to improve the experience that students have while they are studying, and to support and develop those who teach them. Our activities focus on rewarding and recognising excellence in teaching, bringing together people and resources to research and share best practice, and by helping to influence, shape and implement policy - locally, nationally, and internationally.

To request copies of this report in large print or in a different format, please contact the communications office at the Higher Education Academy: 01904 717500 or communications@heacademy.ac.uk

The Higher Education Academy is a company limited by guarantee registered in England and Wales no. 04931031. Registered as a charity in England and Wales no. 1101607. Registered as a charity in Scotland no. SC043946.