Robert Harland

Testing keywords internationally to define and apply undergraduate assessment standards in art and design

Abstract
What language should be featured in assessment standards for international students? Have universities adjusted their assessment methods sufficiently to match the increased demand for studying abroad? How might art and design benefit from a more stable definition of standards? These are some questions this paper seeks to address by reporting the results of recent pedagogic research at the School of the Arts, Loughborough University, in the United Kingdom. Language use is at the heart of this issue, yet it is generally overlooked as an essential tool that links assessment, feedback and action planning for international students. The paper reveals existing and new data that builds on research since 2009, aimed at improving students' assessment literacy. Recommendations are offered to stimulate local and global discussion about keyword use for defining undergraduate assessment standards in art and design.

Keywords: assessment standards, internationalisation, art and design, keywords

Introduction

Students repeatedly say they want more meaningful and constructive feedback (Rae & Cochrane, 2008, p. 145) and they have difficulty learning from feedback (Orsmond et al., 2013, p. 241). As for students who study in a second language, what chance do they have of connecting assessment criteria, standards, feedback, reflection and action planning as parts of an assessment cycle? Unsurprisingly, the link between reflection and action planning is little understood by students (Parkin et al., 2012, p. 969). This paper weaves together issues that illuminate aspects of this problem by reporting the results of action research at Loughborough University in the United Kingdom (UK). This study is set in the context of an emerging "re-internationalisation" agenda in the UK since the early 1990s, driven by economic growth. It challenges the author's previously held assumptions about keyword use in the application of assessment criteria for an international audience. The findings reported in this article raise important questions about how to relate verbal descriptors to class and grade indicators in assessment. Furthermore, different approaches to assessment level indicators at national and international levels are revealed to show considerable variations among universities.

The paper includes a review of recent focus group activities on developing and testing a keyword strategy for assessment standards to support written criteria statements that help guide tutors and tutees towards a collective understanding about levels of achievement. Focus groups have been undertaken in the UK, the Netherlands and Norway, bringing an international dimension to what began in 2009 as an internal evaluation exercise. This research is set in the context of the development of internationalisation, emphasising the need for language use to be more carefully considered and explained as an enabler of learning by international students.

Art and design (a conjoined phrase used here) provides the backdrop for the research. The formative and summative assessment in art and design differs from the "stereotypical" view of assessment that limits the dialogue between the student and the assessor to the student's response to the assessment task (Price et al., 2012, p. 19). Art and design student outputs in the UK are mainly coursework related; it is common for the student and the tutor to hold discussions through critique sessions and informal studio settings. At anytime, the tutor may offer a verbal commentary on the development of a student project, often in the form of

qualitative judgement statements that a student may interpret as an indication of progress and standards. Coursework output in art and design tends to be "divergent" and allows students to demonstrate what Sadler refers to as "sophisticated cognitive abilities, integration of knowledge, complex problem solving, critical opinion, lateral thinking and innovative action" (2009, p. 160). The resulting "artwork" may be assessed in a studio setting through a discussion among lecturers "situated within its disciplinary context", looking for anticipated and unanticipated creative solutions (Orr, 2007). Implicit in this work is the "wow" factor, something difficult to define in assessment criteria but is said to include "creativity, originality, inventiveness, inspiration, ingenuity, freshness and vision" (Gordon, 2004, as cited in Orr, 2007).

Aims of the paper

Letters, numbers, symbols and words are used to code, order and communicate grades in marking systems (Schünemann et al., 2003, p. 677). In assessment criteria, grades align with and are supported by descriptors that characterise levels of achievement. The descriptor provides some explanation and guidance to the student and the tutor about what must be evidenced for attaining the level. This method places high importance on consistent language use if students are to understand assessment criteria, tutors are to use assessment criteria when marking student work and providing feedback, and then students are to develop action plans. In support of this approach, Woolf regards language as central to "a higher level of shared understanding" among "students, tutors and other stakeholders" to fulfil the "educational value" of assessment criteria (2004, p. 479).

The desire for consistent language use becomes more complicated in terms of the aspiration for internationalisation and the need for universities to attract students from abroad. This paper reports on attempts to develop assessment standards that support the links among marking, feedback and action planning for national and international students. The aim is to highlight how keywords used in assessment discourse can assist in the process but at the same time, present problems when applied internationally. Some recommendations are made about how keywords might be used to indicate standards and link assessment criteria, feedback and student response. One intention is to stimulate discussion within art and design about the use of keywords in assessment "rubrics", acknowledging that words such as "qualities", "criteria" and "standards" are used interchangeably (Sadler, 2009, p. 163) to reflect comparative judgement about the work being assessed. In this paper, the term "criteria" means a fixed set of statements within a rubric about knowledge and understanding, subject-specific cognitive skills, subject-specific practical skills and key/transferable skills. "Standards" indicate the level of achievement matched against these criteria. "Qualities" refer to the comparative level of distinction or excellence.

Methods

The research builds on previous work (Harland & Sawdon, 2012) and resembles action research. It utilises "evaluative procedures" in a desire to improve the criteria-based assessment methodology through "continuing professional development" and "behaviour modification" (Cohen et al., 2007). As well as content analysis, focus group activities tested keyword use in the application of assessment criteria in national and international contexts. The focus group method uses "stimuli" (topics and visual aids) provided by the researcher (Silverman, 2005, p. 378) to generate probability samples for making generalisations. The findings reported here are drawn from a simple random data set (Cohen et al., 2007, p. 110–111).

Ways to indicate standards in assessment

As students move from one assessment regime to another, their understanding of assessment must also change. Pass or fail is the simplest way to indicate the outcome of an assessment. A classification, verbal descriptor or literal grade generally confers more detailed attainment levels. Numerals or letters usually indicate class and grade, whereas the nomenclature of "good", "very good", "excellent" or "outstanding" distinguishes among "levels of competence" (Davies, 2012, p. 2). The correlations among class, verbal descriptor and literal grade indicate the way these different codes communicate similar levels of achievement. Table 1 shows how a 2:1 undergraduate degree classification may be the equivalent of the literal grade of A–. Both may be described as "very good". In the same grouping, levels of attainment are indicated by as few as six (class) or as many as 13 (literal grade) options, with verbal descriptors registering seven divisions.

Class	Verbal descriptor	Literal Grade
1	excellent	A+
		A
2:1	very good	A–
		B+
2:2	good	B
		B–
		C+
3	moderate	C
		C–
P	marginal	D+
	pass	D
		D–
F	fail	E

Table 1. Comparison of class, verbal description and literal grade indicators (Brown, 1997, p. 75).

These indicators offer three ways to communicate the achievement level but there are more. Collins (2004, p. 24) identifies five approaches:

> 1) pass or fail (commonly used in competency-based testing),
> 2) letter grades (e.g., A, B, C, D, etc. with and without plus and minus variations),
> 3) numerical grades (e.g., 1 = excellent, 2 = very good, etc.),
> 4) numerical scores (e.g., an achieved score out of a predetermined whole – 12 out of 20) and
> 5) percentage point marks.

In an international review of assessment conventions, Collins discusses regional and national differences. For example, in the UK assessment, conventions are split into "full range percentage marks, grade based marks and what one might call hybrid grade percentage systems" (Collins, 2004, p. 27). The pass threshold is generally set at 40% for undergraduate and 50% for postgraduate studies. Beyond the UK, most European practice is said to be grade based. An exception is Germany where the predominant system is numerical, from 1 (high) to 5 (low), with an accompanying three subdivisions for each number for greater accuracy. Hungary, Sweden and Switzerland have similar systems but in the reverse rank order from 1 (low) to 5 (high). In the European Community, attempts have been made to translate these different approaches by introducing a European Credit Transfer and Accumulation System (ECTS) for Erasmus students wishing to study abroad (European Commission, 2009). The system proposes a five-segment grading scale (top 10%, next 25%, next 30%, next 25% and lowest 10%, corresponding to A, B, C, D and E, respectively), which does not indicate fail or

pass thresholds. Further afield, the United States of America (USA) seemingly favours letter grades, setting the pass threshold at 60% (higher than that of the UK). Specifically, Collins reports that York University in the USA employs a 10-grade letter scale with descriptors (2004, p. 30). At the median, Canada and Australia set the pass threshold at 50%, with the exception of The University of Calgary, The Royal Melbourne Institute of Technology, The University of Technology Sydney, University of New South Wales and the University of South Australia.

These sources suggest at least seven scales for registering achievement levels, as summarised in Table 2. There are more levels if the various configurations of the percentage scale are considered, broken into as little as five divisions to match degree classifications in the UK (e.g., 0–39, 40–49, 50–59, 60–69 and 70–100) and as many as 16 (e.g., 1–19, 20–29, 30–39, 40 pass, 41–43, 44–46, 47–49, 50–53, 54–56, 57–59, 60–63, 64–66, 67–69, 70–71, 72–74 and 75+) reported by Collins (2004, p. 48). Levels range from 5, 6, 10, 13 to 17, with pass thresholds usually just below the mid-point, exceptionally in the middle or slightly above. Some systems are used in combination. For example, percentage scaling may also align with verbal descriptors, classification bands or a points gauge; literal indicators may accompany verbal descriptors or a points gauge. The verbal description is of primary interest in this paper because it is language based and therefore most closely related to what can be read as assessment criteria.

Type	Indicator	Division
Verbal	| Excellent | Very Good | Good | Moderate | Marginal Pass | Fail |	6
Class	| 1 | 2:1 | 2:2 | 3 | P | F |	6
Literal	| A+ | A | A– | B+ | B | B– | C+ | C | C– | D+ | D | D– | E |	13
Percentage	| 100 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 | 0 |	10
Point	| 16 | 15 | 14 | 13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |	17
Fraction	| 1 | 1.3 | 1.7 | 2 | 2.3 | 2.7 | 3 | 3.3 | 3.7 | 4 | 4.3 | 4.7 | 5 |	13
Numerical	| 17–20 | 13–16 | 9–12 | 5–8 | 0–4 |	5
Combined	| 90–100% (A) | 65–89% (B) | 35–64% (C) | 10–34% (D) | 0–9% (E) |	5
	High . . . . Low	

Table 2. Examples of scales for representing achievement levels in assessment.

Alignment of verbal descriptors with percentage levels of achievement

Table 1 shows how "good" corresponds to a 2:2 degree classification or a B literal indicator, but what does this mean to a student? The answers to this question depend on the context and a range of adjectives and synonyms to help clarify the meaning. To a craftsperson, "good" may suggest skilled, a priest may interpret it as virtuous and to a parent, obedient comes to mind. However, the meaning of individual words is less of a concern in this paper. Of more interest are the relationships among words in a hierarchy of standards. What terms help substantiate a word such as "good" when aligned with assessment criteria, regardless of the typology being used to register a mark? The various scaling options available mean that six descriptors – excellent, very good, good, moderate, marginal pass and fail – comprise an insufficient, coarse scale. A finer-grained version is needed to cope with art and design assessment, which may require what Hornby describes as "matters of judgement and interpretation" (2003, p. 439).

Reflecting on establishing a keyword approach to applying assessment criteria

Recent work at Loughborough University School of the Arts extended the range of verbal descriptors in Table 1 – excellent, very good, good, moderate, marginal, pass and fail – by assigning equal percentage divisions, from 0 to 100%, to 10 words. Some of the reasons for this method relate to undergraduate external examiner comments that the full range of marks is under utilised, a common criticism in qualitative assessment using "high validity/low reliability instruments" (Hornby, 2003, p. 439). By establishing 10 words, the intention was also to encourage more consistent use of formative and assessment feedback language among marking tutors (Harland & Sawdon, 2012). This approach meant introducing additional words to further differentiate among underused grade bands in the first class (70% and above) and fail (below 40%) brackets used in UK assessment matrices, representing more than two-thirds of available marks. A working group of six academic staff members developed a set of generic verbal descriptors for marking both written (e.g., essay) and practical (e.g., artefact) outputs by art and design students. As part of the development process, an informal consultation with the staff and the students took place in a small focus group to provide quick feedback. The outcome supported the word recommendations corresponding to a hierarchy of numerical grading. See Table 3. Word options were sourced to support the writing and presentation of generic assessment standards in student handbooks as a guide for the application of assessment criteria across four headings commonly used in the UK: knowledge and understanding, subject-specific cognitive skills, subject-specific practical skills and key/transferable skills. For example, applying assessment criteria for knowledge and understanding in the 60–69% bracket is supported by the statement, "Very good acquisition of knowledge and understanding, with an appropriately critical and controlled approach to your chosen subject".

90–100	Outstanding
80–89	Excellent
70–79	Rigorous
60–69	Very Good
50–59	Good
40–49	Satisfactory
30–39	Marginal
20–29	Insubstantial
10–19	Insufficient
0–9	Deficient

Table 3. A hierarchy of key terms representing 10 percentage divisions (Harland & Sawdon, 2012, p. 79).

A similar approach was adopted after a content analysis review of language use in the art and design assessment criteria at nine universities in the UK. The data contained familiar words (e.g., excellent) and some that could be interpreted as metaphorical (e.g., sound). See Table 4.

	A	B	C	D	E
90–100	Extensive In-depth Coherent Detailed Theoretical	Extensive Deep Excellent Inventive Ambitious	Outstandingly high quality Originality	A Outstanding High Quality Originality Excellent Technical competence and innovation	Outstanding Original Complex Rigorous New Insights
80–89	Intellectual	Exceptional Highly fexible			Publishable Extensive Effective
70–79	Comprehensive Coherent Wide-ranging Specialist techniques Systematic		Exceptional Distinguished Authoritative Methodical Critical Skilled	A Excellent Exceptional Distinguished Very good quality	Confident Appropriate Organised Structured
60–69	Thorough In-depth Advanced Critical Creative	Sound Very good High level Inventive Strong Quite flexible	Very good Greater insight Good qiality Sound Ordering Appropriate	B Greater insight and originality Good quality	Very good Analytical Critical
50–59	Sound Focused	Sound Good Thorough Competent	Insight Satisfactory Relevance Standard Accurate	Sound quality	Satisfactory
40–49	Adequate	Acceptable Competent Moderate Sometimes flexible	Satisfactory Reasonable Competent Familiarity Derivative	D Satisfactory Standard	D Satisfactory Standard
30–39	Limited Unfocused Inadequate Little or no coherence	Limited Partial Basic Unconvincing Limited flexibility	Misinterpreted Misdirected Misunderstood Poor	F Fail	Incomplete Deficient Inadequate
20–29		Insufficient Unacceptable Too narrow Poor Inflexible	Little engagement Irrelevancies Minimal evidence	Clear fail	Misunderstanding Lacks coherence
10–19			Poor quality Minimal effort Little relevance	Very poor fail	
0–9			Virtually nothing of relevance, depth or merit		

Table 4. (Part 1) Assessment criteria keyword analysis from nine higher educational institutions (Harland & Sawdon, 2012, p. 74).

	F	G	H	I
90–100	Significant Original Transcend Transform	Outstanding Exceptionally high standard Trivial defects	Unfamiliar conceptual territories Unified Discernment Technical/craft	High order Appropriate Developed capacity Very high standard
80–89	Flair Thorough	Excellent Very minor defects	Persuasive Compelling Responsibility Well informed Wide ranging Accurate	Rigorous High quality
70–79	Challenging Exceptional Evaluative	Very good Few minor defects	Extensive Risk taking Selective Organised Reflective	Originality
60–69	Coherent Familiarity Evaluation Clarity	Very good – minor defects Generally very good Notable defects	Professional Accurate Checking Testing Finishing	Above average Synthesis Sound
50–59	Satisfactory	Good Creditable Generally sound Number of shortcomings	Good judgement Reflection Adequate References Familiar ideas	Average Clear
40–49	Some knowledge Limited Reiteration Descriptive	Fair Shortcomings Satisfactory On borderline	Awareness Collaborative Independent	Limited Adequate
30–39	 Compensated pass Weaknesses	Unsatisfactory Significant shortcomings	Insufficient Rudimentary Limited knowledge Little judgement Lack clarity Sporadic	Inadequate Very limited Poor
20–29	Confused Inadequate	Very poor Exceedingly poor	Little evidence Poor judgement Very limited Ineffective	
10–19	 Incomplete Inaccurate		No awareness	
0–9	Incomplete Fragmentary Zero response			

Table 4. (Part 2) Assessment criteria keyword analysis from nine higher educational institutions (Harland & Sawdon, 2012, p. 74).

The staff and the students contributed to selecting 10 keywords for use in the application of assessment criteria. The initial process consisted of academic staff who taught practical, historical and theoretical classes, forming a working group from within a larger learning and teaching committee. The group aimed to review the language used in assessment criteria across 9 UK universities, extract useful words, dismiss others and introduce new ones to fit a 10-part percentage division matrix. A mix of familiarity, habit and proposition informed the creation of a new list (with some words in reserve) that could then be shared more widely. The new list, as shown in Table 3, was then tested at a staff-student focus group whose attendees had not previously contributed to the process. Using visual stimuli, the facilitators presented the focus group participants with the new list, which was randomly assembled. See Figure 1. The participants were then asked to rank the words in order from 1 (low) to 10 (high). The outcome of the exercise, whilst using a small sample size, provided quick feedback to the working group, enough for publication in student handbooks as part of a

revised set of assessment standards. (For further reading, see Harland & Sawdon, 2012). However, this initiative provided very limited endorsement of working group recommenddations. Consequently, since then, the same basic exercise had been repeated in three focus groups with national and international audiences. The following section summarises the results.
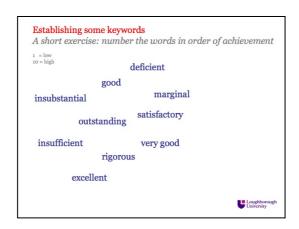

Figure 1. Random keywords for assessment ranking (Harland & Sawdon, 2012).

Testing keywords in national and international contexts

The first focus group was held at the Group for Learning in Art and Design (GLAD) 2012 conference at Kingston University in the UK, with 11 academic staff members as participants. No prior explanation of what the words meant was provided and the participants were left alone to use their own interpretation as they performed the exercise individually. The results showed that most words were ranked one level from the predetermined position, some occasionally higher by two levels. The most consistently misplaced words were "insubstantial" and "insufficient", the former being accurately matched in only four out of 11 responses.

A second focus group was conducted with 45 undergraduate and postgraduate students and five academic staff members at the St Joost Academy of Art in Breda, the Netherlands, in February 2012. The results from the Dutch event presented a more varied data set from respondents who did not speak English as their first language, some of whom were from outside the Netherlands (e.g., Russia). The exact breakdown by nationality is unknown but the majority of the participants were Dutch. The St Joost results revealed a less reliable match between percentage and grade among international participants. This finding was further emphasised by anecdotal feedback during the focus group when some students claimed that certain words do not translate well between assessment cultures. For example, a Russian student confessed that the word "outstanding" may be difficult for Russian speakers as it suggests that the work being assessed stands physically (not intellectually) apart from the rest of the assignments and therefore may not be assessed. The academic staff participants also debated whether the idea of "rigorous" has a Dutch equivalent since it seems to lack a direct translation. This issue clearly suggests potential problems associated with a keyword approach to assessment criteria for international staff and students.

The data from St Joost revealed that building a hierarchy of words with less than 20% variability is difficult in an international context. In fact, there can be as much as 70% difference in the hierarchical placement of words within the predetermined set. The degree of accuracy proved to be very low, compared to the recommendations made by the Loughborough working group. A lack of fluency in English is a possible explanation for this discrepancy. However, it can be assumed that some words (e.g., excellent) are generally understood by most individuals with a basic understanding of English. The reliability of "blindly" ranking

keywords is shown in Table 5; the degree of accuracy varies between the least reliable score of 30% for the word "rigorous" and the most reliable score of 74% for the word "satisfactory". The variability in matching keywords to their respective predetermined rankings cast doubt on the relationship between keywords and the achievement levels they represent, especially for international students who may have limited initial understanding of the application of assessment criteria.

Rank	Verbal descriptor	Reliability
10	outstanding	60%
9	excellent	58%
8	rigorous	30%
7	very good	40%
6	good	62%
5	satisfactory	74%
4	marginal	64%
3	insubstantial	46%
2	insuffiicient	58%
1	deficient	62&

Table 5. Variability of "blindly" ranking keywords according to a predetermined order in an international context.

Most recently, the same keywords were tested with a focus group at a Design Research Society/Cumulus conference "workshop" in Oslo in May 2013. A call for participation attracted seven participants from Australia, Austria, China, England, Iran, Mexico and Venezuela. The results by nationality are shown in Table 6. This micro-sample revealed that "excellent" is most consistently placed in the top two positions, with "outstanding" nearly as recognisable in terms of high attainment levels. "Very good" is similarly ranked one or two levels below the top two words, with "good" or "rigorous" consistently positioned in sixth, seventh or eighth, with the exception of the Iranian participant, who also ranked "outstanding" as a mid-level achievement. "Satisfactory" is consistent in five responses but "marginal", "insubstantial", "insufficient" and "deficient" are the most randomly positioned words. There is less variability here than in that of the St Joost sample, especially with the higher-level keywords, although more so than those in the initial focus group at Loughborough and at the GLAD conference.

Rank	Australia	Austria	China	England	Iran	Mexico	Venezuela
10	Outstanding	Outstanding	Excellent	Outstanding*	Excellent	Excellent	Outstanding
9	Excellent	Excellent	Outstanding	Excellent*	Very good	Outstanding	Excellent
8	Rigorous	Very good	Rigorous	Very good*	Good	Very good	Rigorous
7	Very good	Good	Very good	Good	Satisfactory	Good	Very good
6	Good	Rigorous	Good	Insubstantial	Outstanding	Rigorous	Good
5	Satisfactory	Satisfactory	Satisfactory	Deficient	Deficient	Satisfactory	Satisfactory
4	Marginal	Marginal	Insufficient	Insufficient	Rigorous	Marginal	Insubstantial
3	Deficient	Insubstantial	Insubstantial	Marginal	Insubstantial	Insubstantial	Insufficient
2	Insubstantial	Deficient	Marginal		Marginal	Insufficient	Deficient
1	Insufficient	Insufficient	Deficient	Satisfactory	Insufficient	Deficient	Marginal

Table 6. Keyword ranking responses from seven focus group participants. *Outstanding, excellent and very good were all considered rigorous.

The focus group participants acknowledged the difficulty in establishing 10 keywords that can universally represent standards. Furthermore, they collaborated in small groups to generate alternatives, using the data in Table 4 as reference. Two approaches emerged, one as a direct 10-part alternative and another as keyword combinations across five levels of achievement. See Table 7. The latter approach points in the direction of keyword sets that provide a greater scope to define characteristics associated with a particular level. It may be argued that "very good" and "good" are insufficiently differentiated and may easily stand for the same meaning in everyday language. The respective additions of "rigorous" and "competent" support further differentiation and are arguably better alternatives. The 10-part division situates some words that could be interpreted as synonymous, such as "deficient" and "limited", the former being considered a mid-level achievement.

10-part hierarchy	5-part hierarchy
Exceptional	Excellent/outstanding
Outstanding	Very good/rigorous
Excellent	Good/competent
Appropriate	Satisfactory/pass
Competent	Fail/insufficient
Deficient	
Limited	
Inadequate	
Insubstantial	
Nil	

Table 7. Two suggestions for a keyword hierarchy in assessment criteria.

Clearly, ranking keywords for easy recall by the staff and the students is difficult to achieve with any degree of accuracy. Perhaps this case is truer in art and design due to the nature of "studio and design productions" and "specialised artefacts" that tend to be immeasurable and "open" (Sadler, 2009, p. 160), meaning that limitless possibilities exist. The same can be said for historical, critical and theoretical essays that students may write. However, we should consider matters carefully before dismissing such approaches, especially when keywords are incorporated into a criterion-referenced assessment grid. Students welcome such tools as a "good idea", despite acknowledging that "terminology is open to multiple interpretations by individual staff and students" (Price et al., 2012, p. 32). Moreover, these tools can inspire others to create their own hierarchies. For example, previous work by the author (see Harland & Sawdon, 2012) motivated a colleague to create an alternative version in 2012 for use in a dissertation module by utilising the 10-part division and replacing a previous 20-part standards hierarchy (Barnard, unpublished). With a focus on the acquisition of knowledge and understanding, this revision is shown in Table 8. It incorporates "rigorous" as a property of "excellent", the basic level of achievement for a first-class degree in the UK. It also eradicates potential problems in an international context. Barnard also lifts "good" and "satisfactory" up a band, dispensing with "very good" as an apathetic representation of something better than "good". Furthermore, "adequate" represents the band immediately above the pass threshold, implying something passable but less than the "satisfactory" required for a 2:2 degree.

Classification	Percentage	Keyword criteria statement
1	90–100%	**Exceptional** acquisition of knowledge and understanding: originality of topic and argument; of publishable standard; a model/ideal essay.
	80–89%	**Outstanding** acquisition of knowledge and understanding: demonstrating independent thought and exemplary development of topic.
	70–79%	**Excellent** acquisition of knowledge and understanding: critical; showing rigorously organised argument and well-selected evidence.
2:1	60–69%	**Good** acquisition of knowledge and understanding: convincing display of analytical and reasoning skills; well written.
2:2	50–59%	**Satisfactory** acquisition of knowledge and understanding: some analytical content and argument supported with evidence.
3	40–49%	**Adequate** levels of knowledge and understanding: largely descriptive or narrative; little use of argument, analysis or evidence; adequate use of written English and scholarly apparatus.
Fail	30–39%	**Inadequate** levels of knowledge and understanding: little attention paid to brief or no appropriate topic; descriptive; no reasoned selection and organisation of material.
	20–29%	**Poor** levels of knowledge and understanding: minimal use of argument, evidence or analysis.
	10–19%	**Insufficient** levels of knowledge and understanding: does not answer the question, no use of argument, no evidence collected or used.
	1–9%	**Nil response**: effectively no evidence of knowledge or understanding: irrelevant material; no attempt to answer question; no organisation of material; no structure to writing.

Table 8. Criteria statements for assessing knowledge and understanding in written dissertations (Barnard, unpublished).

Criterion-referenced assessment grids are not new. In the late 1990s, Price and colleagues (2012, p. 29) created one for undergraduate use in the Business School at Oxford Brookes University, which is still employed today. Their grid does not attempt to include verbal descriptors that rank achievement levels, such as "good", but provides "both students and assessors with information about standards applied for each criterion" (2012, p. 29). Some of the values associated with each level of achievement are listed in Table 9. At the lower end of the spectrum, the text tends to resort to partial evidence of higher qualities, specifically the repetition of some words such as "coherent" and "incoherent", or partial demonstration of organisation or logicality. Words such as "inadequate" and "insufficient" do the same in Barnard's (unpublished) text. In Table 9 as well, a common grading scheme introduced at the Robert Gordon University, combining the grade definition and the descriptor, relies similarly on variations on words such as "good", "competent", "satisfactory" and "fail" to differentiate among levels (Hornby, 2003, p. 442–443).

Grade	Keywords
A	Polished, imaginative, comprehensive, confident, challenging, reflective, independent, proactive
B+	Careful, logical, coherent, evaluative, challenging, developmental, autonomous
B	Organised, coherent, logical, focused, dependent
C	Partly organised, logical, focused, dependent, directed, guided
Refer/Fail	Disorganised, incoherent, failing, unable, dependent

(Price et al., 2012, p. 29)

Grade	Keywords
6	Excellent: outstanding performance
5	Commendable/very good: meritorious performance
4	Good: highly competent performance
3	Satisfactory: competent performance
2	Borderline fail: failure, open to compensation
1	Unsatisfactory: fail
0	Very unsatisfactory: abject fail

(Hornby, 2003, p. 442–443)
Table 9. Two examples of values associated with standards of achievement.

Similarly, Sadler's (2005, p. 180) simpler interpretation of words corresponding to letter grades for "objective-based" grading defines the differences among A, B, C and D as clear, substantial, sound or some attainment of course objectives, respectively. Further qualifying terms regarding understanding are differentiated as complete and comprehensive, high-level understanding, mostly understood and basic. There is very limited logic to these words in supporting hierarchies of language use.

The potential for ambiguity associated with the articulation of achievement levels in assessment criteria for "open" outputs clearly represents a challenge for staff and students alike. We may all be familiar with "excellent" but in the UK, this may be used to represent as much as a third of the marking spectrum. As Price and colleagues point out, this issue is further complicated by an international dimension (2012, p. 17). Let us therefore briefly consider internationalisation as a phenomenon in higher education.

Interpretations of Internationalisation
Internationalisation and higher education have been directly linked through the development of research among scholars throughout the history of universities. More recently, the alignment of academic standards for research and teaching is cited as an increasingly important factor (Institutional Management in Higher Education [IMHE], 1999, p. 19) as universities perceive internationalisation as "the concept and the process of integrating an international dimension into the teaching, research and service functions" through "quality assessment and assurance" matters (IMHE, 1999, p. 3). "Globalisation" is therefore an influential factor in the present-day understanding of internationalisation, the incentives being "commercial advantage, knowledge and language acquisition, enhancing the curriculum with international content, and many others" (Altbach & Knight, 2007, p. 290).

Internationalisation has been a priority in Europe since the early 1990s but the contrast between its historical and contemporary interpretation has led some scholars to rename it "re-internationalisation" (Teichler, 2004, p. 6–9). Alternatively, some researchers distinguish between "cooperative internationalisation" and "commercial internationalisation" (Beelen & de Wit, 2012, p. 1), acknowledging increased competition. In the UK, funding for university

education has shifted from the public to the private sector through gradual increases in tuition fees, compensating for the stagnation and recent reduction of government funding. A consequence of this situation has been to seek out more international students willing to pay tuition fees higher than the amounts typically charged for UK-based students of the past. This change has resulted in the need to examine how the curricula are suited to students from overseas and must include assessment and feedback processes; as studies have shown, these assessment and feedback systems differ significantly among institutions in the UK and beyond. Despite the reinterpretation of internationalisation in the guise of economic development, little research appears to have been published on issues that link internationalisation to assessment criteria in art and design.

Discussion and Recommendations

There are numerous methods to indicate achievement levels in assessment through letters, numbers, symbols and words. However, "marks and grades do not in themselves have absolute meaning in the sense that a single isolated result can stand alone as an achievement measurement or indicator that has a universal interpretation" (Sadler, 2005, p. 177). For example, it is hard to communicate "excellence" because of various national and international definitions (Price et al., 2012, p. 17). "Excellence" defines the highest levels of achievement but it is unclear how it can be differentiated from "outstanding" or "exceptional" as definitions are mutually acknowledging. "Excellence" constitutes the first-class band of an undergraduate degree, which in the UK covers as much as 30% of the available marks, difficult to define in terms of standards. Yet "there needs to be a higher level of shared understanding than currently exists (among students, tutors and other stakeholders) of the language in which criteria are couched and the ways in which criteria are applied" (Woolf, 2004, p. 479). The alignment of verbal descriptors with grade indicators seems scarcely considered, especially with international students in mind. This may be due to the fact that the grading schemes have only been established in higher education since the late 1980s (Sadler, 2009, p. 159), which parallels the growth in internationalisation (Teichler, 2004, p. 6–9). When verbal descriptors are used, typically they range from as few as five to not more than 10 keywords. Three sets of keywords have been introduced earlier, as shown in Table 10. Those identified by Collins offer a limited scope of adjectives, heavily relying on too few keywords such as "good", "competent" and "passing", with further qualifying statements. Harland and Sawdon's (2012) hierarchy similarly relies on close similarity words in the distinction between "good" and "very good", as well as incorporates difficult words such as "rigorous" internationally. Barnard's (unpublished) version attempts to define each level independently, drawing on the distinction between "adequate" and "inadequate" at the pass threshold. He also lowers "excellent" but raises "good" and "satisfactory". The single keyword that seems inappropriate is "nil response", which means zero and clearly does not match a mark of 1–9%.

Exceptional	Outstanding	Exceptional
Excellent	Excellent	Outstanding
Very good	Rigorous	Excellent
Good	Very good	Good
Competent	Good	Satisfactory
Fairly competent	Satisfactory (P)	Adequate (P)
Passing	Marginal	Inadequate
Barely passing	Insubstantial	Poor
Marginally passing (P)	Insufficient	Insufficient
Failing	Deficient	Nil response
(Collins, 2004)	(Harland & Sawdon, 2012)	(Barnard, unpublished)

Table 10. Hierarchies of keywords for assessment criteria.

Taking all factors into account, standards in assessment criteria may be represented by a hierarchy of keywords that provide the stimulus for further justification through additional words, phrases and supporting statements. Building on Barnard's (unpublished) developments, based on the earlier work of Harland and Sawdon (2012), Table 11 presents a recommendation of how 10 keywords and secondary qualifiers might function to guide the writing of assessment criteria that make standards explicit. If a student output is considered excellent, why is it so and what further explanation can support it? If rigour and criticality are considered essential for excellence, these words may form the basis of feedback to students and will be determined by the disciplinary perspective. Rigour and criticality may differ for a Fine Art or Textiles undergraduate degree, but each term must define how to ensure that students of both programmes are judged by using the same criteria and standards but in different contexts. One benefit of incorporating keywords is to support (through consistent language) a link between the criteria, written and verbal feedback from tutors, on one hand, and the reflection and action by students as they progress, on the other hand. Clearly, this approach must be inclusive and motivate local and international students to learn equally well. Keywords and supporting statements provide an opportunity to enhance understanding.

Exceptional / Original / Comprehensive
Outstanding / Imaginative / Independent
Excellent / Critical / Rigorous
Good / Focused / Logical
Satisfactory / Organised / Reasoned
Adequate (P) / Predictable / Passable
Marginal / Incomplete / Inadequate
Insubstantial / Vague / Inaccurate
Insufficient / Feeble / Poor
Deficient / Inept / Scant
(P) = pass

Table 11. Hierarchy of keywords and qualifying terms for assessment standards.

The increasing internationalisation agenda in the UK, larger numbers of international students and overseas expansion mean that many more students who use English as a second language are exposed to assessment criteria that are rarely tested in global contexts. The variety of systems used internationally suggests that student migration has not been a major consideration for universities. Nonetheless, a more fundamental issue has been neglected in the litera-

ture on assessment, concerning what international students do with feedback. It is unclear what assessment means to students beyond an indicator of progress. What do students do once they receive their marks? How do they interpret feedback? Assessment and feedback are known to be under-researched topics (Cramp, 2011; Rae & Cochrane, 2008). However, interest is growing (Pitts, 2005) and although research into feedback dates back to the late 1970s (Pokorny & Pickford, 2010), considerable blind spots remain. For example, virtually no studies have been undertaken about first-year undergraduates (Cramp, 2011, p. 114).

Conclusion
The "relational dynamic" between staff and staff, staff and student, student and student (Price et al., 2012, p. 17) and perhaps increasingly, student and parent, allows assessment to be effectual. A consistent and disciplined use of language that defines grades in art and design may help counteract the diverse, often ambiguous range of assessed outputs that display the kind of "tacit knowledge and experience that does not easily lend itself to articulation and explanation" (Price et al., 2012, p. 33). For international students, this approach may be more significant if they are to grasp new assessment systems and align criteria with standards through reflection. This paper shows that although consistent language use may be desirable, it is unlikely. Keywords that clearly differentiate among levels of achievement appear to have been an overlooked aspect, considering the number of times words such as "very", "highly", "mostly" and "partly" are used to substantiate definitions. After testing assumptions that emerged from working group activities in international contexts and reviewing the literature about assessment criteria and standards, the study found some consensus about keywords to define grades. These supplement letters, numbers and symbols and link with descriptions of grades to enhance understanding. In art and design, where students may also experience levels of dyslexia higher than those of other academic disciplines, the recommendations offer a starting point for broader discourse that may extend across universities as well as different levels of education. As students migrate between countries and experience various interpretations of "good", they also mature and have to adapt their notion of "good" as they advance through progressive stages of their education.

Robert Harland
Lecturer
Loughborough University
Email address: r.g.harland@lboro.ac.uk

References

Altbach, P. G. & Knight, J. (2007). The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education, 11*(3–4), 290–305. doi: 10.1177/1028315307303542

Beelen, J. & de Wit, H. (2012). Internationalization revisited: New dimensions in the internationalization of higher education. Amsterdam: Centre for Applied Research on Economics and Management (CAREM), Hogeschool van Amsterdam, 1–3.

Brown, G., Bull, J. & Pendlebury, M. (1997). *Assessing student learning in higher education/George Brown with Joanna Bull and Malcolm Pendlebury*. London: Routledge.

Cohen, L., Manion, L. & Morrison, K. (2007). *Research methods in education*. Abingdon: Routledge.

Collins, P. (2004). *Expressing and reporting the results of assessing in grades and percentages*. Nottingham: Nottingham Trent University Centre for Academic Practice.

Cramp, A. (2011). Developing first-year engagement with written feedback. *Active Learning in Higher Education, 12*, 113–124. doi: 10.1177/1469787411402484

Davies, A. (2012). Learning outcomes and assessment criteria in art and design. What's the recurring problem? *Networks, 18*, 1–8.

European Commission. (2009). *ECTS users' guide*. Luxembourg: Office for Official Publications of the European Communities. Retrieved July 21, 2014, from http://ec.europa.eu/education/tools/ects_en.htm

Harland, R. & Sawdon, P. (2012). From fail to first: Revising assessment criteria in art and design. *Art, Design & Communication in Higher Education, 10*(1), 67–88. doi: 10.1386/adch.10.1.67_1

Hornby, W. (2003). Assessing using grade-related criteria: A single currency for universities. *Assessment & Evaluation in Higher Education, 24*(8), 435–454. doi: 10.1080/0260293032000066254

Institutional Management in Higher Education. (1999). *Quality and internationalization in higher education*. Paris: Organisation for Economic Co-operation and Development.

Orr, S. (2007). Assessment practices in art and design. *Art, Design & Communication in Higher Education, 5*(2), 79–151. doi: 10.1386/adch.5.2.79_2

Orsmond, P., Maw, S. J., Park, J. R., Gomez, S. & Crook, A. C. (2013). Moving feedback forward: Theory to practice. *Assessment & Evaluation in Higher Education, 38*(2), 240–252. doi: 10.1080/02602938.2011.625472

Parkin, H. J., Hepplestone, S., Holden, G., Irwin, B. & Thorpe, L. (2012). A role for technology in enhancing students' engagement with feedback. *Assessment & Evaluation in Higher Education, 37*(8), 963–973. doi: 10.1080/02602938.2011.592934

Pitts, S. E. (2005). "Testing, testing...": How do students use written feedback? *Active Learning in Higher Education, 6*(3), 218–29. doi: 10.1177/1469787405057663

Pokorny, H. & Pickford, P. (2010). Complexity, cues and relationships: Student perceptions of feedback. *Active Learning in Higher Education, 11*. 21–30. doi: 10.1177/1469787409355872

Price, M., Rust, C., O'Donovan, B. & Handley, K. (2012). *Assessment literacy: The foundation for improving student learning*. Oxford: The Oxford Centre for Staff and Learning Development, Oxford Brookes University.

Rae, A. M. & Cochrane, D. K. (2008). Listening to students: How to make written assessment feedback useful. *Active Learning in Higher Education, 9*, 217–230. doi: 10.1177/1469787408095847

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*(2), 175–194. doi: 10.1080/0260293042000264262

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179. doi: 10.1080/02602930801956059Schünemann, H. J., Best, D., Vist, G. & Oxman, A. D. (2003). Letters, numbers, symbols and words: How to communicate grades of evidence and recommendations. *Canadian Medical Association Journal, 169*(7), 677–680.

Silverman, D. (2005). *Doing qualitative research*. London: SAGE Publications Ltd.

Teichler, U. (2004). The changing debate on internationalization of higher education.

Higher Education, 48(1), 5–26. doi: 10.1023/B:HIGH.0000033771.69078.41

Woolf, H. (2004). Assessment criteria: Reflections on current practices. *Assessment & Evaluation in Higher Education, 29*(4), 479–493. doi: 10.1080/0260293031000168904