London Review of Education
Volume 15, Number 3, November 2017
DOI: https://doi.org/10.18546/LRE.15.3.13

Exploring the relationship between validity and comparability in assessment

Victoria Crisp*
Cambridge Assessment

Abstract

This article discusses how comparability relates to current mainstream conceptions of validity, in the context of educational assessment. Relevant literature was used to consider the relationship between these concepts. The article concludes that, depending on the exact claims being made about the appropriate interpretations and uses of the results of an assessment, several comparability concerns fall within the remit of validation. The current exploration supports the addition of comparability to validation studies and may be useful in the context of a growing emphasis on the provision of validity evidence for public examinations.

Keywords: assessment; exams; validity; comparability; validation

Introduction

Validity and comparability are central concepts in educational assessment theory and practice. However, the definitions of both are complex and the relationship between validity and comparability does not appear to have been explicitly determined. Drawing on relevant insights from the literature, this article seeks to make explicit how comparability relates to current mainstream conceptions of validity and to explore the implications for validation studies.

Validity

Definitions of validity have developed over time through a number of conceptualizations: from a simple, measurable notion of whether an assessment really measures what it was intended to (for example, Kelley, 1927; Guilford, 1946), through a triarchic conceptualization of criterion, content and construct validities (AERA, APA and NCME, 1966), to a broader, more unified conception of validity as being about the appropriateness of the interpretations and uses of assessment results embracing multiple evidence types (Cronbach, 1971; Messick, 1989). (See Shaw and Crisp, 2011, and Newton and Shaw, 2014, for historical analyses of the development of validity theory.) The key definition underpinning current mainstream conceptions of validity comes from Messick: 'validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment' (1989: 13). While mainstream conceptualizations see validity as a unified concept, it is also considered multifaceted, and various frameworks have been proposed for structuring the collection of multiple evidence types when conducting

* Email: crisp.v@cambridgeassessment.org.uk

validation (for example, Crooks *et al*., 1996; Frederiksen and Collins, 1989; Linn *et al*., 1991; Kane, 2006). Contemporary validity theory is generally aligned with Messick's key definition (1989), although discussions have increasingly focused on validation, that is, on how to evaluate the appropriateness of the proposed interpretations and uses of assessment outcomes. Kane's (2006; 2013) influential work on validation suggests using an argument-based approach involving constructing an *interpretation/use argument* (setting out the inferences and assumptions that lead from the student work to the decisions made based on results) and a *validity argument* (in which evidence relating to each inference and assumption in the interpretation/use argument is provided). The necessary structure of the interpretation/use and validity arguments depends on the intended interpretations and uses of the assessment outcomes. Therefore, these claims need to be identified, which then influence the inferences and assumptions that need to be evidenced (and might also affect the importance of evaluating comparability concerns). Kane's (2006) work does not provide an explicit steer on incorporating comparability concerns into validation frameworks. However, given that the interpretation/use argument for an assessment will be somewhat different depending on the proposed uses and interpretations of an assessment's outcomes, it is possible that Kane (2006) did not discuss comparability because this was not a relevant issue for the assessment that he used as an example. (Note that while the above represents what is perhaps the 'mainstream' conception of validity, some argue for a more limited definition; see, for example, Borsboom *et al*., 2004.)

Comparability

Comparability of standards between qualifications is a pertinent issue in England, where several awarding bodies offer versions of the same qualification. For example, GCSE (General Certificate of Secondary Education) maths results awarded by two different awarding bodies need to be 'comparable' because results from both will likely be used in the same way. Comparability between awarding bodies has long been a concern in England, with documented studies taking place as early as the 1950s (Elliott, 2011). However, even ignoring the multiple-board system, comparability issues are relevant to most qualifications. For example, the specific exam papers taken in different years or sessions need to be comparable so that relevant stakeholders (for example, admissions tutors and employers) know that students receiving the same grade in different years have reached the same standard in that area of study.

Until recently, theorization of the concept of comparability has tended to be hampered by a lack of distinction between definitions of comparability and methods for achieving or monitoring comparability (Newton, 2010). What is meant by 'comparability' of examination standards can vary depending on how standards are (implicitly or explicitly) defined (Coe, 2007; Coe, 2010; Newton, 2010). Coe (2007) has usefully identified three conceptions of comparability underpinning use of the term. The first of these is *performance comparability*, where the focus of comparison is observed phenomena in relation to specific criteria. Secondly, he defines *statistical comparability*, where the estimated chance of achieving a particular grade is the focus of comparison. The third type is *construct comparability*, where comparison is based on a common linking construct. This third conception allows that two assessments might vary in which is more demanding depending on the construct against which they are being compared. Quotations describing each are given below (Coe, 2010):

- *Performance comparability* – 'Two examinations have comparable standards if candidates' performances that are judged to exemplify the same phenomenon (or set of phenomena) are awarded the same grades on each.' (Coe, 2010: 274)

- *Statistical comparability* – 'Two examinations have comparable standards if the same grades are equally likely to be achieved by comparable candidates in each.' (Coe, 2010: 275)
- *Construct comparability* – 'Two examinations have comparable standards if performances that correspond to the same level of some linking construct are awarded the same grades in each.' (Coe, 2010: 278)

A radically different conceptualization of varying comparability definitions was proposed by Newton (2010): phenomenal definitions (attainments are the same in terms of their characteristics); causal definitions (attainments are the same in terms of their causes); and predictive definitions (attainments are the same in terms of indicating future success). Both Coe's and Newton's categorizations illustrate that definitions used by researchers and practitioners can vary and are not always made explicit. It is suggested that more than one definition of comparability standards may be legitimate, meaning that there may be more than one answer to the question of whether two examinations are comparable (Newton, 2010), and that in practical terms a decision could be made either to try to satisfy all views on comparability, or to prioritize just one (Coe, 2010).

Insights from existing literature on the relationship between validity and comparability

This section considers existing literature that provides insights into the relationship between validity and comparability. In a discussion of various proposed frameworks for validation, Moss noted that many of them included comparability as an important concern: 'These analytic schemes … highlight specific issues that their authors consider important for validity researchers to address. Each balances technical concerns about such issues as reliability, generalizability, and comparability with concerns about the consequences of assessment' (1992: 229). However, Moss and the theorists whose frameworks she discussed were referring specifically to comparability of scores between scorers and across tasks (rather than comparability of standards). The first part of this might more commonly be referred to as 'marking consistency' or 'marking reliability' in the UK – the idea that a candidate should get the same score regardless of which examiner happened to mark their paper. This is clearly a part of validity, as inconsistent marking would make it difficult to argue that scores reflect constructs of interest and can be used in certain specified ways. This fits within the 'scoring' inference of Kane's example interpretation/use argument, where one of the necessary assumptions in the argument would be that 'the scoring rule is applied accurately and consistently' (2006: 24). The second element referred to by Moss – comparability of scores across tasks – relates to how the particular task(s) that the student carried out (for example, which of some optional questions they selected, or which year's exam paper they took) should not affect the result they receive. Again, these issues are part of validity, as fluctuations in outcomes as a result of different tasks being taken in different years, or as a result of different optional questions (or papers) within the qualification being chosen, would likely compromise how the results could be used.

The notion of 'score comparability' was also discussed by Messick (1995), in an article in which he described six aspects of validity, one of which was 'scoring models as reflective of task and domain structure'. This element emphasized that 'the theory of the construct domain should guide … the rational development of construct-based scoring criteria and rubrics' (Messick, 1995: 6–7). The quotation that follows, while dealing mostly with a hypothesized evaluation of likely level of comparability and possible methods for adjusting scores to achieve comparability,

provides additional insight into the meaning of the notion of 'score comparability' being used by Messick and Moss:

> To the extent that different assessments (i.e., those involving different tasks or different settings or both) are geared to the same construct domain, using the same scoring model as well as scoring criteria and rubrics, the resultant scores are likely to be comparable or can be rendered comparable using equating procedures. Otherwise, score comparability is jeopardized but can be variously approximated using such techniques as statistical or social moderation (Mislevy, 1992). Score comparability is clearly important for normative or accountability purposes whenever individuals or groups are being ranked. However, score comparability is also important even when individuals are not being directly compared, but are held to a common standard. Score comparability of some type is needed to sustain the claim that two individual performances in some sense meet the same local, regional, national, or international standard.
>
> (Messick, 1995: 7)

This suggests that scoring comparability is not just about scoring, but also about whether the same constructs are being assessed by different papers. The quotation from Messick starts to make the notion of score comparability sound somewhat broader, but it still appears only to apply to different versions of tests within the same qualification (for example, tasks taken at different times), where the assessments are measuring the same construct, using the same scoring method, and equating or linking (Kolen and Brennan, 2014) can be applied (for example, because the questions all come from a calibrated test battery).

A link between construct validity and comparability was also made by Bachman *et al.* (1988) in a study comparing two different test batteries. They argued that the 'most important aspect of comparability is that of the abilities measured by the two tests. Thus, the examination of comparability must begin with an assessment of the extent to which tests measure the same abilities' (1988: 130). This suggests that elements of validity relating to the constructs assessed are important within comparability studies.

Interestingly, in 1995 Messick wrote as if both 'comparability' and 'reliability' were concepts separate to validity. In discussions of the growing interest in performance assessment, he noted: 'Indeed, it is precisely because of such politically salient potential consequences that the validity of performance assessment needs to be systematically addressed, as do other basic measurement issues such as reliability, comparability, and fairness' (1995: 5). However, some of his later work (for example, Messick, 2000) treats reliability as an element of validity, making it difficult to know whether he still saw comparability as an entirely separate concept to validity. In the later publication, Messick (2000) wrote in depth about validity, reliability and fairness, describing them as principles organized 'within an overarching conception of construct validity' (2000: 18). He considered test fairness to be about impartiality, and related it to certain comparability issues, specifically comparability of scores and of constructs elicited across different individuals, groups and test settings. A somewhat similar link between fairness and validity was made by Kunnan (2000), with reference again being made to comparability for different individuals and groups. He identified three main concerns for fairness: validity, access and justice. For validity, he felt that the focus should be 'on whether test-score interpretations have *equal construct validity* (and reliability) for different test-taker groups as defined by salient test-taker characteristics such as gender, race/ethnicity, field of specialization and native language and culture' (Kunnan, 2000: 3). Fairness for different groups is also mentioned by Kane (2011; 2013) as a point for evaluation in validation studies in relation to the social consequences of the uses of assessment results. The comments from these authors suggest that certain comparability concerns may fall within the remit of validation.

Determining the relationship between comparability and validity

To think through the relationship between validity and comparability, this article will consider the case of a hypothetical new qualification being introduced with the intention that it is equivalent to A levels (Advanced levels) and that results can be interpreted and used in the same way as A level results. Validity is conceptualized as an evaluation of the degree to which specific interpretations of the meaning of assessment outcomes can be supported. These intended interpretations relate to the claims made about the meaning and use of results. For International A level validation exercises, Shaw and Crisp (2012: 12) identified two proposed interpretations of scores/grades:

(1) Scores/grades provide a measure of relevant learning/achievement.
(2) Scores/grades provide an indication of likely future success.

Given that Shaw and Crisp (2012) identified these as the proposed interpretations of outcomes of International A level results, they would likely also be proposed interpretations of results for a new level 3 qualification intended to be equivalent to A levels. There would also be an additional claim that results from the new qualification provide a comparable measure of relevant learning/achievement to A level results and can be used as equivalent. As this additional claim is about the interpretation and use of assessment outcomes, this comparability concern would seem to fall within the remit of validity by Messick's definition, and thus should be evaluated as part of validation studies. Perhaps an appropriate additional proposed interpretation of the new qualification's results would be:

(3) Grades are comparable to grades in A levels.

Note that reference to 'scores' has not been made in this new statement as marks might not be on the same scale as A levels. Proposed interpretation 3 would begin to incorporate 'between qualifications' comparability of standards into validation work, although how this affects the structure to be used in a validation study, and the evidence to be collected, would need to be ascertained (as discussed later).

Claims for comparability within elements of a qualification also need to be considered. Within a qualification, there is usually an implicit claim that any optional questions, optional papers or alternative versions of papers (for example, versions to be taken in different parts of the world), are equivalent. This is perhaps not explicitly stated, except that the exam papers, specifications and other exam board guidance will set out which questions are alternatives and which combinations of units are allowed. However, equivalence is clearly implied given that these are used as alternatives. The issue of equivalence over time is also key to comparability, as the particular year or session in which a student takes a paper should not affect their result. Given that there is an implicit claim that assessment outcomes from different years, different versions of papers, and achieved through different optional questions or papers (if relevant) can be used in the same way, this would seem to fall within the notion of validity. This relates to scoring reliability (as the scores for different alternatives need to be comparable), but also to the constructs assessed (as these should be sufficiently similar). This will be explored further in the next section.

From the discussion so far, it would seem that depending on the claims (explicit and implicit) being made about the meaning and use of assessment outcomes, certain comparability concerns should be considered to be a part of validity, and relevant methods should be used to evaluate comparability as part of validation studies.

Locating comparability issues within a validation study

This section takes as its focus the structure for validation of A levels developed by Shaw and Crisp (2015) in line with Kane's (2006) proposed argument-based approach to validation, and considers how relevant comparability concerns could be incorporated where not already addressed. Shaw and Crisp's framework is based on five inferences, represented by five validation questions. These were designed to act as research questions to structure collection of relevant evidence. Several methods might well be needed to address each validation question. The questions and the related evidence make up the validity argument.

In order to think through where comparability concerns 'fit' in relation to these validation questions, each question will be considered in turn. As mentioned earlier, the case of a hypothetical new level 3 qualification that is intended to be comparable to A level will be used as the focus. This will be referred to as 'Qualification X'. In addition, for the purposes of discussion, the terms 'internal comparability' and 'external comparability' will be used to distinguish between comparability concerns within a qualification (that is, comparability between different optional papers or questions, or different versions of a paper for different sessions) and beyond the qualification (that is, comparability of the qualification of interest with other qualifications with which it is claimed to be equivalent).

Validation Question 1: Do the tasks elicit performances that reflect the intended constructs?

Validation Question 1 relates to construct representation. The tasks used in the assessments need to trigger performances that reflect the constructs to be assessed, both in terms of the topic being on the syllabus and in terms of triggering students to employ relevant skills and processes. If this is not the case, then it is unlikely that the results can safely be used in the ways intended.

Arguably, to assert that a qualification has 'internal comparability' it would be appropriate to establish whether different alternative papers or questions are broadly similar in the constructs elicited. This might not require an exact match of topics assessed, but that questions are drawn from the same pool of topics, test the same kinds of skills and have similar demands. It would also be appropriate to explore the fairness for different individuals and groups in terms of the kind of performances that the tasks elicit. These additional elements could be evaluated as part of a validation exercise by including exam questions from equivalent papers in the analyses (for example, papers from different sessions) and potentially falls within the existing remit of Validation Question 1, although this could perhaps be made more explicit:

> 1b. How similar are the constructs elicited by different optional questions or papers, by different versions of the exam papers (for example, papers in different sessions or for different parts of the world) and for different groups of students?

In terms of comparability with A levels, or 'external comparability', some similar comparison of the constructs would seem appropriate as part of validation, given the claim that Qualification X will be comparable to A levels. This might involve comparison of example papers across boards, and comparison of the content and skills set out in the syllabuses (for example, syllabus mapping exercises). Analysis of this kind is closely related to Validation Question 1, but an additional research question could ensure these elements are investigated:

> 1c. How similar are the constructs elicited to those that A levels intend to assess?

Question 1c represents part of the construct validity analysis that Bachman *et al.* (1988) argued is needed in a study of the comparability of two different tests.

Validation Question 2: Are the scores/grades dependable measures of the intended constructs?

This validation question relates to the translation of student performances into scores/grades that reflect the quality of the performances on the tasks (in relation to the constructs of interest). As such, validity in this respect relates to the appropriateness of marking criteria, accuracy and consistency of their application, and whether aggregation and grading procedures are appropriate (Kane, 2006). Certain comparability issues relate to this validation question because results need to be comparable between markers and between alternative tasks (as previously identified by Moss, 1992, in the notion of 'scoring comparability'). Marking criteria and training are intended to reduce differences between markers, procedures for scaling of examiners are intended to adjust for differences in their internal standards, and grading procedures are intended to adjust for possible differences in difficulty between different versions of exam papers (for example, in different sessions). All these procedures and their outcomes should be analysed in order to evaluate 'internal comparability'. In addition, to address fairness, any differences in how these processes affect different groups of students should be explored. A subsidiary validation question related to scoring might usefully highlight the need for these analyses:

> 2b. Are the scores/grades comparable between different markers, between different optional questions or papers, between different versions of the exam papers (for example, papers in different sessions or for different parts of the world) and between different groups of students?

In terms of comparability of outcomes with A levels, some form of comparison of grades achieved is needed. Unless a substantial number of students are entered for both Qualification X and A level in the same subject (or in similar subjects), methods tend to be somewhat indirect, but there are methods that can provide some relevant evidence (for example, judgemental rank ordering studies, see Bramley, 2007; statistical methods using other data, such as 'common examinee' methods, see Coe, 2007). An additional subsidiary validation question would be needed to emphasize this within the validation framework. For example:

> 2c. Are the grades comparable to those of similarly able candidates achieving similar grades in A level?

Question 2c represents another part of the construct validity analysis that Bachman *et al.* (1988) suggest is needed in a study of the comparability of two different tests.

Validation Question 3: Do the tasks adequately sample the constructs that are set out as important within the syllabus?

Validation Question 3 relates to how well the syllabus has been sampled. Methods for validation tend to involve checking the sampling of content over the last few years and checking the balance of skills assessed. In terms of 'internal comparability', arguably it would be desirable for alternative versions of papers to each provide reasonable sampling of the constructs set out by the syllabus. This is perhaps already addressed by the methods that have tended to be used to answer Validation Question 3 (see Shaw and Crisp, 2012) but an additional subsidiary validation question could emphasize this need:

> 3b. Do the tasks in different optional questions or papers, and in different versions of the exam papers (for example, papers in different sessions or for different parts of the world) sample the constructs that are set out as important within the syllabus equally adequately?

In terms of 'external comparability', it is desirable that A levels are also adequately sampling the constructs set out as important in their specifications. However, this would be the responsibility of the A level awarding bodies and beyond the remit for the validation of Qualification X's interpretation and uses. Nonetheless, given the claims of comparability, it could be of relevance to consider the extent to which the tasks in Qualification X assessments adequately sample the constructs set out as important in the syllabuses of A levels, or in the Department for Education (DfE) subject content and Ofqual subject level conditions and requirements. Comparison to the latter is perhaps more appropriate, as it provides a central point of reference for expectations of a qualification in a particular subject; this would also be a more efficient analysis than considering several awarding bodies' syllabuses. An appropriate validation question might be:

> 3c. Do the tasks adequately sample the constructs that are set out as important by relevant DfE/ Ofqual subject criteria?

Validation Question 4: Do the constructs sampled give an indication of broader competence within and beyond the subject?

This validation question asks whether the scores/grades indicate likely competence beyond the syllabus, including wider competence in the subject beyond the constructs set out in the syllabus and competence beyond the subject. Ideally this should be the case for both Qualification X and A levels, and should be evaluated for Qualification X as part of validation. However, it would seem beyond the remit of the validation of Qualification X to evaluate A levels in this way. There do not seem to be any comparability issues, 'internal' or 'external', that would need to be addressed in relation to Validation Question 4.

Validation Question 5: Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions?

This final validation question considers whether scores/grades give an indication of likely future success and brings us to the focus of evaluating the appropriateness of the intended uses of assessment outcomes. There do not seem to be any 'internal comparability' issues related to this validation question. However, given the claim that imagined Qualification X will be comparable to A levels and accepted by universities, this should be added as an additional validation question and could be worded thus:

> 5b. Do grades give a comparable indication of likely future success to that provided by A level results?

Judgemental methods involving appropriate experts might provide one way to answer this question, or if data could be gathered on performance of students beyond their level 3 qualifications, this would enable use of statistical methods (see, for example, Green and Vignoles, 2012; Gill and Vidal Rodeiro, 2014).

Relationship between comparability definitions and the validation questions

At this point, it is worth returning to definitions of comparability to identify that, or those, which underpin the current discussion. The three key conceptions referred to by Coe (2007; 2010) will be used: performance comparability, statistical comparability and construct comparability.

Taking performance comparability first, this is not identifiable in the themes addressed by the additional subsidiary validation questions. Student performances on the tasks are important, as they should provide evidence of the constructs of interest for each student, and the scores are based on these performances. However, the phenomena exemplified in the performances might not relate wholly to the constructs of interest. So within the current discussion, it is not so much about the performances showing the same phenomena as each other, but that comparable performances should illustrate the same construct to the same extent. If it could safely be assumed that two assessments to be compared are both testing the right constructs without much construct-irrelevant variance in the phenomena exemplified, then perhaps performance comparability is achieved too – but arguably this is not the underpinning definition of interest when focusing on validity.

Given that in the field of validity theory there is considerable focus on the constructs to be assessed, with 'construct validity' as the unifying theme, it is not surprising that 'construct comparability' appears to be prominent in the subsidiary validation questions proposed here. (Note that some validity theorists (for example, Kane, 2012) now avoid use of 'construct validity' as a term and just refer to 'validity'. This is because Cronbach and Meehl's (1955) work, which instigated the notion of construct validity as a unifying concept, was written in the context of the measurement of psychological properties or attributes (for example, ego strength). Such attributes are likely to be less multidimensional than the collection of content and skills that many educational qualifications attempt to assess. However, some authors still make use of the term 'construct' as an overarching label for the content and skills of interest (for example, as explicated in syllabus documents), although it might represent a rather more diverse set of attributes than when used by Cronbach and Meehl. The additional questions falling under Validation Questions 1 and 3 relate to elements within a qualification or between different qualifications being comparable in terms of similar elicitation of the constructs set out in syllabuses or in national subject criteria. Proposed question 5b, relating to decision-making and likely future success, also aligns with the construct comparability definition, but here the linking construct is something much broader, such as 'future potential' or 'preparedness for study/employment'.

Statistical comparability is also represented by the proposed additional validation questions, specifically by those for Validation Question 2 relating to scoring. These proposed questions relate to results being comparable between alternative papers, markers and qualifications, and thus to the notion of similar candidates being likely to achieve the same grades.

The use of two different underpinning conceptions of comparability, and the use of multiple linking constructs within the notion of construct comparability, means that there would not be one simple answer to the question of whether one qualification is comparable to another, and a number of methods would be needed to evaluate comparability issues. However, neither is there usually a simple answer to whether the proposed interpretations and uses of a qualification's assessment results are appropriate when conducting validation studies, and comparability studies to date often take an approach of utilizing more than one method (Pollitt *et al*., 2007).

Conclusion

This article has discussed how the key concepts of validity and comparability relate to each other. For many qualifications there is an implicit claim that versions of the same test (for example, in different years) are comparable and alternative papers or questions are comparable – referred to as 'internal comparability' in this article. This issue is a part of validity, given that it affects the uses of assessment outcomes claimed to be appropriate. If different qualifications are claimed to be equivalent ('external comparability') then this is also important to address as part of validity, as this relates to the interpretations and uses of assessment outcomes claimed to be appropriate.

To consider and exemplify how comparability issues fit within the concept of validity, Shaw and Crisp's (2015) validation questions have been used as the starting point for considering the validation of a hypothetical level 3 qualification. Various subsidiary validation questions could be added to facilitate evaluation of comparability issues within a validation study. Table 1 shows Shaw and Crisp's (2015) validation questions for level 3 qualifications with the addition of the proposed subsidiary validation questions and how these relate to Coe's (2010) comparability definitions. The first question in each group remains the main question, but the subsidiary questions provide prompts for validity researchers to gather comparability evidence as part of a validation study. Note that if the proposed interpretations and uses of assessment outcomes were different, this would affect the validation questions. For example, if no claims were made that the results from hypothetical Qualification X were comparable to A level, then the questions relating to 'external comparability' would not be needed.

Table 1: Validation questions for a hypothetical level 3 qualification claimed to be comparable to A levels

Key validation questions (from Shaw and Crisp, 2015)	Additional subsidiary validation questions	Underpinning comparability conceptions (as defined by Coe, 2010)
1a. Do the tasks elicit performances that reflect the intended constructs?	1b. How similar are the constructs elicited by different optional questions or papers, by different versions of the exam papers (e.g. papers in different sessions or for different parts of the world) and for different groups of students?	Construct comparability (where the constructs are those set out in syllabus)
	1c. How similar are the constructs elicited to those that A levels intend to assess?	
2a. Are the scores/grades dependable measures of the intended constructs?	2b. Are the scores/grades comparable between different markers, between different optional questions or papers, between different versions of the exam papers (e.g. papers in different sessions or for different parts of the world) and between different groups of students?	Statistical comparability
	2c. Are the grades comparable to those of similarly able candidates achieving similar grades in A level?	

Key validation questions (from Shaw and Crisp, 2015)	Additional subsidiary validation questions	Underpinning comparability conceptions (as defined by Coe, 2010)
3a. Do the tasks adequately sample the constructs that are set out as important within the syllabus?	3b. Do the tasks in different optional questions or papers, and in different versions of the exam papers (e.g. papers in different sessions or for different parts of the world) sample the constructs that are set out as important within the syllabus equally adequately?	Construct comparability (where the constructs are those set out in syllabus)
	3c. Do the tasks adequately sample the constructs that are set out as important by relevant DfE/Ofqual subject criteria?	
4. Do the constructs sampled give an indication of broader competence within and beyond the subject?	n/a	n/a
5a. Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions?	5b. Do grades give a comparable indication of likely future success to that provided by A level results?	Construct comparability (where the construct is 'future potential'/ 'preparedness for study/ employment')

The challenge that the additional validation questions present is the additional burden of more data collection and analysis being needed to gather comparability evidence, where needed, as part of validation. Validation studies already tend to be substantial undertakings requiring considerable resourcing (for an example of the extensive work that can be involved in the validation of an assessment, see Chapelle *et al*., 2008, and Shaw and Crisp, 2012). In adding to the workload, the additional data collection and analysis would need to be as efficient as possible. That said, in various assessment contexts comparability issues are already monitored and evaluated, so the incorporation of relevant comparability concerns into validation might, in practice, simply mean synthesizing findings from two or more separate studies.

Acknowledgement

The author would like to thank Stuart Shaw for many previous conversations about validity and validation and for comments on an earlier draft of this article.

Notes on the contributor

Victoria Crisp has been a researcher at Cambridge Assessment for 16 years, exploring a range of issues in relation to examinations and other assessments. Areas of research have included: question difficulty, effects of answer spaces on student responses, comparability issues and methods, validation of general qualifications and judgement processes in assessment.

References

AERA (American Educational Research Association), APA (American Psychological Association) and NCME (National Council on Measurement in Education) (1966) Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Bachman, L.F., Kunnan, A., Vanniarajan, S. and Lynch, B. (1988) 'Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries'. *Language Testing*, 5 (2), 128–59.

Borsboom, D., Mellenbergh, G.J. and van Heerden, J. (2004) 'The concept of validity'. *Psychological Review*, 111 (4), 1061–71.

Bramley, T. (2007) 'Paired comparison methods'. In Newton, P., Baird, J.-A., Goldstein, H., Patrick, H. and Tymms, P. (eds) *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority, 246–94.

Chapelle, C.A., Enright, M.K. and Jamieson, J.M. (eds) (2008) *Building a Validity Argument for the Test of English as a Foreign Language*. New York: Routledge.

Coe, R. (2007) 'Common examinee methods'. In Newton, P., Baird, J.-A., Goldstein, H., Patrick, H. and Tymms, P. (eds) *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority, 331–76.

Coe, R. (2010) 'Understanding comparability of examination standards'. *Research Papers in Education*, 25 (3), 271–84.

Cronbach, L.J. (1971) 'Test validation'. In Thorndike, R.L. (ed.) *Educational Measurement*. 2nd ed. Washington, DC: American Council on Education, 443–507.

Cronbach, L.J. and Meehl, P.E. (1955) 'Construct validity in psychological tests'. *Psychological Bulletin*, 52 (4), 281–302.

Crooks, T.J., Kane, M.T. and Cohen, A.S. (1996) 'Threats to the valid use of assessments'. *Assessment in Education: Principles, Policy and Practice*, 3 (3), 265–85.

Elliott, G. (2011) '100 years of controversy over standards: An enduring problem'. *Research Matters*, Special Issue 2, 3–8.

Frederiksen, J.R. and Collins, A. (1989) 'A systems approach to educational testing'. *Educational Researcher*, 18 (9), 27–32.

Gill, T. and Vidal Rodeiro, C.L. (2014) 'Predictive Validity of Level 3 Qualifications: Extended Project, Cambridge Pre-U, International Baccalaureate, BTEC Diploma' (Cambridge Assessment Research Report). Cambridge: Cambridge Assessment.

Green, F. and Vignoles, A. (2012) 'An empirical method for deriving grade equivalence for university entrance qualifications: An application to A levels and the International Baccalaureate'. *Oxford Review of Education*, 38 (4), 473–91.

Guilford, J.P. (1946) 'New standards for test evaluation'. Educational and Psychological Measurement, 6, 427–38.

Kane, M.T. (2006) 'Validation'. In Brennan, R.L. (ed.) *Educational Measurement*. 4th ed. Westport, CT: Praeger, 17–64.

Kane, M. (2011) 'Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010'. *Language Testing*, 29 (1), 3–17.

Kane, M. (2012) 'All validity is construct validity: Or is it?'. *Measurement: Interdisciplinary Research and Perspectives*, 10 (1–2), 66–70.

Kane, M.T. (2013) 'Validating the interpretations and uses of test scores'. *Journal of Educational Measurement*, 50 (1), 1–73.

Kelley, T.L. (1927) *Interpretation of Educational Measurements*. Yonkers-on-Hudson: World Book Company.

Kolen, M.J. and Brennan, R.L. (2014) *Test Equating, Scaling, and Linking: Methods and practices*. 3rd ed. New York: Springer.

Kunnan, A.J. (2000) 'Fairness and justice for all'. In Kunnan, A.J. (ed.) *Fairness and Validation in Language Assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (Studies in Language Testing 9). Cambridge: Cambridge University Press, 1–14.

Linn, R.L., Baker, E.L. and Dunbar, S.B. (1991) 'Complex, performance-based assessment: Expectations and validation criteria'. *Educational Researcher*, 20 (8), 15–21.

Messick, S. (1989) 'Validity'. In Linn, R.L. (ed.) *Educational Measurement*. 3rd ed. New York: American Council on Education/Macmillan, 13–104.

Messick, S. (1995) 'Standards of validity and the validity of standards in performance assessment'. *Educational Measurement: Issues and Practice*, 14 (4), 5–8.

Messick, S. (2000) 'Consequences of test interpretation and use: The fusion of validity and values in psychological assessment'. In Goffin, R.D. and Helmes, E. (eds) *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at seventy*. Norwell, MA: Kluwer Academic, 3–20.

Mislevy, R.J. (1992) *Linking Educational Assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy Information Center.

Moss, P.A. (1992) 'Shifting conceptions of validity in educational measurement: Implications for performance assessment'. *Review of Educational Research*, 62 (3), 229–58.

Newton, P.E. (2010) 'Contrasting conceptions of comparability'. *Research Papers in Education*, 25 (3), 285–92.

Newton, P.E. and Shaw, S.D. (2014) *Validity in Educational and Psychological Assessment.* London: SAGE Publications.

Pollitt, A., Ahmed, A. and Crisp, V. (2007) 'The demands of examination syllabuses and question papers'. In Newton, P., Baird, J.-A., Goldstein, H., Patrick, H. and Tymms, P. (eds) *Techniques for Monitoring the Comparability of Examination Standards.* London: Qualifications and Curriculum Authority, 166–211.

Shaw, S. and Crisp, V. (2011) 'Tracing the evolution of validity in educational measurement: Past issues and contemporary challenges'. *Research Matters*, 11, 14–19.

Shaw, S. and Crisp, V. (2012) 'An approach to validation: Developing and applying an approach for the validation of general qualifications'. *Research Matters*, Special Issue 3, 1–44.

Shaw, S. and Crisp, V. (2015) 'Reflections on a framework for validation: Five years on'. *Research Matters*, 19, 31–7.