An approach to validation: Developing and applying an approach for the validation of general qualifications

Stuart Shaw Cambridge International Examinations **and Victoria Crisp** Research Division

Acknowledgements

The authors would like to thank Nat Johnson, Paul Newton and Gordon Stobart for discussions about this work and the various examiners, subject experts, higher education tutors, teachers and students who gave up their time to participate in the research.

1. Introduction

Ensuring that educational assessments have high validity is a fundamental aim of all those involved in the development of assessments. Being able to provide evidence for the validity of an assessment is increasingly recognised as important. According to Hughes, Porter and Weir "the provision of satisfactory evidence of validity is indisputably necessary for any serious test" (1988, p.4). Given that the outcomes of high stakes assessments, such as those of general qualifications (e.g. GCSE[1], A level[2]), have an effect on students' futures, it is important to know that the inferences being made from results and the ways that results are used are appropriate. This is the essence of validity. However, providing evidence of validity is complex as Kane notes: "validity is conceptually simple, but can be complicated in practice" (2009, p.40). Validity is multifaceted and thus validation evidence has to be derived from a variety of sources and via various methods. An additional challenge is that the theoretical literature on validity and validation is complex and sometimes contested, and even work intended to give practical guidance on conducting validation often does not achieve this (see Brennan, 1998; Lissitz and Samuelsen, 2007; Lissitz, 2009). Whilst the desire within the educational and psychological measurement and assessment community is for more validation studies, the overwhelming challenge is one of "providing a convincing, comprehensive validity argument" (Sireci, 2009, p.33).

At Cambridge Assessment a programme of research has developed a framework for validation studies of general assessments drawing on relevant key literature, and has gathered validation evidence for two A level qualifications offered by Cambridge International Examinations (CIE). In the first phase of this work a structure was developed, a set of associated methods was designed and this was piloted with International A level Geography. In the second phase of work the validation framework was revisited and reworked based on the experience of the pilot and additional reflection on the literature. This, and a revised set of methods, was then applied to International A level Physics. In this issue of *Research Matters*, the framework development and the validation study for A level Physics are described in detail.

This Special Issue begins with a discussion of the concepts of validity and validation and the perspective taken in the current work. The development of the validation framework and later revision will then be described. This is followed by some information on the assessment context in which the validation studies were conducted and an overview of the methods used to gather validation evidence, including some description of some methods that were piloted with A level Geography but were not used in the later study with A level Physics. The methods, analyses and findings from each method are then described for International A level Physics. The validity argument is then evaluated and a final section provides a conclusion and brief comments on challenges in validation activities.

2. Approach to validity

The approach to validity adopted here is that "all validity is of one kind, namely, construct validity" (Messick, 1998, p.37). This approach has become a mainstay of the modern conception of validity.

Within the psychological measurement and assessment community there is a broad professional consensus over the central tenets of modern validity theory which is grounded in a mature conception of construct validity and the validation of interpretations and inferences from test scores. The points of consensus include:

- validity is not an inherent property of a test but refers to the specified uses of a test for a particular purpose (Messick, 1989; Kane, 2001, 2006, 2009; Sireci, 2007, 2009)
- validity pertains to the intended inferences or interpretations made from test scores (Cronbach, 1971; Messick, 1989; Kane, 2006)
- it is the interpretations and uses of test scores that are validated, and not the tests themselves (Cronbach and Meehl, 1955; Cronbach, 1971; Kane, 2006)
- validity measures integrate diverse sources of evidence in the construction of validity claims (Messick, 1989, 1998)
- notion of discrete kinds of validity has been supplanted by the unified view of validity (Loevinger, 1957; Guion, 1978; Messick, 1989)
- all validity is construct validity (Cronbach, 1971; Guion, 1978; Tenpoyr, 1977; Messick, 1975, 1984, 1989; Embretson, 1983; Anastasi, 1986; Cizek, 2011)
- validity is not expressed as a presence or absence of that characteristic but is a matter of degree (Cronbach, 1971; Messick, 1989; Zumbo, 2007; Kane, 2001)

1 GCSEs (General Certificate of Secondary Education) are qualifications, available in various subjects, taken by most students at age 16 years in England, Wales and Northern Ireland. International GCSEs are also available to students outside of the UK.

2 A levels (General Certificate of Education Advanced Level) are qualifications available in various subjects and taken by many students in England, Wales and Northern Ireland at age 18 years. Again international equivalents are available. The A level is a key qualification used in decisions about entry to Higher Education.

- validation is neither static nor a one-time event but an on-going process that relies on multiple evidence sources (Shepard ,1993; Messick, 1989; Sireci, 2007; Kane, 2006)
- evidence needed for validation depends on claims made about a test and proposed interpretations and uses – different interpretations/uses will require different kinds and different amounts of evidence for their validation (Kane, 2006, 2009)
- recognition that validation processes and validity evidence are value-laden (Messick, 1989)

Cambridge Assessment sees a vital aspect of validity as "the extent to which the inferences which are made on the basis of the outcomes of the assessment are meaningful, useful and appropriate" (2009, p.8) and argues that the concern for validation "begins with consideration of the extent to which the assessment is assessing what it is intended to assess and flows out to the uses to which the information from the assessment is being put" (2009, p.8).

3. Framework development

An early stage of the current research involved reviewing literature on validity and validation, including existing frameworks designed to support validation activities in various contexts. This led to the development of a proposed framework for use with general qualifications drawing on commonalities from previously suggested structures. After the pilot study with A level Geography the framework was revised in the light of further literature and reflection, which was then used as the structure for the main study with A level Physics. Both frameworks involved a list of validation questions, each of which is to be answered by the collection of relevant evidence. The findings of validation studies based on the framework would present '*Evidence for validity*' and any potential '*Threats to validity*'. This section will describe the development of the first and second frameworks.

Initial framework (as applied in the pilot study)

The initial framework for validating general qualifications was developed by reviewing previously proposed frameworks designed for various contexts. Some recent examples, which informed the development of the proposed framework, will now be reviewed. The earliest example of interest is a perspectives-based approach suggested by Cronbach (1988). Cronbach viewed validity as linking concepts, evidence, social and personal consequences, and values. In order to provide a structure for validity enquiries, he organised these into five categories of questions that should be asked about tests:

- the *functional* perspective – about whether there are appropriate consequences for individuals and institutions;
- the *political* perspective – which looks at the role of stakeholders in deciding whether a test is fair;
- the *operationist* perspective – to do with the match of test content to the domain of performance and the fit of this domain of performance with the testing purpose;
- the *economic* perspective – about whether the test score predicts how well a person will perform on a relevant future course, or in a relevant job role;
- and the *explanatory* perspective – whether the reality of the assessment matches up to theoretical ideas.

Around the same time Messick (1989) described six aspects of construct validity that should be addressed in any validation exercise:

- *Content* – relating to an examination of the content relevance and representativeness;
- *Substantive* – relating to the theoretical justifications for the observed consistencies (and inconsistencies). Messick notes the importance of comparability between the cognitive processes underpinning assessment performance and the cognitive processes underlying performance in practice;
- *Structural* – relating to the reliability of scoring procedures and processes;
- *Consequential* – relating to the consequences of the assessment for the individual assessed;
- *Generalisability* – the degree to which score properties or score interpretations can be generalised to (and across) populations, settings, and tasks;
- *External* – the relationship between assessment scores and scores on other assessments designed to measure the same construct.

One shortcoming of Messick's framework is that it does not assist other validators in terms of how, in practice, to draw conclusions about the evidence (see Kane, 2006; Brennan, 1998; Crocker, 2003).

Several frameworks have focused on the authenticity of assessments and on validating performance assessments. Frederiksen and Collins (1989) focused on the concept of "systemically valid tests as ones that induce curricular and instructional changes in education systems (and learning strategy changes in students) that foster the development of the cognitive traits that the tests are designed to measure" (1989, p.27). They present two key characteristics of tests which affect the usefulness of educational assessments as facilitators of educational improvement: the directness of cognitive assessment, and the degree of judgement involved in assigning scores that represent a cognitive skill. Frederiksen and Collins propose that the testing system should address:

- *Directness of measurement* – directness of cognitive assessment, authenticity;
- *Scope* – "the test should cover as far as possible all knowledge, skills and strategies required to do well in the activity" (1989, p.30);
- *Reliability of scoring*;
- *Transparency* – "the terms in which candidates will be judged must be clear to them if a test is to be successful in motivating and directing learning" (1989, p.30) – they should be able to assess themselves.

Linn, Baker and Dunbar (1991) proposed a model focused on performance assessment which was arguably more comprehensive than that of Frederiksen and Collins. Linn, Baker and Dunbar argued a need to expand traditional criteria to evaluate the quality of performance-based, authentic assessments and hence proposed a model for validation with this focus. They argued that evidence in relation to eight areas should be gathered:

- *consequences* – evidence of intended and unintended effects on teaching practices and student learning;
- *fairness* – issues of equitable access for all, avoiding bias, fair scoring unaffected by irrelevant difficulty;
- *transfer and generalisability* – evidence regarding the extent to which performance on specific assessment tasks transfers to other tasks;

- *cognitive complexity* – evidence of whether assessment tasks require students to use the level of complexity of cognitive processes intended;
- *content quality* – relating to whether tasks represent the relevant constructs and match current understanding in the field being tested;
- *content coverage* – the comprehensiveness of the content covered by the assessment;
- *meaningfulness* – the degree to which the tasks are meaningful to the students and provide worthwhile educational experiences;
- *cost and efficiency* of the assessment procedure.

In order to provide a structure for evaluating assessment validity, Crooks, Kane and Cohen (1996) represent validation concerns as a chain of linked stages where a weak link weakens the whole chain. They set out that in a particular validation study the importance of each link in light of the assessment's purpose(s) should first be determined. Then, relevant evidence should be collected and analysed in order to evaluate strengths and threats to validity for each link. The links are:

- *administration*, where threats would include stress affecting performance or inauthentic assessments;
- *scoring*, where one potential threat might be that scores do not capture important qualities of performance;
- *aggregation*, where one possible threat would be giving inappropriate weights to different aspects of assessment;
- *generalisation* from particular tasks to the whole domain of similar tasks, where a threat would be a sample of tasks being too small;
- *extrapolation* from the domain of appropriate tasks to all tasks relevant to the proposed interpretation, where a threat would occur if the tasks do not represent all relevant tasks (this is similar to 'construct under-representation', Messick, 1989);
- *evaluation* of performance or forming judgements about what scores mean, where threats include inadequately supported interpretations of scores;
- *decisions* made on the basis of judgements, where threats would include decisions that are inconsistent with the information on which they are based;
- the *impact* of assessment processes, interpretations and decisions, where threats include the occurrence of negative consequences.

Another interesting approach is that of Mislevy, Steinberg and Almond (2003). Evidence-centred design (ECD) is an approach to constructing assessments in terms of evidence-based arguments intended to support inferences about individuals (Almond, Steinberg and Mislevy, 2002). ECD provides a conceptual design framework for the components of a coherent assessment. It can be applied to support a wide variety of assessment types. ECD deconstructs assessment development into three distinguishable features: the desirable claims to be made about an individual; the accumulation of evidence in support of these claims; and the assessment instruments that elicit observations of the evidence (Bennett *et al*., 2003; Mislevy, 1994; Mislevy, Steinberg and Almond, 2003). The conceptual assessment framework, the core of the evidentiary reasoning argument, is formulated in terms of three models. Each model responds to a key assessment question:

- the *student model* – What complex of knowledge, skills, or other attributes should be assessed?
- the *evidence model* – What behaviours or performances should reveal those constructs, and what is the connection?
- the *task model* – What tasks or situations should elicit those behaviours?

Student models define variables related to the specific knowledge, skills and abilities which are the focus of inferences. *Evidence models* provide a detailed argument relating to why and how observations in a given task situation represent evidence about student model variables. *Task models* offer a framework for constructing and describing the situations in which candidates behave. Task models describe how to configure the types of situations necessary to obtain the evidence required for the evidence models.

The purported benefits of the ECD approach have been used to support the claims of greater validity for the New Generation Test of English as a Foreign Language (NG TOEFL). The development of the new TOEFL outlines how aspects of an ECD approach have been adapted to provide the 'validity argument' for the revised test (Chapelle, Enright and Jamieson 2004, 2008, 2010). This argument, drawing heavily on the work of Kane (2004, 2006), combines 'process with evidence'. Chapelle *et al*. (2010) outline Kane's approach to validation and adapt this to support an 'interpretive argument' setting out the intended inferences and interpretations from test performances. This includes six inferences (evaluation, generalisation, domain description, explanation, extrapolation and utilisation).

In the UK, a recent example framework is the work of Weir (2005) and Shaw and Weir (2007) who proposed a test validation framework based around a unitary concept of validity but including constituent validity elements. These elements reflect the practical nature and quality of an actual testing event and constitute the various types of validity evidence to be collected at each stage in the test development process. The validity elements set out are:

- *test taker characteristics* and whether these are accommodated by the assessment;
- *cognitive validity*, relating to the appropriateness of the cognitive processes required of students to complete tasks;
- *context validity*, relating to the performance setting and conditions under which the assessment is taken;
- *scoring validity*, relating to the dependability of the scoring;
- *criterion related validity*, about the extent to which scores correlate with other measures;
- *consequential validity*, relating to whether social consequences of test interpretation support the intended testing purposes.

This framework has been used as a structure for gathering validity evidence in the context of testing English as a second language (Shaw and Weir, 2007; Khalifa and Weir, 2009) and is perhaps the first comprehensive attempt by a UK examination board to expose the totality of its practice to scrutiny in the public arena.

Although not providing a validation framework, other work in the context of national examinations in the UK by Pollitt, Ahmed and colleagues (e.g. Pollitt, Hutchinson, Entwhistle and de Luca, 1985; Pollitt and Ahmed, 1999; Ahmed and Pollitt, 2007; Crisp, Sweiry, Ahmed and Pollitt, 2008; Ahmed and Pollitt, 2008a, 2008b) provided the basis of some possible validation methods. Their research involved analysing the extent to which exam questions measure appropriate cognitive processes. They argue that in order for scores to be a valid reflection of

relevant constructs, question writers need to be in control of the kind of thought processes that an exam question elicits in a student's mind. They use the following definition: "an exam question can only contribute to valid assessment if the students' minds are doing the things we want them to show us they can do; and if we give credit for, and only for, evidence that shows us they can do it" (Ahmed and Pollitt, 2008a, p.3). This definition overlaps with concerns for 'process models' (Messick, 1989, 1995) or 'cognitive validity' (Shaw and Weir, 2007), and validity in relation to 'scoring' (Messick, 1989, 1995; Crooks, Kane and Cohen, 1996; Shaw and Weir, 2007). Thus, the work of Pollitt and Ahmed takes a more focused view of validity and shows less concern for broader aspects of validity (e.g. the inferences made from test scores, consequences of test use). Whilst some might argue this is a limited view of validity, the substantial advantage of this approach has been to allow in-depth research into one of the central elements of validity – ensuring that scores accurately reflect relevant constructs and that questions function as intended. Consequently their work provides useful insights into appropriate methodologies for exploring these aspects of validity within a wider framework.

Although validity is now considered a unified concept, most theorists have found it necessary to identify different aspects or components of validity or different types of evidence needed to support a claim of validity. Whilst the existing frameworks have their merits, it seemed necessary to synthesise a new structure drawing on these to ensure that all important aspects were included and that the language used was sufficiently accessible to assist in making the framework operational. To assist in this synthesis, existing models were compared to look for overlap and differences in the themes arising.

The models and existing theories of validity described in the literature were used to develop the components of validity within the initial framework, ensuring that common elements from previous frameworks were included. This initial framework is illustrated in Figure 1.

As can be seen, the contributors to validity are grouped by purpose and constructs, sampling and generalisability issues, and impact and inferences. Each contribution to validity was expressed as a validation question. The intention was that appropriate evidence should be collected in relation to each validation question, using them like research questions, thus prompting the consideration of various types of pertinent evidence and providing a wide-ranging evaluation of an assessment's validity. The evidence collected in relation to each validation question would then be summarised in terms of the positive evidence that it provides for validity ('evidence for validity') and in terms of potential 'threats to validity'. Any identified threats to validity might provide advice for the test developers/question writers in future sessions, or might suggest recommendations for changes to an aspect of the qualification, its administration and procedures or associated documentation. This framework was used as the basis for the pilot study in which validation evidence relating to an International A level in Geography was collected.

Revised framework (as applied in the main study with A level Physics)

After the pilot study, the validation framework was revised. This was mostly a result of further examination and reflection on the theoretical literature and in response to various feedback received during external dissemination of the work.

Figure 1: Initial framework for the argument of assessment validation (as applied in the pilot study with A level Geography)

Contribution to validity	*Evidence for validity*	*Threats to validity*
Assessment purpose(s) and underlying constructs* 1.1 What is (or are) the main declared purpose(s) of the assessment and are they clearly communicated? 1.2 What are the constructs that we intend to assess and are the tasks appropriately designed to elicit these constructs? 1.3 Do the tasks elicit performances that reflect the intended constructs? ** By constructs we mean knowledge/understanding/skills that will be reflected in test performance, and about which inferences can be made on the basis of test scores.*		
Adequate sampling of domain, reliability and generalisability 2.1 Do the tasks adequately sample the constructs that are important to the domain? 2.2 Are the scores dependable measures of the intended constructs?		
Impact and inferences 3.1 Is guidance in place so that teachers know how to prepare students for the assessments such that negative effects on classroom practice are avoided? 3.2 Is guidance in place so that teachers and others know what scores/grades mean and how the outcomes should be used? 3.3 Does the assessment achieve the main declared purpose(s)?		

One of the first challenges in a consideration of how the framework needed to be revised was how best to characterise our approach to validity in terms of framing the intended score interpretation: as construct (AERA, APA, and NCME, 1999) or as interpretive argument (Kane, 2006).

The current *Standards for Educational and Psychological Testing* broadly reflects the prevailing views in educational measurement throughout the 1990s, namely, a construct-centered approach to validity (Messick, 1989), a perspective that draws heavily on Messick's view that the construct, through which test scores are interpreted is the foundation for evaluating a test (Messick, 1994).The *Standards* states that "all test scores are viewed as measures of some construct." (AERA, APA, and NCME, 1999, p.174) and frame validity largely in terms of "the concept or characteristic that a test is designed to measure." (AERA, APA, and NCME, 1999, p.5). From this perspective, validity is viewed as "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" (Messick, 1989, p.13).

A recent perspective from Kane calls for an argument-based approach. Kane's work is increasingly gaining credibility as an alternative approach for thinking about validity (see, for example, Kane, 2006; Haertel and Lorie, 2004; Mislevy, *et. al*., 2003; Shaw, Crisp and Johnson, 2012). Central to Kane's approach is the interpretive argument: "Validation requires a clear statement of the proposed interpretations and uses" (Kane, 2006, p.23). The interpretive argument is consistent with the general principles accepted for construct validity that appear in the *Standards*.

The concept of validation adopted here reflects Kane's argument-based approach. Kane defines validation as "the process of evaluating the plausibility of proposed interpretations and uses" (2006, p.17). According to Messick (1989), validation entails ascertaining "the degree to which

multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported" (p.13). It also entails "appraisals of the relevance and utility of test scores for particular applied purposes and of the social consequences of using the scores for applied decision making" (Messick, 1989, p.13).

By setting out a chain of inferences that can be defined in many different ways, a more manageable basis for test score interpretation is provided. Thus test scores are interpreted in terms of constructs and it is the construct-based interpretation of results that needs to be validated. It is the evidence from multiple validation methods that should be used to support a claim of (construct) validity.

The main changes to the framework, therefore, were a greater emphasis on the inferences made from performances and the intended interpretations of results (via the construction of a more explicit 'interpretive argument') and a stronger argument-based underpinning. The framework became more strongly influenced by the recent work of Kane (2006, 2009), whilst continuing to emphasise construct.

Revisions were needed to the framework in order to provide an interpretive argument as well as a validity argument. The interpretive argument for an assessment sets out the proposed interpretations of the assessment outcomes based on a set of inferences from student performance or scores, through to decisions based on scores. It also describes the assumptions supporting each inference. The concept of the interpretive argument proposed by Kane (1992, 1994) and Kane, Crooks and Cohen (1999) emphasises clarity of argumentation, coherence of argument, and plausibility of assumptions. The validity argument involves evaluating the interpretive argument based on relevant evidence which will need to be collected.

Kane is not the first to suggest that validation be conducted in terms of the development of an argument structure. Cronbach (1988) and House (1977) also used the notion of argument in relation to validation efforts. Toulmin (1958/2003) provides a useful model of the structure of argument as involving: a *claim*; a *warrant* that justifies the inference being made from data to support a claim; *backing* evidence; and *rebuttals* (alternative explanations or counter claims). Kane (2006) suggests the use of such an argument-based structure in validation studies as the testing organisation is implicitly or explicitly making a claim that the assessment is valid, which effectively needs to be backed up with evidence through a warrant, and needs to be free from significant rebuttals.

The nature of the interpretive argument that needs to be set out will depend on the purposes of the assessment (Kane, 2006). He gives an example of the interpretive argument for a test used to place students onto courses and sets out the inferences in this situation as: *scoring* (inference from observed performance to an observed score); *generalisation* (from the observed score to universe score, i.e. score for the range of tasks falls within the domain of the qualification); *extrapolation* (from universe score to the level of skill); and *decision-making* (from conclusion about level of skill to placement on a specific course). The revised framework built on this set of inferences for placement tests. However, the aim was still to make the framework more accessible than the literature on validation generally tends to be. Thus, each inference in the interpretive argument was (as in the earlier framework) expressed as a validation question as well as by the kind of labels used by Kane. The revised framework is presented in Figure 2 (reproduced, with permission, from Shaw, Crisp and Johnson, 2012). As with the earlier version of the framework, each of the validation questions is to be answered by the collection of relevant evidence and the findings from validation exercises based on the framework would present '*Evidence for validity*' and any potential '*Threats to validity*'. All five of the validation questions in the revised framework come from the initial framework, although one of the questions on sampling has been split into two to separate sampling of the syllabus (i.e. course content

Figure 2: Revised framework for the argument of assessment validation (as applied in the main study with A level Physics)[3]

Interpretive argument		*Validity argument*	*Evaluation*	
Inference	*Warrant justifying the inference*	*Validation questions*	*Evidence for validity*	*Threats to validity*
Construct representation	**Tasks elicit performances that represent the intended constructs**	1. Do the tasks elicit performances that reflect the intended constructs?		
Scoring	**Scores/grades reflect the quality of performances on the assessment tasks**	2. Are the scores/grades dependable measures of the intended constructs?		
Generalisation	**Scores/grades reflect likely performance on all possible relevant tasks**	3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?		
Extrapolation	**Scores/grades reflect likely wider performance in the domain**	4. Are the constructs sampled representative of competence in the wider subject domain?		
Decision-making	**Appropriate uses of scores/grades are clear**	5. Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?		

Evaluation of claim	**Evidence for validity**	**Threats to validity**
How appropriate are the intended interpretations and uses of test scores? Interpretation 1. Scores/grades provide a measure of relevant learning/achievement Interpretation 2. Scores/grades provide an indication of likely future success		

3 This framework is reproduced, with permission, from: Shaw, Crisp and Johnson (2012) A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy and Practice*, **19**, 2, 159–176.

and skills) from sampling of the wider subject domain. Several questions from the initial framework were not included in the revised framework. Firstly, the first and last questions about assessment purposes were removed from the framework as these should be known in advance of a validation study, and these determine the intended interpretations of assessment outcomes which are central to the overall evaluation of validity. Secondly, the second question in the initial framework was removed and partly subsumed into the following question, because the constructs should be set out in some form in the syllabus rather than needing to be retrieved. Thirdly, the question on guidance for teachers on preparing students for assessments has not been included in the revised framework. Whilst the issues of how assessments influence associated classroom practice are of importance, they are arguably less central to validity if we take Messick's (1989) definition of validity as the appropriateness of inferences and uses of assessment outcomes. However, the exclusion of this question from the revised validation framework, does not prevent parallel investigation of this form of impact alongside a validation study.

The revised framework also provides an 'interpretive argument' with each validation question representing one inference. These inferences (see first column in Figure 2) make up an interpretive chain which runs: from the task to the test performance ('construct representation'); from the test performance to the test score ('scoring', which includes issues of categorising scores into grades); from test score to test competence ('generalisation'); from test competence to domain competence ('extrapolation'); and from domain competence to trait competence ('decision-making'). An associated warrant is provided with each inference (second column in Figure 2). The warrant is a statement that is claimed to be true and justifies the related inference if appropriately supported by evidence via the validity argument.

As with the earlier framework, the validation questions are intended to prompt the collection of relevant evidence, but in the revised framework these questions are strongly linked to an interpretive argument. It should be noted that although the validation questions are phrased such that a 'yes' or 'no' answer is required, this is for reasons of simplicity and it is intended that the evidence collected will provide qualified answers on the extent to which each inference is appropriate.

Below the main table of Figure 2 is a smaller table representing the evaluation of the claim. This table prompts the use of the evidence collected to evaluate the appropriateness of each of two proposed interpretations of A level results (and hence their associated uses).

4. International A level context

A levels are taken by many students at age 18 and are a key qualification for access to higher education study. Whilst primarily a UK qualification, they are also available internationally through some UK-based exam boards as International A levels. The current work was conducted in the context of the latter qualifications. For the pilot study, A level Geography was chosen as the focus because its assessment uses a range of question lengths, because it is a popular subject, and because the nature of the subject provides a reasonable basis from which to consider the generalisability of the validation method to a range of other subjects (e.g. other humanities). The main validation study used International A level Physics as its focus.

A level Physics

The assessment of the Physics A level course is via five exam papers, the first three of which make up the Advanced Subsidiary (AS) assessment. All exam questions are compulsory.

- **Paper 1: Multiple choice**. A one hour exam covering a range of core physics topics assigned for the AS. The exam involves 40 multiple choice items each with four response options and is worth 40 marks.
- **Paper 2: AS Structured questions**. A one hour exam covering a range of core physics topics assigned to the AS. The exam involves structured questions requiring constructed responses. Question parts range in the mark available from 1 mark to around 6 marks. The paper is worth 60 marks in total.
- **Paper 3: Advanced practical skills**. A two hour practical exam involving setting up equipment and taking measurements, and answering various related questions. The exam is worth 40 marks.
- **Paper 4: A2 Structured questions**. An exam lasting one hour and 45 minutes covering a range of physics topics assigned to A2 learning by the syllabus. The paper involves structured questions requiring constructed responses. Question parts range in the mark available from 1 mark to around 6 marks and the paper is worth 100 marks in total.
- **Paper 5: Planning, analysis and evaluation**. An exam lasting one hour and 15 minutes assessing practical skills of planning an experiment and analysing and evaluating experimental data.

The October/November 2009 examinations, as the most recent exam session at the time of the research, were chosen as the focus for the validation study.

For Paper 3, two alternative versions of the exam are available in any session to assist with security where schools/colleges have a large number of entrants and cannot administer the practical examination to all candidates in one day.

As a result of security issues, the use of different exam papers for students in different time zone based areas has been introduced for CIE assessments. For the November 2009 examinations, this meant that there were two time zone areas identified as Zone P and Zone Q. Thus for each exam paper there were two versions of it in this session. For example, for Paper 1 candidates in Zone P took Paper 11, and those in Zone Q took Paper 12. For Paper 3 this resulted in four versions of the test: Paper 31 or 32 in Zone P, and Paper 33 or 34 in Zone Q.

Zone P exam papers 11, 21, 31, 41 and 51 were selected as the main focus of the validation study. The exam papers and mark schemes were obtained along with the syllabus specification. Samples of scripts for papers 21, 31, 41 and 51 were also obtained and item level scores were keyed. Additional papers from the five previous sessions were also obtained for use in one of the methods.

5. Methodological overview

A suite of methods was designed for the pilot study, such that a number of evidence types would be collected in relation to each validity question in the framework. They included a range of quantitative and qualitative methods. The set of methods used involved:

- a series of tasks conducted by senior examiners and external experts, such as identifying assessment constructs, and rating the coverage of Assessment Objective subcomponents;

- document reviews, for example, in relation to scoring procedures;
- statistical analyses of item level data (such as analysis of question difficulty and functioning using scores on questions, factor analysis);
- a multiple re-marking study, involving five markers for each paper, to explore marking reliability;
- questionnaires to teachers and to Higher Education institutions;
- interviews with students after they had answered exam questions.

The set of methods was revised and streamlined for the main study based on the experience of the pilot. Some methods which contributed only limited additional insights were dropped, and two additional methods were used instead. The revised set of methods were mapped onto the revised validation framework. The changes to the methods will be described in more detail shortly.

Thus, we have developed a suite of methods including qualitative and quantitative evidence types. We would not claim that these represent an exhaustive list of all possible types of evidence that may inform regarding an assessment's validity – this would necessarily be highly impractical. However, in both studies, the methods used generated a substantive body of evidence in relation to different contributions to validity. The revisions for the main study streamlined the set of methods by removing those that were less productive, and bridged one or two gaps in the evidence types gathered in the pilot study.

The difficulty of obtaining data on prior or future attainment, or data on concurrent measures of ability, is an issue facing a number of awarding bodies. Some approaches to construct validation have traditionally focused on a range of quantitative methods including factor analysis; correlations between a measure of the construct and the designated construct theory; multitrait-multimethod matrix; correlational analysis of how well performance on an assessment predicts future performance; and, analysis of variance components within a generalisability theory framework (Crocker and Algina, 1986). The success of these techniques is contingent upon a well-articulated construct theory and the availability of appropriate data. However, from a CIE perspective, the international location of candidates makes obtaining such data challenging, making some of these techniques problematic to undertake. Additionally, given the nature of many CIE qualifications, statistical methods such as multitrait-multimethod analyses of convergent and divergent tests are less appropriate than they are for psychometric tests of personality traits, or skills such as verbal fluency or numeracy. The difficulty of obtaining linked assessment data for a CIE qualification led to us not pursuing such methods in the research described here[4].

Methods used in pilot study

The pilot study will not be reported in full in this Special Issue but, to give a feel for the methods used, brief details are given below. Table 1 gives an overview of the methods used in order to address each of the questions in the validity framework.

The experts mentioned in the table of methods were six geography experts. Four were senior examiners involved in setting and marking the A level Geography papers, and two were external geography experts (one was a senior examiner for two UK exam boards and the other was an experienced practicing teacher who has written a number of key textbooks for A level Geography). The experts attended a two day meeting at which some of the tasks (Tasks 1–5) were conducted, and then conducted a further three days work at home to complete the other tasks (Tasks 6–11).

Note that, unfortunately, none of the Higher Education Institutes contacted completed the relevant questionnaire. The recruitment strategy for this method was revisited in the planning of the main study with A level Physics to improve the response rate.

Methods used in main study

Between the pilot and main studies a number of methods were removed or refined to somewhat streamline the validation effort and to exclude methods that were not sufficiently productive given the amount of resourcing that had to be put into them. Some changes were a result of the revisions to the validation framework and further consideration of the literature on validation. Methods relating to retrieval of assessment purposes were removed and replaced with purposes being discerned in advance through consultation with key internal exam board colleagues. The identification of constructs by experts through comparisons of triplets of questions, and associated rating tasks were also removed. This activity was very time-consuming, and the list of constructs produced was slightly problematic (e.g. some overlap between constructs, some differences between experts in the constructs proposed). The activity to identify processes expected to be involved in answering questions and actual processes used in the pilot was also removed. This task was time-consuming, and not very fruitful in terms of insights for validity – it did not provide additional insights into problem questions beyond those that were provided by analysis of performance data. The use of interviews with students is considered a useful activity for validation of the processes involved in answering questions. However, the current context of international A levels meant that the interviewing of students was conducted by teachers rather than the researchers, was very small scale (allowing only a small number of exam questions to be investigated in this way) and often with students for whom English was not their first language. These issues led to the data being less useful in the pilot than had been hoped.

One method used in the main study was new. This looked at the usefulness of A level study as preparation for university. This involved several appropriate Higher Education lecturers being asked to evaluate the importance of elements of the A level syllabus as preparation for university study.

Table 2 gives an overview of the methods used in the main study where the focus was A level Physics. The findings will be presented as a number of separate evidence types labelled 'Validity Evidence 1' and so on. The numbering of the evidence types is shown in the table to help with navigation. The examiners/experts mentioned in relation to several methods were a group of four senior examiners and two other subject experts.

4 Note that there could be future potential for such methods if data were to become available. For example, some data on UK students taking International equivalents of GCSE are now included in the National Pupil Database (a database of qualifications achieved by students in England, Wales and Northern Ireland) which may make this kind of analysis more plausible in the future.

Table 1: Overview of methods used in the pilot study

1. ASSESSMENT PURPOSE(S) AND UNDERLYING CONSTRUCTS	
Question	*Method*
1.1 What is (or are) the declared main purpose(s) of the assessment and are they clearly communicated?	Task 1 conducted by experts: Identifying assessment purposes – Experts recovered the purposes with reference to the syllabus specification and other exam board documents.
	Task 2 conducted by experts: Writing an 'Importance Statement' – Experts wrote around 200 words about what is important in Geography A level.
1.2 What are the constructs we intend to assess and are the tasks appropriately designed to elicit these constructs?	Document review: Retrieving the declared constructs – Researchers retrieved constructs from the syllabus specification and other exam board documents.
	Task 3 conducted by experts: Identifying constructs by comparing questions – Experts looked at triplets of questions (with mark scheme and any figures) in turn and identified differences and similarities in the constructs tested. Then used these to produce a collated list of constructs.
	Task 6 conducted by experts: Rating the extent to which the constructs are triggered by each question – Using a list of constructs synthesised by the researchers from all experts' responses to Task 3, experts rated the extent to which each construct was triggered by each exam question.
	Task 7 conducted by experts: Rating the demands of the tasks indicated by the questions – Experts rated the demands of the questions against a number of demand types (without using mark scheme).
1.3 Do the tasks elicit performances that reflect the intended constructs?	Student data: Interviews with students after they had answered an exam question – Teachers asked students to answer an exam question and then interviewed them about how they answered the question and any difficulties they had.
	Task 5 conducted by experts: Identifying processes expected and apparent in student responses – Experts were asked to look at a selection of exam questions and note the processes that they expect students to have to undertake to answer and to then look at some student responses and infer the processes that students actually used.
	Analysis of performance data: Exploring question difficulty and functioning – Item level score data for a sample of candidates was obtained and analysed using traditional item statistics, Rasch analysis, factor analysis and qualitative analyses of responses on questions of interest.
	Document review: Gathering insights on question answering from examiner reports – Researchers reviewed examiner reports on the exam papers to gain insights into how the questions were answered.
2. ADEQUATE SAMPLING OF DOMAIN, RELIABILITY AND GENERALISABILITY	
Question	*Method*
2.1 Do the tasks adequately sample the constructs that are important to the domain?	Task 8 conducted by experts: Reviewing the coverage of content and skills over three years of the examinations (June 06 to Nov 08) – Experts identified the sub-topic(s) and specific content of each question and the marks available for each Assessment Objective. (Each expert focused on two sessions).
	Task 4 conducted by experts: Rating the extent to which Assessment Objective subcomponents are measured – Experts rated the extent to which each subcomponent of the Assessment Objectives were measured by each question.
2.2 Are the scores dependable measures of the intended constructs?	Task 9 conducted by experts: Rating the extent to which the constructs are reflected by scores – Using a list of constructs synthesised by the researchers from all experts' responses to Task 3, experts rated the extent to which each construct was rewarded by the mark scheme for each question.
	Task 10 conducted by experts: Rating the demands measured by scores – Experts rated the demands measured by scores for each question, as indicated by the mark schemes. Multiple re-marking exercise: Re-marking data to explore marking consistency/reliability – For each exam paper, five markers re-marked 30 scripts.
	Documentary review: Review of marking and scoring procedures – Researchers reviewed documents on marking and scoring procedures.
	Statistical analysis regarding aggregation: Exploring possible aggregation issues – Data on student performances on individual exam units and their aggregated scores for the complete qualification were used to investigate the achieved weightings of components.
3. IMPACT AND INFERENCES	
Question	*Method*
3.1 Is guidance in place so that teachers know how to prepare students for the assessments such that negative effects on classroom practice are avoided?	Documentary review: Review of guidance on teaching – Researchers reviewed various exam board documents and teacher guidance on preparing students for the assessments.
	Teacher questionnaire: Gathering views about guidance on student preparation – Teachers completed a questionnaire, part of which asked about guidance on student preparation.
3.2 Is guidance in place so that teachers and others know what scores/grades mean and how the outcomes should be used?	Documentary review: Review of guidance on score/grade meaning and use – Researchers reviewed various exam board documents and guidance on grade/score meaning and use.
	Teacher questionnaire: Gathering views about guidance on score/grade meaning and use – Teachers completed a questionnaire, part of which asked about guidance on grade/score meaning and use.
	Higher Education Institute questionnaire: Gathering views about guidance on score/grade meaning and use – International Admissions Officers in Higher Education Institutes were asked to complete a questionnaire, part of which asked about guidance on grade/score meaning and use.
3.3 Does the assessment achieve the main declared purpose(s)?	Task 11 conducted by experts: Rating the extent to which assessment purposes are met – Using a list of purposes synthesised by the researchers from the experts' work on Task 1, the experts rated the extent to which each of the purposes identified were/are met by the assessments.
	Teacher questionnaire: Gathering teacher views about whether purposes are met – Teachers completed a questionnaire, part of which involved rating the extent to which assessment purposes are achieved.
	Higher Education Institute questionnaire: Gathering views about whether purposes are met – International Admissions Officers in Higher Education Institutes were asked to complete a questionnaire, part of which involved rating the extent to which assessment purposes are achieved.

Table 2: Overview of methods used in the main study

Validation question	*Method*	*Validity Evidence*
1. Do the tasks elicit performances that reflect the intended constructs?	Analysis of performance data (e.g. item level scores) for a sample of candidates using statistical methods (e.g. Rasch, factor analysis) to explore item functioning and relationships between items.	1, 2 and 4
	Review of examiner reports for insights into how the questions were answered by candidates.	3
	Appropriate examiners/experts rated the extent to which each question appears to elicit each assessment objective set out in the syllabus (using this as a proxy for the constructs).	5
	Appropriate examiners/experts rated the extent to which each question places certain types of cognitive demands on students.	6
	For misfitting items, analysis of the nature of candidate responses to gather insights into any possible sources of construct irrelevant variance.	7
2. Are the scores/ grades dependable measures of the intended constructs?	Review of exam board documents on marking and scoring procedures.	8
	For each paper a number of examiners marked the same exam scripts in a multiple re-marking exercise so that the consistency and reliability of marking could be analysed.	9
	Composite reliability analysis.	10
	Statistical analyses of candidate exam results to explore issues relating to aggregation of test scores and intended and achieved weightings of exam components.	11
3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?	Appropriate examiners/experts identified the topics and sub-topics assessed by each exam question for a number of exam sessions in order to evaluate content and skills coverage.	12
	Appropriate examiners/experts rated, for each exam question, the extent to which the scoring guidelines set out in the mark scheme reward each assessment objective.	13
	Appropriate examiners/experts rated, for each exam question, the cognitive demands rewarded by each question, as reflected in the mark scheme.	14
4. Are the constructs sampled representative of competence in the wider subject domain?	Higher education representatives reviewed the syllabus content in relation to the preparation it provides for further study.	15
5. Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?	A questionnaire to stakeholders (e.g. higher education providers) to gather their views on guidance on score/grade meaning and uses and gather insights on how they use scores/grades.	16
	A questionnaire to teachers to gather their views on guidance on score/grade meaning and uses and gather insights on how they use scores/grades.	17
	Review of guidance documents relating to score meaning and use.	18

6. Constructing an interpretive argument for International A level Physics

Issues arising from the pilot study

As already described, reflection on the experience of the pilot, further reading and feedback in response to dissemination led to a revised validation framework which was applied in the main validation study with A level Physics. In addition, the set of methods was reviewed and revised as described in the overall methodology section. A number of other changes of approach also arose. A greater focus on the work of Kane (2006, 2009) developed which influenced the adjustments made. Chapelle *et al.'s* (2008, 2010) work also informed the assumptions underlying the interpretive argument.

One of the areas that we felt needed to be done differently from the pilot study was that of identifying the assessment purposes. These ought to be known in advance of a validation study rather than having to be retrieved. The assessment purposes for International A level were somewhat implicit and assumed in documentation, apart from the strong function of the qualifications as providing recognition for university entrance. Thus the purposes did need to be drawn out but we came to the view that this should be determined internally at the exam board, rather than needing to be retrieved by external experts. Key internal staff were therefore consulted and a set of purposes determined and linked to the interpretations to be made from scores. This links to a second issue of the need to set out the interpretive argument for an assessment. According to Kane, this is an important part of validation and an argument for validity needs to be set out against this. The interpretive argument should set out the ways of interpreting scores/grades that are claimed to be valid, along with the inferences that are claimed to be appropriate and the assumptions on which these are based.

Validation, according to Kane, also requires that the interpretive argument and the validity argument (which includes the evidence gathered) are evaluated. In other words, the appropriateness of the score/grade interpretations that are claimed need to be evaluated. How to conduct this overall evaluation in a more detailed way than we achieved with the pilot needed to be ascertained.

Because of the points raised above, a number of additional issues were addressed before beginning the validation study of A level Physics. In this section, the proposed uses (or purposes) of A level Physics are set out as this is the logical starting point for a validation exercise, and the interpretive argument is set out.

Proposed uses and interpretations

Validity is a property of the argument for interpreting outcomes from a given assessment in a certain way and, therefore, for using those outcomes in a certain way. According to Kane, "to validate a proposed interpretation or use of test scores is to evaluate the rationale for this interpretation or use. The evidence needed for validation necessarily depends on the claims being made. Therefore, validation requires a clear statement of the proposed interpretations and uses" (2006, p.23). Cronbach (1971) distinguishes between test interpretation (using a test to describe a person) and test use (using a test to make decisions about the person).

The assessment purposes of A levels were identified in advance of the validation study by consulting key exam board personnel, and keeping in mind the distinction between claims made by the exam board for appropriate purposes/uses, and uses that others might make of the data.

A key point to note here relates to whether validation efforts should focus only on the intended inferences (where the particular inferences derive from the intended uses/purposes), or whether unintended actual inferences and uses should also be evaluated. The validation research reported here focused mainly on intended inferences, but with some consideration of anticipated inappropriate inferences. The underlying assumption here is that CIE has a responsibility to provide some explanation of what inferences can and cannot be drawn, but that there is a limit on the purposes to which we as an organisation might choose to claim as appropriate. If an outside organisation chose to use our assessments for a purpose that we do not claim as appropriate and have not validated (e.g. use of grades in school league tables), arguably it is the responsibility of that organisation to validate that purpose/use, although we might wish to advise on the appropriateness of this usage based on evidence available to us.

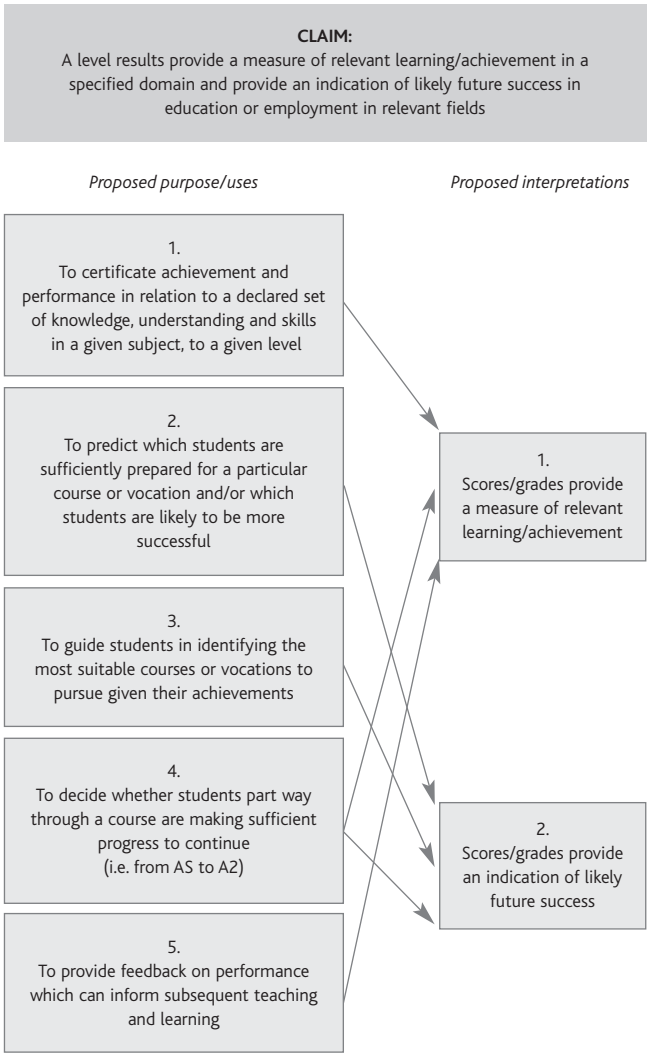
The claims we wish to make about International A level Physics are that:

1) scores provide a measure of relevant learning/achievement – this can be thought of in terms of an 'attainment' construct, for example, whether a student has made satisfactory progress in relation to a specific curriculum.
   - If a student gets a high score, is it legitimate to infer that this student has very good knowledge, understanding and skills in the field?
2) scores provide an indication of likely future success – this can be thought of in terms of an 'aptitude' construct, for example, readiness for studying Physics (or another subject) at a higher level or aptitude for a career teaching Physics (or a related subject).
   - Is it legitimate to infer that this student will do well in a further course in this field, or in a job in this field?

It is necessary to provide a validation argument for all proposed uses of results. A validation exercise would need to provide grounds for interpreting (the same set of) results in terms of 'readiness' for studying a related course at a higher level or in terms of a significantly broader construct of which 'Physics attainment' is only a part. As an illustration, whilst a student with grade B has attained a higher level (in school-level Physics) than a student with grade C, it is less certain that the B grade student has a higher aptitude (for college-level study) than the C grade student. Different sources of validity evidence might be needed to determine how accurately inferences concerning aptitude and inferences about attainment, can be drawn.

Figure 3 shows the overall claim made for A level Physics in the top box. This consists of the two elements presented as (1) and (2) above. These represent two proposed interpretations of assessment outcomes as shown in the boxes in the right column of the diagram. Five proposed uses/purposes to which these interpretations relate are shown in the left column along with lines to indicate the relation between uses/purposes and interpretations. There might be other valid uses/purposes for the assessment results but we are not necessarily making claims for these and thus it is the responsibility of those who might use or interpret results in other ways to ensure they validate the use of results for these additional purposes.

Figure 3: Proposed uses and interpretations



Interpretive argument

Kane (2006, 2009) conceptualises validation as the assembly of a wider argument or justification for the claims that are made about an assessment. He proposes that validation exercises require the provision of an interpretive argument as well as a validity argument. The interpretive argument for an assessment attempts to spell out the proposed interpretations and uses of test results. It includes both the theoretical and empirical backing for the intended interpretation and use.

The interpretive argument is illustrated in Figure 4 for international A level Physics (and probably applicable to all A levels). It describes the network of inferences and supporting assumptions which lead from scores to conclusions and decisions and attempts to bridge the gap between what a test is actually designed to measure and the inferences that need to be drawn from results in order to support different kinds of decision. This gap needs to be bridged through a process of argument which involves logic, common sense and empirical evidence. The interpretive argument comprises statements of claimed inferences from assessment outcomes, and the warrants which justify the inferences.

Kane suggests that there are at least four major inferential leaps in bridging the gap. 'Construct representation' has been added due to the importance of this in the context of general academic qualifications.

Kane separates the interpretive argument into two parts:

- The descriptive part of the argument involves a network of inferences leading from scores to descriptive statements about individuals.
- The prescriptive part involves the making of decisions based on the descriptive statements.

An inference begins with 'Grounds', a term used by Toulmin, Rieke

Figure 4: Interpretive argument for International A level: summary of the inferences, warrants and assumptions

Inference	*Warrant justifying the inference*	*Assumptions underlying warrant*
Construct representation (task → test performance)	Tasks elicit performances that represent the intended constructs	1 Constructs (knowledge, understanding and skills) relevant to the subject can be identified 2 It is possible to design assessment tasks that require these constructs 3 Task performance varies according to relevant constructs and is not affected by irrelevant constructs
Scoring (test performance → test score/grade)	Scores/grades reflect the quality of performances on the assessment tasks	1 Rules, guidance and procedures for scoring responses are appropriate for providing evidence of intended constructs (knowledge, understanding and skills) 2 Rules for scoring responses are consistently and accurately applied 3 The administrative conditions under which tasks are set are appropriate 4 Scaling, equating, aggregation and grading procedures are appropriate for differentiating performance in relation to intended constructs
Generalisation (test score/grade → test competence)	Scores/grades reflect likely performance on all possible relevant tasks	1 A sufficient number of tasks are included in the test to provide stable estimates of test performances 2 The test tasks provide a representative sample of performance 3 Task, test and scoring specifications are well defined enabling construction of parallel test forms
Extrapolation (test competence → domain competence)	Scores/grades reflect likely wider performance in the domain	1 Constructs assessed are relevant to the wider subject domain beyond the qualification syllabus
Decision-making (domain competence → trait competence)	Appropriate uses of scores/grades are clear	1 The meaning of test scores/grades is clearly interpretable by stakeholders who have a legitimate interest in the use of those scores i.e. admissions officers, test takers, teachers, employers

and Janki (1984) to denote the basis for making a claim.[5] The inference connects the grounds to the claim (shown as an arrow in Figure 4). The inference allows for a conclusion (or claim) – "a conclusion whose merits we are seeking to establish" (Toulmin, 1958/2003). An interpretive argument, therefore, specifies the interpretation to be drawn from the grounds to a claim by an inference. For example, for the 'construct representation' inference, the task itself forms the 'grounds', and from this we can infer or conclude something about test performance. Test performance in turn becomes the grounds for the next inference.

Kane attempts to connect the inferences in the interpretive argument through the use of two types of statements: warrants and assumptions. A warrant used to justify the inference from data to claim. A warrant can be thought of as a law, or a generally held principle, or established procedure. The interpretive argument for International A level is shown in Figure 4. For each inference there is an associated warrant (column 2). For example, the intended score interpretation is based on the 'Construct representation' inference, which has a warrant that the tasks elicit performances that represent the intended constructs. Also shown in Figure 4 are assumptions which need to be generated (column 3) by the validation researcher to guide the validity research. Each of the warrants shown is made on the basis of assumptions. In Kane's terminology, the interpretive argument lays out the intended inferences and warrants.

It should also be remembered that each of the inferences is used to move from grounds to a claim; each claim becomes grounds for a subsequent claim. For example, a generalisation inference connects the grounds of an observed score, which reflects the relevant aspects of performance, with a claim that the observed score reflects the expected score across tasks, occasions and raters. That is, the generalisation inference connects the test score (mark total across Physics components) to test competence (overall achievement in Physics). Test competence then becomes the grounds for the next inference.

Figure 5 provides additional explanation of the interpretive argument by setting out the details of each inference, including details of what it is that is inferred in each inference and including, where necessary, definitions of terms.

Each of the five inferences in the A level interpretive argument prompts a particular set of investigations. An interpretive argument consisting of different types of inferences provides guidance as to the types of research needed. The evidence gathered can provide 'backing' for the assumptions. The backing is expressed through statements that summarise findings that support inferences, and the validity argument is composed of these statements within an overall argument leading to the intended conclusion (claim). Such an argument might include rebuttals, which would weaken the strength of the inferences. A rebuttal constitutes alternative explanations, or counter claims to the intended inference. Kane's approach to the validity argument provides a place for counterevidence. Even when the warrant is supported with backing, exceptions may be relevant or other circumstances may undermine the inference, thereby rebutting the force of the interpretive argument.

Figure 6 shows an extended representation of the 'Scoring' inference, as an example of how each inference relates to its associated warrant and assumptions. The warrant for the scoring inference is based on four assumptions and five types of evidence or analysis (i.e. five different potential sources of 'backing'); these are also shown in the diagram.

5 Toulmin (1958/2003) also used the term 'data' to refer to the same functional unit of the argument.

Figure 5: Explanation of the interpretive argument

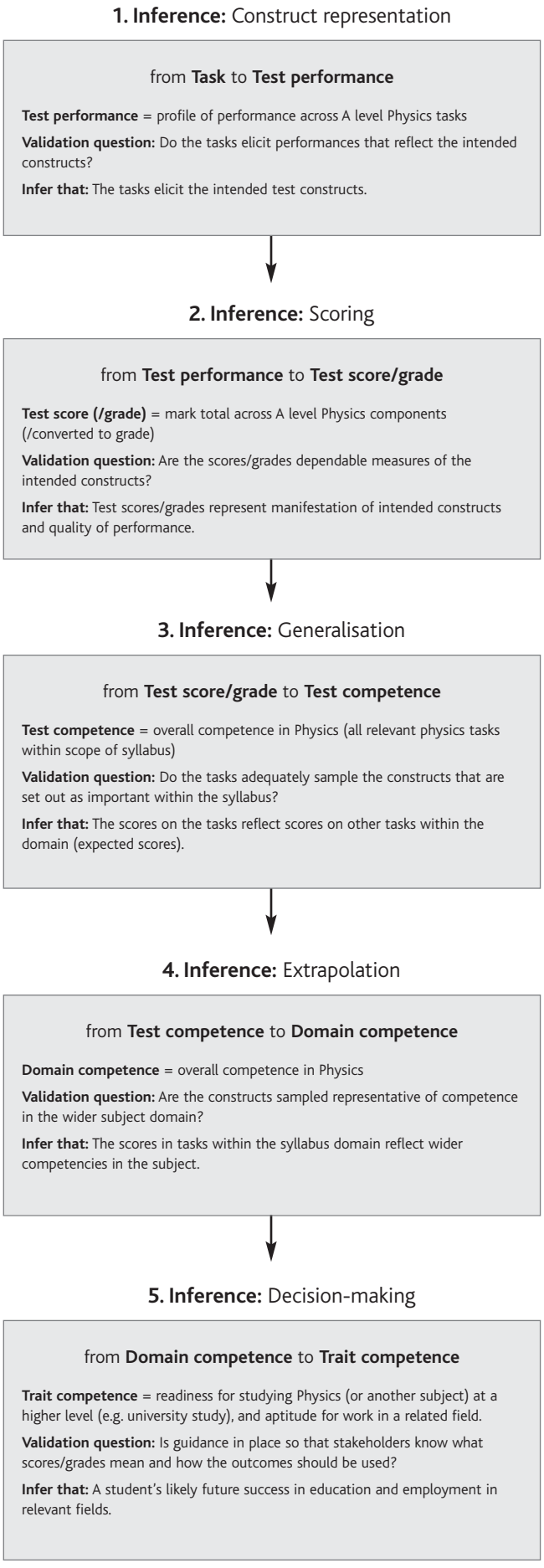
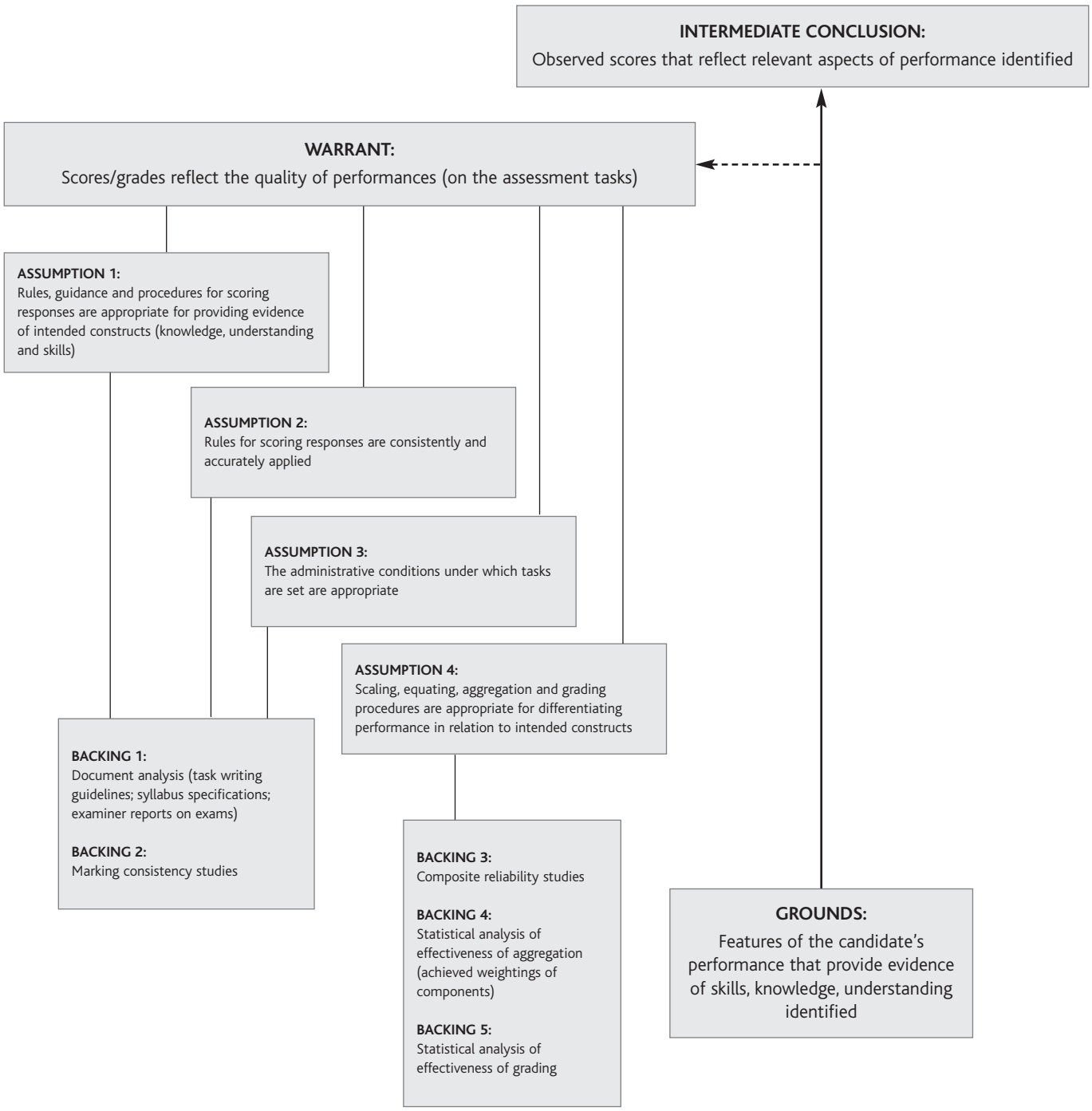


Figure 6: Example schematic for 'Scoring' inference with assumptions and backing



To interpret this diagram the reader should begin with 'grounds' (bottom right corner) such as the features of the candidate's performance that provide evidence of skills, knowledge, understanding identified. The arrow extending from the grounds to the claim represents the scoring inference. The inference allows for a conclusion, which, in the example, is the claim that the observed scores reflect relevant aspects of performance identified. Here, the interpretive argument specifies the interpretation to be drawn from the grounds to a claim by an inference.

The quality of the scoring inference rests on the assumption that "the criteria used to score the performance are appropriate and have been applied as intended and second, that the performance occurred under conditions compatible with the intended score interpretation" (Kane, Crooks and Cohen, 1999, p.9). In other words, if the test administration conditions are not favourable or if they vary for candidates, then the intended interpretation of a candidate's score may not be supported.

7. Gathering evidence to construct a validity argument for International A level Physics

This section presents evidence to support the validity argument for A level Physics in the form of subsections relating to particular evidence types. A section contents list is provided to assist readers with navigation. Each piece of evidence relating to validity reflects a data source and method of analysis in response to a validation question within the framework. For each validity evidence type a brief description of the method used is given. Some examples of the data and analysis are also provided and then the evidence for validity and threats to validity from that evidence type are summarised.

The validity argument provides an evaluation of the interpretative argument by providing and considering appropriate evidence. If the interpretive argument is sound, its inferences and assumptions can be evaluated using such evidence: "To claim that a proposed interpretation or use is valid is to claim that the interpretative argument is coherent, that its inferences are reasonable, and that its assumptions are plausible" (Kane, 2006, p.23).

Section 7 contents:

Validity Evidence 1:
Traditional analyses of item level data

VALIDATION QUESTION 1

Do the tasks elicit performances that reflect the intended constructs?

Method

Data from a sample of scripts for each paper were used to calculate various traditional test and item statistics such as facility values and point biserial correlations. (A much larger sample was possible for Paper 11 given that it was a multiple choice paper.)

Item level data were obtained for the samples shown in Table 3.

Table 3: Samples of student item level data

Paper	*No of candidates*
Paper 11	4600
Paper 21	319
Paper 31	275
Paper 41	210
Paper 51	259

Samples were selected at random and to represent a range of entering countries. As far as possible, samples were selected to align with the mean and standard deviation of total marks for each cohort.

The item level data for each paper were analysed separately to provide various traditional statistics to describe question difficulty and functioning for both whole questions and question parts. Key statistics calculated included facility values (the proportion of available marks scored across all candidates), correlations between marks on an item and total marks on the paper (R_tot), and correlations between marks on an item and total marks on the rest of the test (R_rest). Facility values indicate difficulty level (higher values represent easier questions) and the correlations provide an indication of whether an item measures similar knowledge, understanding and skills to the rest of the test (an item with a low correlation might be subject to construct-irrelevant variance in scores).

Findings for Paper 21

Some of the data output for Paper 21 are displayed (see Tables 4 to 6 and Figures 7 and 8) to illustrate the nature of the analyses for all five papers.

Tables 4 and 5 show that the sample of Paper 21 scripts used for the analysis was fairly similar in mean and standard deviation to the overall cohort of students, thus suggesting that the sample is representative of the cohort. In order to evaluate the internal consistency (reliability) of the exam Cronbach's Alpha was calculated. Higher values indicate better reliability and the value of 0.86 for Paper 21 suggests a good level of reliability.

The distribution of the total marks for Paper 21 is shown in Figure 7.

In Table 6 various statistics are presented for each question part. There appears to be no excessively easy or excessively difficult items in Paper 21, and no items with very low correlations between item scores and total scores on the paper.

Item Characteristic Curves based on quartiles along with histograms (see examples for Question 3ai and 3aii in Figure 8) were used to visually explore how items were functioning.

Table 4: Test Statistics (based on sample of 319 scripts)

Number of sub-questions:	33
Number of whole questions:	7
Test total mark:	60
Mean mark:	29.7
Standard deviation of mark:	10.2
Mean mark (%):	49%
Standard deviation of mark (%):	17%
Standard error of the mean:	0.57
Standard error of measurement:	3.80
Cronbach's Alpha:	0.86

Table 5: Cohort Statistics

Mean mark:	29.3
Standard deviation mark:	10.7

Table 6: Sub-question Statistics for Paper 21

Item	*Max Mark*	*Omit*	*Fac*	*R_tot*	*R_rest*
p21_q1ai	2	0.00	0.87	0.29	0.25
p21_q1aii	1	0.00	0.87	0.18	0.14
p21_q1bi	1	0.00	0.78	0.38	0.34
p21_q1bii	1	0.00	0.45	0.23	0.19
p21_q2ai	2	0.00	0.18	0.51	0.46
p21_q2aii	1	0.00	0.38	0.44	0.40
p21_q2b	3	0.00	0.63	0.65	0.56
p21_q2ci1	1	0.00	0.91	0.37	0.34
p21_q2ci2	1	0.00	0.88	0.29	0.26
p21_q2cii	1	0.00	0.69	0.46	0.42
p21_q3ai	2	0.00	0.60	0.48	0.42
p21_q3aii	2	0.00	0.13	0.49	0.45
p21_q3bi	2	0.00	0.85	0.49	0.45
p21_q3bii	2	0.00	0.56	0.66	0.60
p21_q3c	2	0.00	0.60	0.57	0.50
p21_q4ai	2	0.00	0.70	0.40	0.33
p21_q4aii1	1	0.00	0.88	0.51	0.49
p21_q4aii2	1	0.00	0.79	0.47	0.44
p21_q4b	3	0.00	0.71	0.70	0.63
p21_q5a	2	0.00	0.28	0.44	0.39
p21_q5bi	1	0.00	0.59	0.33	0.28
p21_q5bii1	1	0.00	0.23	0.24	0.20
p21_q5bii2	1	0.00	0.43	0.30	0.26
p21_q5c	6	0.00	0.24	0.64	0.56
p21_q6a	2	0.00	0.18	0.35	0.30
p21_q6b	3	0.00	0.36	0.56	0.48
p21_q6ci	2	0.00	0.94	0.22	0.18
p21_q6cii	3	0.00	0.37	0.55	0.45
p21_q6d	1	0.00	0.25	0.38	0.34
p21_q7a	1	0.00	0.13	0.24	0.21
p21_q7b	2	0.00	0.42	0.31	0.23
p21_q7c	2	0.00	0.31	0.53	0.48
p21_q7d	2	0.00	0.29	0.45	0.39

Fac = Facility value: proportion of available marks scored across all candidates
R_tot = correlation between item marks and total marks
R_rest = correlation between item marks and the total marks on the rest of the paper (not including mark on the item)

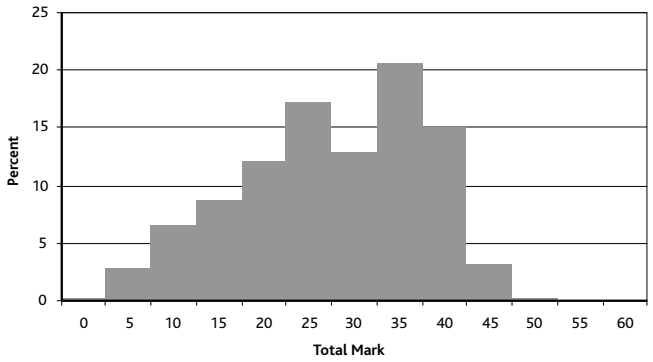


Figure 7: Score Distribution Chart for Paper 21

Evidence for validity

For Papers 21, 31, 41 and 51, there were no whole questions that were either extremely difficult or extremely easy, thus suggesting that all questions were appropriate to the ability of the candidates. All correlations between whole question marks and total marks on the rest of the paper were good. For Paper 11, where each multiple choice item constitutes a whole question, the items represented a reasonable spread of difficulty and no items were excessively easy or difficult.

For all papers the vast majority of items (question parts) were not extreme in their level of difficulty or easiness and had reasonable correlations between item marks and total marks on the rest of the paper. Additionally, for the multiple choice questions (Paper 11) the correlations between response choice and total marks suggested that almost all items functioned well with positive correlations for correct responses and negative correlations for incorrect responses.

Threats to validity

For a small number of items/question parts, there were low correlations between item marks and total marks on the rest of the paper (or for multiple choice options, low or negative correlations between response option chosen and total marks).

Items with very low positive correlations or negative correlations (<0.10):

– Paper 11, Questions 13 and 26;
– Paper 31, Questions 1a, 1b(1), 2bi, 2ci, 2d(1) and 2d(3);
– Paper 41, Question 3bi and 12cii;
– Paper 51, Question 1(4).

Point biserial correlations that are low or negative may indicate that these items measured something different to most items, and thus might potentially have introduced construct-irrelevant variance. However, further consideration is needed to confirm whether or not these constitute threats to validity.

If the items with low or negative correlations were also found by the Rasch analysis (see Validity Evidence 2) to behave unexpectedly they were investigated further by analysing student responses (see Validity Evidence 7).

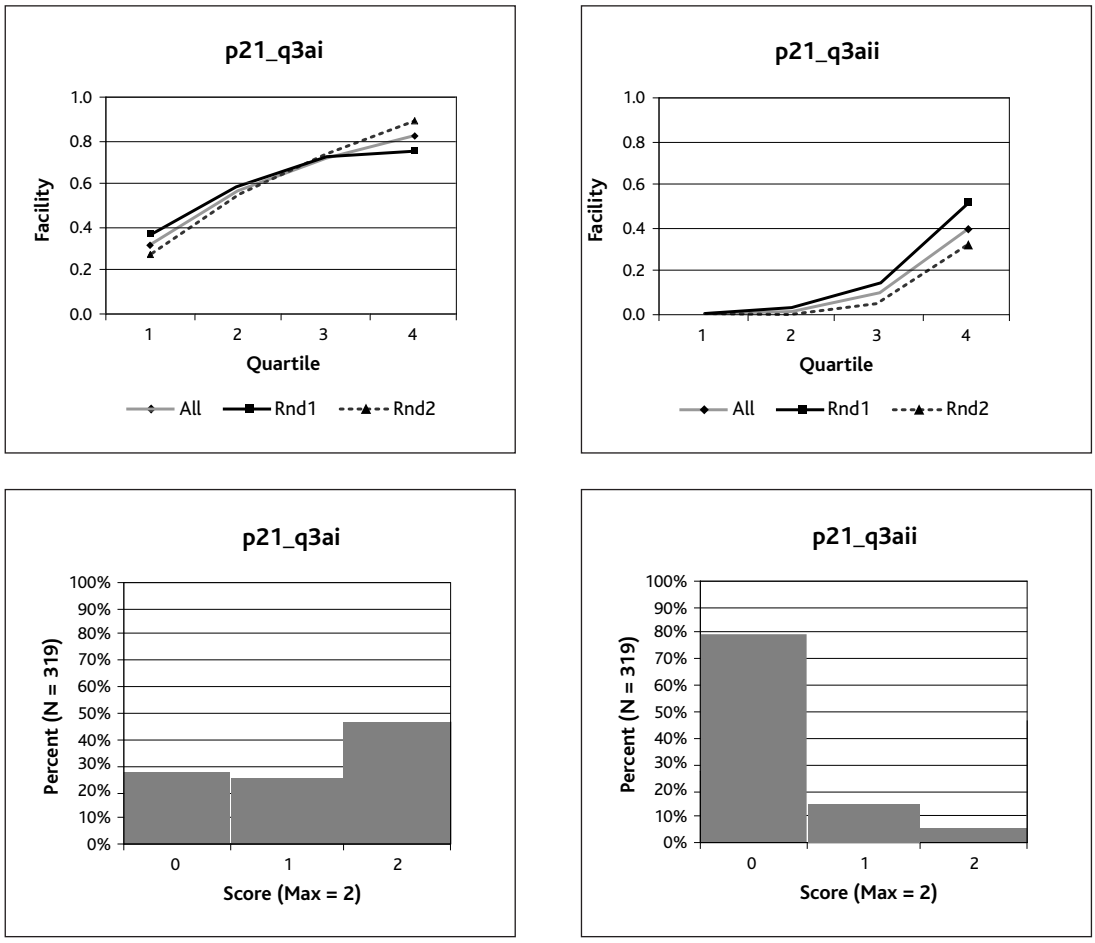


Figure 8: Example Item Characteristic Curves based on quartiles and histograms of marked scores

Validity Evidence 2:
Rasch analysis of item level data

VALIDATION QUESTION 1

Do the tasks elicit performances that reflect the intended constructs?

Method

Item level data from a sample of scripts for each paper were analysed using the Rasch partial credit model. (A much larger sample was possible for Paper 11 given that it was a multiple choice paper.) Item level data for the samples shown in Table 7 were used. These samples are almost identical to those used in the traditional statistical analyses.

Table 7: Samples of student item level data

Paper	*No of candidates*
Paper 11	4590
Paper 21	319
Paper 31	275
Paper 41	210
Paper 51	259

The data for Papers 21, 31, 41 and 51 were analysed together using the Rasch partial credit model (e.g. Wright and Masters, 1982). This is a latent trait model which assumes that the probability of success on a question depends on two variables: the difficulty of the question and the ability of the candidate. Rasch modelling produces estimates of difficulty that take into account the ability of the students answering particular questions. Difficulty estimates are reported on a logarithmic scale with a mean difficulty of 0 logits (log odds units). Due to overlap in students between papers, the total number of candidates in this data set was 402. Data for Paper 11 (multiple choice) were analysed separately to allow additional analysis of the performance of the distractor response choices.

Rasch analysis also provides statistics that can indicate whether items discriminated appropriately between students. Two different statistics provided by the software used (RUMM 2020) are useful for this and will be described later: fit residuals, and chi square probability. Analysis was at the item (question part) level.

Rasch analysis can also allow various graphs (e.g. Item Characteristic Curves) to be drawn modelling the functioning of each item in relation to difficulty and student ability. These were used to confirm findings suggested by the fit residuals and chi square probability statistics.

Findings for Paper 11

Some of the Rasch data output and graphs for Paper 11 are shown to illustrate the nature of the analyses (see Figures 9 and 10 and Table 8). Figure 9 shows summary statistics and indicates that the data fitted the Rasch model well.

The item map (Figure 10) shows difficulty of each item (shown on the right by question number) against the logit scale. The distribution of candidates by ability is shown by the bars on the left.

Various item level statistics for Paper 11 are shown in Table 8.

The difficulty estimates shown in Table 8 are derived on a standardised logits (log odds units) scale where the average has been set at 0. Difficulty estimates above 0 represent items that are more difficult than the average. Difficulty estimates below 0 represent items that are less difficult than average. No items appear to be excessive in difficulty or ease and all fall within the range of ability of the candidates.

Figure 9: Summary test-of-fit statistics (Paper 11)

ITEM-PERSON INTERACTION

ITEMS

	Location	*Fit Residual*
Mean	0.000	-0.073
SD	1.037	6.763
Skewness		0.902
Kurtosis		0.786
Correlation		0.000

PERSONS

	Location	*Fit Residual*
Mean	0.372	-0.125
SD	0.874	1.069
Skewness		0.193
Kurtosis		-0.323
Correlation		-0.392
Sum of Squared Std Resid =		356467.5

ITEM-TRAIT INTERACTION

Total Item Chi Squ	1147.915
Total Deg of Freedom	360.000
Total Chi Squ Prob	0.000000

RELIABILITY INDICES

Separation Index	0.813
Cronbach Alpha	0.810

POWER OF TEST-OF-FIT

Power is GOOD
[Based on SepIndex of 0.813]

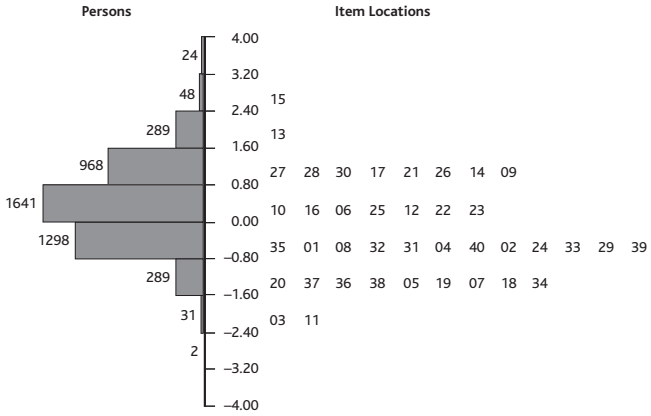


Figure 10: Item map (Paper 11)

Fit residuals provide a measure of unpredictability of scores on questions, or in other words they indicate how well the scores on a question part could be predicted by student ability in relation to the constructs being assessed. High positive values indicate question parts on which students scored unpredictably (i.e. scores on the item do not reflect the ability of students very well). Such an item may not be measuring the intended construct appropriately and is described as 'misfitting', or more specifically, as 'under-fitting'. Items with strong negative fit residuals are likely to indicate that scores on an item are highly discriminated with able students scoring even better than the model would predict and less able students scoring even less well than would be predicted ('overfit'). This can happen where several questions test very similar skills, or where answering a question correctly is dependent on answering another item.

The chi square probability statistic is calculated using the discrepancies between the observed means and the values expected according to the Rasch model, for a number of class intervals (i.e. groups of students based on ability). Significant probability values (e.g. below 0.01) indicate that the item may not be discriminating appropriately.

Table 8: Individual Item fit (Paper 11)

QNo	Itm Code	Type	Difficulty Estimate	Std Error	Fit Residual	DF	ChiSqu	DF	Prob
Q1	I0001	MC	-0.679	0.035	-3.641	4474.28	12.056	9	0.21017
Q2	I0002	MC	-0.285	0.033	-6.447	4474.28	19.122	9	0.02418
Q3	I0003	MC	-1.876	0.047	-3.232	4474.28	8.282	9	0.50595
Q4	I0004	MC	-0.379	0.033	-1.117	4474.28	4.975	9	0.83647
Q5	I0005	MC	-1.024	0.037	-4.305	4474.28	12.587	9	0.18218
Q6	I0006	MC	0.332	0.032	5.086	4474.28	15.469	9	0.07884
Q7	I0007	MC	-0.909	0.036	-4.519	4474.28	19.700	9	0.01986
Q8	I0008	MC	-0.457	0.033	-4.012	4474.28	9.210	9	0.41810
Q9	I0009	MC	1.379	0.035	0.159	4474.28	11.986	9	0.21408
Q10	I0010	MC	0.088	0.032	-3.099	4474.28	9.596	9	0.38413
Q11	I0011	MC	-1.761	0.045	-0.297	4474.28	1.688	9	0.99550
Q12	I0012	MC	0.567	0.032	**14.384**	4474.28	63.011	9	***0***
Q13	I0013	MC	1.739	0.037	**8.323**	4474.28	67.400	9	***0***
Q14	I0014	MC	1.363	0.035	3.501	4474.28	15.748	9	0.07234
Q15	I0015	MC	2.464	0.045	0.605	4474.28	14.952	9	0.09226
Q16	I0016	MC	0.262	0.032	-1.684	4474.28	6.530	9	0.68593
Q17	I0017	MC	1.279	0.034	6.203	4474.28	23.003	9	***0.00619***
Q18	I0018	MC	-0.861	0.036	-1.358	4474.28	3.484	9	0.94199
Q19	I0019	MC	-0.969	0.037	-1.680	4474.28	6.192	9	0.72053
Q20	I0020	MC	-1.496	0.042	1.850	4474.28	7.707	9	0.56388
Q21	I0021	MC	1.290	0.034	-2.078	4474.28	19.852	9	0.01885
Q22	I0022	MC	0.768	0.032	3.478	4474.28	13.890	9	0.12631
Q23	I0023	MC	0.773	0.032	**10.775**	4474.28	41.812	9	***0.00004***
Q24	I0024	MC	-0.215	0.033	-1.485	4474.28	4.695	9	0.86005
Q25	I0025	MC	0.509	0.032	7.524	4474.28	17.620	9	0.03985
Q26	I0026	MC	1.339	0.034	**20.71**	4474.28	337.630	9	***0***
Q27	I0027	MC	0.968	0.033	6.380	4474.28	15.645	9	0.07468
Q28	I0028	MC	1.041	0.033	-2.129	4474.28	20.856	9	0.01331
Q29	I0029	MC	-0.132	0.032	0.305	4474.28	5.113	9	0.82437
Q30	I0030	MC	1.101	0.033	**10.658**	4474.28	57.342	9	***0***
Q31	I0031	MC	-0.382	0.033	-7.892	4474.28	31.491	9	***0.00024***
Q32	I0032	MC	-0.404	0.033	*-12.464*	4474.28	67.522	9	***0***
Q33	I0033	MC	-0.159	0.032	-4.695	4474.28	12.513	9	0.18589
Q34	I0034	MC	-0.830	0.036	-0.408	4474.28	3.203	9	0.95568
Q35	I0035	MC	-0.735	0.035	*-8.569*	4474.28	42.596	9	***0.00003***
Q36	I0036	MC	-1.109	0.038	*-8.430*	4474.28	46.670	9	***0***
Q37	I0037	MC	-1.181	0.038	-6.880	4474.28	39.479	9	***0.00009***
Q38	I0038	MC	-1.065	0.037	-4.196	4474.28	8.613	9	0.47371
Q39	I0039	MC	-0.019	0.032	-7.725	4474.28	26.380	9	***0.00177***
Q40	I0040	MC	-0.335	0.033	-0.513	4474.28	2.291	9	0.98598

(Note that chi square stats were calculated on basis of a random sample of 1000 candidates)

KEY:

Bold – high fit residual (over +8)
Italic – strong negative fit residual (below -8)
Bold and italic – significant chi square probability values (below 0.01)

(It should be noted that the chi square statistic is particularly affected by sample size and thus discretion must be used in its interpretation. Thus, chi square probability values were calculated on the basis of a random sample of 1000 candidates rather than the full available sample to reduce the effect of sample size on these values.)

Where statistics indicated inappropriate item functioning Item Characteristic Curves and Distractor Curves (for multiple choice items) were used to explore further. An example of each is shown for Paper 11 Question 12 (see Figures 11 and 12). Item Characteristic Curves (ICCs) show the marks expected to be achieved ('expected value') by candidates of varying ability ('person location') according to the Rasch model. This is shown by the curve. The dots show the average marks actually scored by candidates for a number of class intervals (i.e. students grouped by ability). If the dots form a shallower line than the modelled curve then this indicates underfit. If the dots form a steeper line than the modelled curve then this indicates overfit.

Distractor curves also show the Rasch modelled curves of the ICCs, but also display the response options selected by candidates across the ability range. For all items there were four response options (A, B, C & D). 'E' represents non-response.

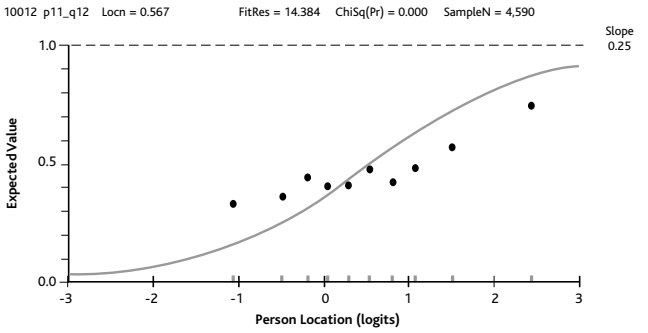


Figure 11: ICC Item 0012 – Paper 11 Question 12

The ICC for Question 12 in Paper 11 (Figure 11) confirms underfit. The distractor curves for Question 12 in Paper 11 (see Figure 12) show that the correct and incorrect options work quite well. As ability ('person location') increases the correct option, B, is more likely to be chosen and the incorrect options are less likely to be selected.

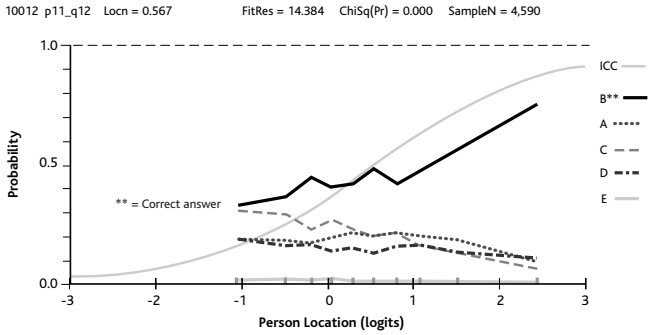


Figure 12: Distractor curves for Item 0012 – Paper 11 Question 12

Evidence for validity

For Paper 11 there were no items with difficulty levels inappropriate to the ability of the candidates. For Papers, 21, 31, 41 and 51, the vast majority of items were appropriate in difficulty to the ability of the candidates.

Many items functioned well according to the Rasch fit statistics.

Threats to validity

For a number of items reverse thresholds (where a particular partial mark is not the most likely outcome for a student of any ability) and overfit were identified. This is not ideal and suggests some inter-dependence of items, and that some of the available marks were less useful than others at discriminating between students. However, these points are not a strong threat to validity.

Several items were found to underfit the Rasch model. Underfitting items may be evidence of construct-irrelevant variance in scores and hence warrant further investigation.

Underfitting items:
– Paper 11, Questions 12, 13, 17, 23, 26 and 30;
– Paper 21, Questions 1aii and 7b;
– Paper 31, Question 2bi;
– Paper 41, Questions 9b and 11.

Underfit and overfit in different ability ranges:
– Paper 21, Question 1bii.

Underfitting items were investigated further using analysis of student responses (see Validity Evidence 7).

Validity Evidence 3: Document review of examiner reports

VALIDATION QUESTION 1
Do the tasks elicit performances that reflect the intended constructs?

Method

For each exam paper, the Principal Examiner (PE) writes a report on student performance. These reports for the November 2009 exam papers were reviewed for any insights on how the questions were answered and any threats to validity indicated by these. This was conducted for the main papers used in the A level Physics validation study (Papers 11, 21, 31, 41 and 51 from November 2009).

Findings

PAPER 11

Question 6: The PE commented on the unusual presentation of the graph given in the question. The graph shows how the displacement of a bouncing ball varies with time. However, presentation of displacement, though accurately portrayed, is somewhat counter-intuitive in that the magnitude of successive peaks is shown in a negative direction, that is, pointing downwards.

PAPER 31

The PE reported that a few centres were unable to obtain the exact springs recommended for Question 1 and noted that "Any deviation between the requested equipment and that provided to the candidates should be given in the Supervisor's Report … so that Examiners can adjust the mark scheme appropriately…. Where Centres had used springs with a different spring constant, and a sample set of results was provided in the Supervisor's report, Examiners took this into account in deciding a suitable range of values for k."

PAPER 41

The PE noted that "Candidates should be advised to give explanation of all their working."

Evidence for validity

In reviewing the reports for insights into any possible questions which might not be measuring the intended constructs appropriately, none were identified for any of the papers in the November 2009 session.

Threats to validity

None

Validity Evidence 4: Factor analysis of item level data

VALIDATION QUESTION 1
Do the tasks elicit performances that reflect the intended constructs?

Method

Exploratory factor analysis in SPSS was conducted on item level score data from samples of scripts to explore the traits underlying the test scores. For example, are most items testing the same trait, or are certain items testing unrelated traits? Factor analysis allows exploration of the relationships between scores on items (question parts) in terms of a smaller number of underlying variables. It can provide insights into whether all items contribute to the measurement of a single trait, or whether any items are measuring a different trait, perhaps reflecting a different type of skill. The latter might or might not be a trait or skill that is relevant to the domain being assessed.

The use of factor analysis as a method for construct validation is well established. Traditionally, construct validity has been investigated by "determining the relationship between the empirical (patterns of scores on the test) and the theoretical (proposed explanatory concepts), so, for example a factor analysis may be undertaken to identify the number of factors (or constructs) in the test data and their relationship with one another" (Davies *et al*., 1999, p.33). Cronbach and Meehl (1955) discuss the prominence of factor analysis in construct validation suggesting that factors function as constructs. However, the use of factor analysis for evaluating construct validity is not without its critics (e.g. Delis *et al*., 2003).

The data for Papers 11 (n=4600), 21 (n=319), 31 (n=275), 41 (n=209) and 51 (n=259) were analysed separately. In this analysis the method used for performing the factor analysis was *principal components analysis*. Principal components analysis considers the total variance and attempts to explain the maximum amount of variance by the minimum number of underlying factors (variables that can explain the variability in the original data). As it produces more factors to explain all the variance, some factors explain significantly more variance than others.

Once the factors have been identified the analysis performs a rotation to get the clearest and simplest way of associating the original variables to the factors. As a rule of thumb it is often taken that a variable makes a significant contribution to a factor if the loading is 0.3 or greater. The rotation method performed for this analysis is Varimax with Kaiser Normalization. The Varimax rotation method rotates the factors in such a way that when the final factors are produced they are not correlated (i.e. orthogonal) with each other.

The findings for Paper 21 are shown.

Findings for Paper 21

PAPER 21 (n = 319)

Table 9 shows an abridged factor analysis output table showing variance explained. The 'Total' column shows the eigenvalues we are interested in. The *% of Variance* column shows how much variance each individual factor can explain. The *Cumulative %* column shows the amount of variance accounted for by each consecutive factor added together. Eleven components had eigenvalues greater than 1 (an eigenvalue of 1 means that the factor can explain as much variability in the data as a single original variable) but the total variance explained by the first three components were the most contributory. Thus, the main criterion for inclusion was the variance explained by each factor, set at 5% (though here we have included Component 3 as the variance is close to 5%). If a factor cannot explain as much as this it is not worth including as an important underlying factor.

Table 9: Total variance explained (> 5%)

Component	*Initial Eigenvalues*		
	Total	*% of Variance*	*Cumulative %*
1	6.735	20.408	20.408
2	1.981	6.004	26.411
3	1.589	4.815	31.226

Three factors explained a cumulative 31.226 % of the variance of the data for Paper 21.

Figure 13 shows the scree plot for the analysis and Table 10 shows the rotated component matrix.

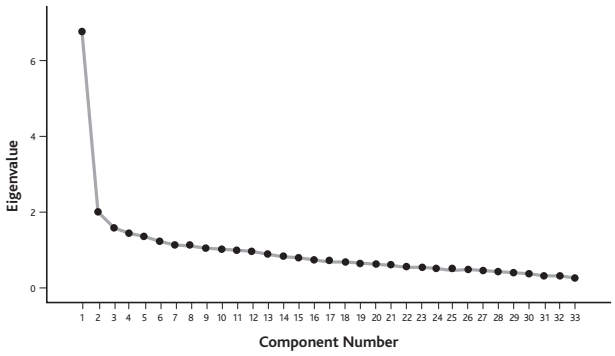


Figure 13: Scree plot

The rotation allowed questions relating to each factor to be considered for commonalities. This resulted in the following inferences about the meaning of the factors:

- Factor 1 questions appear to measure understanding and applying equations (sometimes using data).
- Factor 2 questions appear to measure knowledge and understanding of physics concepts.

There were no apparent similarities between what appears to be measured by the questions within factor 3. Therefore we can discern that there are only two factors assessed by Paper 21.

Evidence for validity

The factor analyses showed generally good coherence of the factors assessed by the papers. In terms of underlying traits, we may surmise that for:

Table 10: Rotated Component Matrix (Principal Component Analysis, Varimax with Kaiser Normalization)

	Component		
	1	*2*	*3*
p21_q1ai			
p21_q1aii			
p21_q1bi			
p21_q1bii			
p21_q2ai		.504	
p21_q2aii		.501	
p21_q2b	.418		
p21_q2ci1			
p21_q2ci2			
p21_q2cii			
p21_q3ai		.508	
p21_q3aii		.543	
p21_q3bi	.422		
p21_q3bii			
p21_q3c	.491		
p21_q4ai			.639
p21_q4aii1	.735		
p21_q4aii2	.716		
p21_q4b	.758		
p21_q5a	.443		
p21_q5bi			
p21_q5bii1			
p21_q5bii2			
p21_q5c			
p21_q6a		.662	
p21_q6b			
p21_q6ci			
p21_q6cii			.647
p21_q6d			
p21_q7a			
p21_q7b			
p21_q7c			.453
p21_q7d			

PAPER 11

- Factor 1 involves a single underlying 'physics' trait

PAPER 21

- Factor 1 involves understanding and applying equations (sometimes using data)
- Factor 2 involves knowledge and understanding of physics concepts

PAPER 31

- Factor 1 involves using instruments, conducting calculations, drawing graphs appropriately, evaluating methods
- Factor 2 involves taking measurements

PAPER 41

- Factor 1 involves a single underlying 'physics' trait

PAPER 51

- Factor 1 involves interpreting/reading off graphs
- Factor 2 involves plotting graphs (e.g. error bars, best fit, worst fit), safety considerations
- Factor 3 involves plotting graphs (e.g. error bars, best fit), planning an experiment (e.g. defining the problem, method of analysis)

Some factors on certain papers (Papers 11, 41 and 51) were problematic to interpret and suggest overlap with other factors on the same paper or a single underlying physics trait.

Threats to validity

None

Validity Evidence 5:
Assessment Objectives elicited by the exam questions

VALIDATION QUESTION 1

Do the tasks elicit performances that reflect the intended constructs?

Method

Six experts (four senior examiners and two other subject experts) rated the extent to which each exam question was thought to assess each Assessment Objective subcomponent on a scale of 0 (not assessed at all) to 5 (strongly assessed). Ratings were made at the whole question level. They were asked not to refer to the mark scheme. Frequencies of questions receiving each rating were calculated and then these were weighted by the maximum mark available. This method assumes that the Assessment Objectives can be used as a representation of the underlying constructs.

Findings

The weighted frequencies (across all five exam papers used as the focus of this study) are shown as a percentage in Table 11.

Evidence for validity

For all components of the Assessment Objectives, there were some exam questions that were rated as eliciting them.

Threats to validity

A possible threat to validity is that some components of the Assessment Objectives were elicited less frequently than others, even when the marks available on each question have been taken into account by weighting the frequencies. For example, A1 (scientific phenomena, facts, laws, definitions, concepts, theories) was judged as being assessed much more frequently than A5 (scientific and technological applications with their social, economic and environmental implications) or B9 (demonstrate an awareness of the limitations of physical theories and models). However, there may be legitimate reasons for differences in frequencies if certain skills are considered more important than others.

Table 11: Weighted frequencies of ratings of Assessment Objectives elicited by questions (without reference to mark scheme)

Assessment Objectives		*Rating*					
		0 'not assessed at all'	*1*	*2*	*3*	*4*	*5 'strongly assessed'*
A1	%	17.8	8.8	11.4	9.8	8.7	43.5
A2	%	14.9	5.8	15.4	12.3	10.4	41.2
A3	%	61.4	7.3	3.2	7.7	6.2	14.3
A4	%	47.0	5.6	5.1	10.7	8.0	23.6
A5	%	75.9	7.2	6.5	2.6	6.5	1.2
B1	%	64.4	2.2	7.6	3.0	7.2	15.6
B2	%	42.3	8.3	7.3	12.5	3.5	26.0
B3	%	22.2	1.5	6.5	15.1	9.6	45.1
B4	%	59.8	4.9	5.6	19.3	3.1	7.4
B5	%	62.2	7.2	4.7	12.9	9.3	3.8
B6	%	87.7	5.5	4.5	1.0	1.3	0.0
B7	%	48.0	7.4	17.9	9.6	4.5	12.6
B8	%	78.6	5.9	7.4	5.6	0.1	2.5
B9	%	94.8	3.1	1.8	0.4	0.0	0.0
C1	%	81.7	0.1	2.5	0.0	3.7	12.0
C2	%	82.0	0.4	2.2	1.9	1.2	12.3
C3	%	67.2	4.6	4.3	8.6	2.0	13.4
C4	%	81.4	1.9	1.9	3.1	3.1	8.6
C5	%	79.6	0.9	2.2	3.1	2.2	12.0

Validity Evidence 6:
Ratings of the cognitive demands placed on students by the exam questions

VALIDATION QUESTION 1

Do the tasks elicit performances that reflect the intended constructs?

Method

In this task six experts (four senior examiners and two other subject experts) rated the cognitive demands placed on students by each exam question for each of five types of demand (from Hughes, Pollitt and Ahmed, 1998) on a scale of 1 (low demand) to 5 (high demand). Ratings were made at the whole question level. Experts were asked not to refer to the mark scheme, and were informed that they should focus on what the students have to do. Frequencies of questions receiving each rating were calculated and then these were weighted by the maximum mark available on each question.

Experts were provided with explanatory information defining the concept of assessment demands, the types of cognitive demands to be rated, and the rating scale.

The five demand types or 'dimensions' from Hughes, Pollitt and Ahmed (1998) are:

- *Complexity* – the number of components or operations or ideas involved in a task and the links between them
- *Resources* – the use of data and information (including the student's own internal resources)
- *Abstractness* – the extent to which the student must deal with ideas rather than concrete objects

- *Task strategy* – the extent to which the student devises (or selects) and maintains a strategy for tackling the question
- *Response strategy* – the extent to which students have to organise their own response

Findings

These weighted frequencies (across all five exam papers used as the focus of this study) are shown as percentages in Table 12.

Table 12: Percentage weighted frequencies of ratings of the cognitive demands elicited by exam tasks across November 2009 Papers 11, 21, 31, 41 and 51 (without reference to mark scheme)

Demand type		*Ratings of demand*				
		1 low demands	*2*	*3*	*4*	*5 high demands*
Complexity	%	2.4	24.0	23.3	27.7	22.6
Resources	%	7.6	13.5	32.3	28.5	18.0
Abstractness	%	35.2	29.1	23.1	11.9	0.7
Task strategy	%	5.4	17.8	42.7	26.7	7.5
Response strategy	%	4.3	19.1	36.5	22.5	17.6

Evidence for validity

For all five types of demands (complexity, resources, abstractness, task strategy and response strategy) the questions ranged in demand according to expert ratings, suggesting a good spread of cognitive demand types and levels are placed on candidates across the questions.

Threats to validity

None

Validity Evidence 7: Analysis of student responses

VALIDATION QUESTION 1
Do the tasks elicit performances that reflect the intended constructs?

Method

For items identified as misfitting by the Rasch modelling (see Validity Evidence 2) the question and mark scheme were considered and responses were analysed. For Paper 11 the statistical analyses of the response options were used.

For misfitting items in Papers 21, 31 and 41 (there were no misfitting items in Paper 51), the responses of 40 candidates (selected to represent a representative range of candidates in terms of total marks awarded) were analysed and similar answers were categorised together. This was used, along with consideration of the question and mark scheme, to inform possible explanations for misfit.

Findings for Paper 41

To illustrate the analyses undertaken, the analysis of responses for two items on Paper 41 that were identified as misfitting is shown.

PAPER 41 QUESTION 9B

Question 9b in Paper 41 involved working out the percentage change in the size of a crack in a wall from change in resistance in the wire of a strain gauge. Students needed to state the equation they used (change in resistance divided by original resistance, multiplied by 100), show the correct use of the values given and then state their answer. Three marks were available. The responses from 40 candidates are categorised in Table 13.

Table 13: Categorisation of responses from 40 candidates to Question 9b

Response	*Frequency*
Correct calculation and answer (2.24%) – with or without appropriate equation (2 or 3 marks, depending on inclusion of equation)	28
Mostly correct, equation included, an error in calculation (2 marks)	2
Calculation partly correct, no equation (1 mark)	4
Attempt at response but equation incorrect if given, and calculation incorrect (0 marks)	4
Omit	2

Most candidates carried out the calculation correctly though some did not include the equation they used. Others made errors in their calculations or struggled to carry out an appropriate calculation.

One possible partial explanation for the underfit of this item is that some candidates did not include the equation, even though they must have had a concept of this equation in order to correctly conduct the calculation, thus leading to some very able students scoring only 2 out of 3 marks. However, the misfit also affected the lower range of ability.

This analysis does not suggest construct irrelevant variance in scores and hence does not suggest that this item is a threat to validity.

PAPER 41 QUESTION 11

Question 11 in Paper 41 was slightly misfitting. This question was about the principles of use of magnetic resonance to obtain diagnostic information about internal body structures. The question asked for a description of how magnetic resonance works. The question was slightly harder than average but not excessively difficult. The responses from 40 candidates are categorised and described in Table 14.

Table 14: Categorisation of responses from 40 candidates to Question 11

Response	*Frequency*
Full detailed correct response (5/6 marks)	5
Partially correct response, e.g. some mention of magnetic field, radio pulses, hydrogen atoms (1–4 marks)	12
Includes some of the right kinds of concepts but understanding is muddled/imprecise (0 marks)	6
Confusion with X-rays/radioactive substances (0 marks)	4
Confusion with electromagnetic waves in general – non-specific response (0 marks)	2
Describes why MRI is used (e.g. non-invasive means of diagnosis) (0 marks)	1
Repeats part of question/incomplete response (0 marks)	1
Omit	9

Only 5 of the 40 candidates gave strong responses, but a further 12 included some correct points. A further six seemed to have learnt about the topic, mentioning some of the right concepts but without sufficient precision. Some responses indicated that students had little knowledge of this topic and used ideas relating to x-rays, or electromagnetic waves in general, in their responses.

This question had reverse thresholds and looking at the category probability curve there is only a small part of the ability range where

marks other than 0 or 6 are most likely. It would appear that some students had learnt this topic well whereas others had not or could not recall the details. Perhaps this is a topic that some teachers do not prioritise.

The underfit of this question is minor and there is no evidence to suggest that the question is a threat to validity in terms of the constructs measured.

Evidence for validity

For 8 of the 12 items found to be underfitting, the analyses of student responses did not suggest any threats to validity in relation to the constructs elicited.

Threats to validity

For 4 of the 12 underfitting items, worth one mark each, there was some indication of threats to validity in the question. For two questions in Paper 11 (Q13 and 26), it seems that many students, regardless of ability, omitted an initial or final step in generating their answer which led to performance on the item not being a good representation of Physics ability. For an item in Paper 21 (Q1bii), writing skills as well as Physics understanding appeared to influence performance. Finally, for one item in Paper 31 (Q2bi) the precision with which experimental equipment should be arranged was not transparent and appeared to affect performance for reasons other than ability in Physics.

However, the potential threats to validity identified here would have been difficult to anticipate and affect only 4 marks out of 270 raw marks available across all five papers.

Validity Evidence 8:
Document review on marking and scoring procedures

VALIDATION QUESTION 2
Are the scores/grades dependable measures of the intended constructs?

Method

A document review was conducted focusing on the marking/scoring procedures relating to International A levels. The code of practice and documents providing instructions for examiners were reviewed.

Findings

SUMMARY OF FINDINGS FROM REVIEW OF THE CODE OF PRACTICE

The code of practice sets out details of various aspects of exam procedures and includes two chapters, one on *Setting of question papers and mark schemes* (including a section on mark schemes and a section on the Question Paper Evaluation Committee) and a chapter on *Marking* (including types of marking and types of examiner; the standardisation process; and monitoring).

The section about mark schemes in the question setting chapter covers such issues as conformity with the question paper and syllabus, facilitating reliable marking and discrimination, and the development and format of the mark schemes.

The chapter about question setting also describes the Question Paper Evaluation Committee (QPEC) which evaluates each question paper and mark scheme to ensure that it meets the requirements of the code of practice.

The chapter on marking relevant to this syllabus describes:

- *Types of marking and types of examiner* – the level of expertise to mark the questions
- *Reporting lines* – lines of examiner reporting
- *Allocation of marking to examiners* – the apportionment of scripts to ensure reliability and minimal bias
- *The standardisation process* – ensuring that examiners understand the mark scheme through a standardisation meeting
- *Marking of scripts* – ensuring examiners use the mark scheme in a transparent way
- *Checking the accuracy of the recording of marks* – minimising the risk of transcription or arithmetic error
- *Monitoring examiners and sampling their marking* – procedures for quality control of marking
- *Loss, absence or late arrival of evidence* – procedures for missing work
- *Principal Examiner's report* – reports on candidate performance
- *Review of marking after issue of results* – responding to school enquiries and appeals after the issue of results
- *Retaining evidence* – retaining scripts in case of enquiries about results

In addition, the code of practice contains a chapter on Grading which includes such issues as: what the grading procedure entails; how standards will be maintained; and the contribution of professional judgement.

SUMMARY OF FINDINGS FROM REVIEW OF EXAMINER INSTRUCTIONS

Appropriate instructions are given to Assistant Examiners, Team Leaders and Principal Examiners. These provide instructions to examiners on many of the procedures described in the code of practice.

Assistant Examiners

The instructions for Assistant Examiners include guidance on administrative aspects of marking, preparation for and conduct during the standardisation meeting, subsequent marking (under supervision), procedures for completion and despatch of mark sheets and return of marks, the use of checkers and help with how to complete the Assistant Examiner's Report Form.

Team Leaders

The instructions for Team Leaders include guidance on administrative aspects of supervising marking, the Team Leader role within the standardisation process, sampling the marking of each of their examiners, making scaling recommendations, liaising with the Principal Examiner over examiner marking anomalies, evaluating examiner performance, and help with how to feed back to the Principal Examiner on the content of the Assistant Examiners' reports.

Principal Examiners

The instructions for Principal Examiners include guidance on administrative procedures, co-ordinating and leading the Team Leaders' Meeting and Standardisation Meeting, selecting scripts for co-ordination, producing a revised mark scheme, sampling the marking of their Team Leaders and any allocated Assistant Examiners, making scaling decisions, evaluating examiner performance, writing their Examiner's report to

schools on the performance of candidates on the exam paper, recommending grade thresholds, contributing to grade review and results enquiries and preparation for developing the next exam paper and mark scheme.

Evidence for validity

Marking procedures are comprehensively documented at all levels of marking and in conformity with appropriate assessment codes of conduct.

Threats to validity

None

Validity Evidence 9:
Marker agreement analyses of multiple marking data

VALIDATION QUESTION 2
Are the scores/grades dependable measures of the intended constructs?

Method

In order to look at the reliability of marking, a multiple marking exercise was conducted. This involved five markers from each of papers 21, 31, 41 and 51 marking copies of 30 students' scripts. Paper 11 was not included in this exercise because it is a multiple choice paper.

For each of the four exam papers, 30 scripts were selected at random but ensuring that the mean and standard deviation of total marks were similar to those for the overall cohort of candidates who took the paper. A further 10 scripts were selected in the same way for use in a standardisation exercise before the marking. Clean copies of the scripts were created with all marks, annotations and comments removed. Markers were asked to mark the 10 standardisation scripts first, return these to the Principal Examiner and await feedback before continuing to mark the main set of 30 scripts.

The pursuit of high reliability should be a continuing goal of all test construction. The essential concern explored here is whether a student would receive a different mark if his or her examination paper were marked by a different examiner using the same mark scheme. The multiple marking exercise attempted to address this issue by exploring examiner agreement as a statistical indicator of a set of marks, that is, multiple observations of the same performances by a group of examiners (inter-examiner reliability) and as an indicator of 'Gold Standard agreement'. The Principal Examiner's mark (representing the 'standard' for a particular performance) was used as the comparator, against which all other examiners' marks are compared.

Marking agreement was explored using a range of quantitative methodologies. Agreement is used here in three senses:

- as an indicator of a set of marks, that is, multiple observations of the same performances by a group of examiners (inter-rater reliability)
- as an indicator of 'Gold Standard agreement', the Principal Examiner's mark used as the comparison mark against examiner marks
- as an indicator of examiner severity/leniency and fit (Rasch analyses)

To illustrate the analyses, the findings for Paper 31 are shown below.

Findings for Paper 31

Table 15 shows descriptive statistics for each marker for Paper 31. The Principal Examiner is shown as 'PE'. There were some differences in the mean total mark awarded across the 30 scripts but these were small.

The issue of whether differences in mean values are statistically significant was investigated using Analysis of Variance (ANOVA). The total marks given by the five examiners are not statistically significantly different ($F = 0.027$, d.f. = 4, 145, $p = 0.999$).

Table 15: Descriptive statistics for Paper 31 from multiple marking

	Mean	*Std. Deviation*	*N*
PE	28.23	6.15	30
Ex1	28.03	5.83	30
Ex2	28.13	6.13	30
Ex3	27.77	6.37	30
Ex4	28.17	5.92	30

Inter-rater reliability indices were calculated for questions within each of the papers. A Fisher Z transformation then allowed the correlations to be averaged. The Pearson inter-rater correlation between total scores was 0.85 for Paper 31 suggesting a good level of agreement. This is a measure of inter-rater reliability, which is defined as the "level of consensus between two or more independent markers in their judgements of candidates' performance" (Davies, Brown, Elder, Hill, Lumley and McNamara, 1999, p.88).

For the analysis of 'Gold standard agreement', the Principal Examiner's marking was used as the basis against which to compare the other examiners' marking. The plots in Figure 14 show the differences in total mark for each script. The differences are small.

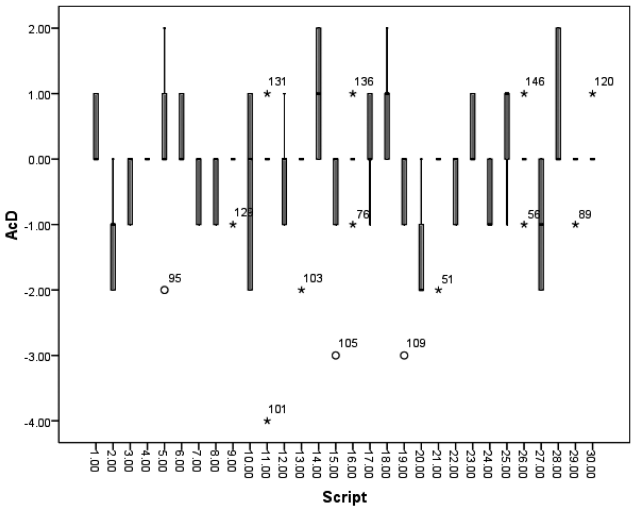


Figure 14: Actual differences in total mark awarded to each script (Paper 31)

Figure 15 shows the total mark differences by examiner. (Note that the Examiner labelled Examiner 1 is the Principal Examiner against which the other examiners have been compared.) The horizontal black lines indicate the median values and the shaded boxes indicate where 50% of the data fall, that is, the data that lie between the twenty-fifth and seventy-fifth percentile. The remaining 'whiskers' indicate the lowest and highest values except for any outliers which are shown as a circle. Again, any differences are generally small although Examiners 4 and 5 varied a little more than the others from the Principal Examiner's marks.

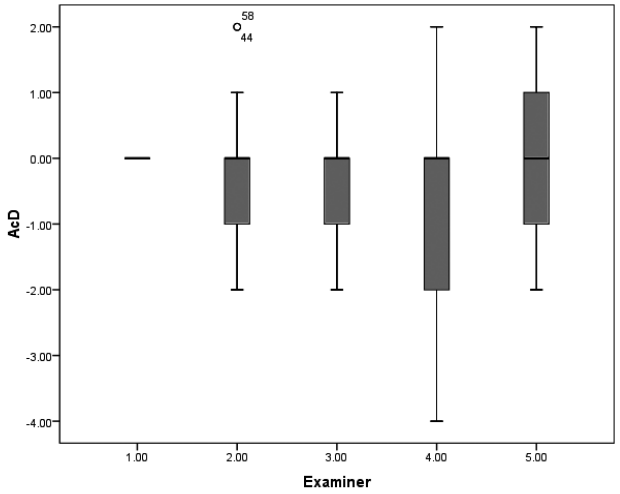


Figure 15: Examiners' actual differences in total marks (Paper 31)

Figure 16 and 17 show further graphical representations of the multiple marking data for Paper 31. Both suggest that the marking was quite consistent.

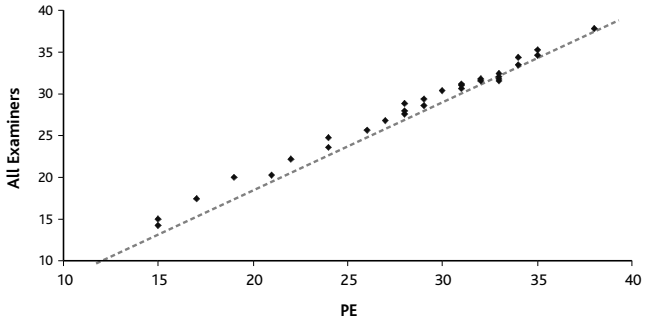


Figure 16: Comparison of total marks awarded by the PE and average of other examiners

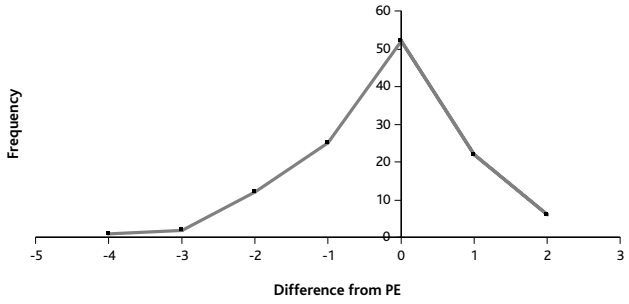


Figure 17: Differences between marks given by PE and marks given by other examiners (at the item level)

A Multi-facet Rasch Measurement approach using FACETS software (Linacre, 1989, 2005) allows inter-examiner reliability to be investigated from a different perspective. FACETS models the examiners as 'independent experts'. Figure 18 shows a graphical overview of the results for Paper 31 as the output of the FACETS program. The scale along the left represents the logit scale, which is the same for the two FACETS of interest here: 'candidates' and 'raters'. Each script is represented by an asterisk and is ordered with the highest level of performance at the top and the lowest level at the bottom. The other facet, 'raters', is ordered so that harsher examiners are closer to the top and more lenient examiners are closer to the bottom. The most likely scale score (the mark) for each ability level is shown in the rightmost column. Figure 18 suggests that there are no clear differences in the severity of the markers.

The FACETS output provides measures of fit or consistency: the infit and the outfit values. The infit is the weighted mean-squared residual which is sensitive to unexpected responses near the point where decisions are being made, while the outfit is the unweighted mean-squared residual and is sensitive to extreme scores. For ease of interpretation, the two sets of fit statistics are expressed either as a mean square fit statistic or as a standardised fit statistic, usually a z or t distribution.

```
-------------------------------------------
|Measr|+Candidates|-Raters              |Scale|
-------------------------------------------
+   6 + *          +                    +(38) +
|     |            |                    |     |
|     |            |                    |     |
|     |            |                    | --- |
+   5 +            +                    +  37 +
|     |            |                    |     |
|     |            |                    | --- |
|     |            |                    |  36 |
+   4 +            +                    + --- +
|     | *          |                    |     |
|     |            |                    |  35 |
|     | *          |                    | --- |
+   3 + *          +                    +     +
|     |            |                    |  34 |
|     | *          |                    | --- |
|     |            |                    |  33 |
+   2 + *          +                    + --- +
|     | ***        |                    |  32 |
|     | **         |                    | --- |
|     | **         |                    |  31 |
+   1 + **         +                    + --- +
|     |            |                    |  30 |
|     | *          |                    |  29 |
|     | ***        |                    |  28 |
*   0 * **         * E1  E3  E2  PE  E4 *  27 *
|     | *          |                    |  26 |
|     | *          |                    |  25 |
|     | *          |                    |  23 |
+  -1 + *          +                    +  22 +
|     | *          |                    |  21 |
|     | *          |                    |  20 |
|     |            |                    |  19 |
+  -2 +            +                    +  18 +
|     | *          |                    | --- |
|     |            |                    |  17 |
|     |            |                    |     |
+  -3 +            +                    + --- +
|     |            |                    |     |
|     |            |                    |     |
|     |            |                    |     |
+  -4 + *          +                    +  15 +
|     |            |                    |     |
|     |            |                    |     |
|     |            |                    |     |
+  -5 + *          +                    + --- +
|     |            |                    |     |
|     |            |                    |     |
|     |            |                    |  14 |
+  -6 +            +                    +(12) +
-------------------------------------------
|Measr| * = 1      |-Raters              |Scale|
-------------------------------------------
```

Figure 18: Facets vertical summary report

The infit and outfit values for the Paper 31 examiners are shown in Table 16. There are different views on what fit index is actually acceptable (see for example, Lunz and Wright, 1997; Wright and Linacre, 1994). Operational experience would suggest lower and upper bound limits of 0.7 and 1.6 respectively for mean squares to be useful and acceptable for practical purposes. Table 16 indicates that the examiners 'over-fit' since the infit and outfit mean squares are below 0.7.

One index of inter-rater reliability is the proportion of exact agreements (Cohen's Kappa): the 'exact observed agreement' statistic.

Table 16: Rater Measurement Report (arranged by N)

Obsvd Score	Obsvd Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Exact Agree. Obs %	Exact Agree. Exp %	N Raters
847	30	28.2	29.63	-.06	.11	.14	-4.7	.17	-4.4	1.62	43.3	23.7	1 PE
841	30	28.0	29.44	.01	.11	.20	-4.0	.24	-3.8	1.63	37.5	23.7	2 E1
844	30	28.1	29.54	-.02	.11	.16	-4.4	.18	-4.2	1.58	38.3	23.7	3 E2
833	30	27.8	29.20	.10	.11	.40	-2.5	.49	-2.0	1.48	35.8	23.4	4 E3
845	30	28.2	29.57	-.03	.11	.25	-3.6	.33	-3.0	1.34	31.7	23.7	5 E4
842.0	30.0	28.1	29.48	.00	.11	.23	-3.9	.28	-3.5				Mean (Count: 5)
4.9	.0	.2	.15	.05	.00	.09	.8	.12	.9				S.D. (Populn)
5.5	.0	.2	.17	.06	.00	.10	.8	.13	1.0				S.D. (Sample)

Model, Populn: RMSE .11 Adj (True) S.D. .00 Separation .00 Reliability (not inter-rater) .00
Model, Sample: RMSE .11 Adj (True) S.D. .00 Separation .00 Reliability (not inter-rater) .00
Model, Fixed (all same) chi-square: 1.2 d.f.: 4 significance (probability): .88
Model, Random (normal) chi-square: .9 d.f.: 3 significance (probability): .82
Rater agreement opportunities: 300 Exact agreements: 112 = 37.3% Expected: 70.9 = 23.6%

If examiners must agree on the exact value of the ratings, then it is necessary to use a Cohen's-Kappa type of inter-examiner reliability index. Cohen's Kappa is where chance is determined by the marginal category frequencies and is given by the formula:

$$\text{Cohen's Kappa} = \frac{\text{(Observed agreement \% – Chance agreement \%)}}{\text{(100 – Chance agreement \%)}}$$

In Rasch terms, this translates to the 'Expected Agreement %' for an adjustment based on 'chance + rater leniency + rating scale structure' (see Linacare, 2005, pp.106–7). Then the Rasch-Cohen's Kappa becomes:

$$\text{Rasch-Cohen's Kappa} = \frac{\text{(Observed agreement \% – Expected agreement \%)}}{\text{(100 – Expected agreement \%)}}$$

Rasch-Cohen's Kappa for this paper was 0.18. Under Rasch-model conditions ideally this should be close to 0, indicating that inter-rater reliability is within the acceptable range.

Evidence for validity

The levels of marking reliability estimated for this qualification were found to be high. Pearson's correlation coefficients for Papers 21, 31, 41 and 51 were 0.82; 0.85; 0.82; and 0.80 respectively. These findings were corroborated using an estimated Rasch-based Kappa statistic.

There were some differences between individual examiner judgements and the Principal Examiner's mark but these were very minor.

In terms of Rasch range of severity (the difference between the most severe examiner and the least severe examiner) examiners appeared to be behaving largely the same and over a very narrow and acceptable range.

Rasch analyses for each paper indicated a generally well-fitting Rasch model with no instances of mis-fitting examiners, suggesting similar individual variability.

Threats to validity

There was a general tendency amongst examiners for all four papers to exhibit insufficient variability in their scores (though this was less pronounced for Paper 51).

Validity Evidence 10: Composite reliability estimation

VALIDATION QUESTION 2

Are the scores/grades dependable measures of the intended constructs?

Method

Most CIE exams contain several components or papers and A level Physics is no exception. For many purposes the most useful index of reliability to report would relate to the total syllabus score rather than the individual papers. Approaches to estimating the reliability of a composite test are discussed in Feldt and Brennan (1989, pp.116 –117) and Crocker and Algina (1986, pp.119–121). The method illustrated below is from Crocker and Algina. The reliability of a composite test is defined as:

$$\rho_{cc} \geq \frac{k}{k-1}\left[1 - \frac{\sum \sigma_i^2}{\sigma_c^2}\right]$$

where the ≥ indicates that this is a lower-bound estimate

k = the number of papers

$\sum \sigma_i^2$ = the sum of the variances of each paper

σ_c^2 = the total test variance

This is in fact the equation for Cronbach's Alpha, and this is helpful in understanding Alpha (i.e. any test may be regarded as a composite and each item as a subtest). The reliability of A level Physics as a composite test was calculated using this approach.

Findings

Table 17 illustrates this approach with data from the November 2009 administration of A level Physics. The data required to estimate Alpha are shown on the left.

The estimated syllabus reliability of A level Physics based on these five papers is: Alpha ≥ 0.868.

Table 17: Calculation of Cronbach's Alpha

Variance Paper 11 (σ = 6.18)	38.19
Variance Paper 21 (σ = 10.72)	114.92
Variance Paper 31 (σ = 6.11)	37.33
Variance Paper 41 (σ = 17.95)	322.20
Variance Paper 51 (σ = 5.53)	30.58
Sum of paper variance	543.22
Syllabus variance (σ = 42.18)	1779.15

$$\rho_{cc} \geq \frac{k}{k-1}\left[1-\frac{\sum\sigma_i^2}{\sigma_c^2}\right] \qquad \frac{5}{5-1}\left[1-\frac{543.22}{1779.15}\right] = 0.868$$

Evidence for validity

The estimated syllabus reliability of A level Physics is high.

Threats to validity

None

Validity Evidence 11:
Analysis of achieved weightings of components

VALIDATION QUESTION 2

Are the scores/grades dependable measures of the intended constructs?

Method

The syllabus specification sets out the intended weightings of the exam papers and the scores are combined in these proportions. However, if there is a smaller spread of marks on one paper compared to another, this can potentially lead to a situation where the achieved weightings of the different components do not match the intention. This could pose a threat to validity with regard to whether the scores are dependable measures of the intended constructs. In the current validation study, intended and achieved weightings of components were compared because large differences from the intended weightings would compromise the claims about the syllabus in terms of the relative importance of different papers, and thus of different topics or aspects of the qualification. If a particular paper has a lower weight than intended, this will be due to lower variance in marks, and suggests that better use needs to be made of the range of marks available.

In order to investigate the effectiveness of the aggregation of marks from the five exam papers making up the A level Physics, assessment component scores for candidates were obtained for those completing their A or AS level in June or November 2009. As described earlier, the A level Physics qualification is assessed via five examination papers, the first three of which (the AS papers) can be taken earlier than the others. Because the grade boundaries and measurement properties for the AS papers in different sessions may vary, the analysis of achieved weights had to be conducted separately for different combinations of sessions in which papers were taken.

The covariance of marks on components with total marks on the qualification (aggregate mark) can be used to estimate the achieved weight of each examination component (Fowles, 1974). Achieved weights for the components forming an aggregate will sum to 1.00 and may be interpreted as the proportion each contributes to the total. This method was applied to the current data such that the covariance of scores on each component with the total mark across papers was calculated for each combination and then expressed as a proportion of the total variance.

It is the intention that the papers contribute to the total aggregated mark in the following proportions: Paper 1 (15%); Paper 2 (23%); Paper 3 (12%); Paper 4 (38%); and Paper 5 (12%). Both Zone P and Zone Q papers were analysed. Some Zone P grading options in November 2009 had low numbers of candidates and thus could not be analysed.

Findings for aggregation in November 2009

The achieved weightings for combinations of papers aggregated in November 2009 are presented in Tables 18 to 21. The values can be interpreted as the actual importance of each component in contributing to students' overall outcomes.

Evidence for validity

Papers 1, 2 and 5 were found to contribute as intended to overall scores.

Threats to validity

Paper 3 contributed too little to variance in overall scores (e.g. as part of the complete A level it often contributed only about 0.07 of the variance instead of the intended 0.12). This in turn led to Paper 2 contributing too much when considered in terms of its contribution to AS level scores. In addition, Paper 4 often contributed too much to variance in total scores, especially when scores from AS papers were carried over from a previous session. When the relative contributions of AS and A2 papers were considered, the AS papers tended to contribute less than 50% of the variance in overall scores.

Validity Evidence 12:
Coverage of content and learning outcomes for Papers 1, 2 and 4 across six sessions

VALIDATION QUESTION 3

Do the tasks adequately sample the constructs that are set out as important within the syllabus?

Method

The six experts (four senior examiners and two other subject experts) conducted a task to identify subject content and learning outcomes coverage for exams across six sessions (June 2007 to November 2009) for Papers 1, 2 and 4. For sessions after the introduction of different papers for different time zone based areas, Papers 11, 21 and 41 were used. Papers 3 and 5 were not included in this task because they relate to practical skills and data analysis and interpretation rather than to topic content.

Table 18: Achieved weightings in November 2009 Zone P – All papers from November 2009

	Paper 1	Paper 2	Paper 3		Paper 4	Paper 5	N
Combination	*P_11 (Nov09)*	*P_21 (Nov09)*	*P_31 (Nov09)*	*P_32 (Nov09)*	*P_41 (Nov09)*	*P_51 (Nov09)*	
1	0.13	0.24	0.08		0.45	0.10	1701
2	0.12	0.25		0.06	0.46	0.10	1098
Intended weightings	0.15	0.23	0.12		0.38	0.12	

Table 19: Achieved weightings in November 2009 Zone Q – All papers from November 2009

	Paper 1	Paper 2	Paper 3		Paper 4	Paper 5	N
Combination	*P_12 (Nov09)*	*P_22 (Nov09)*	*P_33 (Nov09)*	*P_34 (Nov09)*	*P_42 (Nov09)*	*P_52 (Nov09)*	
3	0.13	0.25	0.07		0.45	0.11	408
4	0.13	0.23		0.07	0.45	0.12	237
Intended weightings	0.15	0.23	0.12		0.38	0.12	

Table 20: Achieved weightings in November 2009 Zone Q, with AS carried forward

	AS papers		Paper 4	Paper 5	N
Combination	*Comp_66 (Jun09)*	*Comp_67 (Nov08)*	*P_42 (Nov09)*	*P_52 (Nov09)*	
5	0.39		0.51	0.10	1145
6		0.33	0.55	0.12	317
Intended weightings	0.50		0.38	0.12	

Table 21: Achieved weightings in November 2009 Zones P & Q, AS level awarded

9	Paper 1		Paper 2		Paper 3				N
Combination	*P_11 (Nov09)*	*P_12 (Nov09)*	*P_21 (Nov09)*	*P_22 (Nov09)*	P_31 (Nov09)	P_32 (Nov09)	P_33 (Nov09)	P_33 (Nov09)	
7	0.30		0.53		0.17				1036
8		0.31	0.50			0.19			670
9		0.28		0.54			0.17		2719
10		0.29		0.55				0.16	1370
Intended weightings	0.31		0.46			0.23			

Due to the size and detailed nature of the task, each expert was asked to consider the content and learning outcomes covered in two sessions. For each paper in turn, the experts identified the content point or points which apply to the question. They were then asked to identify the Learning Outcome or Learning Outcomes which apply to the question. They continued with this process for all the questions on all three papers for their two allocated sessions.

Findings

CONTENT AREAS – SUMMARY OF COVERAGE

The content areas identified as assessed were summarised across experts. For Paper 1 (and 11), the subject content areas that were either not covered or only marginally covered are shown in Table 22.

Table 22: Subject content areas not covered or only marginally covered for Paper 1/11

Section	*Part*
II Newtonian Mechanics	7. Motion in a circle 8. Gravitational Field
III Matter	11. Ideal Gases 12. Temperature 13. Thermal Properties of Materials
IV Oscillation and Waves	14. Oscillations
V Electricity and Magnetism	18. Capacitance 21. Magnetic Fields 22. Electromagnetism 23. Electromagnetic Induction 24. Alternating Currents
VI Modern Physics	25. Charged Particles 26. Quantum Physics
VII Gathering and Communicating Information	28. Direct Sensing 29. Remote Sensing 30. Communicating Information

Across all three papers, a number of content areas were less frequently covered over the three years of sessions. These were:

- III Matter: Part 12 – Temperature
- IV Electricity and Magnetism: Part 21 – Magnetic Fields
- VI Modern Physics: Part 25 – Charged Particles

This evidence could imply that ideally these topics need to be covered more frequently, if all content areas are considered equally important. However, as there is some coverage of all content areas these data are not a significant concern for validity.

LEARNING OUTCOMES – SUMMARY OF COVERAGE

Learning outcomes are used in the A level Physics syllabus in order to specify the content and learning as precisely as possible and also to emphasise the importance of skills other than recall. Each part of the syllabus is specified first by a brief contents section followed by detailed learning outcomes.

The data suggest that the coverage of learning outcomes across the six sessions is good. However, a small number of content areas were only partly assessed, and a few section parts were appreciably under-represented, in terms of the underlying skills (see Table 23).

Table 23: Under-represented section parts of the syllabus

Section	*Part*
III Matter	12. Temperature
V Electricity and Magnetism	22. Electromagnetism 24. Alternating Currents
VI Modern Physics	25. Charged Particles
VII Gathering and Communicating Information	28. Direct Sensing 29. Remote Sensing

Evidence for validity

Analysis revealed that the main subject content topics were equally represented overall. All specific content areas within main subject content topics were covered over the six sessions.

Coverage of learning outcomes across the six sessions was good.

Threats to validity

There were a few specific topics sparsely covered over the six sessions according to experts' judgements. This evidence could imply that ideally these topics need to be covered more frequently, if all content areas are considered equally important. However, as there is some coverage of all content areas these data do not constitute a great threat to validity. Some content areas are only partly assessed in terms of learning outcomes.

Validity Evidence 13:
Ratings of the Assessment Objectives measured by the exam questions

VALIDATION QUESTION 3

Do the tasks adequately sample the constructs that are set out as important within the syllabus?

Method

This exercise involved the experts judging the extent to which they felt each subcomponent of the Assessment Objectives was measured by each of the questions on the papers from November 2009.

The six experts (four senior examiners and two other subject experts) were asked to look at the list of Assessment Objectives and the subcomponents of these and think about which of these the question assesses and to what extent. They then rated the extent to which each exam question was thought to reward each Assessment Objective subcomponent on a scale of 0 (not assessed at all) to 5 (strongly assessed). Ratings were made at the whole question level. Experts were asked to use the question paper and the mark scheme and to focus on the knowledge/understanding/skills rewarded in the mark scheme. The instructions to the experts about the task noted that they might find that a particular question only assesses a handful of the Assessment Objective sub-components, thus many of the ratings would be 0 for that question. Frequencies of questions receiving each rating were calculated and then these were weighted by the maximum mark available. In this method the Assessment Objectives are being used as a representation of the underlying constructs.

Differences between ratings given by each expert for the same question with and without the mark scheme were calculated and frequencies found. Some random variation is to be expected but a stronger difference in one direction may suggest that the mark scheme does not reward the same knowledge/understanding/skills as the question alone would suggest.

Findings

These weighted frequencies are shown as a percentage in Table 24. In interpreting these data, it would be desirable to see each subcomponent being assessed to varying degrees by a number of questions. Thus, ideally there should not be an excessively high frequency of '0' ratings, and a spread of other ratings. However, it should also be noted that some Assessment Objective subcomponents may be more important than others and hence likely to be tested more frequently, and more strongly. Others might be less important, or intrinsically only likely to form one smaller element of a question.

Table 24: Weighted frequencies of ratings of Assessment Objectives elicited by questions (with reference to the mark scheme)

Assessment Objectives		*Rating*					
		0 'not assessed at all'	*1*	*2*	*3*	*4*	*5 'strongly assessed'*
A1	%	26.9	6.0	10.9	13.8	16.2	26.2
A2	%	17.5	6.9	23.1	17.1	16.2	19.3
A3	%	73.9	2.0	4.5	5.4	6.6	7.7
A4	%	51.5	4.3	4.7	15.6	15.0	8.9
A5	%	82.5	5.2	5.2	2.6	3.0	1.5
B1	%	69.0	4.3	5.8	7.9	2.5	10.6
B2	%	54.6	4.3	9.3	13.5	7.5	10.9
B3	%	21.4	4.0	7.2	17.1	17.8	32.5
B4	%	65.6	4.2	14.8	8.6	6.3	0.6
B5	%	64.7	2.8	6.8	5.9	16.0	3.9
B6	%	90.6	5.0	2.5	1.6	0.0	0.2
B7	%	68.6	6.6	4.1	13.0	4.6	3.2
B8	%	93.1	3.6	1.0	0.7	1.6	0.0
B9	%	98.3	0.7	0.0	0.9	0.0	0.0
C1	%	82.4	0.0	0.0	3.1	3.7	10.8
C2	%	80.4	3.0	0.0	0.9	1.2	14.5
C3	%	70.2	3.5	2.7	5.6	3.1	14.9
C4	%	82.2	0.4	2.2	3.7	1.2	10.2
C5	%	79.6	1.0	1.9	0.9	4.9	11.7

According to the experts' ratings, many of the Assessment Objective subcomponents are tested to varying degrees by a number of the questions. This is a positive finding suggesting a good coverage of knowledge, understanding and other skills are being assessed. For a few subcomponents it can be noted that these appear to be tested less frequently, and less strongly.

Table 25 shows the frequency with which experts' ratings shifted when considering the mark scheme compared to when they had rated the Assessment Objectives elicited with reference only to the question. (For simplicity, the frequencies of differences in ratings have not been weighted to take into account the marks available.)

Table 26 shows the total positive and negative differences and the overall change.

Evidence for validity

For all components of the Assessment Objectives, there were some questions that were considered to reward them. Considering the differences in ratings when made with and without the mark scheme,

Table 25: Percentage frequencies of differences in ratings of Assessment Objectives with and without reference to the mark scheme (ratings with mark scheme minus ratings without mark scheme)

Assessment Objectives		*Change in ratings*										
		-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A1	%	2.4	0.5	2.9	6.1	13.5	52.1	12.2	7.1	2.1	0.3	0.8
A2	%	1.3	0.5	9.3	10.3	10.3	42.3	17.7	6.9	1.1	0.3	
A3	%	1.3	0.5	1.6	1.6	5.6	84.1	2.4	2.4		0.5	
A4	%	1.3		4.5	7.9	7.4	65.6	7.7	2.9	1.1	1.6	
A5	%	0.3	1.6	0.5	1.9	5.8	81.7	5.8	1.9	0.5		
B1	%	2.9	0.8	0.3	3.4	7.4	71.4	8.2	3.2	2.1	0.3	
B2	%	3.4	1.1	4.0	6.3	7.7	64.3	6.1	4.5	1.9	0.3	0.5
B3	%	0.5	0.3	1.1	8.2	13.5	63.8	6.3	3.4	1.9	0.5	0.5
B4	%	0.8	0.8	4.5	3.7	10.6	67.2	4.5	4.0	3.7	0.3	
B5	%		0.8	1.6	1.6	5.6	78.8	4.5	3.4	1.9	1.6	0.3
B6	%		0.3	0.8	1.6	4.0	90.7	1.6	0.3	0.5		0.3
B7	%	1.3	0.8	3.4	11.9	11.4	60.8	3.7	4.5	1.3	0.8	
B8	%		0.3	1.9	3.4	8.2	84.4	1.3	0.5			
B9	%			0.3	0.5	1.3	97.1	0.5	0.3			
C1	%				0.8	0.3	98.4	0.3		0.3		
C2	%			0.3	0.3	0.5	96.8	1.3		0.8		
C3	%		0.3	0.3	4.5	4.0	86.0	1.3	2.4	1.1		0.3
C4	%	0.3		0.3	0.8	1.1	94.7	1.1	1.3	0.5		
C5	%	0.3		0.3	0.8	1.1	95.8	0.5	0.5	0.5		0.3

Table 26: Totals of positive and negative differences in ratings for each Assessment Objective

Assessment Overall	*Total negative change*	*Total positive change*	*Objectives change*
A1	-183	143	-40
A2	-255	135	-120
A3	-84	35	-49
A4	-164	87	-77
A5	-71	42	-29
B1	-124	83	-41
B2	-203	92	-111
B3	-139	89	-50
B4	-146	93	-53
B5	-63	93	30
B6	-40	19	-21
B7	-209	75	-134
B8	-82	9	-73
B9	-12	4	-8
C1	-7	4	-3
C2	-7	14	7
C3	-56	40	-16
C4	-18	20	2
C5	-18	17	-1

in most cases there was not a strong change. This suggests that the questions give a reasonable impression of the skills that will be rewarded.

Threats to validity

A possible threat to validity is that some components of the Assessment Objectives were rewarded more frequently than others, even when frequencies were weighted to take into account the marks available. Whether this constitutes a threat to validity relates to whether there are differences in the intended importance of each Assessment Objective.

When differences in ratings with and without the mark scheme were considered there was a general tendency for ratings to be lower once the mark scheme was considered and there were a few Assessment Objectives (e.g. B7) for which this difference was fairly strong. This suggests that the skills that will be rewarded are not always entirely clear.

Validity Evidence 14:
Ratings of cognitive demands as rewarded by the mark schemes

VALIDATION QUESTION 3

Do the tasks adequately sample the constructs that are set out as important within the syllabus?

Method

Six experts (four senior examiners and two other subject experts) rated the cognitive demands rewarded by the mark scheme for each exam question for each of five types of demand (from Hughes, Pollitt and Ahmed, 1998) on a scale of 1 (low demand) to 5 (high demand). Ratings were made at the whole question level. Experts were asked to focus on the mark scheme and demands that, if met, were rewarded. This task was conducted for the five Zone P exam papers from November 2009 which formed the focus of most other analyses, but also for the alternative practical paper (Paper 32) and for all six Zone Q papers (Papers 12, 22, 33, 34, 42 and 52) from November 2009. This task was similar to the task reported in 'Validity Evidence 6' but with the focus on what is rewarded by the mark scheme. The experts were asked not to refer to their previous ratings when conducting this task and were asked to leave at least a delay of one day between the two tasks. As before, experts had explanatory information defining the concept of assessment demands, the types of cognitive demands to be rated, and the rating scale available to them.

Frequencies of questions receiving each rating were calculated and then these were weighted by the maximum mark available on each question. Frequencies and weighted frequencies were compared between zones. For Papers 11, 21, 31, 41 and 51, differences between ratings given by each expert for the same question with and without the mark scheme were calculated and frequencies found. Some random variation is to be expected but a stronger difference in one direction may suggest that the mark schemes do not reward the same demands as the questions appear to require.

Findings

The weighted frequencies of ratings for Papers 11, 21, 31, 41 and 51 are shown as a percentage in Table 27.

Table 27: Percentage weighted frequencies of ratings of cognitive demands reflected in scores as indicated by mark scheme across November 2009 Papers 11, 21 31, 41 and 51 (with reference to the mark scheme)

Demand type		*Ratings of demand*				
		1 low demands	*2*	*3*	*4*	*5 high demands*
Complexity	%	3.9	21.6	29.4	29.3	15.7
Resources	%	6.0	18.1	30.6	23.9	21.4
Abstractness	%	38.7	27.5	24.3	8.8	0.8
Task strategy	%	2.0	21.2	38.3	26.9	11.5
Response strategy	%	3.3	23.8	36.5	19.8	16.7

Table 28 shows the frequency with which experts' ratings shifted when considering the mark scheme compared to when they had rated the required demands based only on the question. (For simplicity, the

Table 28: Differences between demand ratings made with and without reference to the mark scheme for November 2009 Papers 11, 21, 31, 41 and 51

Demand type		*Changes in ratings of demand**							
		-3	*-2*	*-1*	*0*	*+1*	*+2*	*+3*	*+4*
Complexity	%	0.5	1.1	14.8	73.0	10.3	0.3		
Resources	%	1.1	1.9	18.3	62.2	14.6	1.6	0.3	0.3
Abstractness	%		0.5	18.8	66.7	13.2	0.8		
Task strategy	%	0.5	1.1	12.2	65.3	19.3	1.3	0.3	
Response strategy	%		1.6	23.3	61.9	12.4	0.8		

*Values were calculated by subtracting ratings without mark scheme from ratings with mark scheme.

Table 29: Percentage weighted frequencies of ratings of cognitive demands rewarded by the mark scheme for all November 2009 Zone P exam papers (Papers 11, 21, 31, 32, 41, 51) and all November 2009 Zone Q exam papers (Papers 12, 22, 33, 34, 42, 52)

Demand type		*Ratings of demand*				
		1 low demands	*2*	*3*	*4*	*5 high demands*
Complexity – P	%	3.4	24.2	26.7	29.8	15.9
Complexity – Q	%	7.0	25.7	24.0	26.3	16.9
Resources – P	%	5.3	19.0	29.9	21.9	24.0
Resources – Q	%	5.8	19.8	35.3	17.6	21.5
Abstractness – P	%	42.3	28.2	21.1	7.6	0.7
Abstractness – Q	%	42.1	27.6	20.7	9.0	0.6
Task strategy – P	%	1.7	19.6	37.7	27.7	13.3
Task Strategy – Q	%	1.9	23.2	36.4	23.6	14.9
Response strategy – P	%	2.8	20.7	36.1	19.4	21.0
Response strategy – Q	%	3.2	24.1	35.0	17.5	20.2

frequencies of differences in ratings have not been weighted to take into account the marks available).

Table 29 shows the weighted frequencies of ratings for each demand type for all Zone P papers (including Paper 32) and all Zone Q papers in order to allow comparison between zones.

Evidence for validity

For all demand types, the questions varied across the range of the demands rating scale. There were fewer questions rated as high in demands for 'abstractness' but this may be a consequence of the nature of physics rather than indicating an inappropriate lack of spread in demands.

Considering the differences in ratings when made with and without the mark scheme, it was found that in most cases there was not a strong change. This suggests that the questions give a reasonable impression of the demands that will be rewarded. When differences in ratings with and without the mark scheme were considered there was a slight tendency for ratings to be lower once the mark scheme was considered but these differences do not seem to be large enough to be problematic.

Comparison of the demands of the Zone P and Zone Q papers suggested that the profiles of demands overall were very similar. This suggests comparability of the task demands between the two time zones.

Threats to validity

None

Validity Evidence 15:
Views from Higher Education experts on the importance of various aspects of the syllabus

VALIDATION QUESTION 4

Are the constructs sampled representative of competence in the wider subject domain?

Method

Five Higher Education representatives were asked to evaluate the importance of different elements of the Physics A level syllabus. They were asked to identify whether each syllabus aim, each assessment objective subcomponent, each content area and each learning outcome was 'very important', 'quite important' or 'not important' as preparation for university study. The views from the experts were then summarised for each element that was rated (for example, if an element was rated as 'very important' by two or three experts, and 'quite important' by the remaining two or three experts, it was categorised as 'very important/quite important' overall). For the Assessment Objectives and content, the experts were also asked to list anything they felt was missing from the syllabus that is important as preparation for further study in the domain.

Summary of findings

Table 30 shows the number of elements for which different views were given.

Table 30: Frequency of views on the importance of aspects of the syllabus

Summary of views	*Aims*	*Assessment Objectives*	*Content*	*Learning Outcomes*
Very important	5	11	57	126
Very important/quite important	9	6	33	85
Quite important	1	0	5	30
Not important/quite important	1	0	1	8
Not important	0	0	1	7
Very important/not important	0	1	0	4
Mixed views	2	1	7	45
Total	**18**	**19**	**104**	**305**

A relatively small number of additional suggestions for knowledge, understanding and skills and content points that are important to preparation for university were given.

Evidence for validity

The majority of the aims, Assessment Objective subcomponents, content areas and learning outcomes were evaluated as very important or quite important. Only a small number of additional knowledge and skill types (i.e. Assessment Objectives and content) were suggested as necessary for good preparation for Higher Education.

Threats to validity

A small number of rated elements were considered unimportant or received mixed evaluations from the experts (suggesting that their value is less clear). For example, content on the mobile phone network in relation to communication of information was generally considered to be of low importance. Some experts suggested a number of possible additional content areas that they felt were important as preparation for

Higher Education. Sometimes these were extra points that could be added to existing content areas whilst sometimes they constituted additional topics. Examples of extra topics or points included the need for linked maths skills, and topics such as particle physics and cosmology.

Validity Evidence 16:
Questionnaire to Higher Education representatives

VALIDATION QUESTION 5
Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?

Method

In order to explore how Higher Education Institutions understand and use grades from International A levels using the available guidance on grade/score meaning, a questionnaire was designed. The questionnaire drew on ideas from the literature. The questionnaire was sent by email to senior tutors within the Physics department of 17 universities in the UK known to have substantial numbers of international students. Tutors were invited to complete and return the questionnaire by email. Only three universities completed the questionnaire. The five Higher Education representatives who completed the task of evaluating the importance of various aspects of the syllabus also completed a version of this questionnaire, but only including questions on the preparation that A level provides for further study.

Summary of findings

Tutors did not know which of their students had taken International A level Physics nor do they perceive any difference between different Physics A level qualifications in terms of the knowledge, understanding and skills that students bring with them when they begin their degree studies.

In terms of preparation for university study, most tutors felt that students lack the necessary mathematical skills. Generic laboratory skills such as problem solving, error assessment and note taking are also thought to be inadequate. Content areas perceived to be missing from A level Physics courses related to aspects of modern physics such as biophysics, nanotechnology and Large Hadron Collider (LHC) physics.

According to the questionnaire respondents, A level Physics grades are not a particularly good predictor of academic success at university. Some tutors believe that A levels test different skills compared to university exams. Following an initiative by one Physics department, an analysis of student performance was conducted resulting in the conclusion that the strongest predictor of success on a university Physics course is the A level Mathematics grade: students with A grades almost never failed the first year according to their analysis.

For the most part, students' A level Physics grades reportedly compare well with the tutor's own assessment of their Physics ability. This is especially true in early years of university study.

Nearly all of the tutors considered a current grade A in A level Physics to be at a lower standard than in previous years though several qualified their answers by suggesting that grades over time are based on different forms of assessment. The 'decline' in standards that they reported was not perceived to be linear across grades, for example, the standard at grade C appears to be maintained according to some of the respondents.

A level Physics courses are perceived to be sufficiently modern though there are some notable contemporary omissions including recent developments in particle physics and in cosmology.

Most tutors, though not all, consider A level Physics courses to provide students with good preparation for Higher Education study. However, given that there has been a 'retreat' from mathematics in the present Physics A level, many students reportedly now arrive at university perceiving maths and physics as quite separate disciplines, unaware of the connections between them.

Tutors are unaware of any forms of guidance on the meaning of A level grades and on how grades should be used.

Evidence for validity

Higher Education tutors felt that A level Physics courses generally provide good preparation for Higher Education study.

Threats to validity

Some tutors felt that A level Physics does not prepare learners with some of the skills and content knowledge valuable to further study. (These comments were general to A level Physics qualifications and not specific to the International A level course that was the focus of the validation study.)

Validity Evidence 17:
Teacher questionnaire

VALIDATION QUESTION 5
Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?

Method

A questionnaire for teachers was designed to provide evidence regarding validity with respect to decision-making in the framework. This relates to whether guidance is in place about how to prepare students for exams, whether guidance is in place on the meaning of grades/scores, and issues relating to progression from IGCSE to A level. The questionnaire was sent by email to 258 schools/colleges known to teach the Physics A level syllabus. The email explained the purpose of the research and invited teachers to complete and return the questionnaire by email. Fifty-three schools/colleges returned a completed questionnaire (representing an approximate 20% response rate).

Summary of findings

TEACHER CHARACTERISTICS

The responding teachers varied in teaching experience (2 to 49 years) and in experience teaching CIE Physics A level (1 to 37 years). Some had previously taught other Physics A level syllabuses including some from the UK. They were teaching in 17 different countries including Africa (e.g. Kenya, Zimbabwe) and Asia (e.g. India, Singapore).

GUIDANCE ON TEACHING

A number of the questions asked teachers about guidance helpful to teaching. Firstly they were asked whether they were aware of a number of different possible sources of guidance on teaching such as the CIE website, Teacher Support site (website), schemes of work, training events, the syllabus, revision checklists etc. The resources that more teachers were aware of were the CIE website (n=47), Examiner reports (n=47),

the syllabus specifications (n=46) and the Teacher Support website (n=46). When asked which resources they found most useful and why, the most popular responses were the Teacher Support site (n=19) and the syllabus specifications (n=15). The Teacher Support site was felt to be most useful because it enhances teaching and enriches facts; provides many teaching resources such as syllabus, past papers, schemes of work, practical skills, lesson plans; and, provides FAQs. The syllabus specification was found useful because it provides good guidance for students; provides contents that are divided into AS and A2 Core as well as applications and emphasises the importance of skills other than recall; and offers guidance on width and depth of required teaching and learning.

The questionnaire then asked teachers to describe how the guidance available affects their teaching strategies. According to the teachers, the guidance appears to help with suggesting class activities/tasks, teaching strategies and course structure; determining practice questions; facilitating lesson planning, course structuring (and learning outcomes) and practicals. The available guidance also provides a valuable teaching resource including identifying what constitutes a 'good' answer and discerning trends in questions.

Teachers were then asked whether they used the schemes of work as a basis for the lessons that they teach. Responses were mixed: approximately the same number of teachers answered 'never' (n=13) as 'always' (n=11). Some teachers suggested that the schemes of work provide clear delineation of subtopics, topics and learning outcomes and a range of student activities. Two teachers were unaware that CIE prepare schemes of work.

When asked whether they used the content set out in the syllabus as a basis for the lessons that they teach, the overwhelming response was either 'often' or 'always'. Reasons given were categorised under two groups: learning outcomes; application of content. Content set out in the syllabus provides guidance on the width and breadth of topics; enables a teacher to make notes/tests based on syllabus content; provides guidance on teaching and lesson plan preparation; offers useful guidelines for students as a summary of course content, and ensures students are within the scope set by CIE.

Teachers were asked whether they used the content of key textbooks as a basis for the lessons that they teach. Most of the teachers (n=42) either used them 'sometimes', 'often' or 'always'. Textbooks seem to provide questions for student practice and helpful definitions of quantities. They also help students organise themselves and practice problems and questions. One teacher was not aware of key textbooks.

The teachers were asked whether the guidance helps them to teach the knowledge, understanding and skills that the course requires. Only three answered 'no' suggesting that: the contents in the syllabus are too vague; that Paper 4 requires mainly application on the part of the student; and that the syllabus is simply the content required.

Teachers were also asked the same question in relation to schemes of work. The vast majority (n=38) answered 'yes'. Five teachers claimed not to have used CIE schemes of work (or did not perceive a need for them) and a further two were unaware of them.

Asked if they felt that preparing students for the assessments has a positive influence on learning, many responded with 'strong positive' (n=38) suggesting that assessment engenders greater student motivation and provides an evaluative dimension through feedback. Other reported positive influences include self-evaluation and application of knowledge beyond the classroom/assessment. When asked whether preparing students for the assessments has a negative influence on learning, 31 teachers reassuringly suggested 'no negative influence'. Assessment was not without its detractors, however. Some teachers suggested that assessment promotes rote learning, and discourages creativity and deep learning.

Teachers were then asked if they use Assessment Objectives with their students when preparing them for the exams. Most did (n=46) indicating that Assessment Objectives enhance question familiarity, ensure content coverage, and help prepare students adequately for examinations.

Teachers were also asked how they currently prepare students for their examinations and whether the guidance available enabled them to do this. Nearly every response referred to the importance of using past papers for student preparation.

The last question in this section asked whether teachers believed there is sufficient guidance in place so that they know how to prepare students for the assessments in a way that encourages positive backwash on classroom practice. Forty two teachers answered 'yes'. More importantly, however, 11 answered 'no'. Some of these teachers believed that the Applications support could be a little more explanatory and that Guidance for Paper 5 (designing and planning) is insufficient. Also, marking schemes provide too little information.

GUIDANCE ON SCORE/GRADE MEANING AND USE

In the first question in this section on guidance on score/grade meaning and use, teachers were asked whether the guidance helped them to understand what an exam grade/score meant. An overwhelming number answered 'yes'. Those answering 'no' suggested that they would prefer marks to grades. Moreover, they did not consider the grading process to be transparent. Teachers expressed concern that the range of marks for grades and the meaning of grades are unclear to universities and employers and that more guidance is required.

About half of responding teachers felt that the guidance on grades/scores helped them when advising their students on further education and/or future employment. Additionally, two-thirds of respondents felt that guidance does not provide information on how exam grades should be used by Higher Education Institutes and employers. Some teachers considered the choice of future education "a much bigger picture" and "beyond the scope of A level subject teaching".

When asked whether examination grades/scores informed teaching, only 7 teachers answered 'no' believing that their role is to teach physics and not how to pass an exam.

The grades/scores achieved by students taking AS exams at the end of the first year of A level study appear to inform subsequent teaching of students in a number of ways: supplying information as to whether students should have access to an A2 programme of learning; identifying student (and teacher) strengths and weaknesses; and, advising students whether or not to re-sit AS.

For many teachers exam scores/grades informed their teaching practices. When asked how final exam grades/scores (across all papers) inform the teaching of future A level classes, teachers indicated that paper breakdown is highly informative in that it helps to evaluate the efficacy of their own teaching and provides a means for modifying future teaching practice especially in relation to the different skills assessed. Not all teachers were as encouraging, however, with some suggesting that final grades provide only limited information.

Teacher responses when asked whether sufficient guidance is in place

so that they and others know what scores/grades mean and how the outcomes should be used suggest that some of the information available is useful in terms of score/grade meaning and use but that this may be an area where CIE could increase the assessment validity in terms of impact by providing more guidance on intended score/grade meaning and uses.

PROGRESSION FROM IGCSE TO A LEVEL

Teachers were asked to consider whether the learning encouraged in the IGCSE prepared students well for success at A level. Just under half of those who answered thought it did and considered A level to be a natural and seamless extension of IGCSE, though requiring a greater depth of understanding. Additionally, their experience suggested that students who are successful at IGCSE tend to be (though not always) successful at A level. However, one teacher suggested that critical thinking and the synoptic skills of linking ideas across topics is not yet fully developed at the end of the IGCSE programme.

A third of teachers considered the transition from IGCSE to A level to be huge. IGCSE is thought to be shallow by comparison and is content based whereas the AS course requires in-depth thinking especially in relation to the mathematical skills required. High IGCSE grades can mislead students by erroneously influencing their A level subject choice.

Teachers were entirely split in terms of whether they thought IGCSE grades a good predictor of how well students perform at A level though several teachers answering 'no' were unsure. Teachers found a large variation in performance across IGCSE and A level. Some teachers thought that AS and A2 papers were unnecessarily difficult especially in the demands they made of students mathematically. The disparity in difficulty makes prediction of future success at A level on the basis of IGCSE grades somewhat uncertain.

Evidence for validity

The questionnaire data suggested that most teachers found resources such as the CIE website, examiner reports, the syllabus specifications and the Teacher Support website particularly useful in their teaching and in preparing students for their exams. This provides support for assessment validity in relation to impact. Available guidance appears to affect teaching strategies positively and most teachers use the content set out in the syllabus as a basis for the lessons they teach. Guidance also helps teachers to teach the knowledge, understanding and skills that the course requires. Assessment appears to have a strong positive influence on learning.

Threats to validity

None

Validity Evidence 18:
Document review on guidance on score/grade meaning and use

VALIDATION QUESTION 5

Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?

Method

This document review considered the guidance for stakeholders on what scores/grades mean and how they should be used. The CIE website provides general information on the recognition of A levels. In addition the 'Standards booklet' for A level Physics, which is available to centres and others, was reviewed.

Findings

REVIEW OF CIE WEBSITE

Information is provided on CIE's website about the recognition of A level qualifications (http://www.cie.org.uk/qualifications/academic/uppersec/alevel/recognition). It states that International A levels have the same value in admitting students to universities as the UK equivalent. Admission to the world's major Anglophone universities is usually contingent upon good A level grades.

The information on the website states that A level grades A to E are passing grades and then provides information on university admission criteria in relation to these grades. The recognition search tool (http://recognition.cie.org.uk/), a 'Recognition Brochure' (http://www.cie.org.uk/docs/recognition/cie_recognition_brochure_row.pdf), and some of the 'Frequently Asked Questions' (http://www.cie.org.uk/qualifications/recognition/faqs) on the CIE website also provide information on this area.

REVIEW OF STANDARDS BOOKLET

CIE produce a 'Standards booklet' for A level Physics which is available to centres and others via their publications department. The latest version – currently in print – includes exemplars of candidate exam responses from November 2008. Examples are given for a selection of questions from each of the four written papers: Paper 2 (AS Theory); Paper 31 (Advanced Practical Skills); Paper 4 (A2 Theory); Paper 5 (Planning, Analysis and Evaluation).

For each paper, sample responses are given for a number of questions. In each case, responses are shown from each of three candidates:

- Responses typical of a student who receives an eventual grade A in the overall International A level Physics qualification.
- Responses from candidates who, although also very good, fall just short of the standard expected from a grade A student.
- Responses from candidates who demonstrate only just sufficient knowledge to gain an eventual pass mark.

In each case, candidates' responses are accompanied with an examiner's commentary, explaining where answers fall short of the standard expected, and giving suggestions for how students' answers could be improved.

The responses shown are genuine answers given by candidates, though in some cases these come from an amalgam of different scripts. The answers have been rewritten, to ensure anonymity.

Evidence for validity

There is information available on recognition of A level grades for entrance to university courses.

Threats to validity

The documentary review suggested that there is limited further detail on the meaning of A level grades and how they should be used. There are no grade descriptors for the qualification (since International A levels are intended to align with their UK counterparts). There may be a threat to validity here in that only limited information is available about meaning and use of grades.

8. Summary of A level Physics validation findings and evaluation of the argument

Table 31 summarises the evidence relating to each validity question for A level Physics.

The evidence collected as the backing for the assumptions supports the A level validity argument. The types of validity evidence collected here have been designed to support the main inferences and assumptions in the interpretive argument and are dependent upon the proposed interpretations and uses of the test scores.

The validation process broadly entails three stages: the development stage (constructing the interpretive argument); the appraisal stage (gathering evidence in order to construct the validity argument); and the evaluation stage. Thus, after the interpretive and validity arguments have been specified they need to be evaluated.

In Table 32, the validity argument is presented, including the inferences, warrants and assumptions of the interpretive argument. The evidence for validity that was collected via the methods is presented as the backing for the validity argument.

Any argument for validity is open to challenge – what Kane refers to as refuting the argument, or the presence of counterevidence. In other words, it is possible to "challenge the appropriateness of the proposed goals and uses of the testing program…the adequacy of the interpretive argument, or the plausibility of the inferences and assumptions" (Kane, 2004, p.166). Counterevidence can be identified as 'rebuttals' which are added to the validity argument. Rebuttals constitute alternative explanations, or counter claims to the intended inference. The counterevidence for the A level Physics validity argument are shown in the final column of Table 32.

Within the evaluation stage, criteria are applied to evaluate the interpretive argument and the validity argument. The criteria used are based on theories of the evaluation of informal and practical arguments (e.g. Blair and Johnson, 1987; Toulmin, 1958/2003; Verheij, 2005).

This section will now present the evaluation of the validity argument for International A level Physics. In order for the validity arguments to be evaluated, three questions need to be addressed (these are drawn from Kane, 2006).

(a) Does the interpretive argument address the correct inferences and assumptions?

The first question relates to whether the correct inferences and assumptions are addressed in the interpretive argument. The main point of the interpretive argument is to make the assumptions and inferences in the interpretation as clear as possible: "If some inferences in the argument are found to be inappropriate, the interpretive argument needs to be either revised or abandoned" (Kane, 2001, p.331). For the interpretive argument to be well structured the underlying assumptions must support, or lend credence to, the claim through an implicit or explicit inference from another, already accepted belief (the warrant).

Blair and Johnson (1987) have proposed a set of criteria that are required of argument premises. These include *acceptability* (the truth status of the premises); *relevance* (whether the assumptions warrant the conclusion), and *sufficiency* (whether the assumptions provide enough evidence considering everything known).

We would argue that the inferences and assumptions for A level Physics are made clear in the interpretive argument set out in Section 2. We argue that these inferences and assumptions are appropriate and logical given the identified assessment purposes/uses. In preparing them we built on Kane's (2006) exemplification of the inferences for a particular assessment type and tailored this strategy to be applicable to A levels. We drew on how Kane (2006) and Chapelle, Enright and Jamieson (2010) set out assumptions in setting out those for A levels.

We are satisfied that:

- the inferences and assumptions are logical and reasonable (they are *acceptable*);
- the assumptions, if shown to be true, can warrant the claims (the inferences and assumptions are *relevant*);
- the assumptions, if shown to be true, justify the claimed inference (they are *sufficient*).

(b) Are the inferences justified by the evidence collected?

The second question to be addressed is whether the inferences are justified. In order to answer this question it is necessary to evaluate whether the evidence presented is plausible and whether the inferences are coherent. When arguments are evaluated in formal logic, it has to be decided whether an argument is valid or invalid. Toulmin (1958/2003) employs an 'evaluation status' to determine this (see Verheij, 2005; Wools, Eggen and Sanders, 2010). This requires two steps, as described and addressed below.

Step 1:
Evaluate the assumptions and statements included in the argument individually and decide whether each statement or assumption is *accepted*, *rejected*, or *not investigated*.

Table 33 gives the evaluation status for each assumption, as judged using the evidence from the validation research.

Step 2:
Assign an evaluation status (Verheij, 2005) to the inference as a whole: *justified*, *defeated*, or *unevaluated*.

- The evaluation status is *justified* when the warrant(s) and backing(s) are accepted and the rebuttal(s) are rejected.
- The evaluation status is *defeated* when a warrant or backing is rejected or when a rebuttal is accepted.
- The evaluation status is *unevaluated* when some statements are not investigated and it is still possible for the inference to become justified.

Evaluations of each inference are given in Table 34.

(c) Is the validity argument as a whole plausible?

The answer to this question – which is contingent upon the first two questions being satisfied (i.e. that the right inferences were chosen and it is also established that the inferences are justified) – addresses the outcomes of the validation process. The purpose of this evaluation is to determine whether the validity argument as a whole is plausible. To answer this we need to take all evidence into account to decide whether the argument is strong enough to convince us of the validity of the assessment. Table 32 summarised the full validity argument and evidence and hence can assist with this judgement. The table includes a summary of the backing that has been collected for the assumptions underlying the inferences of *Construct representation*, *Scoring*, *Generalisation*, *Extrapolation*, and *Decision-making*. It also illustrates the counterevidence as rebuttal which is added to the validity argument.

Table 31: Summary of validation evidence for A level Physics

Validation questions	*Evidence for validity*	*Threats to validity*
1. Do the tasks elicit performances that reflect the intended constructs?	Statistical analyses of item level data (traditional statistics and Rasch analysis) found that for all papers the vast majority of items (question parts) were not extreme in their level of difficulty or easiness, were appropriate in difficulty for the ability of the candidates and functioned well. A review of examiner reports did not suggest that any questions were measuring unintended constructs. Factor analysis was able to identify groups of items that contributed to measuring a number of key traits. Ratings by experts suggested that for all components of the Assessment Objectives, there were some questions that elicited them. Ratings of demand types by experts suggested that a good spread of cognitive demands were placed on candidates across the questions. For 8 of 12 items found to be underfitting across all papers, analyses of student responses did not suggest any threats to validity in relation to the constructs elicited.	For a small number of items/question parts, there were low correlations between item marks and total marks on the rest of the paper and/or possible concerns about functioning. This may be an indication that these items measured something different to most items. According to experts' ratings, some components of the Assessment Objectives were elicited less frequently than others, even when the marks available on each question were taken into account by weighting the frequencies.
2. Are the scores/grades dependable measures of the intended constructs?	A document review indicated that marking procedures conform to good assessment practice. Marking procedures were comprehensively documented. The estimated syllabus reliability was high. Statistical analysis of effectiveness of aggregation (achieved weightings of components) indicated that Papers 11, 21 and 51 contributed as intended to overall scores. Multiple marking exercises in which, for each paper, five examiners marked 30 scripts, showed encouragingly high marking reliability statistics. Rasch analysis of the data found very few instances of misfitting examiners, suggesting similar individual variability in marking. Examiners were similar in their level of severity.	There was a general tendency amongst examiners to exhibit a lack of variability in their scores. Analysis of weightings suggested that Paper 3 contributed too little to variance in overall scores. Paper 2 contributed too much when considered in terms of its contribution to AS level scores. Paper 4 often contributed too much to variance in total scores, especially when scores from AS papers were carried over from a previous session. When the relative contributions of AS and A2 papers are considered, the AS papers tended to contribute less than 50% of the variance in overall scores.
3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?	When experts reviewed the coverage of syllabus content across six sessions the main subject content topics were equally represented overall. All specific content areas within main subject content topics were covered over the six sessions. Coverage of learning outcomes across the six sessions was good. Expert ratings of the extent to which the Assessment Objectives were rewarded by the mark schemes suggested that all Assessment Objective subcomponents were rewarded by some questions. Considering the differences in ratings when made with and without the mark scheme, in most cases there was not a strong change. This suggests that the questions give a reasonable impression of the skills that will be rewarded. Experts' ratings of the extent to which different questions measure cognitive demand types suggested that the questions varied in level of demands. Comparison of the demands of the Zone P and Zone Q papers suggested that the profiles of demands overall were very similar. This suggests comparability of the task demands between the two time zones.	A few specific topics were sparsely covered over the six sessions according to experts' judgements. Expert ratings suggested that some components of Assessment Objectives were rewarded more frequently than others, even when frequencies were weighted to take into account the marks available. Whether this constitutes a threat to validity relates to whether there are differences in the intended importance of each Assessment Objective.
4. Are the constructs sampled representative of competence in the wider subject domain?	The majority of the aims, Assessment Objective subcomponents, content areas and learning outcomes from the syllabus were evaluated as very important or quite important by Higher Education representatives as preparation for university study.	Higher Education representatives suggested a small number of additional knowledge and skill types (i.e. Assessment Objectives and content areas) as necessary for good preparation for Higher Education study in the domain. A small number of aims, Assessment Objective subcomponents, content areas and learning outcomes in the syllabus were considered unimportant or received mixed evaluations from experts (suggesting that their value is less clear).
5. Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?	A document review showed that CIE provides information on the recognition of A level grades for entrance to university courses. Teacher questionnaire data suggested that most teachers found resources such as the CIE website, examiner reports, the syllabus specifications and the teacher support website particularly useful in their teaching and in preparing students for their exams.	The document review and questionnaires to teacher and Higher Education representatives suggested that there is limited detail available on the meaning of A level grades and how they should be used.

Table 32: The International A level Physics validity argument including backing for inferences in the interpretive argument

Inference	*Warrant justifying the inference*	*Assumptions underlying warrant*	*Evidence for validity (Backing/supporting evidence)*	*Threats to validity (Rebuttals)*
Construct representation (task → test performance)	W1 Tasks elicit performances that represent the intended constructs.	A1 Constructs (knowledge, understanding and skills) relevant to the subject can be identified. A2 It is possible to design assessment tasks that require these constructs A3 Task performance varies according to relevant constructs and is not affected by irrelevant constructs.	B1 For Papers 21, 31, 41 and 51, there were no whole questions that were either extremely difficult or extremely easy, thus suggesting that all questions were appropriate to the ability of the candidates. All correlations between whole question marks and total marks on the rest of the paper were good. For Paper 11, where each multiple choice item constitutes a whole question, the items represented a reasonable spread of difficulty and no items were excessively easy or difficult. B2 For Paper 11 there were no items with difficulty levels inappropriate to the ability of the candidates. For Papers 21, 31, 41 and 51, the vast majority of items were appropriate in difficulty to the ability of the candidates. B3 A review of examiner reports did not suggest that any questions were measuring unintended constructs. B4 Factor analysis was able to identify groups of items that contributed to measuring a number of key traits. B5 For all components of the Assessment Objectives, there were some questions that were considered to elicit them. B6 For all five types of demands (complexity, resources, abstractness, task strategy and response strategy) the November 2009 exam questions ranged in demand according to expert ratings, suggesting a good spread of cognitive demands are placed on candidates across the questions. B7 For 8 of the 12 items found to be underfitting, the analyses of student responses did not suggest any threats to validity in relation to the constructs elicited.	R1 Some low point biserial correlations might potentially introduce construct-irrelevant variance. R2 Instances of reverse thresholds and item overfit suggest inter-dependence of items, and that some of the available marks were less useful than others at discriminating between students. Instances of item underfit may be evidence of construct-irrelevant variance in scores. R3 Some components of the Assessment Objectives elicited less frequently than others, even when the marks available on each question have been taken into account by weighting the frequencies.
Scoring (test performance → test score/ grade)	W1 Scores/ grades reflect the quality of performances on the assessment tasks.	A1 Rules, guidance and procedures for scoring responses are appropriate for providing evidence of intended constructs (knowledge, understanding and skills). A2 Rules for scoring responses are consistently and accurately applied. A3 The administrative conditions under which tasks are set are appropriate. A4 Scaling, equating, aggregation and grading procedures are appropriate for differentiating performance in relation to intended constructs.	B1 A document review indicated that marking procedures conform to good assessment practice. Marking procedures were comprehensively documented. B2 The multiple marking exercises in which five examiners marked 30 scripts for each paper showed encouragingly high marking reliability statistics, which compare favourably with those estimated for other qualifications. Rasch analysis of the data found very few instances of misfitting examiners, suggesting similar individual variability in marking. Examiners were similar in their level of severity according to Rasch analysis. B3 The estimated syllabus reliability was high. B4 Statistical analysis of effectiveness of aggregation (achieved weightings of components) indicated that Papers 11, 21 and 51 contributed as intended to overall scores.	R1 General tendency amongst examiners to exhibit lack of variability in their scores. R2 Paper 3 contributed too little to variance in overall scores. Paper 2 contributed too much when considered in terms of its contribution to AS level scores. Paper 4 often contributed too much to variance in total scores, especially when scores from AS papers were carried over from a previous session. When the relative contributions of AS and A2 papers are considered, the AS papers tended to contribute less than 50% of the variance in overall scores.
Generalisation (test score/ grade → test competence)	W1 Scores/grades reflect likely performance on all possible relevant tasks.	A1 A sufficient number of tasks are included in the test to provide stable estimates of test performances. A2 The test tasks provide a representative sample of performance. A3 Task, test and scoring specifications are well defined enabling construction of parallel test forms.	B1 The main subject content topics were equally represented overall. All specific content areas within main subject content topics were covered over the six sessions. Coverage of learning outcomes across the six sessions is good. B2 For all components of the Assessment Objectives, there were some questions that were considered to reward them. Considering the differences in ratings when made with and without the mark scheme, in most cases there was not a strong change. This suggests that the questions give a reasonable impression of the skills that will be rewarded.	R1 A few specific topics were sparsely covered over the six sessions according to experts' judgements. R2 Some components of Assessment Objectives rewarded more frequently than others, even when frequencies were weighted to take into account the marks available. Whether this constitutes a threat to validity relates to whether there are differences in the intended importance of each Assessment Objective.

Table 32: *continued*

Inference	*Warrant justifying the inference*	*Assumptions underlying warrant*	*Evidence for validity (Backing/supporting evidence)*	*Threats to validity (Rebuttals)*
			B3 For all demand types, the questions varied across the range of the demands rating scale. There were fewer questions rated as high in demands for 'abstractness' but this may be a consequence of the nature of physics rather than indicating an inappropriate lack of spread in demands. B4 Comparison of the demands of the Zone P and Zone Q papers suggested that the profiles of demands overall were very similar. This suggests comparability of the task demands between the two time zones.	
Extrapolation (test competence → domain competence)	W1 Scores/grades reflect likely wider performance in the domain.	A1 Constructs assessed are relevant to the wider subject domain beyond the qualification syllabus.	B1 The majority of the aims, Assessment Objective subcomponents, content areas and learning outcomes were evaluated as very important or quite important.	R1 Only a small number of additional knowledge and skill types (i.e. Assessment Objectives and content areas) were suggested as necessary for good preparation for Higher Education in the domain. A small number of aims, Assessment Objective subcomponents, content areas and learning outcomes were considered unimportant or received mixed evaluations from experts (suggesting that their value is less clear).
Decision-making (domain competence → trait competence)	W1 Appropriate uses of scores/ grades are clear.	A1 The meaning of test scores/grades is clearly interpretable by stake holders who have a legitimate interest in the use of those scores i.e. admissions officers, test takers, teachers, employers.	B1 There is information available on recognition of A level grades for entrance to university courses. B2 Teacher questionnaire data suggested that most teachers found resources such as the CIE website, examiner reports, the syllabus specifications and the Teacher Support website particularly useful in their teaching and in preparing students for their exam. This provides support for assessment validity in relation to impact.	R1 There is limited further detail on meaning of A level grades and how they should be used.

We have developed a considered argument by exploring both backing for the validity argument and counter evidence.

We then need to relate the evaluation back to the claims made and the proposed interpretations of assessment outcomes. The claim we make for International A level Physics is that the test scores provide:

a) a measure of relevant learning/achievement in a specified domain and;

b) an indication for likely future success in education or employment in relevant fields.

The two proposed interpretations, indicated in Table 35 in relation to the inferences in the validation framework, are that:

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

According to Kane, "validation requires a clear statement of the proposed interpretations and uses" (2006, p.23) and "if the interpretations and uses are not clearly specified, they cannot be adequately evaluated" (Kane, 2006, p.29).

In validating International A level Physics, we have:

- specified the inferences included in the interpretations and uses;
- evaluated these inferences and their supporting assumptions using appropriate evidence;
- considered plausible alternative interpretations.

Validation, Kane argues, "always involves the specification (the interpretive argument) and evaluation (the validity argument) of the proposed interpretations and uses of the scores" (2006, p.22).

In terms of the interpretive argument, we have demonstrated the clarity and coherence of the arguments as well as the plausibility of the

Table 33: Evaluation status for each assumption in the interpretive argument (Step 1)

Inference	*Assumptions*	*Evaluation status*
Construct representation	1 Constructs (knowledge, understanding and skills) relevant to the subject can be identified	Accepted
	2 It is possible to design assessment tasks that require these constructs.	Accepted
	3 Task performance varies according to relevant constructs and is not affected by irrelevant constructs.	Accepted
Scoring	1 Rules, guidance and procedures for scoring responses are appropriate for providing evidence of intended constructs (knowledge, understanding and skills)	Accepted
	2 Rules for scoring responses are consistently and accurately applied	Accepted
	3 The administrative conditions under which tasks are set are appropriate	Accepted
	4 Scaling, equating, aggregation and grading procedures are appropriate for differentiating performance in relation to intended constructs	Accepted
Generalisation	1 A sufficient number of tasks are included in the test to provide stable estimates of test performances	Not investigated
	2 The test tasks provide a representative sample of performance	Accepted
	3 Task, test and scoring specifications are well defined enabling construction of parallel test forms	Accepted
Extrapolation	1 Constructs assessed are relevant to the wider subject domain beyond the qualification syllabus	Accepted
Decision-making	1 The meaning of test scores/grades is clearly interpretable by stakeholders who have a legitimate interest in the use of those scores i.e. admissions officers, test takers, teachers, employers	Accepted

Table 34: Evaluation status for each assumption in the interpretive argument (Step 2)

Inference	*Evaluation status*
Construct representation	Justified
Scoring	Justified
Generalisation	Justified
Extrapolation	Justified
Decision-making	Justified

Table 35: Validation of proposed interpretations

Interpretive argument		*Validity argument*	*Proposed interpret-ations*
Inference	*Warrant justifying the inference*	*Validation questions*	
Construct represent-ation	**Tasks elicit performances that represent the intended constructs**	1. Do the tasks elicit performances that reflect the intended constructs?	1, 2
Scoring	**Scores/grades reflect the quality of performances on the assessment tasks**	2. Are the scores/grades dependable measures of the intended constructs?	1, 2
General-isation	**Scores/grades reflect likely performance on all possible relevant tasks**	3. Do the tasks adequately sample the constructs that are set out as important within the syllabus?	1, 2
Extra-polation	**Scores/grades reflect likely wider performance in the domain**	4. Are the constructs sampled sampled representative of competence in the wider subject domain?	2
Decision-making	**Appropriate uses of scores/grades are clear**	5. Is guidance in place so that stakeholders know what scores/grades mean and how the outcomes should be used?	2

inferences and assumptions. The plausibility of each assumption underpinning the warrants justifying the interpretive inferences have been judged in terms of all of the evidence for and against the assumption. Cronbach (1971) drew attention to the place and significance of plausible rival hypotheses or alternative explanations of test scores (p.464). According to Cronbach (1971, 1982, 1988), assumptions in the interpretive argument which are most open to challenge are more likely to provide the most useful information particularly in relation to potential threats to validity. Cronbach (1971) suggested that the evaluation of the validity of the proposed test score interpretations and uses necessitates a consideration of all of the evidence for and against the proposed interpretation or use and, wherever possible, any other evidence pertinent to credible counter-interpretations and decision procedures. By highlighting alternate interpretations, the role of evaluating counterhypotheses and the need to collect evidence in support of such hypotheses, Cronbach (1971) raised the potential for identifying aspects of the theory that were either deficient or weak. The rebuttals shown in Table 32 (final column) satisfy Cronbach's call for a balanced argument though they potentially undermine or weaken the force of the interpretive argument. Future

Table 36: Overall evaluation of the validity argument for international A level Physics

Evaluation of claim	*Evidence for validity*	*Threats to validity*
How appropriate are the intended interpretations and uses of test scores? *Interpretation 1.* Scores/grades provide a measure of relevant learning/achievement *Interpretation 2.* Scores/grades provide an indication of likely future success	The interpretive argument is clear, coherent and plausible, and the validity argument is backed up by the supportive nature of much of the validity evidence collected. Thus, both proposed interpretations of scores/grades, and by connection their associated uses, are likely to be highly appropriate based on the available evidence.	The evidence suggests some minor threats to validity in certain areas which should be addressed in order to be further enhance the validity of the proposed interpretations and uses of assessment outcomes.

research might potentially extend the validity narrative to include both the identification of more backing evidence and further challenges to the interpretive argument. Thus the evaluation of new evidence and future alternative explanations will play an integral role in the on-going nature of validation effort.

The validity argument for A level Physics has provided the necessary evidence to evaluate the interpretive argument, including expert judgement, empirical studies and value judgements. The evidence needed for validation depends on the claims being made about the assessment. Given that each inference has been argued as justified in the discussion above based on the evidence, each of the two proposed interpretations are arguably justified. The overall evaluation of the validity argument is shown in Table 36.

9. Conclusions

The validation study presented here marks one of a small number of substantive attempts by any awarding body to describe the theoretical issues and methodological practices surrounding a construct validation exercise. It is hoped that this work provides some guidance on how to validate the use of an assessment for a specified purpose particularly as "those who are actually responsible for validation almost always require detailed and concrete guidance for conducting validation activities" (Brennan, 1998, p.7). The framework and methods described here could provide a useful model for awarding bodies to inform future validation exercises for similar qualifications. However, given that collecting some types of evidence is very time-consuming and requires intensive analysis, it is likely that the use of some of the types of evidence we have used will not be possible for every examination year-on-year. Instead, it might be that a full approach might be appropriate in the development of a new qualification, where a challenge has been made about a particular qualification, or as ongoing monitoring by validating one or more qualifications each year. In addition, an approach using a smaller range of routinely available data types might be adequate to provide a reasonable analysis of an assessment's validity in other cases. As well as post-hoc validation exercises such as the approach described here, validity needs to be built into the process of qualification and assessment design and revision.

The approach to validity taken here assumes a unitary view as articulated by Messick. However, the construct model as a unifying framework for validation has been subject to sustained criticism.

For example, Lissitz (2009) asserts that the unified view "is not just historically unsettled at a theoretical level, but it offers little in the way of usable advice to people working on test construction efforts in the field" (2009, p.5). Even Kane (2006) has criticised validity theory as being 'quite abstract' and called for a more pragmatic approach to validation. He has also commented that long lists of different kinds of validity evidence in the literature, and the suggestion that every possible kind of evidence is needed to claim adequate validation of an assessment, have led to a perception that validation is too difficult (Kane, 2009).

There is a well-established disjunction between modern validity theory and modern validity practice (e.g. Jonson and Plake, 1998; Hogan and Agnello, 2004; Cizek, Rosenberg and Koons, 2008) and there have been explicit criticisms of the capacity of a unified concept of validity to provide sufficient guidance on how to validate the use of a test. Despite this, Moss (2007) contends that "a unitary conception of validity is in no way inconsistent with the provision of substantial guidance, nor does it preclude the making of well-reasoned, practical judgements about what can and should be undertaken before (and after) a test is put into operational use" (p.470).

Whilst acknowledging the difficulties of conducting such studies, Kane cautions that the task of validation "is not one that can be shirked, without potentially serious consequence" (2009, p.61). However, validation is demanding. It entails consideration of multiple 'foci' (e.g. the extent of additional validation required for distinct subgroups of the population); multiple constructs (e.g. attainment, aptitude); multiple uses of results (e.g. student monitoring, selection, placement, comparability); and multiple measures (e.g. decisions taken on the basis of a number of sources of evidence). Further to this, it is difficult to determine how much evidence and what kinds provide a compelling and sufficient validity argument. Sireci (2009) asserts that "most validation practices still fall far short of providing a convincing, comprehensive validity argument" (p.33). The implementation of validation studies raises challenging issues relating to, for example, the quantity, relevance and necessity of validation evidence collected; the frequency with which validity evidence should be collected; and the defensibility of the evidence collected.

The revised framework for the argument of assessment validation presented here provides a methodical template for validating a wide range of examinations and for demonstrating evidence of the validity of test tasks. Formulating an argument for validity can support the design and operation of examinations. A framework that demonstrates adequate evidence in support of the claims relating to the usefulness of an assessment for its intended purpose (and the reasoning underpinning the claims) provides a systematic, transparent and defensible mechanism for confronting those who wish to challenge or refute such claims. We hope that the case study presented here contributes towards coherent, useful and meaningful guidance to those wishing to undertake their own validation studies.

References

Ahmed, A. & Pollitt, A. (2007). Improving the quality of contextualized questions: an experimental investigation of focus. *Assessment in Education: Principles, Policy and Practice*, **14**, 2, 201–232.

Ahmed, A. & Pollitt, A. (2008a). *It's judging the evidence that counts*. A presentation at the Conference of the Chartered Institute of Educational Assessors, London, May 2008.

Ahmed, A. & Pollitt, A. (2008b). *Evaluating the evidence in assessment*. A paper presented at the International Association for Educational Assessment Annual Conference, Cambridge, UK, September 2008.

Almond, R.G., Steinberg, L.S., & Mislevy, R.J. (2002). A four-process architecture for assessment delivery, with connections to assessment design. Online article at http://www.education.umd.edu/EDMS/mislevy/papers/ProcessDesign.pdf.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA & NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, **37**, 1–15.

Bennett, R.E., Jenkins, F., Persky, H., & Weiss, A. (2003). Assessing complex problem solving performances. *Assessment in Education: Principles, Policy and Practice*, **10**, 3, 347–359.

Blair, J.A. & Johnson, R.H. (1987). Argumentation as dialectical. *Argumentation*, **1**, 41–56.

Brennan, R.L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issue and Practice*, **17**, 1, 5–9.

Cambridge Assessment. (2009). *The Cambridge Approach: Principles for designing, administering and evaluating assessment*. Cambridge: A Cambridge Assessment Publication.

Chapelle, C.A., Enright, M.K. & Jamieson, J.M. (2004). *Issues in developing a TOEFL validity argument*, draft paper presented at LTRC, March 2004.

Chapelle, C.A., Enright, M.K. & Jamieson, J.M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Abingdon: Routledge.

Chapelle, C.A., Enright, M.K. & Jamieson, J.M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, **29**, 1, 3–13.

Cizek, G.J. (2011). *Reconceptualizing validity and the place of consequences*. Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA. April 7–11.

Cizek, G.J., Rosenberg, S.L. & Koons, H.H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement*, **68**, 3, 397–412.

Crisp, V., Sweiry, E., Ahmed, A. & Pollitt, A. (2008). Tales of the expected: the influence of students' expectations on question validity and implications for writing exam questions. *Educational Research*, **50**, 1, 95–115.

Crocker, L. (2003). Teaching for the test: Validity, fairness, and moral action. *Educational Measurement: Issue and Practice*, **22**, 3, 5–11.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich.

Cronbach, L.J. (1971). Test validation. In: R.L.Thorndike (Ed.), *Educational measurement*, (2nd edition). Washington D.C.: American Council on Education.

Cronbach, L.J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Cronbach, L.J. (1988). Five perspectives on validity argument. In: H. Wainer & H. Braun (Eds.), *Test Validity*. 3–17. Hillsdale, NJ: Lawrence Erlbaum.

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

Crooks, T.J., Kane, M.T. & Cohen, A.S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy and Practice*, **3**, 3, 265–285.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). Dictionary of Language Testing. *Studies in Language Testing*, Volume 7, University of Cambridge Local Examination Syndicate/Cambridge University Press.

Delis, D.C., Jacobson, M., Bondi, M.W., Hamilton, J.M., & Salmon, D.P. (2003). The myth of testing construct validity using factor analysis or correlations with normal or mixed clinical populations: lessons from memory assessment. *Journal of the International Neuropsychological Society*, **9**, 936–946.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, **93**, 179–197.

Feldt, L.S. & Brennan, R.L. (1989). Reliability. In: R. L. Linn (Ed), *Educational Measurement*, 3rd edition. American Council on Education, Macmillan.

Fowles, D.E. (1974). CSE: two research studies. *Schools Council Examinations Bulletin 28*. London: Evans/Methuen.

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

Guion, R.M. (1978). 'Content validity' in moderation. *Personnel Psychology*, **31**, 205–213.

Haertel, E. H. & Lorie, W. A. (2004). Validating standards-based test score interpretations. *Measurement: Interdisciplinary Research and Perspectives*, **2**, 2, 61–103.

Hogan, T.P. & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement*, **64**, 4, 802–812.

House, E.R. (1977). *The logic of evaluation argument*. Los Angeles: Center for the Study of Evaluation.

Hughes, A., Porter, D. & Weir, C.J. (1988). *Validating the ELTS test: a critical review*. Cambridge: The British Council and the University of Cambridge Local Examination Syndicate.

Hughes, S., Pollitt, A., & Ahmed, A. (1998). *The development of a tool for gauging the demands of GCSE and A Level exam questions*. A paper presented at the British Educational Research Association Annual Conference, Belfast.

Jonson, J.L. & Plake, B.S. (1998). A historical comparison of validity standards and validity practices. *Educational and Psychological Measurement*, **58**, 5, 736–753.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin*, **112**, 527–535.

Kane, M.T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, **64**, 3, 425–461.

Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, **38**, 4, 319–342.

Kane, M. T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives*, **2**, 3, 135–170.

Kane, M.T. (2006). Validation. In: R.L. Brennan (Ed.), *Educational Measurement* (4th ed.). Westport: American Council on Education/Praeger.

Kane, M.T. (2009). Validating the interpretations and uses of test scores. In: R.W. Lissitz (Ed.), *The concept of validity*. Charlotte, NC: Information Age Publishing.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, **18**, 2, 5–17.

Khalifa, H. & Weir, C.J. (2009). Examining Reading: Research and practice in assessing second language reading. *Studies in Language Testing*, Volume 29. Cambridge: University of Cambridge Local Examination Syndicate/Cambridge University Press.

Linacre, J.M. (1989). *Many-faceted Rasch Measurement*. Chicago: MESA Press.

Linacre, J.M. (2005). A User's Guide to FACETS Rasch-Model Computer Programs. Available at: www.winsteps.com

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

Lissitz, R.W. (2009) (Ed.). *The concept of validity: Revisions, new directions, and applications*. USA: Information Age Publishing.

Lissitz, R.W. & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, **36**, 8, 437–448.

Loevinger, J. (1957). "Objective tests as instruments of psychological theory", *Psychological Reports*, **3**, 635–694.

Lunz, M.E., & Wright, B.D. (1997). Latent Trait Models for Performance Examinations. In: Rost & Langeheine (Eds), *Applications of Latent Trait and Latent Class Models in the Social Sciences*. http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/ltlc.htm

Messick, S. (1975). The standard problem: meaning and values in measurement and evaluation. *American Psychologist*, **30**, 955–966.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, **21**, 3, 215–237.

Messick, S. (1989). Validity. In: R. Linn (Ed.) *Educational Measurement*. 13–103. New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, **23**, 2, 13–23.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, **50**, 741–749.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, **45**, 35–44.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika*, **59**, 439–483.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, **1**, 3–67.

Moss, P. A. (2007). Reconstructing Validity. *Educational Researcher*, **36**, 8, 470–476.

Pollitt, A. & Ahmed, A. (1999). *A New Model of the Question Answering Process*. A paper presented at the International Association for Educational Assessment Annual Conference, Bled, Slovenia, May 1999.

Pollitt, A., Hutchinson, C., Entwistle, N. & de Luca, C. (1985). *What makes examination questions difficult?* Edinburgh: Scottish Academic Press.

Shaw, S., Crisp, V. & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy and Practice*, **19**, 2, 159–176.

Shaw, S.D. & Weir, C.J. (2007). Examining Writing: Research and Practice in assessing second language writing, *Studies in Language Testing*, Volume 26, Cambridge: University of Cambridge Local Examination Syndicate/Cambridge University Press.

Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education*, **19**, 405–450.

Sireci, S.G. (2007). On validity theory and test validation. *Educational Researcher*, **36**, 8, 477–481.

Sireci, S.G. (2009). Packing and unpacking sources of validity evidence: history repeats itself again. In: R.W. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions, and Applications*. 19–37. USA: Information Age Publishing.

Tenopyr, M.L. (1977). Content-construct confusion. *Personnel Psychology*, **30**, 47–54.

Toulmin, S.E. (1958/2003). *The uses of argument*. Cambridge: Cambridge University Press.

Toulmin, S.E., Rieke, R., & Janki, A. (1984). *An Introduction to reasoning*. New York: Macmillan.

Verheij, B. (2005) Evaluating arguments based on Toulmin's scheme. *Argumentation*, **19**, 347–371.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *Cadmo*, **18**, 1, 63–82.

Wright, B. & Linacre, J. (1994). Reasonable Mean-square Fit Values. *Rasch Measurement Transactions*, **8**, 3, 370.

Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis*. Chicago: Mesa Press.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In: C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics*. 45–79. Amsterdam, Netherlands: Elsevier Science.

Afterword

It has been eight years since the publication of this special issue exemplifying 'An approach to validation' (and closer to ten years since the work it describes was conducted). Validation studies continue to be demanding activities, not helped by considerable variety in views about what validation should involve, what it can achieve and whom it should serve (Newton & Shaw, 2016). One thing is clear, however. There is an increasing demand for awarding bodies to demonstrate the quality of their qualifications and meeting this demand is no mean feat.

Our main motivation for publishing this work was to provide a practical example for would-be validators by describing the framework (based on Kane, 2006) and methods that we applied in a validation study of International A level Physics. More detailed description of some elements of the study can be found in Crisp and Shaw (2012). An earlier pilot validation study is described in Shaw and Crisp (2010a; 2010b), whilst use of the literature to develop the framework is described in Shaw, Crisp and Johnson (2012).

Since the study on A level Physics, researchers at Cambridge Assessment have conducted validation studies for a variety of other qualifications (e.g. IGCSEs and International A levels). Some elements of these studies have been reported in various publications and at conference (e.g. Greatorex & Shaw, 2012; Greatorex et al., 2013a, 2013b). Over time, we have made some adjustments to the set of methods used, jettisoning a small number of validation methods that were resource-intensive but provided minimal additional validity evidence, and adjusting or extending others – for example, to gather validity evidence about speaking assessments. Where methods have been changed, care has been taken to ensure that the revised set of methods still provides evidence in relation to each validation question.

Five years after implementing the original validation framework, a number of issues emerged which prompted us to review and revise the framework, as described in Shaw and Crisp (2015). We believe that these changes have strengthened the theoretical structure underpinning the framework. The changes were made to validation questions 4 and 5.

Validation question 4 relates to the *Extrapolation* inference and was previously phrased in terms of whether the constructs sampled are representative of competence in the wider subject domain. We broadened the question to include related competence beyond the subject. In the revised framework (Shaw and Crisp, 2015) it appears as:

> *Do the constructs sampled give an indication of broader competence within and beyond the subject?*

The *Decision-making* inference was revised to better reflect current thinking (e.g. Kane, 2013). Appropriate decisions can only be made if the meaning of test scores is clearly interpretable by a range of relevant, credible stakeholders. However, the previous wording of the validation question for this inference focused too much on providing guidance to stakeholders on the meaning and uses of results, and not enough on whether scores and grades indicate students' potential. Validation question 5 appears in the revised framework as:

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

Since the revision of the framework, a number of new methods have been explored in order to address the changes. For example, we have used the size of the correlation between student results in a specific IGCSE and later performance in AS or A levels to provide evidence relating to validation question 5 for that IGCSE.

As described in the 'Conclusions' section of the Special Issue, a key challenge with validation work is the breadth and depth of evidence needed when conducting a study of the kind described. Thus, we alluded to how it might be appropriate to implement the full validation approach to a small number of qualifications, and to apply a more streamlined, *operational* approach to validation to some further qualifications. This was discussed in Shaw and Crisp (2011) and has since been implemented for a range of qualifications. The *operational* approach uses the same validation framework but uses only existing, operationally-available data such as marking (scoring) data and documentary evidence (as opposed to data generated through experimental work). This *operational* approach may not be able to address each of the validation questions as robustly as the original *experimental* approach. However, conducting some studies of each type allows awarding bodies to provide validation evidence for a wider range of qualifications.

Following on from the development of the *operational* approach, a hybrid of the two approaches (operational and experimental) has since been trialled. This involved routinely available evidence plus gathering some new data using a small number of methods from the experimental approach. Relevant stakeholders selected those methods of particular interest to their assessment context. Whilst a hybrid approach will not be as substantive as a full experimental study, it may nevertheless yield targeted validity information in a more time and cost effective way. Balancing the robustness of evidence against the resources involved in its collection continues to be an ongoing debate in the implementation of validation studies.

Given that the quality of qualifications needs to be ensured, we would still argue that "the challenge of validation – no matter how great, should not impede its continuing execution" (Shaw & Crisp, 2015, p.36).