An approach to understanding validation arguments

October 2017

Ofqual/17/6293

Contents

Author

This report was written by Paul E. Newton, from Ofqual's Strategy, Risk and Research directorate.

Acknowledgements

Too many people have contributed to the thinking that is synthesised in this report for me to be able to thank them all individually; but their support is gratefully acknowledged all the same. Particular thanks to those who originally encouraged me to develop this line of work and who have provided very helpful feedback along the way, including Glenys Stacey, Sally Collier and Michelle Meadows, and to many other of my colleagues at Ofqual. I am very grateful to members of Ofqual's various expert groups for their incisive questions, suggestions, and recommendations; including members of the Standards Advisory Group, Vocational Advisory Group, Research Advisory Group, and the Ofqual Board.

For granting permission to reproduce figures, I am very grateful to:

- Michael and Cynthia Eraut (Figure 1)
- Robert Mislevy and copyright owner CRESST, the National Center for Research on Evaluation, Standards, and Student Testing, UCLA (Figure 4)
- Stuart Shaw, Vicki Crisp and copyright owner Cambridge Assessment (Figure 5).

Preface

The approach to understanding validation arguments that is set out in the present report represents a collection of ideas that I have been working on since joining Ofqual in October 2014. In July 2014, the (former) Chief Regulator had announced Ofqual's intention to change its approach to regulating vocational qualifications, by 'tearing up' the old 'rule books' and by putting validity at the heart of what we do (Stacey, 2014). Bearing in mind that I had spent the previous few years immersed in the literature on validity and validation (eg Newton, 2012; Newton and Shaw, 2014), it made sense for me to commit time to helping Ofqual to develop a clear, comprehensive and consistent account of these elusive concepts; to provide a technical point of reference for subsequent discussions with awarding organisations concerning our new approach to regulation. The present report provides a synthesis of that work.

This report is more of a scholarly exposition than a conventional regulatory document. It attempts to explain, at a high level, the criteria according to which qualifications (and educational assessments more generally) are designed, developed, delivered and reviewed. Its principal organising concepts are:

- **validity** – the fundamental technical criterion for evaluating qualifications; and
- **validation argument** – an approach to structuring the evaluation of any particular qualification, or group of qualifications.

Although the first two of Ofqual's five statutory objectives do not mention the term explicitly, they are essentially all about validity.[1] Similarly, the majority of the conditions that we require organisations to comply with, for them to continue being recognised to award qualifications in England, can also be traced back to this core concept.[2] It is therefore absolutely right that validity is at the heart of what we do, and at the heart of what any assessment organisation does, including all of the organisations that we recognise. Rather than setting out regulatory requirements or expectations, the present report attempts to explain what it means to put validity at the heart of what we do and what might be involved in being able to demonstrate this. It aims to help practitioners to grapple with the core concepts of validity and validation argument, and thereby to help them to appreciate more fully what might be involved in planning a validation research programme.

Finally, as a scholarly exposition, some of the ideas and terms in this report are new. Over the past few years, they have been refined in conversation with various of

[1] The qualifications standards objective and the assessment standards objective.
[2] Our *General Conditions of Recognition* (see https://www.gov.uk/government/publications/general-conditions-of-recognition, accessed 27/07/2017).

Ofqual's expert groups, and through the process of drafting successive versions of journal submissions (eg Newton, 2016; Newton, 2017a; Newton, 2017b). Yet, some may still need to be refined further and some may not stand the test of time. So, this report is very much 'work in progress' and any feedback would be very welcome.

Dr Paul E. Newton
Research Chair, Ofqual
paul.newton@ofqual.gov.uk

Introduction

To express Ofqual's statutory objectives in a nutshell: we regulate so that qualifications are sufficiently valid and are trusted. Validity is at the heart of what we do; and the same is true for the organisations whose qualifications we regulate, which have direct responsibility for designing, developing, delivering and reviewing qualifications. But what is validity? And how is it possible to judge whether a qualification has sufficient validity? Unfortunately, there are no definitive answers to either of these questions. On the one hand, it can be tricky to pin validity down, because people often mean quite different things when using the term. On the other hand, validation – the work of investigating validity – is not a precise science. Subtle arguments, based upon many different sources of empirical evidence and logical analysis, are required in order to conclude that a qualification has sufficient validity. This report is intended to help readers to understand what is meant by, and what might be involved in, constructing a validation argument of this sort.

Importantly, this is not a manual for constructing validation arguments. It is an introductory overview, written to explain, at a fairly high level, the principles and practices of validation argument. It is important to develop a solid understanding of these principles and practices because no two validation arguments are likely to be exactly the same. This is partly because different qualifications have different purposes; and, for each purpose, a slightly different argument will need to be constructed. But it is also because no validation argument will ever be as complete as it possibly could be: it will always be possible to gather additional sources of evidence and analysis; and the particular combination of evidence and analysis relied upon for a particular qualification will depend upon all sorts of considerations. In other words, validation argument is not clerical exercise, involving little more than box-ticking. It is a professional exercise, involving insight, judgement and understanding.

When Ofqual says that it regulates so that qualifications are sufficiently valid and trusted, we do not mean to exclude other forms of large-scale educational assessment that are not traditionally described as 'qualifications', eg national tests administered at the end of primary schooling. In our strapline, and in the present report, the term 'qualification' is used generically, to include any large-scale educational assessment that involves implementing a specified assessment procedure – typically the same procedure from one session to the next – in order to deliver accurate and useful assessment results. Informally, we tend to refer to the validity of a particular qualification. More formally, we tend to refer to the validity of the assessment procedure that operationalises the qualification. Just as we will use the term 'qualification' generically we will also use the term 'candidate' generically to refer to those who are assessed.

Although it is true that people use the word 'validity' to mean different things, many assessment professionals would agree that it boils down to something like: assessing the right thing, in the right way, to provide accurate and useful assessment results. At its core, then, is the idea of educational measurement: quantifying people in terms of their level of (a certain kind of) proficiency. Any qualification that delivers a summative assessment result – a result that summarises attainment in an overall score, level or grade – supports measurement of this kind: whether that involves ranking candidates in terms of their level of proficiency, which is true of many school examinations; or whether that involves classifying candidates in terms of whether or not they are sufficiently proficient, which is true of many vocational and occupational qualifications. Fundamentally, then, validity is concerned with measurement quality, ie how well a particular proficiency is measured through a particular qualification.

Already, we have encountered quite a few technical concepts, including purpose, measurement, validity, and assessment procedure. The following sections will shed light on these and other technical concepts before confronting the issue of validation argument directly.

Purpose

What is the purpose of a qualification? In fact, there are all sorts of qualification purposes. When considering validation arguments, it is helpful to begin by distinguishing between three major kinds, ie the purpose is:

1. to **measure;**
2. to make **decisions; and**
3. to achieve **impacts**.

These distinctions can be illustrated, in turn, using the example of a national test in reading comprehension, designed to satisfy each of the following purposes:

1. to rank and classify pupils in terms of their level of attainment in reading comprehension at the end of primary school;
2. to enable secondary school teachers to decide whether incoming pupils have mastered the primary curriculum in sufficient depth to be allowed to begin working on the secondary curriculum (or else to place them in a catch-up teaching group); and
3. to ensure that primary school teachers align their reading instruction with the national curriculum for reading.

Not only is it true that any particular qualification might be intended to satisfy different kinds of purpose, it is also true that within each of these kinds the qualification might be intended to satisfy more than one sub-purpose. For instance, when reading test results are aggregated to the level of a school, they are often combined with other test results and with school inspection judgements in order, such as:

1. to measure the educational effectiveness of the school;
2. to decide whether the school should be put in special measures or closed; and
3. to motivate less effective schools to become more effective.

In each of these cases, fitness-for-purpose can be investigated by following the same three steps:

- specifying a critical claim;
- constructing an argument to support that claim; and
- evaluating the strength of that argument (and modifying the claim, if necessary).

Naturally, each kind of purpose involves a different kind of critical claim, for instance:

1. it is possible to measure accurately by using our assessment results – the measurement claim;
2. it is possible to make (more) accurate decisions by using our assessment results (than if they were not used) – the decision-making claim; and
3. it is possible achieve positive impacts by implementing our assessment policy – the impact claim.

Generally speaking, each kind of claim will require a different kind of supporting argument: a measurement argument; a decision-making argument; or an impact argument. Often, though, an impact argument will subsume a decision-making argument; and a decision-making argument will subsume a measurement argument. For instance, Lorrie Shepard (2012) developed an impact argument to theorise the use of test results to improve national educational effectiveness, based upon the following proposition from Eric Hanushek (2011): if the bottom 7-12% of teachers are fired and replaced with average teachers, then over a 13-year period the USA would end up attaining at the level of Finland. Her argument consisted of the following sub-claims:

1. student achievement is the key value or goal of schooling, and constructing teacher evaluation systems around student growth will focus attention on this valued outcome.
2. student achievement is accurately and authentically measured by the assessment instruments in use.
3. teacher contributions to growth are accurately quantified by Value-Added Modelling (VAM).
4. the poorest teachers can be eliminated on the basis of VAM results and sufficient numbers of teachers with average student growth are available to replace those who are fired.
5. improved instruction and higher levels of achievement will result.
6. unfortunate unintended consequences are minimal.

Notice that sub-claim 2 involves measurement claims; concerning the measurement of maths and reading. Notice also that sub-claim 3 is a decision-making claim; concerning the decision over which teachers to fire, ie the poorest 7-12%, as determined from the VAM analyses. Sub-claim 5 represents the key impact claim; with sub-claim 6 as an important caveat.

The purpose of this impact argument is to unpack the logic of the mechanism by which a particular assessment policy is assumed to achieve its ultimate goal; so that it can be scrutinised thoroughly. Each of the sub-claims within the impact argument will need to hold true for it to be considered strong. In fact, each sub-claim can be broken down into its own constituent argument and sub-claims. For instance: sub-claim 2 can be broken down into separate measurement arguments for reading and maths; while sub-claim 3 can be broken down into a decision-making argument for firing the poorest teachers on the basis of VAM analyses. Notice how this decision-making argument subsumes measurement arguments, just as the overarching impact argument subsumes both.

The reason why we have begun this report with what might seem like a slightly laboured deconstruction of fitness-for-purpose is because much of the confusion in the literature on validity and validation can be attributed to a failure to distinguish clearly and consistently between these three different, albeit interrelated, evaluation objectives. This is most evident in the debate over how best to use the word ‘validity’ whereby: some scholars argue that validity ought to be treated as a measurement concept (and *purely* as a measurement concept); while others argue that it ought to be treated as a decision-making concept (which subsumes measurement); while yet others argue that it ought to be treated as an impact concept (which subsumes both decision-making and measurement).

Why does this debate matter? It matters because it is important that awarding organisations are clear what they mean when they claim ‘sufficient validity’ for any of their qualifications. Does it mean that the qualification supports good measurement, or good decision-making, or good impacts, or any combination of these? This is a critical question because it is quite possible for a qualification to support good measurement but not necessarily good decision-making; especially when there is a significant ‘gap’ between the proficiency measured by the qualification and the use to which qualification results are put.[3] For example, imagine deciding to hire a maths teacher purely on the basis of their distinction grade in a computing qualification. The qualification might enable us to measure computing proficiency very accurately. But it is hard to see how it could be said to enable us to measure aptitude for teaching maths with the same degree of accuracy. Indeed, the organisation responsible for the computing qualification might go so far as to insist that this would represent a **misinterpretation** and therefore a **misuse** of results. This is part of the reason why awarding organisations are sometimes resistant to talking about validity as though it were *essentially* a decision-making concept; because it then becomes a matter of how results are used, over which they have less than complete control. An alternative, and even more extreme, perspective is that the ultimate purpose of any qualification is to make a (specified) decision and that this will also have (intended) impacts; so the concept of validity *ultimately* resides at the level of decisions and/or

[3] Equally, it is quite possible for a qualification to support good decision-making but also to have bad consequences.

impacts. To be fair, there are pros and cons to all sides of this debate (see Newton and Baird, 2016).

As will become clear in subsequent sections, the present report discusses validity as though it were *fundamentally* a measurement concept, tantamount to measurement quality. This is not to trivialise the fact that qualifications are designed to support decision-making and to have impacts. Indeed the way in which we explicate the validity concept very clearly links it to real-world decisions and impacts. However, it is to acknowledge that there are different objectives when evaluating different kinds of qualification purpose, that these are logically separable, and that the approach to evaluating measurement quality is foundational. The focus of the present report, *An Approach to Understanding Validation Arguments*, is therefore upon understanding measurement arguments.

Measurement

Qualifications judge individuals in terms of their level of (a certain kind of) proficiency:

- either in terms of ‘more than’ or ‘less than’ – where the comparison is relative to other people (eg a percentile rank)
- or in terms of ‘enough’ versus ‘not enough’ – where the comparison is against an absolute standard (eg a passing grade).

In fact, most qualifications fall somewhere between these extremes, showing elements of both. However, the point is that, in all cases, the summative judgement is quantitative, in the sense of referring to an **amount** of proficiency. When we refer to ‘measurement’ in the present report, this is the sense in which we mean it; the fairly weak sense of ‘more than/less than’ or ‘enough/not enough’.

Some people object to the idea of educational measurement on the basis that the target proficiencies in question – the things that we need to measure in educational settings – are just too complex and nuanced to be reduced to a simple quantity. This is an important possibility. However, since qualification results *do* reduce those proficiencies to a simple quantity, ie to weakly-defined measurements, it makes sense for them to be evaluated *as though they were* weakly-defined measurements. If the corresponding measurement argument simply cannot be supported on the basis of evidence and analysis, then, clearly, we should give up on the idea of educational measurement, and on the idea of simple quantitative summaries of proficiency. Conversely, the stronger the measurement argument we are able to construct, the more plausible the measurement hypothesis becomes.

Other people object to the idea of educational measurement on the basis that the target proficiencies in question – the things that we need to measure in educational settings – bear no resemblance to the kind of things that ‘real scientists’ measure, ie physical properties like length. To measure the length of an object, you first define your unit, eg centimetre, and then count the number of units into which the object can be divided. In other words, the measurement scale is no more than the sum of its units. This is clearly not true in the case of educational measurement because there are no natural units. Furthermore, the difference between one end of a proficiency scale and the other is not simply quantitative but also qualitative. An expert juggler who can juggle 10 balls is not simply (or even) five times better than a novice juggler who can only juggle two balls. The expert’s knowledge, skill and understanding of how to juggle balls is of a different level of sophistication to the novice’s, but also of a different kind. It is both qualitatively and quantitatively different. Yet, that is no reason to prevent us from judging the two in purely quantitative terms, ie to conclude that the expert has a far higher level of juggling proficiency than the former. Nor is it a reason to prevent us from making real-world decisions on the basis of this classification, eg which of the two to bet on during a juggling competition. The more general point is that metrologists – real-world measurement scientists – get stuck into measuring all sorts of unwieldy phenomena; from hardness, to windspeed, to baldness. This is regardless of whether those phenomena can be said to have natural units.

The question is not so much *whether* it is appropriate to measure in education, but *how*, ie what kind of **measurement model** is most appropriate for us to think in terms of. Models of growth, or decay, may have particular utility, here. A useful analogy might be O'Tar Norwood's description of standards for classification of male pattern baldness (Norwood, 1975). Norwood identified seven points on scale of baldness – from not at all bald to completely bald – with the third point representing the minimal extent of hair loss considered sufficient to constitute baldness. Quantifying baldness is not simply a matter of counting redundant follicles. It is a matter of pattern matching, to identify which particular hair-loss-stage the person in question has reached. This is all the more apparent when it is appreciated that baldness is quantified differently for women versus men, ie against quite different baldness pattern scales. This kind of pattern matching approach resonates strongly with Eraut's summary of the Dreyfus model of progression (see Eraut, 2008, p.3), which seems particularly pertinent to vocational and occupational qualifications (see Figure 1).

Again, the question is not so much *whether* it is appropriate to measure in education, but *how*, ie what kind of measurement model is most appropriate for us to adopt. Assessment designers are very familiar with specifying **proficiency constructs**, ie the elements of knowledge, skill and understanding into which a target proficiency (the 'thing' that needs to be measured) can be decomposed. The idea of growth or progression models reminds us that it is equally important, from a measurement perspective, to specify **proficiency scales**, ie the features that characterise having different levels of a particular proficiency.

Figure 1. Eraut’s summary of the Dreyfus progression model (Eraut, 2008)[4]

Level 1 **Novice**
Rigid adherence to taught rules or plans
Little situational perception
No discretionary judgement

Level 2 **Advanced Beginner**
Guidelines for action on attributes or aspects (aspects are global characteristics of situations recognisable only after some prior experience)
Situational perception still limited
All attributes and aspects are treated separately and given equal importance

Level 3 **Competent**
Coping with crowdedness
Now sees actions at least partially in terms of longer-term goals
Conscious deliberate planning
Standardised and routinised procedures

Level 4 **Proficient**
See situations holistically rather in terms of aspects
See what is most important in a situation
Perceives deviations from the normal pattern
Decision-making less laboured
Uses maxims for guidance, whose meaning varies according to the situation

Level 5 **Expert**
No longer relies on rules, guidelines or maxims
Intuitive grasp of situations based on deep tacit understanding
Analytic approaches used only in novel situations, when problems occur or when justifying conclusions
Vision of what is possible

[4] Incidentally, the present report often refers to ‘the proficiency’ when describing the thing that a qualification needs to measure, where other reports might refer to ‘the construct’ or ‘the attribute’. In Figure 1, the term ‘proficient’ is used differently, to refer to a particular point on the proficiency scale.

Validity

> If one were to select a sample of psychometricians from each of the last five to ten decades and gather them together in, say, a bar, it is quite likely that all would drink a toast to validity as the paramount concept in the field of testing. However, a mêlée would ensue if they were asked to define what validity *is*.

(Fast and Hebbler, 2004, p.11)

This quotation sums up the problem with validity very succinctly. Everyone agrees that it is the most important concept in the field of educational assessment. However, it is impossible to formulate a definition of validity that will satisfy everyone who works in this field. The reasons for this lack of consensus are tricky to pinpoint. It is a complex debate that does not yet seem to be close to resolution (see Newton and Baird, 2016).

Having said that, one particular definition of validity is widely regarded, and is the closest there is to a consensus definition. It is located in the validity chapter of the *Standards for Educational and Psychological Testing*, which is a consensual statement of the North American measurement professions, now in its sixth edition. The definition goes like this:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.

(AERA, APA and NCME, 2014, p.11)

There are all sorts of reasons why this particular version has become the preferred definition in North America. Part of the explanation is the recognition that test results can be used for multiple purposes and, as we have already seen, a test fit for one purpose may not be fit for another. As such, the idea that a *test instrument* can be declared either valid or invalid is inappropriate. Instead, since results from a single test can be interpreted in different ways for different purposes, it is more appropriate to refer to the validity of a particular *interpretation of test results*; at least, so proponents of this consensus definition claim. Consider the earlier example of a distinction grade in a computing qualification. If the qualification had been effectively designed, and if its assessment procedure had been correctly implemented, then the interpretation ‘high attainment in computing’ should have high validity; whereas the interpretation ‘high aptitude for teaching maths’ would (presumably) not.

The downside of the North American consensus definition is that it is quite nebulous, especially in one very important sense. Particularly when this definition is interpreted alongside the rest of the content of the chapter in which it appears, it can be interpreted in a variety of different ways: either to imply that validity is fundamentally a measurement concept; or to imply that validity is essentially a decision-making concept; or to imply that validity is ultimately an impact concept. Indeed, this nebulousness may well be another part of the reason why this particular version has become the preferred formulation in North America. In other words, it allows people to continue to mean quite different things, whilst upholding the centrality of the concept.

There is no single, correct way to define validity. There will be pros and cons of any particular formulation. We, at Ofqual, recently decided to use a version that is slightly more explicit than the North American definition, adopting the stance that validity is fundamentally a measurement concept. This was partly to foreground the technical aspects of qualification design, development, delivery, and review. But it was also partly in recognition of the fact that our statutory objectives are framed primarily in terms of promoting and ensuring those technical aspects, and that our regulatory oversight does not extend to *all* aspects of qualification impact. In other words, our decision on how to explicate validity was at least partly pragmatic.

As noted earlier, we acknowledge that validity is a matter of assessing the right thing, in the right way, to provide accurate and useful assessment results. However, to formalise this idea, and to turn it into a technical point of reference for subsequent discussion, we opted for the following formulation:

> The validity of a particular qualification is the degree to which it is possible to measure whatever that qualification needs to measure by implementing its assessment procedure.

This formulation incorporates three critical aspects. First, it foregrounds the idea that validity is fundamentally a measurement concept, and that a qualification is judged first and foremost in terms of its potential to support accurate measurement. This idea of *potential* to support accurate measurement is important; partly for theoretical reasons and partly for practical ones. Theoretically, it helps to remind us that there is a useful distinction to be drawn between the accuracy of measurement interpretations and the validity of a measuring procedure. A qualification can legitimately be described as having high validity – the potential to support accurate measurement – even though a substantial proportion of candidates will receive inaccurate results during each delivery phase, because measurement inaccuracy can never be eliminated entirely. Practically, it helps to remind us that measurements, ie qualification result interpretations, occur in the real world and are only partially under the control of the organisation responsible for designing, developing and delivering the qualification. That qualification must have the *potential* to support accurate measurement – and it is the responsibility of the awarding organisation to ensure that it does – but measurement interpretations are actually drawn by those who use qualification results in the real world, who may fail to interpret results accurately even when attempting to use them for the purpose for which they were designed.

Second, the Ofqual formulation foregrounds the importance of each and every qualification measuring the thing that it *needs* to measure. This, of course, begs the question of how to determine what any particular qualification needs to measure, which turns out to be a very tricky question to answer. As noted above, the North American literature tends to emphasise the decision that will need to be taken on the basis of test results. This works very well for many occupational qualifications. For instance, if the purpose of a qualification is to earn a licence-to-practise in plumbing, then it needs to measure the learning outcomes required for safe and competent plumbing. Critically, if what it means to be a safe and competent plumber changes over time, in response to changes in the nature of plumbing or in the nature of societal expectations of plumbers, then the target proficiency for that qualification – the proficiency that the qualification needs to measure – will need to be reconstructed accordingly. By focusing on what a qualification needs to measure, we ensure that measurement interpretations will not only be accurate but also useful. Nowadays, for many general qualifications, eg school-leaving examinations, results tend to be used for a multiplicity of purposes, making it much harder to specify the target proficiency with clarity. In such cases, the specification of what needs to be measured will typically be driven by a curriculum statement or qualification syllabus. However, the specification process will also need to bear in mind a variety of additional considerations, stemming from alternative perspectives on qualification purposes (see Newton, 2017a).

Third, the Ofqual formulation foregrounds the centrality of an assessment procedure to judgements of validity. It is useful shorthand to refer to the validity of a qualification. Yet is more helpful, from a technical perspective, to think in terms of the validity of its assessment procedure. The next section explains in more detail what we mean by an assessment procedure; but, in a nutshell, it means everything that an awarding organisation puts in place (ie standardises) for a particular qualification, to ensure that measurement interpretations will be as accurate and useful as possible. By focusing on the validity of an assessment procedure, the implication is that validity resides neither in the measuring instrument (alone) nor in the measurement interpretation (alone), but in everything that is put in place to ensure the potential for accurate and useful measurement.[5]

[5] Although those who use results from a particular assessment procedure may attempt to interpret them in different ways, the procedure will typically have been designed to support a very specific measurement interpretation, ie more or less of the target proficiency that has been specified. Indeed, it is helpful to think of this intended interpretation as a critical element of the assessment procedure. Validity is therefore judged primarily in relation to this intended measurement interpretation. If a different kind of interpretation were to be drawn from results, then a separate validation argument would be required, judged in relation to a differently specified target proficiency; and the assessment procedure would inevitably have somewhat less validity in relation to this new target proficiency, not having been specifically designed to support it. In essentially the same way, it is helpful to think of the specification of the target population of candidates (and of the broader context within which they will be measured) as part of the assessment procedure. The judgement of validity is therefore relative to a specified population in a specified context. If the validity claim needed to be extended to other candidates in other contexts, then additional evidence and analysis would be required.

An assessment procedure

A qualification, like any other large-scale educational assessment, is operationalised through an assessment procedure: the (general) procedure through which (particular) measurement interpretations are generated. The idea of a procedure is that certain features and processes are standardised, ie held constant, each time results are delivered for interpretation and subsequent use. In fact, we can think of an assessment procedure as the **set of specifications** that govern the entire activity of measuring, which makes explicit the features and processes that ought not to change from one occasion to the next.

Although qualifications vary widely in the number and kind of features and processes that are standardised, specifications typically govern things like:

- the nature of the proficiency that needs to be measured (in order to satisfy specified purposes);
- the nature of the candidate population (and the broader context within which candidate proficiency needs to be measured);
- the processes involved in developing and administering tasks to elicit evidence of proficiency;
- the processes involved in evaluating evidence of proficiency from task performances;
- the processes involved in combining and transforming performance evaluations into measurement results; and
- the ways in which those results should (and should not) be interpreted.

Awarding organisations exercise direct control over their assessment procedures. Sometimes, however, they devolve control of certain elements to centres, eg when they allow schools and colleges to design their own approaches to eliciting assessment evidence. Under these circumstances, the awarding organisation will establish additional processes – as part of its assessment procedure – to ensure that this devolution of control does not unduly compromise validity. For example, it might require each centre to appoint a suitably qualified assessment expert to take overall responsibility for assessment in the centre; perhaps requiring them to submit an assessment strategy for the qualification that sets out the features and processes that the centre will establish for the elements under its control.

Qualification lifecycles

There are different ways of thinking about qualification lifecycles, some more operational and some more conceptual. From an operational perspective, it is helpful to think about the qualification lifecycle in terms of four practical stages:

1. **Design**. The assessment procedure is designed.
2. **Development**. The apparatus for measuring candidates are developed.
3. **Delivery**. A measurement result is delivered for each candidate.
4. **Review**. The assessment procedure is evaluated.

This stage-based model is cyclical in the sense that, following review, insights into how to improve the assessment procedure will feed into re-design, to improve the accuracy and usefulness of results during the next implementation phase. The basic structure, here – design, implement, review – is generic. The distinction between development and delivery is especially useful for qualifications that operate on a largely external basis, ie they devolve very little control over critical assessment elements (eg evidence elicitation, performance evaluation) to centres. For qualifications like these, development and delivery are clearly demarcated stages, and the outputs from the development stage will typically be used for candidates in all centres. The distinction is less clear-cut and therefore perhaps less useful for qualifications that operate on a largely internal basis, ie they devolve a lot of control over critical assessment elements to centres. For qualifications like these, development and delivery are less clearly demarcated stages, and development outputs will differ somewhat across centres and sometimes across candidates within centres.

From a conceptual perspective, it is helpful to think about the qualification lifecycle in terms of five logical steps:

1. **Clarification**. Measurement objectives are clarified.
2. **Elicitation**. Multiple performances are elicited from each candidate (via tasks) to provide a sample of evidence of proficiency.
3. **Evaluation**. Each performance in the sample is evaluated in terms of what it implies about candidate proficiency.
4. **Combination**. The set of performance evaluations, for each candidate, is combined into an overall measurement result.
5. **Interpretation**. The measurement result is interpreted by those for whom it has been provided.

These steps identify what is involved in measuring people on the basis of qualification results. The steps are defined at a high level to ensure their applicability across any kind of large-scale educational assessment. Notice, first, that this step-based model extends beyond what is traditionally viewed as the delivery stage – and beyond the four walls of the awarding organisation – because it is framed in terms of the interpretation of a measurement result by someone for whom it has been provided. This will include the candidate, of course; but it is actually directed more toward a different stakeholder, the user, who uses the result to make a decision. This might, for instance, be an employer, who reads the qualification result from an application form and interprets it to mean one thing or another. This point is an important one: assessment results do not measure people; people measure people, using assessment results. Ultimately, the interpretation of the result *is* the measurement. The role of an awarding organisation is to empower people to measure: by providing them with accurate results; but also by enabling them to interpret those results accurately.

Also notice that this step-based model begins by emphasising the clarification of measurement objectives. This is to ensure that results are not simply accurate but are useful too. The critical challenge, here, is to specify the target proficiency; the thing that needs to be measured.

The step-based lifecycle highlights the fact that there is a series of intermediate outputs (from steps 1 to 4) on the way to the final output (from step 5). These include:

1. a proficiency specification;
2. a set of task performances for each candidate (a performance profile);
3. a set of evaluations for each candidate (an evaluation profile);
4. an overall result for each candidate; and
5. an interpretation of the result for each candidate.

This lifecycle can be thought of as a production line, in which various participants take responsibility for producing a series of outputs. The very first output in this production line – the principal output of the design stage – is a **proficiency specification**. This is a representation of the target proficiency, which is the thing that needs to be measured. As noted earlier, this involves two dimensions:

- representing the proficiency construct (including the elements of knowledge, skill and understanding into which the target proficiency can be decomposed); and
- representing the proficiency scale (the features that characterise having different levels of the target proficiency).

A paper by Edward Haertel (1985), entitled *Construct validity and criterion-referenced testing*, is well-worth reading for its insights into developing proficiency specifications for educational assessments. He used the following description to illustrate a core component of a specification of Functional Literacy for North American high school graduates:

> Functional literacy represents a point along the continuum of reading skill acquisition, typically attained sometime during the middle school or high school years of instruction. It entails sufficient reading skill to comprehend the main ideas in a typical newspaper article, to respond appropriately to the kinds of forms and applications most adults encounter in day-to-day life, to understand operating instructions for unfamiliar household appliances and on the labels of household products, and to read and enjoy a contemporary popular novel. It therefore implies some familiarity with the organizational schemes of these different prose forms, including both narrative and expository writing; a reasonable vocabulary as well as skill in inferring word meanings from context; and some special conventions, for example, ways of representing dates on applications and regulations governing the labeling of supermarket items. The functionally literate high school graduate will be able to summarize orally a newspaper article he or she has read; fill out credit card, job, or license applications, the 1040-E, and so forth; learn to operate appliances by reading the accompanying instructions; make informed choices among newly encountered products at the supermarket; and choose and read popular literature according to his or her tastes. (p.37)

Outputs from steps 2 to 4 occur during the delivery stage. The output from step 2 is a **set of task performances** for each candidate. If, for example, a qualification in Everyday Numeracy is based upon a single test booklet containing 100 questions, then the set of task performances – the candidate's performance profile – would consist of their responses to those questions. For certain questions this might include ticks alongside multiple-choice options; for other questions this might include short or extended written responses. Ultimately, the candidate is responsible for this step, ie for demonstrating her true level of proficiency. Of course, her demonstration will have been scaffolded by a team of facilitators – including process designers, task developers, administrators, etc – with responsibility for manufacturing the conditions that enable her to represent her true level of proficiency in her responses (via tasks, response booklets, administration conditions, and so on).

The output from step 3 is a **set of evaluations** for each candidate. Extending the above example, the set of evaluations would include marks awarded for each of the candidate's 100 responses. Ultimately, it is the responsibility of an assessor to ensure that each evaluation, ie mark, reflects the true quality of each performance. Once again, this task will have been scaffolded by a team of facilitators – including process designers, mark scheme developers, quality assurers, etc – with responsibility for manufacturing the conditions that enable the assessor to represent the true quality of each of the candidate's responses in his marks (via mark schemes, mark capture mechanisms, and so on).

Step 4 constructs an **overall result** for each candidate. This involves combining the discrete evaluations and transforming them into a single outcome. This might be as simple as totalling the marks awarded for each response. Typically, though, raw marks will need to be transformed into a reporting metric. This might require a standard setting panel, to determine a cut-off point between passing and failing, according to which the candidate's raw mark can be classified (as pass or fail). As before, these tasks will have been scaffolded by a team of facilitators – including process designers, various developers, technicians, panellists, etc – with responsibility for manufacturing the conditions that enable the aggregator to represent the true value of the assessor's mark profile in the overall result (via aggregation rules, checklists, databases, standard setting procedures, and so on).

Lastly, the output from step 5 is an interpretation of the result for each candidate, ie a **measurement interpretation**. This happens in the real world, so there may be many interpretations drawn for each candidate depending on how many times the result is used. For instance, if the candidate reports her Everyday Numeracy pass grade in an application for a checkout assistant job, the store manager might interpret this to mean that she has sufficient numeracy skill to be able to deal with the demands of everyday life and work, including sufficient skill for a job that requires accurate counting of money. In exactly the same way as for previous steps, the task of the result user will have been scaffolded by a team of facilitators – including process designers, certificate developers, communications teams, etc – with responsibility for manufacturing the conditions that enable interpreters to represent the true meaning of the overall result in their interpretation of it (via reporting mechanisms, targeted communications, and so on).

Outputs from these same five steps can be illustrated in terms that resonate more strongly with the provision of vocational qualifications in England:

1. in England, the dominant approach to representing target proficiencies for vocational qualifications involves the identification of a set of Learning Outcomes (LOs) within which are nested Assessment Criteria (AC);

2. for any particular qualification, LOs might be assessed in a variety of ways. For instance, certain LOs might be assessed holistically via a written test, whilst ACs for the remaining LOs are assessed individually via work-based observational assessment. The performance profile for a qualification like this would include the subset of responses to the test questions, plus the subset of performances observed at work;

3. extending this example, the evaluation profile would include the subset of marks for the test responses, plus the subset of pass/fail judgements for the observations. As work-based observations tend to be undertaken 'when ready'

the evaluation profile for the subset of work-based judgements is likely only to record passes;

4. the overall result in this example is a passing grade. This is based on a non-compensatory aggregation principle, ie a pass is required on all of the ACs observed at work and a pass is also required on the test. The passing mark for the test might have been determined by a standard setting panel, using the Angoff method; and

5. the passing grade might be interpreted as meaning that the candidate is safe and competent to practise the function stated in the qualification title, eg bricklaying.

The point of teasing apart these steps is to demonstrate that educational measurement is a **representational** process:

- the task of the designer is to represent the true nature of the target proficiency in the proficiency specification;
- the task of the candidate (with the support of her facilitators) is to represent her true level of proficiency in her performance profile;
- the task of the assessor (with the support of his facilitators) is to represent the true quality of each of the candidate’s responses in his evaluation profile;
- the task of the aggregator (with the support of their facilitators) is to represent the true value of the assessor’s evaluation profile in the overall result; and
- the task of the user (with the support of their facilitators) is to represent the true meaning of the overall result in the final measurement interpretation.

Finally, it is worth underlining the importance of representing the target proficiency – the thing that needs to be measured – as faithfully as possible, via the proficiency specification. This is because it is the point of reference for everything else that occurs during the design and development phases.

Figure 2 represents the five-step qualification lifecycle graphically. Its dotted lines illustrate how the proficiency specification is the point of reference for everything else that follows. Figure 3 illustrates how the four-stage lifecycle and the 5-step lifecycle can be integrated, to provide an even more comprehensive graphical representation.

Figure 2. The 5-step qualification lifecycle

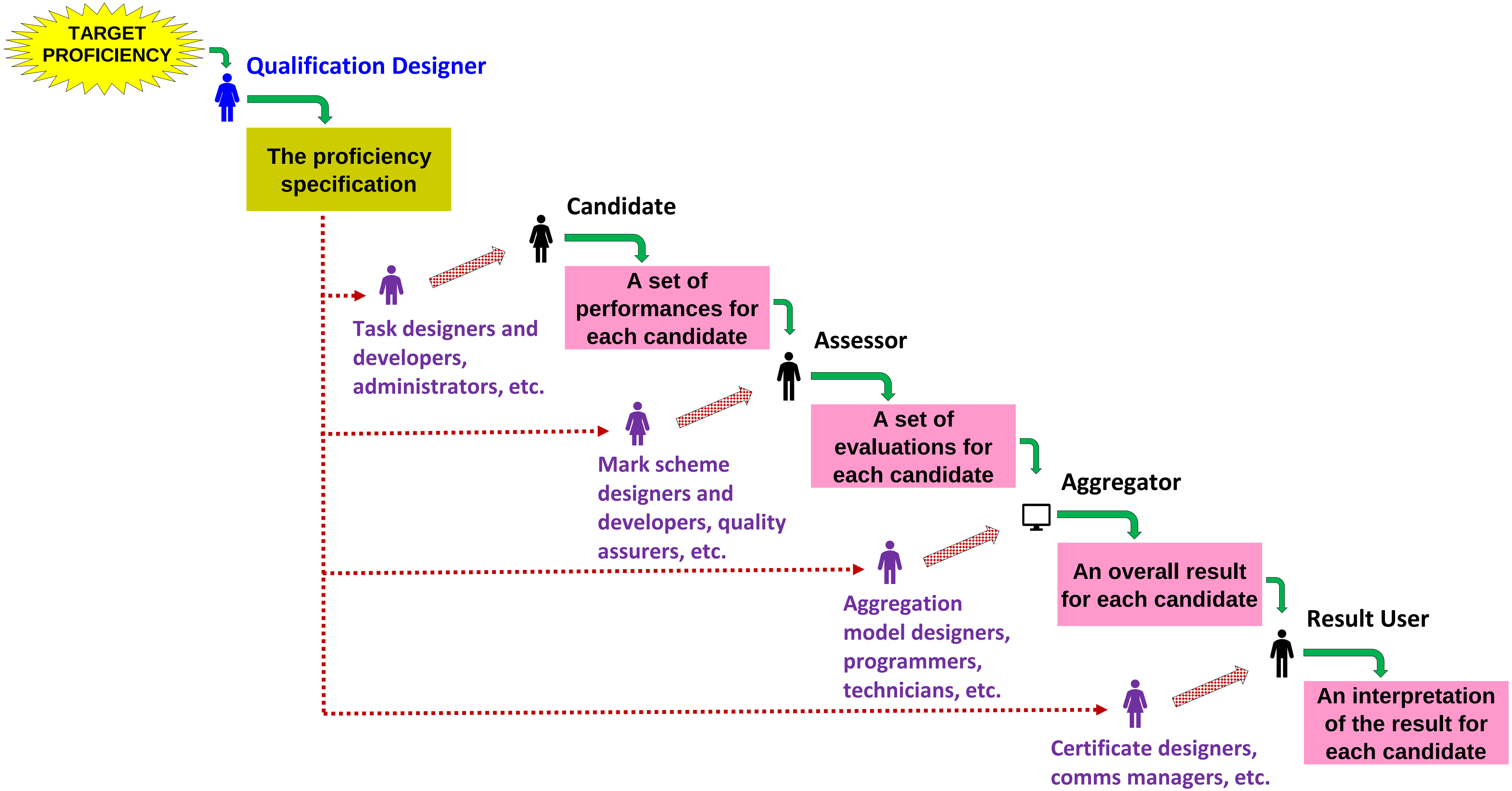
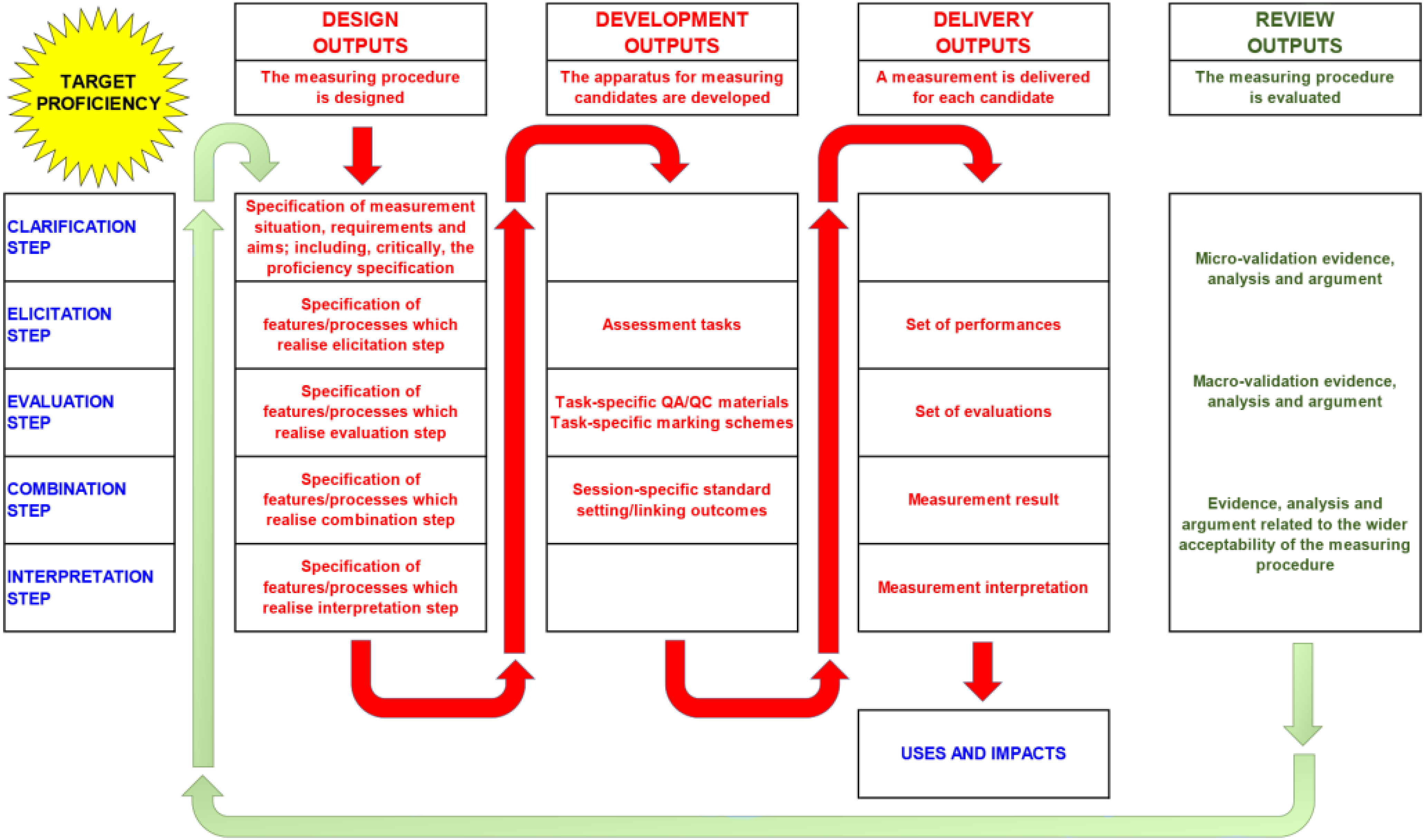

TARGET PROFICIENCY
Qualification Designer
The proficiency specification
Candidate
A set of performances for each candidate
Task designers and developers, administrators, etc.
Assessor
A set of evaluations for each candidate
Mark scheme designers and developers, quality assurers, etc.
Aggregator
An overall result for each candidate
Aggregation model designers, programmers, technicians, etc.
Result User
An interpretation of the result for each candidate
Certificate designers, comms managers, etc.

Figure 3. The 5-step lifecycle (vertical plane) alongside the 4-stage lifecycle (horizontal plane)



TARGET PROFICIENCY	DESIGN OUTPUTS	DEVELOPMENT OUTPUTS	DELIVERY OUTPUTS	REVIEW OUTPUTS
	The measuring procedure is designed	The apparatus for measuring candidates are developed	A measurement is delivered for each candidate	The measuring procedure is evaluated
CLARIFICATION STEP	Specification of measurement situation, requirements and aims; including, critically, the proficiency specification			Micro-validation evidence, analysis and argument
ELICITATION STEP	Specification of features/processes which realise elicitation step	Assessment tasks	Set of performances	
EVALUATION STEP	Specification of features/processes which realise evaluation step	Task-specific QA/QC materials Task-specific marking schemes	Set of evaluations	Macro-validation evidence, analysis and argument
COMBINATION STEP	Specification of features/processes which realise combination step	Session-specific standard setting/linking outcomes	Measurement result	Evidence, analysis and argument related to the wider acceptability of the measuring procedure
INTERPRETATION STEP	Specification of features/processes which realise interpretation step		Measurement interpretation	

USES AND IMPACTS

Validation argument

During the first half of the twentieth century, particularly in North America, there was a tendency for:

- tests that were used ‘to predict’ (eg personnel selection tests) to be validated **empirically**, ie by correlating their results against a criterion measure (eg subsequent performance in a job), and for
- tests that were used ‘to measure’ (eg school maths tests) to be validated **logically**, ie by decomposing their target proficiency into its constituent elements (eg calculation, number, algebra, calculus) and then checking that the test instrument sampled those elements relevantly and representatively.

The fact that the former came to be known as ‘criterion validity’ and the latter came to be known as ‘content validity’ helped to reinforce two caricatured ideas: (i) that these were quite different *kinds* of validity; and (ii) that a *single* criterion validity study was sufficient to validate a personnel selection test, just as a *single* content validity study was sufficient to validate a school maths test. As time went by, the fallacy of these ideas became increasingly evident. Scholars began:

- to challenge the idea that prediction and measurement could be so neatly separated;
- to reject the idea of single-study validation, acknowledging a far wider variety of sources of validation evidence and analysis, with relevance to any kind of test;
- to emphasise that there is just one kind of validity – ‘construct validity’ or nowadays simply ‘validity’ – which requires a scientific approach to validation; and
- to use validity as the organising framework for anything to do with measurement quality (ie to embrace concepts like reliability, comparability and bias).

All of these changes turned validation into a much more complex, wide-ranging activity. The idea that a test could be validated via a one-off empirical or logical study was gradually replaced with the idea of validation as a **scientific programme of research**, involving theory-based hypothesis-testing and provisional conclusions.

Validation as argumentation

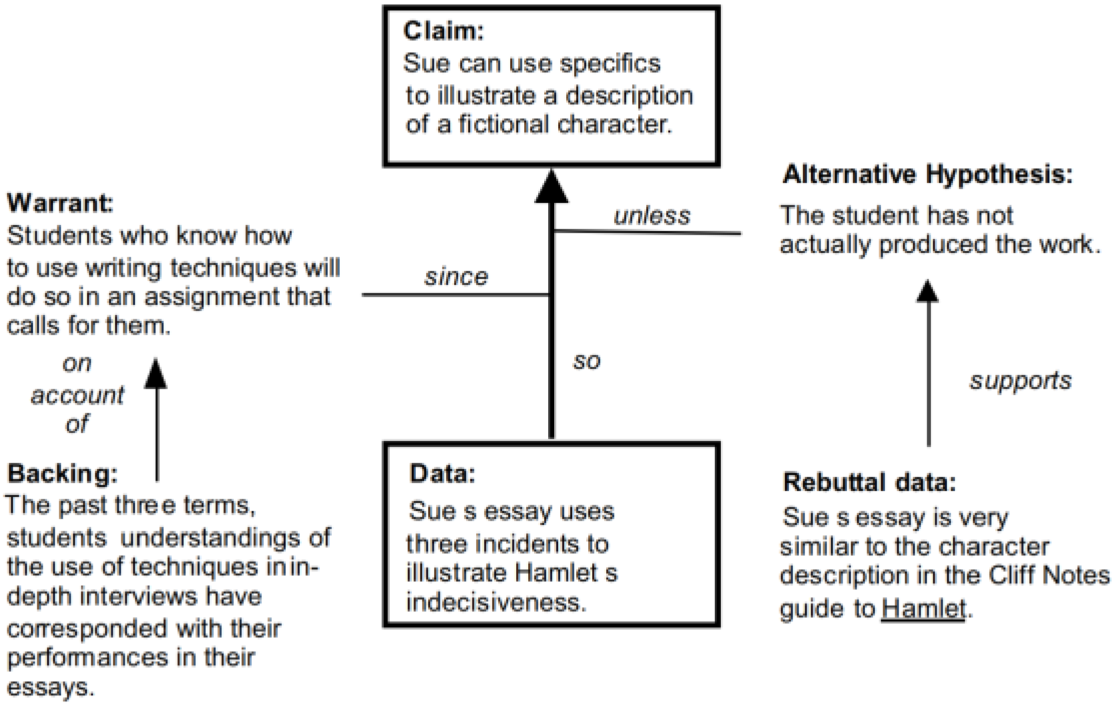
The most recent progression in thinking about validation involves the recognition that sources of evidence and analysis relevant to validity need to be organised into a persuasive measurement **argument**. As noted earlier, this involves three steps:

- specifying the measurement claim (ie that it is possible to measure the target proficiency accurately using assessment results);

- constructing an argument to support that claim; and
- evaluating the strength of that argument (and modifying the claim, if necessary).

Much of the recent literature is based upon the model of informal argumentation developed by Stephen Toulmin (1958). He proposed that everyday arguments can be characterised in terms of drawing a **claim** from **data** (evidence) on the basis of a **warrant** which is supported by **backing**. However, that claim might be **rebutted** if there is a plausible alternative explanation for the data, with its own supporting evidence. Figure 4 reproduces a figure from a report by Robert Mislevy, et al (2002, p.13) to illustrate how this kind of analysis can be used to characterise an assessment argument. In this instance, the claim is that Sue can be described as having a particular ability (to use 'specifics' to illustrate a description) and the data comes from Sue's essay.

Figure 4. An assessment-based Toulmin Diagram (Mislevy, et al, 2002)


Claim:
Sue can use specifics to illustrate a description of a fictional character.
Warrant:
Students who know how to use writing techniques will do so in an assignment that calls for them.
since
unless
Alternative Hypothesis:
The student has not actually produced the work.
so
on account of
supports
Backing:
The past three terms, students understandings of the use of techniques in in-depth interviews have corresponded with their performances in their essays.
Data:
Sue s essay uses three incidents to illustrate Hamlet s indecisiveness.
Rebuttal data:
Sue s essay is very similar to the character description in the Cliff Notes guide to Hamlet.

Although the claim-data-warrant model is at the heart of the recent literature on constructing validation arguments, different scholars have used it in different ways.

A psychometric argument

One of the most influential scholars in this field is Michael Kane (eg Kane, 2013) who has recommended thinking of validation argument as a chain of reasoning that allows you to generalise from performance in an assessment situation to performance in the real world. The key inferences in his generic measurement argument are the:

- scoring inference – the test score makes a claim on the basis of the observed performance;
- generalisation inference – the universe score makes a claim on the basis of the test score[6]; and
- extrapolation inference – the score interpretation makes a claim on the basis of the universe score.

In each case, it is the responsibility of the evaluator to consider: (i) whether there is a persuasive warrant, with effective backing, to justify making the claim on the basis of available data; and (ii) whether plausible alternative hypotheses can be ruled out. The scoring inference concerns whether the test score is a faithful representation of performance on a particular test. The generalisation inference concerns whether it would be safe to generalise this conclusion across legitimate replications, eg across different versions of the test or across different assessors. The extrapolation inference concerns whether it would be safe to extrapolate this conclusion to the candidate's performance in the real world.

Stuart Shaw and Vicki Crisp (2012; 2015) have used this framework to construct validation arguments for general qualifications in England. Following Kane's idea of an Interpretation *and Use* argument, they concluded with a decision-making inference. Departing slightly from Kane's approach, they began with a construct representation inference. It is important to note that Kane does not insist upon any particular argument structure. He merely insists that the argument should be **coherent**, that its inferences should be **reasonable**, and that its assumptions should be **plausible**. Figure 5 reproduces a figure from Shaw and Crisp (2015, p.36, Figure 3) to illustrate how this kind of framework can be used to identify key questions for a research programme.

[6] A 'test score' is the actual score that a candidate receives; after having been assessed via a particular set of questions (from a pool of questions that might possibly have been asked), a particular assessor (from a pool of assessors that might possibly have evaluated the question responses), on a particular day (from a pool of possible administration days), and so on. Just by chance, that candidate might have received a different score, if they happened to have sat the test on a different day (from the pool of admissible administration days), or if they happened to have faced a different set of questions (from the pool of admissible questions), or if their responses had been evaluated by a different assessor (from the pool of admissible assessors), and so on. It therefore seems reasonable to define the score that any particular candidate 'deserves' to receive – given their level of attainment rather than their test performance, per se – as the average of the scores that they would have achieved had they been assessed repeatedly via all possible combinations of admissible alternatives. This is their (hypothetical) universe score, as opposed to their (observed) test score.

Figure 5. Argument framework (adapted from Shaw and Crisp, 2015)

Inference	**Warrant justifying the inference**	**Validation questions**
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 give an indication of likely wider performance	4. Do the constructs sampled give an indication of broader competence within and beyond the subject?
Decision-making	Scores/grades give an indication of likely success in further study or employment	5. Do scores/grades give an indication of success in further study or employment such that they can be used to make appropriate decisions?

Kane's argument structure is 'psychometric' in the sense that it is based upon a number of important psychometric concepts, including the principle of drawing successively richer conclusions on the basis of primary assessment evidence and, in particular, the notion of a universe score. Its logic might therefore be immediately apparent to someone steeped in the psychometric literature, as many test evaluators are. However, its accessibility for practitioners who may lack this grounding has been questioned (Knorr and Klusmann, 2015) and concerns have been raised that its lack of accessibility might risk evaluators underemphasising or overlooking important validation research questions and therefore important evidence and analysis (Ferrara, 2007). The following section proposes an alternative argument structure – concerning the functioning of an assessment procedure – that was developed in an attempt to provide a more generally accessible approach.

A functional argument

A somewhat different approach to constructing validation arguments follows from the step-based lifecycle model discussed in the previous section.[7] Recognising that educational measurement is essentially a representational process, and developing the analogy of a production line in which successive steps operate on the output of preceding ones, we arrive at the following validation argument:

- IF the target proficiency is faithfully represented by the proficiency specification (step 1 – clarification);
- AND IF the proficiency specification is faithfully represented in the performance profile (step 2 – elicitation);
- AND IF the performance profile is faithfully represented by the evaluation profile (step 3 – evaluation);
- AND IF the evaluation profile is faithfully represented by the measurement result (step 4 – combination);
- AND IF the measurement result is faithfully represented by the measurement interpretation (step 5 – interpretation);
- THEN those measurement interpretations will be both accurate and useful.

Notice how the measurement interpretation is a representation of the measurement result, the measurement result is a representation of the evaluation profile, and so on back through the production line. There is a slight disjunction between step 1 and step 2 as we transition from the principal design output to the four delivery outputs. What needs to be ensured during step 2 is that the proficiency specification is faithfully represented in (rather than by) the performance profile. In other words, we

[7] This functional approach – which considers the functioning of each of the features and processes that comprise an assessment procedure, both individually and as an ensemble – resonates with a number of validation frameworks that have been proposed in recent years; including, the ‘temporal’ framework proposed by Cyril Weir (eg Weir, 2005; Shaw and Weir, 2007). It is also very similar to the framework originally proposed by Steve Ferrara (2007) and developed by Ferrara and Lai (2016). They identified seven critical steps – assessment policies and principles, design and development, implementation, response scoring, technical analyses, score feedback, interpretation and use – and illustrated the kind of claims that need to be supported at each step. The 5-step lifecycle model is essentially the same; although it was designed to be slightly more generic, ie to be directly applicable across a wider range of assessment procedures. It also provides a more explicit argument structure. The functional approach and the psychometric approach are two ways of solving the same problem, ie how to structure a validation argument. They differ primarily in terms of focus and emphasis. One of the most accessible resources for teaching validation argument is a paper by Terry Crooks, Michael Kane and Allan Cohen (1996), entitled *Threats to the valid use of assessments*. It describes an approach that uses “time sequences of assessment processes as [its] organising schema” (p.267), and that might be described as half-way between a functional approach and a psychometric one.

need to ensure that the kind of evidence actually elicited reflects the kind of evidence that ought to be elicited, as determined by the proficiency specification.

Step 1 has the same structure as the other steps; although, of course, it operates on the target proficiency, per se. This is a fascinating step because, although it presumes some kind of reality for the target proficiency, it is actually *through* the proficiency specification that the target proficiency is characterised. This step literally constructs a version of the target proficiency. The important point, from a validation perspective, is that even this step can, and often does, go wrong! The effectiveness of this step needs to be interrogated just as much as, if not more so than, any of the subsequent steps.

This illustrates how the argument presented above establishes a framework for validation research by identifying a series of claims that need to be justified in order to be confident in the conclusion. For each step, the degree to which representations are likely to be faithful (across our target population of candidates) is determined by the effectiveness of the features and processes built into the assessment procedure in order to operationalise the step. In other words, this validation argument concerns the effective design of the assessment procedure. If the procedure has been designed effectively, then the target proficiency will be faithfully represented by the proficiency specification, the proficiency specification will be faithfully represented in the performance profile, and so on.

The idea of **validity-by-design** – of building validity into an assessment procedure during the design stage – is not new; but it has been given a new lease of life via a number of detailed, systematic frameworks, which have been produced in North America in recent years, to transform the art of assessment design into a robust technology, grounded in a more scientific approach.[8] The validation argument presented above introduces the idea of **validation-of-design**. Perhaps the most attractive feature of the functional model is that the structure of the evaluation process mirrors the structure of the design process. The validation argument considers each of the features and processes built into the assessment procedure (to operationalise each of the five steps) and considers their contribution to the validity of the assessment procedure overall. This functional approach emphasises how validation and design ought to begin at the same time; that is, from the point at which the logic that underpins the design of the assessment procedure begins to be made explicit.

Finally, the most important insight from the production line metaphor can be brought home via the metaphor of a bucket brigade, which was introduced by Alastair Pollitt and Ayesha Ahmed (2009). The point of this bucket brigade is to extinguish flames in a village using water that comes from a river nearby. Each bucket is filled with water

[8] For example, Evidence-Centered Design (eg Mislevy, 2007) and related approaches.

and then each one is passed along a chain of villagers until it reaches the village and can be thrown over the fire. The challenge for the bucket brigade, of course, is to prevent splashes and leaks. If the members of the brigade are not careful, then there may be insufficient water in their buckets by the time they reach the flames. In this metaphor, water represents validity and the bucket brigade represents an assessment procedure. If the assessment procedure has been designed effectively, and is implemented carefully, then there will be sufficient validity left by the end of step 5. The important insight is that, once a bucket has lost its water, even if that occurs right at the beginning of the chain, that water cannot be put back in during a subsequent step. Each step needs to work effectively – independently and as an ensemble – for the assessment procedure to have sufficient validity. If a single step fails, then the entire assessment procedure fails.

Validation evidence and analysis

Empirical evidence and logical analysis are used to test the claims in a validation argument in order to evaluate its overall strength. From a scientific perspective, we might think of this in terms of whether it is possible to *falsify* any of the claims, and therefore to undermine the measurement argument. If the argument is robust to efforts to undermine it, then we can be confident in it; we can conclude that the validation argument is strong, and that results are likely to be accurate and useful. In practice, we often think of validation research in terms of collating evidence and analysis in *support* of each of the claims in the measurement argument. Whichever perspective we prefer to adopt, at any particular point in time, the first thing that we need to do is to identify the *kinds* of evidence and analysis that can be used either to falsify or to support a measurement argument. Over the years, a variety of frameworks have been proposed. However, as is true of most of the ideas in this field, all of these frameworks have their strengths and weaknesses, and none represents the last word on the subject.

The 'five sources' framework

The validity chapters from each of the six editions the North American *Standards for Educational and Psychological Testing* (AERA, et al, 2014) have strongly influenced international thinking on sources of evidence and analysis for validation research. The current edition identifies five major sources:

1. test content;
2. response processes;
3. internal structure;
4. relations to other variables; and
5. consequences of testing.

For each source, the basic research question concerns the degree to which the evidence or analysis that is collated is consistent with the overarching measurement claim (that it is possible to measure the target proficiency accurately using assessment results). **Test content** analysis generalises the earlier notion of 'content validity' and is concerned with the degree to which the proficiency specification – and therefore the target proficiency – is faithfully represented via apparatus (eg test questions, mark schemes). For example, it might involve scrutinising the set of questions that comprises an exam paper, to consider the degree to which the 'content' of those questions is relevant to and representative of the components of the target proficiency, as articulated via the proficiency specification.

Response process evidence goes one step further by investigating the extent to which the proficiency specification is faithfully represented via responses (eg candidate performances, evaluations of those performances). For example, it might involve setting up a ‘cognitive laboratory’ experiment, in which candidates are asked to externalise their thought processes whilst answering test questions. Sometimes it may turn out to be possible for a candidate to answer a question correctly without necessarily engaging the thought process that the question is supposedly testing; for instance, if the question appears to be testing problem solving, yet candidates are able simply to recall the correct answer. Other times it may turn out to be impossible for candidates to answer a question correctly even though they would normally be able to engage the thought process that the question is supposedly testing; for instance, if the question is so confusingly worded that they do not understand what it is asking them to do. Cognitive laboratory experiments can help an awarding organisation to identify serious threats to validity that may cause results to be inflated or deflated for substantial numbers of candidates. It can be particularly helpful for identifying problems specific to subgroups of the candidate population, eg ethnic minorities.

Internal structure analyses investigate candidates’ performances across the separate elements of an assessment. If, for instance, a test were designed to assess proficiency in spelling everyday words – by asking candidates to spell a sample of 50 words which are read aloud to them – then each word that a candidate spells could be thought of as an independent sample of evidence of their proficiency in spelling everyday words. If it is legitimate to think of proficiency in spelling everyday words as essentially unidimensional – meaning that there is no reason to think that it ought to be decomposed into different kinds of spelling proficiency – then we might predict a strong correlation between how accurately a candidate spells any particular word and how accurately that candidate spells all other words in the test. Correlations can be computed to investigate the extent to which observed patterns of performance across questions are consistent with those that would be predicted, on the assumption that the test is measuring what it is supposed to be measuring (eg using Cronbach’s Alpha statistic to investigate relationships between performances across items).

The analysis of evidence concerning **relations to other variables** embodies a similar logic, in the sense that it investigates the degree to which observed patterns are consistent with those that would be predicted, on the assumption that the test is measuring what it is supposed to be measuring. However, instead of investigating patterns across elements *within* an assessment, it examines patterns *between* results from the assessment and other outcome variables (eg concurrent assessments, experimental results, future outcomes). This source generalises the earlier notion of ‘criterion validity’ and is concerned with relationships between results from the qualification which is being evaluated and different kinds of outcomes. For instance, imagine that an awarding organisation had designed a qualification in Employability Skills. Clearly, we would predict that those who passed the qualification ought to be more employable than those who failed. If that were put to the test, by asking a panel of employers to rate the basic employability of a sample of candidates who had passed or failed, via a generic job interview, then we would expect there to be a substantial correlation between their ratings and qualification results. If the correlation turned out to be low, then this would cast some doubt upon the claim that the qualification was actually measuring what it was supposed to be measuring.

The inclusion of **consequences of testing** as a legitimate source of validation evidence or analysis has been hotly debated. This is because many assessment professionals believe that the consequences of an assessment policy – impacts arising from the decision to implement the assessment procedure and to use its results for particular purposes – ought to be investigated entirely independently of the validity of the assessment procedure, ie its potential to support high quality measurement. Although this is not an unreasonable stance, it actually misses a crucial point. Evidence from the consequences of implementing an assessment procedure can often be used to judge the validity of that assessment procedure; even when that is understood purely in the sense of its potential to support high quality measurement. The logic is exactly the same as for the other sources: are the observed consequences consistent with what we would predict, if it were true that the qualification is actually measuring what it is supposed to be measuring? So, whereas the previous example set up a criterion validation experiment, asking a panel of employers to rate the employability of a sample of candidates, we could investigate essentially the same hypothesis via consequential evidence; that is, by considering how candidates who passed the qualification fared in the job market in comparison with candidates who failed.

Other frameworks

When it comes to categorising different kinds of validation evidence and analysis, there are all sorts of frameworks, but none of them is perfect. For this reason, is it helpful to mention two additional frameworks, by way of contrast.

In recent years, Ofqual (and the Qualifications and Curriculum Authority, prior to that) has made good use of its five ‘common criteria’ for evaluating qualifications:

1. validity;
2. reliability;
3. comparability;
4. minimising bias; and
5. manageability.

Each of these can be understood as an important source of evidence of validity. This is even true of the least technical of these criteria, manageability. For instance, if an assessment is not **manageable**, ie not practically viable, then ultimately it will not be possible to measure whatever that qualification needs to measure.

Similarly, it is important both to the validity and to the credibility of a qualification that it is as free as possible from **bias**. Bias indicates measurement problems for particular groups of candidates, eg those who are consistently over-estimated and those who are consistently under-estimated. In other words, it is the degree to which assessment results are systematically less accurate for certain subgroups of candidates, which is especially significant with respect to subgroups defined within equality legislation. Bias is typically evidenced via differences between subgroups in (averaged) assessment results that cannot plausibly be explained in terms of differences in (averaged) levels of attainment. It means inappropriately favouring certain subgroups whilst inappropriately penalising others and should therefore always be minimised as far as possible.

Because England has a tradition of different organisations awarding equivalent qualifications under the same qualification title, **comparability** has always been a central concern. Comparability is the degree to which assessment results, derived from separate assessments, embody the same standard. There is an expectation of comparability whenever two or more qualifications award results under the same qualification title, for example: when an organisation delivers results for the same qualification from one session to the next; or when two organisations deliver results under the same qualification title within a single session.

Relatedly, **reliability** concerns the degree to which assessment results are reproducible; that is, the likelihood that learners would receive the same assessment result if the assessment procedure were to be replicated, ie implemented for them a second time.

Although these common criteria are useful for highlighting particularly important sources of evidence and analysis, one of the problems with this kind of framework is that the categories lack mutual exclusivity, ie they tend to overlap conceptually, which can be confusing. For instance, in one sense, a lack of comparability is also a particular kind of bias, where one set of results is consistently inflated or deflated in comparison with another. In another sense, a lack of comparability is also a particular kind of unreliability, where ‘session’ or ‘organisation’ is specified as a dimension across which results ought to be replicable. Finally, the fact that the international literature tends nowadays to use validity as a framework for organising anything to do with measurement quality – thereby embracing concepts like reliability, comparability and bias – means that specifying it *alongside* those other concepts is a little puzzling: exactly what it is presumed to add to the other criteria is unclear.

In the context of vocational and occupational qualifications in England – more specifically, in the context of training for assessors and internal quality assurers who work for assessment centres, and external quality assurers who work for awarding organisations – a slightly different framework has gained traction. The following five categories come from Ros Ollin and Jenny Tucker’s *Vocational Assessor Handbook* (Ollin and Tucker, 2012, p.60):

1. the type of assessment used and the evidence provided should be fit for the purpose for which it is intended (validity);
2. the assessment should be consistent and reliable (reliability);
3. the evidence being assessed should be sufficient for the assessor to make a judgement on the learner's knowledge and/or skills against specified criteria (sufficiency);
4. there should be no doubt that the evidence is genuine and has been produced by the candidate (authentication[9]); and
5. the evidence can prove that the candidate is up to date on current methods, skills and knowledge in the chosen vocational area (currency).

Once again, the above definition of validity seems large enough to subsume the other categories. But notice, this time, how the last three sources highlight particularly important threats to validity for vocational and occupational qualifications:

- sufficient sampling: which is particularly challenging for qualifications that utilise large, complex, integrated, performance tasks (as opposed to qualifications that utilise small, simple, discrete, written tasks), but that can also be particularly problematic when candidates are required to compile their own portfolios of evidence (as opposed to sitting a test that has been specifically designed to ensure sufficient sampling);
- authentication of performances: which is particularly challenging for qualifications that utilise portfolios of evidence (whereby, under the guise of formative feedback, the portfolio can sometimes end up being as much the product of a teacher/trainer as the candidate; and because portfolios tend also to be more susceptible to plagiarism); and
- qualification currency: which is particularly challenging at the qualification level in occupational areas that change rapidly over time (where a qualification can end up measuring what needed to be measured ten years ago, but not today), but that can also be particularly problematic at the individual candidate level (if, for instance, a candidate were to seek recognition of prior learning for competencies demonstrated, say, seven years previously).

[9] They actually use the term 'authenticity' which tends to have a quite different meaning in the international literature on performance assessments, which is why it has been changed to 'authentication' here. The term 'authenticity' is generally used to refer to the degree to which an assessment task resembles the kind of situation in which element(s) of knowledge, skill or understanding (supposedly tapped by the task) would be deployed in the real world.

The important point to note from this cursory review of these two English frameworks is that they tend to foreground sources of evidence and analysis of particular significance to the contexts within which they were originally designed to be used. For instance, the fact that comparability is foregrounded within the five common criteria would come as no surprise to anyone who has worked in the context of school tests and examinations in England – for which the common criteria were originally developed – where comparability has remained an enduring fixation of educational discourse for over a century (Newton, 2007). Conversely, in the Ollin-Tucker framework, comparability is not foregrounded, but different criteria – with particular relevance to the context of vocational and occupational qualifications in England – are foregrounded. Clearly, comparability cannot be ignored in the context of vocational and occupational qualifications; in just the same way that sufficiency, authentication, and currency cannot be ignored in the context of school tests and examinations. So it is important to recognise two major limitations that affect the categories that comprise both of these frameworks:

1. they are not mutually exclusive (as they tend to overlap conceptually); and

2. they are not collectively exhaustive (as they only foreground certain sources).

Of course, they are both still useful! They *do* foreground important concerns. Importantly, though, they do not paint an entirely comprehensive picture of validation evidence and analysis.

Exactly the same kind of criticism can be levelled at the five sources framework from North America. It was created for a context that has been dominated by the use of multiple-choice testing for nearly a century. It is therefore not surprising that it is especially useful for planning a validation research programme for tests that comprise a large number of small, simple, discrete, written tasks that are administered to large cohorts of candidates – the multiple-choice test being a classic example. In this kind of context, common questions, and responses to those common questions, become a natural focus for validation research: Cronbach’s Alpha; item facility indices; Differential Item Functioning statistics; item-test correlations; factor analyses; item-objective congruence judgements; candidate ‘think aloud’ studies; and so on. Clearly, this focus is problematic for assessments that comprise a small number of large, complex, integrated, performance tasks that are administered to a small cohort of candidates – the workplace simulation being a classic example. In this kind of context, it is often not useful, and sometimes simply impossible, to focus validation research upon common questions and responses. This emphasises the importance of considering alternative sources of evidence and analysis, which might not fit quite so neatly within the North American framework.

A different way of considering the classification of validation evidence and analysis relates to the idea of scrutinising assessment procedures. Like evidence from test content, this generalises the earlier notion of ‘content validity’ yet not in a way that could simply be accommodated by the addition of a new category to an existing framework. This is to introduce a distinction between macro-validation research and micro-validation research; which, in fact, is actually more about the way in which sources of evidence and analysis are used than about the nature of those sources (see Newton, 2016).

Macro- vs. micro-validation

To introduce this distinction by way of an analogy: there are two perspectives from which the work of a restaurant chef can be judged. As they are preparing the meal, an expert chef can observe what the restaurant chef is doing, asking questions like:

- have they followed the right recipe?
- have they added all the right ingredients?
- have they combined them in the right way?

Once the meal is prepared, a food critic can consume it, asking questions like:

- does it look like it ought to?
- does it taste like it ought to?
- am I going to be sick?

It is often said that the proof of the pudding is in the eating. But that is not entirely true. The pudding might taste and look great despite using unhealthy, or even dangerous, ingredients. Or maybe the restaurant chef followed a lax procedure, perhaps not washing any of the ingredients, which just happened not to impair the quality of the meal this time, but that might well impact negatively upon future meals. The fact of the matter is that, when judging the work of a restaurant chef, it is important to ask both product-related questions and process-related ones. The same is true when judging the work of an awarding organisation. The procedure that is followed when preparing the meal is analogous to the assessment procedure. The meal is analogous to the qualification result; including its interpretations, uses, and consequences.

As discussed earlier, we have stipulated that the principal focus for validation research is the validity of an assessment procedure; because it is the effective design of an assessment procedure that underpins accuracy and usefulness, *each time qualification results are delivered*. Now, when focused upon the validity of the procedure, the food critic and the head chef will have quite different perspectives. By the time the food critic gets involved, the procedure is basically complete; although, to extend the analogy with assessment, the chef might recommend a process for eating the meal, which emphasises that the food critic is like the qualification user in completing the production line. But the important point is that the food critic's perspective on evaluating the procedure is essentially **holistic**. The procedure has been implemented and, by reflecting upon its outcome, the food critic passes judgement upon the effectiveness of the procedure as a whole. In contrast, the expert chef focuses upon each and every element of the procedure as it is being implemented. This means that the expert chef's perspective on evaluating the procedure is essentially **atomistic**; narrow, and targeted on the various features and processes that comprise the procedure.

In essentially the same way, validation research can be conducted from an holistic perspective; drawing conclusions about the effectiveness of the assessment procedure as a whole, typically on the basis of evidence from results, interpretations, uses, and consequences. Let's call this the **macro-validation perspective**. Or it can be conducted from an atomistic perspective; drawing conclusions about the effectiveness of the various features and processes that comprise the assessment procedure. Let's call this the **micro-validation perspective**.

Rather than casting this as a clear-cut distinction between perspectives, it is probably more helpful to think of it as characterising two ends of a continuum. In other words, certain investigations will be closer to the macro-validation end, evaluating the procedure overall; while other investigations will be closer to the micro-validation end, evaluating its constituent elements (and their interactions). Response process analysis is a good example of an approach that is close to the micro-validation end, because it is focused on the effectiveness of features and processes put in place to elicit evidence of proficiency. The analysis of evidence concerning relations to other variables is a good example of an approach that is close to the macro-validation end, because it is focused on the effectiveness of the assessment procedure overall, by judging its outcomes; more specifically, by investigating whether its results are predictably related to other outcomes.

The point of distinguishing between macro- and micro-validation is to emphasise just how important the micro-validation perspective can be; and, in doing so, *to highlight all sorts of sources of evidence and analysis that extant validation frameworks typically fail to foreground*. These are sources related to the effective design of the many features and processes that comprise an assessment procedure, which might include:

- **routine formative analyses**, which are undertaken during the development stage, to hone the effectiveness of a particular feature or process (eg fairness reviews, whereby stakeholders pass judgement on the quality of assessment tasks, some of which will be revised as a consequence)

- **quality control metrics**, which are undertaken during the delivery stage, to monitor the effectiveness of a particular process (eg marker-moderator consistency statistics, printing error statistics)

- **auxiliary investigations**, which are undertaken during the review stage, to evaluate the effectiveness of a particular feature or process, and ideally also to improve understanding its mechanism so as to feed back into the re-design stage (eg investigation into comparability of standards for equivalent qualifications offered by different awarding organisations).

To summarise this section, we can conclude that *any* source of evidence or analysis that helps to establish a case for or against the overarching measurement claim (that it is possible to measure the target proficiency accurately using assessment results) should be considered a legitimate source; whether or not it seems to fit neatly within any of the established frameworks.

Validation research

Our core validation research question concerns the degree to which evidence and analysis is consistent with the overarching measurement claim: that it is possible to measure the target proficiency accurately using assessment results. Macro-validation research tends to confront that claim directly and holistically. Micro-validation research, on the other hand, necessitates the kind of scaffolding that is provided by a validation argument, comprising a series of sub-claims which concludes with the overarching measurement claim. Design-centred functional arguments, of the sort described earlier, are particularly useful for this purpose. Each of the sub-claims in the validation argument is tested one-by-one in order to evaluate its overall strength and therefore the strength of its conclusion.[10] The following sections characterise macro-validation and micro-validation research in greater detail, before discussing their respective argument structures, and how to go about planning a validation research programme.

Macro-validation research

Macro-validation research focuses directly upon the overarching measurement claim; hence the idea of an holistic perspective. Certain kinds of evidence and analysis tend to be aligned more closely with macro-validation research; particularly those concerned with the evaluation of results, interpretations, uses and consequences. This includes the two sources of evidence/analysis from the North American framework that focus upon relationships between and within results:

- relations to other variables – based on overall results (eg test-criterion correlations, test-indicator correlations, multi-trait multi-method correlations, theory-based predictions)
- internal structure – based on scores for component tasks (eg reliability statistics, factor analyses, component correlations).[11]

[10] According to this model, strength and validity are distinct concepts. The validity of a particular qualification is the degree to which it is possible to measure whatever that qualification needs to measure by implementing its assessment procedure. From this perspective, high validity is tantamount to high quality measurement, ie measurement that can relied upon to deliver high levels of accuracy and usefulness. The more evidence and analysis that we collate, and the more persuasively we organise that evidence and analysis within our validation argument, the stronger that argument is likely to become. However, as additional evidence and analysis is integrated, we may need to modify the conclusion of the argument, to reflect humbler claims concerning the degree of accuracy and usefulness that is possible. In other words, it is quite possible that we might end up with a strong argument that concluded with an overarching measurement claim concerning only moderate levels of accuracy and usefulness; that is, a moderately valid assessment procedure.

[11] See AERA, et al (2014) for more details concerning these sources of evidence.

It also includes all sorts of evidence/analysis related to interpretations, uses and consequences, including:

- consequences and side-effects (eg progression routes, rejection rates)
- misuses of results (since widespread unintentional misuse raises questions concerning a variety of potential problems – from misspecification to misinterpretation – without necessarily being able to pinpoint the root cause of the misuse)
- customer satisfaction (eg uptake/sales figures, general feedback)
- public opinions (eg public confidence surveys).

It is important to appreciate that all of these sources of evidence/analysis are used, during validation, simply to challenge or to bolster the measurement claim; and that none of these sources could ever be considered definitive in its own right. Having said that, it is likely that these sources will differ in their evidential/analytical **power**; particularly, their power to undermine the measurement claim. For instance, uptake/sales figures provide a relatively weak source of evidence concerning the validity of an assessment procedure. Just because a qualification proves to be unpopular, that does not necessarily cast doubt over its validity. Indeed, sometimes validity can even contribute to unpopularity; for instance, if the standard of a competitor qualification were to drift downwards over time, (inappropriately) making it more attractive to customers. The point is simply that evidence of unexpectedly low uptake, just like evidence of unexpectedly high uptake, ought to prompt an awarding organisation to ask itself whether this might have anything to do with the validity of the qualification, and potentially to conduct further investigations that might help to arbitrate the matter.

Other sources of evidence can prove to be far more powerful, for example, ‘parallel forms’ reliability studies, which administer two versions of an assessment to a single group of candidates, to investigate the degree to which measurement results are replicable. Even highly standardised assessment procedures do not and could not control each and every aspect of the delivery process; and the particular elements that do vary from one administration to the next, eg the particular questions that are asked within an exam paper, will have some bearing upon the degree to which results are replicable. The assumption, here, is that each candidate’s level of proficiency is stable (ie the same) from one administration to the next; which is usually a reasonable assumption to make within a small time interval. So any inconsistency between candidates’ results, from first to second administration, provides an estimate of unreliability and therefore an indicator of invalidity. If an assessment procedure has been designed effectively, then the elements that vary from one administration to the next should not cause too much inconsistency in measurement results. The greater the inconsistency observed, the less valid the assessment procedure, assuming that candidate proficiency genuinely does remain stable. Evidence from this kind of reliability study can be particularly powerful when it indicates a very low level of consistency. When unreliability falls below a certain level, it becomes implausible to conclude that anything at all is being measured.

Importantly, though, even very high levels of consistency are far less powerful indicators than very low levels of consistency. This is because reliability is a necessary criterion for validity, but not a sufficient one. There are all sorts of reasons why high reliability can be demonstrated despite low validity. For example, in step 1 (clarification), the qualification designer might have specified an easy-to-measure proficiency, but not the proficiency that actually needed to be measured. Alternatively, in step 2 (elicitation), the designer might have decided only to develop tasks that cover easy-to-assess elements of the proficiency specification, excluding important but tricky-to-assess elements. Incidentally, even the most rigorous reliability studies are almost always based upon measurement *results*, not measurement *interpretations*. So they will always (substantially) underestimate the unreliability associated with any particular assessment procedure, because they fail to incorporate unreliability arising during step 5 (interpretation), when the measurement interpretations are actually drawn by qualification result users.

Macro-validation research is important for testing the overarching measurement claim directly, and ought to be included within any validation research programme. However, it is not as powerful as sometimes assumed, and it needs to be complemented by micro-validation research, which scrutinises the assessment procedure in detail through a series of lenses with far finer grain sizes. Indeed, contrary to the earliest conceptions of validation, micro-validation presents itself as the natural foundation for any comprehensive validation programme. This is because micro-validation commences as soon as we begin to design an assessment procedure, and then extends from qualification design into qualification development, delivery and review. Macro-validation, on the other hand, can only commence once we have delivered a set of results; even pilot results from developmental trialling studies emerge fairly late into the process.

Micro-validation research

Micro-validation research focuses upon lower-level claims within a validation argument; hence the idea of a more atomistic perspective. It is especially compatible with functional validation arguments, which concern the effective design of an assessment procedure. In other words, micro-validation research investigates the degree to which an assessment procedure has validity built into it by design. This can be judged in relation to principles and practices that have been refined by assessment scholars and practitioners over a period spanning considerably more than a century – which we might refer to as ‘the literature’ on educational measurement.

Various questions can be used to interrogate the **design logic** and **design efficacy** of an assessment procedure from a micro-validation perspective, for instance:

1. Does the procedure include all of the features and processes that seem (from the literature) to be associated with high validity?
2. Do each of the features and processes that comprise the procedure possess characteristics that seem (from the literature and from how they operate in practice) to be associated with high validity?
3. Are the constituent features and processes appropriately integrated and coordinated in a manner that seems (from the literature and from how they operate in practice) to be associated with high validity?
4. Is it safe to assume that the procedure can be and will be implemented correctly each time it is implemented?

This fourth question is particularly important. Framing validity at the level of assessment procedures – on the assumption that they are implemented correctly – invites the fair criticism that a procedure might be valid, but implemented incorrectly. We therefore risk not paying due attention to the threat of incorrect implementation. Whilst this is true, concerns can partially be alleviated by establishing additional, high-level controls. In other words, it is often possible and generally appropriate to build controls into processes within assessment procedures to help ensure that they are implemented correctly and to help ensure that, when incorrect implementation does occur, it is identified and rectified before it can have significant impact. For instance, in England, awarding organisations are expected to establish an appeals process, *as part of the assessment procedure for each qualification they offer*, which allows candidates or centres to appeal against results that they believe to be incorrect.

Just as for macro-validation, certain kinds of evidence and analysis tend to be aligned more closely with a micro-validation perspective; particularly those that focus on the mechanisms through which results and interpretations are delivered. This includes the two sources of evidence/analysis from the North American framework that focus upon apparatus and responses, ie test content and response processes. Importantly, though, the distinction between micro- and macro-validation does not turn on the *type* of evidence/analysis collated, but on the *use* to which it is put. If the focus is upon a particular lifecycle step, and upon a particular feature or process that helps to realise that step, then this is to adopt a micro-validation perspective, regardless of what kind of evidence/analysis is brought to bear.[12]

In addition to the sources of evidence/analysis discussed earlier, the following examples help to illustrate the diverse nature of micro-validation research:

- step 1 – clarification: survey-based evidence of stakeholders' views concerning the degree to which a proficiency specification represents what that qualification really needs to measure;
- step 2 – elicitation: experience-based and literature-informed analysis of the process for authenticating candidates' work; evidence from question paper quality control logs;
- step 3 – evaluation: experience-based and literature-informed analysis of processes for assessor training and standardisation; experimental evidence of inter-assessor consistency;
- step 4 – combination: logical analysis of the aggregation model in terms of an underlying 'theory' of the target proficiency; evidence from data entry quality control logs; and
- step 5 – interpretation: opportunistic social media evidence concerning widespread misinterpretation of assessment results.

These sources were chosen purely to illustrate the diversity of potential evidence and analysis relevant to each step in a validation argument, rather than to suggest that they have particular significance or power. Ultimately, they are all just sources of evidence/analysis alongside many other such sources.

Before attempting to interrogate design logic and design efficacy in terms of the four questions raised above, it is helpful to begin by simply describing the features and processes that are going to be (or that have already been) put in place to deliver each step in the qualification lifecycle. Being able to **describe** the assessment procedure in detail helps us to spot the omission of potentially important aspects; and it also helps us to shine a light upon aspects that are so commonplace as to be taken for granted.

[12] For instance, when individual item scores are correlated with the aggregate of all item scores, the intention is to investigate how well particular items have functioned and, by extension, to evaluate an aspect of the item development process, which is of relevance to a particular sub-claim within the validation argument (step 2). So this would be a micro-level analysis. Whereas, when individual item scores are correlated with each other via Cronbach's Alpha, the intention is to investigate the overall reliability of results, and thereby to evaluate the overarching measurement claim directly. The alpha coefficient provides a (partial) thumbs-up or thumbs-down in relation to the assessment procedure overall. So this would be a macro-level analysis.

It is important not to overlook validation research for commonplace features and processes, even those with robust backing from the literature and with longstanding precedents. This is partly because times change: features and processes that were capable of supporting validity in the past might no longer be able to do so in the present day. This can occur when results become very much more important for candidates for one reason or another: if the assessment procedure is to retain validity, then it may need to be made more resistant to cheating, for instance. It is also because delivering any step in an assessment procedure will involve a particular configuration of features and processes – from a multitude of possible configurations – and different configurations will have different strengths and weaknesses. For instance, a very tightly prescribed marking scheme might (perhaps) improve the overall accuracy of results for a team of novice assessors; whereas it might (conceivably) reduce the overall accuracy of results for a team of highly expert assessors.

Considering the evaluation step, just by way of example, a variety of step-specific questions might be identified to help flesh-out a thorough description of operational features and processes, including:

- What steps are taken to ensure that success criteria are correctly applied by assessors?
    - that they are applied in the right way (eg written instruction sheets);
    - that they are applied in a conducive environment (eg freedom from discomfort and distractions, sufficient time);
    - that assessors have a robust understanding of success criteria (eg mark schemes, training sessions);
    - that assessors are made aware of, and fully understand, any modifications that need to be made to success criteria (within a session, or from one session to the next);
    - that assessors are aware of potential judgemental biases and of strategies for ameliorating them;
    - that adequate mechanisms exist for ensuring consistency of application (eg standardisation mechanisms, moderation mechanisms); and
    - that assessors exercise diligence in capturing/recording judgements (eg proformas, checking procedures).

In addition to step-specific questions, like these, various generic, process-related questions will be relevant to any of the 5 steps, including the following examples:

- What professional standards, principles or guidelines are followed when delivering this process?

- How are different roles coordinated and managed to ensure the effective delivery of this process?

- What steps are taken to ensure that resources required to deliver this process are effectively anticipated and provided?

- How are the credentials of those responsible for delivering this process assured, both in terms of expertise and integrity (eg qualifications, experience, track record), and how is expertise updated over time?

- How are the steps involved in delivering this process formally documented, and how are those documents controlled and managed?

- What training, guidance and supervision is provided for those responsible for delivering this process?

- How is delivery progress and delivery quality monitored on an ongoing basis, and how is this monitoring information used to ensure that progress and quality targets are met?

- What other steps are taken to quality control and/or assure the delivery of this process?

- What safeguards are in place to prevent human error (ie maladministration) during the delivery of this process, what steps are taken to mitigate its impacts when it occurs, and how are negative impacts recorded and reviewed?

- What safeguards are in place to prevent subversion (ie malpractice) during the delivery of this process, what steps are taken to mitigate its impacts when it occurs, and how are negative impacts recorded and reviewed?

- What steps are taken to avoid conflicts of interest arising during the delivery of this process, and what steps are taken to prevent negative impacts from arising when conflicts are impossible to avoid?

- Where electronic data are captured/stored during the delivery of this process, what steps are taken to ensure accuracy, completeness, security and confidentiality of those data?

- What steps are taken to ensure the security and confidentiality of hard copies of materials and data during the delivery of this process?

- Where control over delivering any aspect of this process is devolved to an external agency or agent, what mechanisms are established:
    - to ensure that they have the expertise and integrity to deliver to an appropriately high standard?
    - to secure an enforceable agreement to follow specified delivery procedures?
    - to provide sufficient guidance on procedural compliance and delivery standards?
    - to monitor procedural compliance and delivery standards?
    - to take corrective action if necessary?
- How are deliverables from this process fed into subsequent processes?

Generic questions like these are not actually specific to assessment processes. Yet, the lack of a good answer to any of these questions is likely to reveal a significant threat to validity.[13] If, for instance, assessment results are used to judge teaching effectiveness, and those teachers are responsible for assessing their own students, and if there are no safeguards in place to prevent subversion during the evaluation step, then this highlights a major threat to the validity of results arising from conflict of interests. It is likely that at least some of those teachers will ‘play the system’ by unduly inflating outcomes for their own students.

Of course, description is just part of the foundation for micro-validation research. At a deeper level, micro-validation involves understanding the features and processes that comprise an assessment procedure in the way that an engineer understands the features and processes that comprise an engine, including their:

- core function (the role they play in the assessment procedure)
- core characteristics (how their design details help to secure validity)
- key vulnerabilities (predictable threats to validity)

[13] Bear in mind that a ‘good answer’ might, conceivably, involve *not* having established a particular feature or process. For instance, if there is no explicit mechanism for establishing that assessment tasks are completed by the right candidates, then this raises a potential threat to validity and an onus of responsibility to be able to justify this lack of control. However, there might well be a plausible justification. For example, there may be no need for a formal authentication process when assessment tasks are administered by candidates’ teachers who know them well and who can be trusted. If so, then this justification becomes a key component of the argument supporting the elicitation step claim.

- key safeguards (how additional controls help to minimise those threats)
- compromises (indicating where to keep a watchful eye during the delivery stage).

From this perspective, the essence of micro-validation is the ability to **justify** why a particular feature or process has been incorporated within the assessment procedure, in terms of its contribution to validity. As explained above, the task of justification is both empirical and logical:

- Where is the evidence of design efficacy? What features and processes have been established, and what information is available concerning their operation (eg routine formative analyses, quality control metrics, auxiliary investigations)?
- What are the warrants underpinning the inclusion of those features and processes – ie what are their underlying design logics – and where is the backing for those warrants (eg what does the literature have to say)?

Complementary perspectives on validation argument

As noted above, our core validation research question concerns the degree to which evidence and analysis is consistent with the overarching measurement claim: that it is possible to measure the target proficiency accurately using assessment results. It can be helpful to think of micro-validation research constructing and interrogating a validation argument from the bottom-up (perhaps scaffolded using the 5-step lifecycle model) and to think of macro-validation research interrogating the same argument from the top-down. These complementary perspectives on validation argument are embodied in the following propositions:

- if the assessment procedure has been designed effectively, then it ought to be possible to measure the target proficiency accurately (micro-validation); and
- if the target proficiency has been measured accurately, then certain implications – concerning assessment results, interpretations, uses and consequences – ought to follow (macro-validation).

Micro-validation therefore considers the degree to which evidence and analysis is consistent with a claim concerning effective design (and, by implication, accurate measurement); whereas macro-validation (directly) considers the degree to which evidence and analysis is consistent with a claim concerning accurate measurement.

Figures 6 and 7 incorporate the kind of Toulmin Diagram presented in Figure 4 to illustrate the ways in which evidence and analysis might be organised within a macro-validation argument (Fig. 6) and a micro-validation argument (Fig. 7). For the sake of illustration, they consider the example of a test that has been designed to measure reading comprehension for pupils at the end of their primary school phase.

It is important to appreciate that these examples are presented at a very high level, simply to illustrate the idea of constructing an argument network on the basis of component arguments. For instance, there are only two component arguments within each of these figures; so, the fact that the long, blue horizontal line projects far to the right of each figure suggests that many more component arguments might be added, each one providing additional evidence and analysis. Similarly, within each component argument, only a single datum, warrant, and potential rebuttal is elaborated; whereas, in practice, each one of the component arguments could be broken down into its own argument network. Consequently, the point of these figures is to illustrate the most important principles of validation argument construction:

- validation involves bringing a variety of sources of evidence and analysis to bear on the overarching measurement claim; and
- each one of these sources needs to be critically evaluated, rather than simply taken for granted.

Incidentally, it is not the schematic nature of these figures that is important, here – in other words, it is not that it is necessary to *present* outcomes from validation research like this – it is simply that we need to *think* of validation in this way, ie as constructing a network of arguments to support an overall judgement of validity.

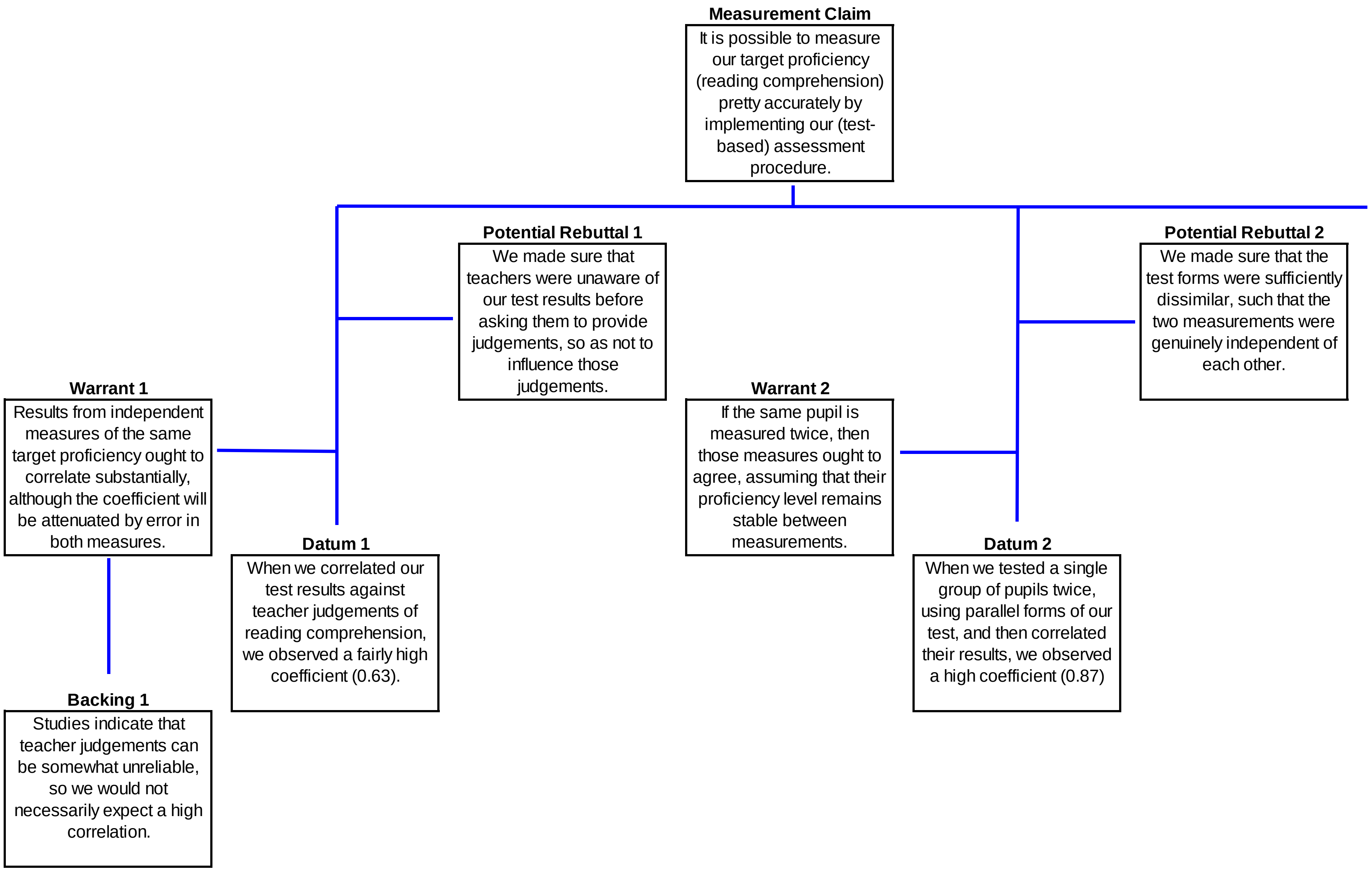
On the basis of the evidence and analysis presented in Figure 6, the provisional measurement claim concludes that it is possible to measure reading comprehension pretty accurately on the basis of results from our test. This is a provisional conclusion in the sense that it seems to be consistent with the evidence and analysis presented so far; although subsequent evidence and analysis might force us to conclude differently. We conclude that it is possible to measure ‘pretty’ accurately, on the basis of a high coefficient from the reliability study and a reasonably high coefficient from the concurrent validation study (acknowledging that our concurrent validation criterion measure, teacher judgement, was not as reliable as we would have liked it to have been). If, instead, we had observed somewhat lower correlation coefficients from both studies, then we might have modified our conclusion; perhaps to include the term ‘moderately’ rather than ‘pretty’ accurately. For both component arguments within Figure 6, the Potential Rebuttal boxes indicate that alternative explanations for positive outcomes from each study were anticipated, and that steps were taken to rule them out in advance, by introducing effective methodological controls.

On the basis of the evidence and analysis presented in Figure 7, the provisional claim for step 2, the elicitation claim, concludes that the reading comprehension proficiency specification is fairly faithfully represented in the performance profile, for candidates from our target population. Again, this is merely a provisional conclusion, because all sorts of additional evidence and analysis could be brought to bear on the elicitation claim. We conclude that the proficiency specification is ‘fairly’ faithfully represented, because the evidence elicited by test questions seemed to be relevant to the reading proficiency construct, albeit not entirely representative of it. To elaborate upon the information provided within the Datum 1 box: the test failed to include any non-fiction texts; whereas the proficiency specification placed substantial weight upon the non-fiction genre.

Notice that Potential Rebuttal 2, from Figure 7, was not entirely addressed: we were only able to conduct this investigation with a relatively small number of pupils from high-attaining schools, so our results might not generalise to pupils in low-attaining schools. A concern like this might motivate an additional evidence-gathering exercise; for example, we might seek consequential evidence concerning the teaching of reading comprehension in low-attaining primary schools. If it turned out that teachers in these schools were routinely drilling their pupils in how to answer reading comprehension questions correctly by applying formulaic strategies, ie without applying the cognitive processes that question writers intended to target, then this would seriously challenge the elicitation claim; at least for pupils from this subset of the population. In the short term, one way of dealing with this challenge, again using an idea from Toulmin, would be to **qualify** the elicitation claim conclusion: the reading comprehension proficiency specification is fairly faithfully represented *in the performance profile of pupils from high-attaining schools, but not necessarily in the performance profile of candidates from low-attaining schools*. Ultimately, though, evidence of this sort should motivate the testing agency to redesign their assessment procedure, to make it more resistant to this kind of strategic subversion.

Finally, note that the macro-validation measurement claim within Figure 6 is also provisional in the sense of still needing to be reconciled with the full set of micro-validation arguments. This is required in order to make an **integrated evaluative judgement** on the basis of all of the relevant evidence and analysis collated to date. For instance, the fact that micro-validation research revealed significant construct underrepresentation (in step 2) might require us to modify the provisional macro-validation conclusion, which asserted the potential for ‘pretty’ accurate measurement; perhaps, by downgrading it to the potential for ‘moderately’ accurate measurement.

Figure 6. High-level (partial) characterisation of a macro-validation argument structure



Measurement Claim

It is possible to measure our target proficiency (reading comprehension) pretty accurately by implementing our (test-based) assessment procedure.

Potential Rebuttal 1

We made sure that teachers were unaware of our test results before asking them to provide judgements, so as not to influence those judgements.

Potential Rebuttal 2

We made sure that the test forms were sufficiently dissimilar, such that the two measurements were genuinely independent of each other.

Warrant 1

Results from independent measures of the same target proficiency ought to correlate substantially, although the coefficient will be attenuated by error in both measures.

Warrant 2

If the same pupil is measured twice, then those measures ought to agree, assuming that their proficiency level remains stable between measurements.

Datum 1

When we correlated our test results against teacher judgements of reading comprehension, we observed a fairly high coefficient (0.63).

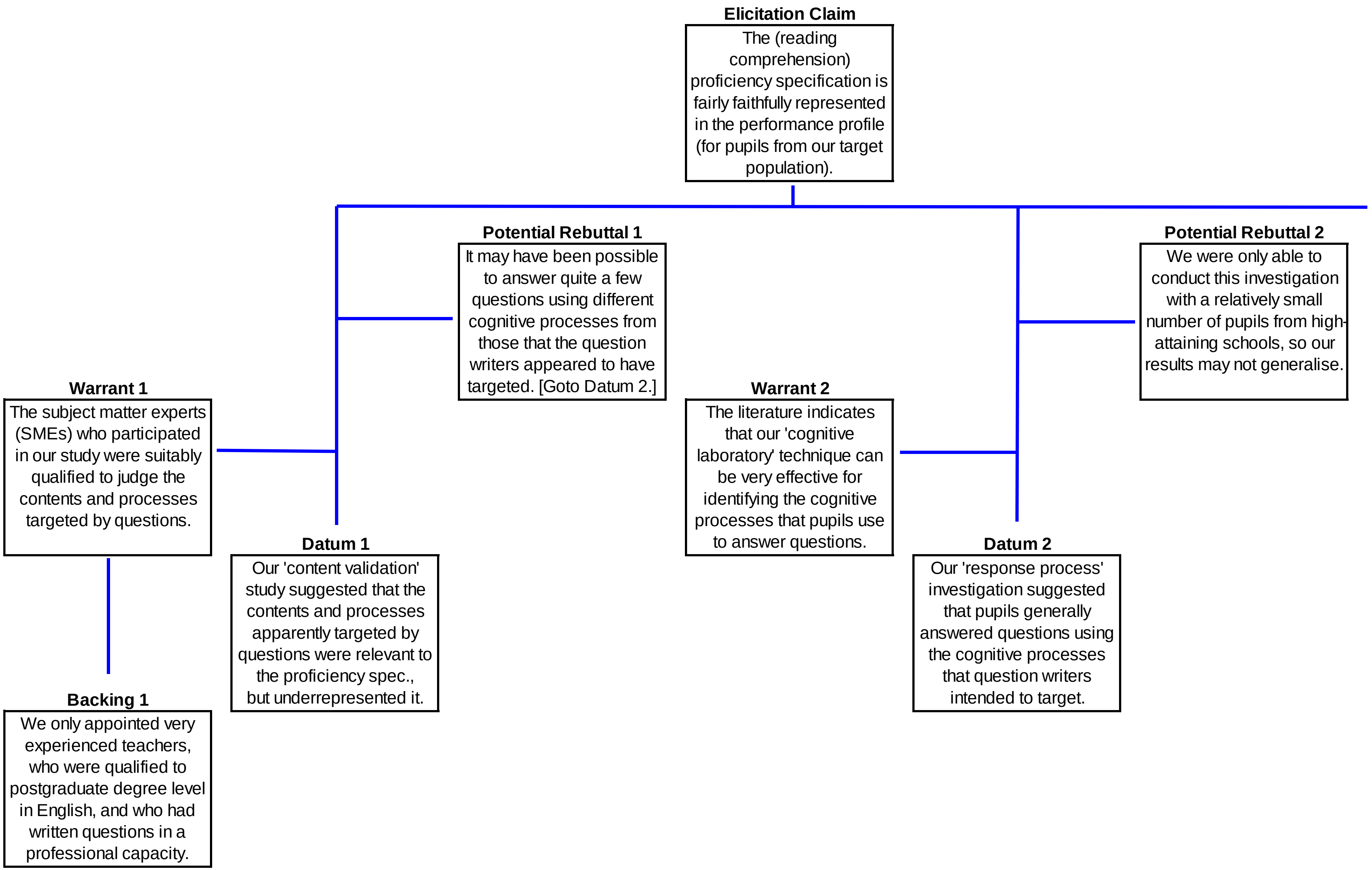
Datum 2

When we tested a single group of pupils twice, using parallel forms of our test, and then correlated their results, we observed a high coefficient (0.87)

Backing 1

Studies indicate that teacher judgements can be somewhat unreliable, so we would not necessarily expect a high correlation.

Figure 7. High-level (partial) characterisation of a micro-validation argument structure



Elicitation Claim

The (reading comprehension) proficiency specification is fairly faithfully represented in the performance profile (for pupils from our target population).

Potential Rebuttal 1

It may have been possible to answer quite a few questions using different cognitive processes from those that the question writers appeared to have targeted. [Goto Datum 2.]

Potential Rebuttal 2

We were only able to conduct this investigation with a relatively small number of pupils from high-attaining schools, so our results may not generalise.

Warrant 1

The subject matter experts (SMEs) who participated in our study were suitably qualified to judge the contents and processes targeted by questions.

Warrant 2

The literature indicates that our 'cognitive laboratory' technique can be very effective for identifying the cognitive processes that pupils use to answer questions.

Datum 1

Our 'content validation' study suggested that the contents and processes apparently targeted by questions were relevant to the proficiency spec., but underrepresented it.

Datum 2

Our 'response process' investigation suggested that pupils generally answered questions using the cognitive processes that question writers intended to target.

Backing 1

We only appointed very experienced teachers, who were qualified to postgraduate degree level in English, and who had written questions in a professional capacity.

Planning a research programme

A comprehensive validation research programme would address the overarching measurement claim, for any particular qualification, from both a micro-validation perspective and a macro-validation one; which represents a substantial undertaking. Indeed, a comprehensive validation research programme is appropriately understood as an ongoing, ie never-ending, one: partly because it takes time to address each of the links in the argument chain, making validation a gradual, cumulative activity; but also partly because each link in the argument chain will need to be revisited, from time to time, as the contexts of qualification delivery change.

The most obvious practical challenge, when envisaging a comprehensive validation research programme like this, is that resources will always be limited. Even the most high profile educational assessment is unlikely to receive anywhere near the kind of funding that would support a Rolls Royce programme on a truly comprehensive scale. However, that is not an excuse for conducting no validation research! Nor is it an excuse for conducting only the cheapest research projects, or for choosing research priorities arbitrarily, or for focusing research on sources of evidence/analysis least likely to undermine the measurement claim. Perhaps the most important function of a validation argument is to help the evaluator to avoid accusations of randomness or bias in planning validation research.

The good news is that the argument-based approach to validation recommends a variety of useful heuristics for prioritising research, for instance:

1. document what you have already done;
2. build upon the literature;
3. capitalise on common features/processes;
4. invest in powerful evidence/analysis;
5. target likely weaknesses;
6. target novelty; and
7. avoid classic fallacies.

Each of these heuristics is explained in more detail below.

Document what you have already done

Even if only implicitly, micro-validation research commences as soon as we begin to design an assessment procedure, and then extends from qualification design into qualification development, delivery and review. If the procedure has been designed effectively, then it will have sufficient validity. However, if neither its logic nor any available evidence/analysis concerning its efficacy is routinely documented, then the substantive basis for justifying that claim will remain implicit and inaccessible. From a validation perspective, this is clearly a missed opportunity.

It is good practice to document any procedure that will need to be replicated from one occasion to the next. It avoids undue reliance upon human memory, including the risk of key players leaving an organisation without passing on their expertise, and it also facilitates accountability by making the procedure transparent. Documentation for micro-validation purposes goes beyond simply documenting the procedure, to explaining why the procedure takes its particular form, linked to any available evidence/analysis in support of that explanation. This kind of documentation is necessarily more time-consuming in the short run; but in the longer run it further facilitates accountability and makes it far easier for future designers to reengineer the qualification in response to changing contexts.

Build upon the literature

When it comes to micro-validation research, there is no need to reinvent the wheel. Certain principles and practices of assessment design are well-established and well-understood. For principles and practices like these, the foundation for a persuasive argument is likely to be provided by an authoritative text, within which the hard work of interrogating design logic via evidence/analysis of design efficacy has already been done. There will always be a certain amount of adaptation to the particular context in which the principle or practice is applied. And this adaptation will require additional justification, beyond the foundation provided by the literature. However, it will almost always be a matter of building upon the literature, rather than starting from scratch.

It is critically important, therefore, that any awarding organisation is able to access that literature when necessary. Unfortunately, when it comes to educational measurement, many of the most authoritative texts are North American and reflect a context of application that has traditionally been characterised by high reliance on testing, multiple-choice testing in particular. Consequently, some of the principles and the practices explicated via these texts do need to be adapted for alternative contexts. Three of the most authoritative North American texts include:

> American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing* (6th edition). Washington, DC: American Educational Research Association.
>
> Brennan, R.L. (2006) (Ed). *Educational Measurement* (4th edition). Washington, DC: American Council on Education/Praeger.
>
> Lane, S., Raymond, M.R. and Haladyna, T.M. (2016) (Eds.). *Handbook of Test Development* (2nd edition). New York: Routledge.

All of these texts provide insights into the design, development, delivery and review of a wide range of assessment formats, despite the test-dominated context. The second and third provide a large number of references to other authoritative texts.

Capitalise on common features/processes

The suggestion that validation research is required for every single qualification offered by an awarding organisation might sound extreme, if not implausible, bearing in mind the scope of a comprehensive validation programme. Although it is true that at least some research will be required for every single qualification – for instance, it is hard to see how an awarding organisation could develop a valid qualification without properly researching its target proficiency – it is also true that many features and processes will be common across a wide range of qualifications. For features and processes like these, it should be possible to specify a common design logic and to produce evidence/analysis of design efficacy by sampling across qualifications, rather than for each one individually. Capitalising on this kind of generic research project would be similar to building upon the literature. Likewise, extending conclusions from generic research to specific qualifications might require additional justification, but it provides a more tractable solution than conducting the research anew for every single qualification.

Invest in powerful evidence/analysis

Power can be understood as the potential of a source of evidence/analysis to support or to undermine the overarching measurement claim. In general, the potential of any source to undermine a claim (falsification) will the greater than its potential to support one (verification). This is because there is more to high quality measurement than can be established using a single source of evidence/analysis; which means that supporting evidence/analysis can only ever provide a partial thumbs-up. Conversely, single sources of evidence/analysis do sometimes have the potential to provide a total thumbs-down, by demonstrating failure to satisfy critical criteria.

Once upon a time, the principal empirical indicators of technical quality, for educational measurement, were defined like this:

> Reliability has been regarded as the correlation of a given test with a parallel form. Correspondingly, the validity of a test is the correlation of the test with some criterion.

(Gulliksen, 1950, p.88)

There seems to be an implication, here, that the combination of these two sources – parallel forms reliability and criterion validity – would be sufficient to determine the technical quality of a test, which would make them incredibly powerful sources, indeed.

Even today, it still seems fair to conclude that both of these approaches are capable of providing powerful evidence; at least in theory. Concurrent validation, which is a kind of criterion validation, is particularly promising. It involves correlating results from the qualification under scrutiny with results from the 'best possible' measure of its target proficiency, which is therefore treated as a yardstick. A perfect correlation would strongly suggest that the qualification was measuring exactly the same thing as the yardstick measure, ie the target proficiency, in exactly the same way. This comes as close as possible to being able to provide a total thumbs-up on the basis of validation research. Of course, the problem – which perhaps seems more obvious now than it did in the 1920s/30s when it was the preferred validation technique – is that the validity of the criterion measure puts an upper limit on the power of this kind of evidence. Indeed, the constant challenge of low validity criterion measures – which became known as the criterion problem – has been widely acknowledged in the literature since the 1940s. Although concurrent validation does have the potential to provide fairly powerful evidence, when a plausible criterion measure can be found, it is often not possible to identify plausible criterion measures, particularly in educational contexts. As such, the approach is more powerful in theory than in practice.

High reliability coefficients – arising from parallel forms reliability studies, but also from approaches that approximate them, such as the split-half technique and Cronbach's Alpha – are often assumed to provide a powerful thumbs-up for an assessment procedure. However, this is not really true, because the easiest route to high reliability is by compromising other characteristics that are critical to validity, such as construct representation; in other words, high reliability often indicates low construct representation. Conversely, very low reliability coefficients do provide a powerful thumbs-down for an assessment procedure. They indicate that it is not possible to measure anything, by implementing the procedure.

Both concurrent validation studies and reliability analyses are closer to the macro-validation end of the continuum. Yet, many sources of micro-validation evidence/analysis are also potentially powerful. For instance, evidence of a widespread lack of support for the qualification amongst key stakeholders, indicating their belief that the proficiency specification fails to represent what that qualification really needs to measure, potentially undermines the measurement claim; providing powerful evidence that the qualification will not enable those who use its results to measure what needs to be measured, ie the target proficiency. In a similar way, evidence of widespread cheating amongst candidates potentially undermines the measurement claim; providing powerful evidence that qualification results are unlikely to be accurate.

Target likely weaknesses

The argument-based approach to validation reminds us that any argument is only as strong as its weakest link. It therefore makes sense for validation research to target likely weak links in an argument chain, to make sure that they are not so weak as to undermine it. It might sound counter-intuitive for an awarding organisation to target research upon likely weak links – assuming that it wants to 'validate' its qualifications – but from a cost-effectiveness perspective it does make sense. After all, if the research suggests that the weak link does undermine the validation argument, then it suggests that the qualification is not effective; which means that the organisation is selling a sub-standard product each time it delivers results for that qualification. More constructively, the kind of micro-validation research that is used to investigate potential weak links will typically also provide diagnostic insights into how that feature or process can be made more effective, facilitating its redesign.

The literature is probably the best place to look for insights into potential weak links. Different approaches to designing an assessment procedure will be associated with different strengths and weaknesses. For example, multiple-choice tests tend to be stronger in terms of sampling breadth and weaker in terms of authenticity; whereas the converse tends to be true for complex performance assessments. It is also important to appreciate that an assessment procedure might have different strengths and weaknesses when administered with different groups of candidates, or when administered in different contexts. Tasks with a high reading load will be especially challenging for candidates with dyslexia, and might indicate a weak link in the argument chain that needs to be addressed, eg via extra time. Similarly, task controls that are satisfactory under low stakes conditions might prove to be unsatisfactory under high stakes conditions. And so on.

Target novelty

The more novel an assessment feature or process, the less information it will be possible to glean from the literature concerning its strengths and weaknesses. It therefore makes sense for validation research to target novelty. Although it is obviously important to evaluate novel features or processes, eg the use of vlogs for capturing assessment evidence, it is just as important to evaluate novel configurations of well-established features and processes. For example, if a traditional mastery test, with a high pass mark, were to be transformed into a graded test, then serious consideration would need to be given to its redesign. It would not be a matter of simply adding additional cut-scores above the pass mark. The test blueprint would need to be designed quite differently. Similarly, the interpretation of test grades might need to be conceived quite differently; if, for instance, it was no longer feasible to interpret the pass mark as a mastery threshold. Ultimately, the role of the test within the qualification as a whole would need to be re-evaluated.

Avoid classic fallacies

Michael Kane's important article on *Validating the interpretations and uses of test scores* (Kane, 2013) identified a number of classic reasoning fallacies, relevant to the task of prioritising projects within a validation research programme. The begging-the-question fallacy occurs when critical assumptions or inferences in the argument chain are simply taken for granted. In the context of validation research, it is often simply taken for granted that it is legitimate to extrapolate from conclusions based upon assessment evidence to conclusions concerning proficiency in the real world. However, the less authentic the assessment evidence elicited, the less likely this may be, and the more the question of extrapolation is begged. The extrapolation inference corresponds to a specific sub-claim within the generic argument structure popularised by Kane (following the scoring inference and the generalization inference). In the 5-step lifecycle model, extrapolation legitimacy is primarily a function of the faithfulness of the proficiency specification to the target proficiency (step 1), and the faithfulness of the performance profile to the proficiency specification (step 2).

Another classic fallacy that ought to be avoided is known as gilding-the-lily, ie unnecessary embellishment of a particular claim within the validation argument, by presenting more evidence and analysis than is needed. Although this might seem harmless enough, it can be problematic when it creates a spurious impression of argument strength and, of course, because it wastes precious validation resources. It is particularly pernicious when it masks the fact that other links in the argument chain are weak. In the context of validation research, it can be tempting to fill a report with all sorts of technical analyses bearing upon essentially the same source of evidence, eg multiple indices of reliability, whilst totally ignoring other sources of evidence, eg concerning construct representation. This would be particularly pernicious if, during the design phase, construct representation had intentionally been sacrificed for the sake of reliability.

How to begin

How does an ant eat an elephant? One mouthful at a time. The same principle applies to validation research. If, as suggested, a comprehensive validation programme is a never-ending one, at least as long as a qualification continues to be delivered, then the idea of beginning to plan such a programme will inevitably seem daunting, very daunting. Under circumstances like these, the only sensible advice is to start small and to scale up. But do start!

Sufficient validity

There is no such thing as perfect validity. No assessment procedure can be relied upon to produce results that are completely accurate and entirely useful. Assessment is a technology that helps us to make everyday attributions concerning levels of proficiency as unambiguously as possible. But it is an inherently imprecise technology all the same. Although steps can be taken to identify and to remedy a reasonable proportion of the 'human errors' that occur during the delivery stage – for example, by establishing a mechanism for appealing against results – a substantial amount of the 'measurement inaccuracy' that arises during the delivery stage will not be detectable and will therefore remain (see Newton, 2005). Within any set of assessment results, a sizeable proportion will be inaccurate.

Just as importantly, because assessment is an everyday technology, which has to operate in the context of unavoidable real-world constraints, assessment procedures are never designed to maximise validity. Instead, the pragmatic objective underlying assessment design is to **optimise** validity; typically, to accommodate a broad profile of intended purposes, and to recognise a wide range of operational constraints (Newton, 2017b). Indeed, from a regulatory perspective, Ofqual's goal is simply to ensure that qualifications have sufficient validity.

From this perspective, it should be clear that assessment design is fundamentally concerned with trade-off and compromise. All sorts of trade-offs are typically made during assessment design, for instance:

- the desire to increase the reliability of results by increasing the number of assessment tasks, versus the ability of candidates to sustain concentration and effort when the duration of an assessment event is too long;
- the desire to measure complex skills authentically using tasks that mirror real-world situations, versus the ability of assessors to evaluate complex performances with consistent accuracy; and
- the desire to measure all of the elements identified within a proficiency specification, versus the ability of an assessment community even to reach consensus over criteria for judging certain elements.

Designers make different trade-offs and compromises for different qualifications. However, whatever trade-off or compromise is made, it is important that an awarding organisation is able to rationalise each decision, and to understand its likely impact upon validity.

So, how much validity is sufficient validity? Although this might sound like an obvious question to ask, and one that ought to have a straightforward answer, there are all sorts of reasons why its answer is very far from straightforward. First, although we characterise validity quantitatively, as a property that comes in degrees, we can only quantify it impressionistically, using categories like ‘very low’ or ‘low’ or ‘moderate’ or ‘high’ (or other such terms). Although it might seem reasonable to describe degrees of validity using terms like these, any decision over which of these terms to apply to a particular assessment procedure will be highly complex. As noted earlier, it requires what scholars have described as an integrated evaluative judgement of all of the very many different sources of evidence and analysis that can be brought to bear. And bear in mind that different evaluators might well reach different judgements, even on the basis of the same corpus of evidence and analysis.

Second, although the present report characterises validity as fundamentally technical – a measurement concept, tantamount to measurement quality – the idea of *sufficient* validity is fundamentally social. In other words, the grounds for deciding how much validity is sufficient validity are primarily consequential; concerned with the positive and negative impacts arising from implementing an assessment procedure. There will always be negative impacts arising from the use of any set of assessment results; if only because some of those results will inevitably be inaccurate, and those inaccurate results will typically lead to incorrect decisions, and those incorrect decisions will typically have inappropriate consequences. Equally though, there will be positive impacts for individuals, organisations and society more generally, when accurate results lead to correct decisions and appropriate consequences. Part of sufficient validity, then, is weighing-up the likelihood of correct decision-making against the likelihood of incorrect decision-making. However, the severity of the consequences of those decisions needs also to be taken into account; and we might even choose to weight the consequences of incorrect decisions more heavily than the consequences of correct ones.

Third, as the decision concerning sufficient validity is essentially a judgement of values, this raises the question of whose values ought to be taken into account. In other words, who gets to judge sufficient validity, and through what due process? There are no absolute right answers to questions like these. For regulated qualifications in England, it seems reasonable to conclude that Ofqual is legally empowered to make decisions concerning sufficient validity, and part of our due process requires that judgements are made with regard to our *General Conditions of Recognition* alongside associated guidance.[14] Ultimately, judgements concerning sufficient validity must be influenced by the court of public opinion, to the extent that qualification systems are established to serve the interests of society.

Fourth, sufficient validity is not simply a matter of degree, it is also a matter of the way(s) in which an assessment procedure departs from the (unattainable) ideal of

[14] See https://www.gov.uk/government/publications/general-conditions-of-recognition (accessed 26 July 2017)

perfect validity. We might, for instance, conclude that a particular qualification has a pretty high level of validity, all things considered. Yet, if it failed to represent a certain element of its target proficiency, and if that element turned out to be especially important for making a certain kind of decision, then it might be unreasonable to conclude that the qualification had sufficient validity; at least, in relation to that that kind of decision. In other words, whether it is acceptable to use qualification results for a particular purpose is not simply a matter of how valid that qualification is, in an overall sense; it is also a matter of exactly how that qualification lacks validity.

Finally, referring back to the three examples of trade-offs listed above, note how they could all be described as *intra*-validity trade-offs; because they involve trading-off a validity facilitator against a validity threat. It is important to appreciate that other kinds of trade-off are also made during assessment design, which could be described as *extra*-validity trade-offs; because they involve trading-off a validity facilitator against a pragmatic constraint. For example:

- the desire to increase the reliability of results by embedding a large number of assessment tasks in lesson time, ie presenting them as work-of-the-course, versus the time that those coursework tasks would take away from teaching and learning
- the desire to maximise the discriminative power of test/exam results by eliminating questions that many candidates would get right, versus the credibility of low pass marks (that hard tests/exams tend to require) and the negative experience that hard tests/exams inflict on weaker candidates
- the desire to increase the reliability of results by evaluating the same performances multiple times, and averaging those marks, versus the inability to appoint enough qualified assessors.

The inevitability of trade-offs like these remind us that validity is only part of the story. It is just one dimension of the overall acceptability of a qualification, which is the focus of the next section.

Beyond validity

In terms of criteria for judging educational assessments, validity is just one criterion within a multiplicity of concerns, which might be grouped under the broad heading of **acceptability**. Validity focuses our attention on the assessment procedure itself, and its potential to support accurate and useful measurement; that is, whether or not the qualification *can* be used to measure a target proficiency. Acceptability focuses our attention on the decision to implement that assessment procedure; that is, the decision concerning whether or not the qualification *should* be used. We might define acceptability as the strength of the argument in favour of implementing a particular assessment procedure, ie delivering a particular qualification, when considered from a broad, societal perspective. In other words, all things considered, is it ok for an awarding organisation to offer a qualification like this?

Whereas validity is the primary **design driver**, acceptability – or, more specifically, the threat of *un*acceptability – foregrounds a series of **design constraints**. Thus, once again, the pragmatic objective underlying assessment design is not to maximise validity, but to optimise it, in relation to constraints like these. Good practice in assessment design involves **prospective evaluation** of how the procedure is likely to operate in the real world; in particular, to anticipate likely negative impacts. For instance, might implementing the procedure:

1. exert (unacceptable) pressure on limited resources?
2. breach any laws?
3. have (unacceptable) negative educational consequences?
4. have (unacceptable) negative political consequences?
5. have (unacceptable) negative ethical consequences?
6. undermine its own credibility, or the credibility of other procedures/systems?

Resource availability

Resource availability is partly an economic issue, concerning direct and indirect costs. As with any product, higher quality (ie higher validity) is normally associated with higher costs; so it is important to question whether the quality that is desired is achievable within limited budgets. Resource availability also concerns other resources too, including time and expertise. Sometimes, these concerns can be reduced to economic issues by increasing the availability of funds; but not always.

Example. A critical question for any assessment designer is whether their preferred design is cost-effective. For instance, the increased authenticity of complex, performance assessments is desirable, from a validity perspective, when compared with multiple-choice tests. Yet, basing large-scale educational assessments upon short-answer questions and essay questions – as opposed to more ‘efficient’ formats that can be machine-scored – raises serious resource challenges; particularly in a context whereby teachers/trainers are not legally required to mark such tests or exams as part of their teaching contract. The challenges are legion, in fact. Given the standard requirement for rapid turn-around of results, there is limited time available to achieve high quality marking. This is exacerbated by the need to build-in time for appealing against results. Yet, to mark short-answer and essay questions is very time consuming, and requires many experienced markers. Experienced markers constitute a very limited resource in their own right. They are not simply ‘there’ to be enticed by sufficiently high pay, even though higher pay might help to increase their availability in the longer term.

Legal compliance

Beyond sufficient validity, awarding organisations in England need to ensure that their qualifications are compliant with all sorts of legislation. Fairly obviously, they need to comply with the Equalities Act 2010 (so as not to disadvantage members of protected groups) and the Data Protection Act 1998 (so as not to compromise the security or confidentiality of personal data). Less obviously, they need to comply with health and safety legislation as well as competition law.

Example. In England, it would be illegal for a qualification, such as a professional licensing test, to cause harm by presenting unnecessary barriers to people from protected groups, including people with disabilities. It would *not* be illegal for a driving instructor licensing test to include a performance assessment that required candidates to observe a learner driver; even though this would present an obvious barrier to a blind person. This is because being able to observe a learner driver is a necessary requirement of being a driving instructor. However, if a piano tuner licensing test were to include a written exam, then this might be a different matter. Even if the exam tested knowledge, skill and understanding of direct relevance to piano tuning – and might otherwise be considered an appropriate mechanism for tapping those elements of the target proficiency – the written format would present an obvious barrier to a blind person, *requiring competence beyond that necessary for being a piano tuner*. To comply with the Equalities Act, this assessment format challenge might need to be addressed via access arrangements or reasonable adjustments, eg by providing a Braille version of the task and by permitting an alternative mode of responding.

In this example, the barrier also impacts upon validity; because providing the accommodation would also improve result accuracy for blind people. However, if there were only small numbers of blind candidates, then the overall impact on validity would be negligible. The primary issue here is one of fairness and, in this case, the law.

Educational alignment

Qualifications do not operate in a vacuum, independently of other educational concerns. The four pillars of education – curriculum, teaching, learning and assessment – need to operate in synergy with each other. It is especially important that assessment design decisions – however sensible from a validity perspective – do not impact unduly upon curriculum, teaching or learning in such a way as to threaten the acquisition of the very learning outcomes that the qualification is supposed to certify.

Example. There are all sorts of challenges associated with educational alignment, which are often described in terms of the **backwash** of assessment upon curriculum, teaching and learning. In English primary schools, the impact of testing science at the end of primary schooling provides a good example of positive, intended backwash; whereby it effectively engaged teachers with teaching national curriculum science, in response to concerns that science was not being taught effectively in primary schools. However, there are many examples of negative, unintended backwash, which indicate a lack of alignment between assessment design and broader educational concerns.

For instance, certain design features associated with vocational assessment in England have recently been criticised for their negative impacts upon curriculum, teaching and learning; notably the way in which target proficiencies tend to be specified using long lists of learning outcomes and associated criteria, all of which need to be achieved for the award of the qualification. Specifying long lists of learning outcomes is good, from a validity perspective, in terms of ensuring that all elements of the target proficiency are covered. However, Alison Wolf has argued that the need to satisfy *all* assessment criteria drives down curriculum standards, because each criterion needs to be accessible to all candidates (Wolf, 2011). On a different note, Doug Richard has argued that the atomistic approach to specifying target proficiencies for apprenticeships leads to an atomistic approach to assessment, which leads to an atomistic approach to teaching and learning. This can result in apprentices having each of their assessment criteria ticked off, by the end of their apprenticeship, but still not being fully competent or genuinely employable. This is problematic, from an assessment perspective, but it is highly problematic from a learning perspective and from a broader societal perspective. On a similar note, Richard has also argued that apprentices spend too much of their ‘training time’ being assessed, and not enough being trained (Richard, 2012).

Policy alignment

Just as qualifications do not operate in an educational vacuum, they also do not operate in a wider political vacuum. From a broad, societal perspective, it is important that policy and practice in relation to qualifications is aligned with policy and practice elsewhere.

Example. Even remaining within the field of education, it is possible to see this criterion in action, beyond the pillars of curriculum, teaching and learning. Because assessment for formal certification in England has traditionally been the responsibility of external awarding organisations, any proposal to devolve some of this responsibility to teachers/trainers working in schools/colleges – however useful that might be from a validity perspective – would raise issues of workload. The greater the proposed transfer of assessment burden to teachers/trainers, the louder their unions would object, and not without reason when those teachers/trainers are already working at or beyond capacity. Education policy makers are very sensitive to the workload issue, and it is quite possible to see how validity-based assessment design decisions may come into conflict with broader political concerns, eg related to workload.

Moral reputability

Not all instances of unfairness will be addressed by the law, so it is important to establish that qualification uses and impacts are morally acceptable, as well as legally so.

Example. Technical criteria for judging quality often seem to be quite utilitarian, ie framed in terms of the greatest good for the greatest number. For instance, this is implicit in the use of statistical concepts like the mean, mode and median (and derivative statistics) to judge how well questions and question papers function. Even the question of sufficient validity can be answered from a utilitarian perspective. For instance, we might conclude that a qualification has sufficient validity if it is more likely than not to return results which are accurate (although, this is actually quite a low hurdle, which might not pass the test of credibility). But there are other ways of answering the question of sufficient validity, including ethical positions that put substantially more weight on negative impacts than positive ones; thereby raising the sufficient validity hurdle considerably.

Beyond the question of sufficient validity, ethical questions can arise concerning the morality of using results for certain kinds of decision-making. They can arise when it can be argued that the assessment is technically fit to be used for that kind of decision-making; and they can arise when it becomes clear that the assessment will be used routinely for purposes for which fitness has not been demonstrated. In recent years, international awarding organisations have grappled with how to respond to increasing demand for tests that can be used for access, integration, and citizenship. Tests like these do not simply present technical challenges; they also present ethical ones (ALTE, 2016; Council of Europe, 2014).

Public credibility

Credibility and validity often go hand-in-hand, and we might expect the validity of a qualification to be the primary determinant of its credibility. But this is not necessarily so, which is why credibility needs to be considered separately. Like any kind of currency, qualification results only have value when there is widespread consensus within the community that uses them that they are, indeed, valuable. Credibility, therefore, ultimately trumps validity.

Example. In recent years, results from many qualifications have become increasingly high stakes, as they have been put to all sorts of accountability uses. This creates or exacerbates a variety of perverse incentives. If candidates or their teachers/trainers succumb to these incentives – subverting the system via one form of malpractice or another – this raises all sorts of threats. It challenges the moral reputability of the system, as it embodies unfairness, par excellence. It potentially undermines legal compliance too. Furthermore, to the extent that it becomes a social media story, it undermines the credibility of the system in the eyes of the public.

The use of results to judge the effectiveness of schools/colleges and teachers/trainers presents a particular threat to public credibility when teachers/trainers participate in the assessment process, eg when they are responsible for providing coursework marks. Under such circumstances, they are, in a very real sense, being required to judge themselves. It can be hard to build subversion resistance into a procedure that devolves a lot of control over assessment processes to teacher/trainers in schools/colleges.

An approach to understanding validation arguments

This report has outlined an approach to understanding validation arguments. It has outlined 'an approach' in two key senses. First, although there has been considerable scholarly debate on validity during the past few decades, it has tended to be fairly piecemeal and unintegrated. For instance, there has been a lot of debate over the best way to use the word 'validity' but relatively little debate over how the validity concept – howsoever defined – relates to other technical concepts, eg reliability.[15] In addition, some scholars have come to this debate being more interested in defining validity, whereas others have come to it being more interested in supporting validation. Perhaps inevitably, these scholars have tended to be most clear and consistent on matters closest to their own interests and least clear and consistent on matters furthest away. In short, there are few truly comprehensive accounts, which is problematic when so many of the theoretical 'details' can be treated quite differently.

In response to challenges like these, the present report has aimed to present a clear, comprehensive and consistent account of both validity and validation. In other words, it has attempted to draw together the most important insights from the dispersed and divergent literature and to integrate them within a single, coherent narrative. It starts from the idea of educational assessment as measurement; locates measurement at the heart of validity; casts validity as a property of assessment procedures; explains how validation arguments can be framed in terms of the effective design of assessment procedures; recommends a broad perspective on validation evidence and analysis; and finally situates validity within the broader concept of acceptability. Whether or not you have found this presentation entirely persuasive, hopefully you will have found it reasonably coherent!

Second, on that note, you are welcome to take it or leave it. It is just 'an approach' to understanding validation arguments, amongst others. Standing firmly on the shoulders of the giants of the literature, including Lee Cronbach, Samuel Messick, and Michael Kane, it ploughs a slightly different furrow, particularly in its emphasis upon validation-of-design. However, if you do not find this approach useful, then try a different one. The final section on 'resources' identifies a variety of authoritative texts, each of which provides important insights into validity and validation arguments; some of which suggest quite different approaches.

Ultimately, when it comes to validity and validation, no single text contains all of the right answers. Indeed, there are no absolute right answers in this business. Texts

[15] By way of example, the North American *Standards for Educational and Psychological Testing* seems to buy-into a unitary view of validity, implying that validity subsumes reliability; yet, it is structured as a chapter on validity followed by a chapter on reliability. Much of the literature is like this, presenting mixed, or unclear, messages on how the core technical concepts are supposed to relate to each other.

need to be judged in terms of their usefulness, which puts an onus of responsibility upon readers: to read widely; to put a variety of insights into practice; and to decide which ones prove to be most and least useful to them.

References

AERA, APA, and NCME (2014). *Standards for Educational and Psychological Testing* (6th edition). Washington, DC: American Educational Research Association.

ALTE (2016). *Language Tests for Access, Integration and Citizenship: An outline for policy makers.* Cambridge, UK: Association of Language Testers in Europe.

Council of Europe (2014). *Integration Tests: helping or hindering integration*? Resolution 1973 (2014). Final version. (http://www.assembly.coe.int/nw/xml/XRef/Xref-DocDetails-EN.asp?FileID=20481&lang=EN, accessed 28/07/2017)

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

Eraut, M. (2008). How professionals learn through work. (http://learningtobeprofessional.pbworks.com/w/page/15914995/Michael%20Eraut, accessed 28/07/2017)

Fast, E. and Hebbler, S. with ASR-CAS Joint Study Group on Validity in Accountability Systems (2004). *A Framework for Examining Validity in State Accountability Systems*. Washington, DC: Council of Chief State School Officers.

Ferrara, S. (2007). Our field needs a framework to guide development of validity research agendas and identification of validity research questions and threats to validity. *Measurement: Interdisciplinary Research and Perspectives*, 5 (2–3), 156–164.

Ferrara, S. and Lai, E. (2016). Documentation to support test score interpretation and use. In S. Lane, M.R. Raymond and T.M. Haladyna (Eds.). *Handbook of Test Development* (2nd edition) (pp.603-623). New York: Routledge.

Gulliksen, H. (1950). *Theory of Mental Tests*. New York: John Wiley and Sons, Inc.

Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research*, 55 (1), 23-46.

Hanushek, E.A. (2011). Valuing teachers. *Education Next*, 11 (3), 41–45.

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50 (1), 1–73.

Knorr, M. and Klusmann, D. (2015). The trouble with validity: what is part of it and what is not? *Medical Education*, 49 (6), 548–555.

Mislevy, R.J. (2007). Validity by design. *Educational Researcher*, 36 (8), 463–469.

Mislevy, R.J., Wilson, M.R., Ercikan, K. and Chudowsky, N. (2002). *Psychometric Principles in Student Assessment*. CSE Technical Report 583. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Newton, P.E. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal*, 31 (4), 419-442.

Newton, P.E. (2007). Contextualising the comparability of examination standards. In P.E. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.). *Techniques for Monitoring the Comparability of Examination Standards* (pp.9-42). London: Qualifications and Curriculum Authority.

Newton, P.E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives*, 10 (1–2), 1–29.

Newton, P.E. (2016). Macro- and micro-validation: Beyond the 'five sources' framework for classifying validation evidence and analysis. *Practical Assessment, Research & Evaluation*, 21 (12). (http://pareonline.net/getvn.asp?v=21&n=12, accessed 28/07/2017)

Newton, P.E. (2017a). There is more to educational measurement than measuring: The importance of embracing purpose pluralism. *Educational Measurement: Issues and Practice*, 36 (2), 5–15.

Newton, P.E. (2017b). Assessment dilemmas. *Research Intelligence*, 133, 18–20.

Newton, P.E. and Baird, J. (2016). Editorial: The great validity debate. *Assessment in Education: Principles, Policy & Practice*, 23 (2), 173–177.

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

Norwood, O.T. (1975). Male pattern baldness: Classification and Incidence. *Southern Medical Journal*, 68 (11), 1359–1365.

Ollin, R. and Tucker, J. (2012). *The Vocational Assessor Handbook* (5th edition). London: Kogan Page.

Pollitt, A. and Ahmed, A. (2009). *The Importance of Being Valid*. Paper presented at the 10th Annual Conference of the Association for Educational Assessment – Europe. Valette, Malta. November 5th – 9th.

Richard, D. (2012) *The Richard Review of Apprenticeships*. (https://www.gov.uk/government/publications/the-richard-review-of-apprenticeships, accessed 28/07/2017)

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

Shaw, S. and Crisp, V. (2015). Reflections on a framework for validation – Five years on. *Research Matters: A Cambridge Assessment Publication*, 19, 31–37.

Shaw, S.D. and Weir, C.J. (2007). *Examining Writing: Research and practice in assessing second language writing. Studies in Language Testing, 26*. Cambridge: Cambridge University Press.

Shepard, L. (2012). *Evaluating the Use of Tests to Measure Teacher Effectiveness: Validity as a Theory-of-Action Framework*. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, April 14.

Stacey, G. (2014). *Putting Validity at the Heart of What We Do*. Speech to the Federation of Awarding Bodies National Conference, 13-14 October. Marriott Hotel, Leicester.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Weir, C.J. (2005). *Language Testing and Validation: An evidence-based approach*. Hampshire, England: Palgrave Macmillan.

Wolf, A. (2011). *Review of Vocational Education – The Wolf Report*. (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/180504/DFE-00031-2011.pdf, accessed 28/07/2017)

Resources

Successive editions of the ‘bible’ of the educational measurement profession, entitled *Educational Measurement*, have all included a chapter on validity/validation. The three most recent chapters are all classics, in their own way. The most recent chapter is the most accessible. The chapter by Samuel Messick is perhaps the single most important contribution to validity theory… but it is very hard to read!

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.). *Educational Measurement* (4th edition) (pp.17–64). Washington, DC: American Council on Education/Praeger.

Messick, S. (1989). Validity. In R.L. Linn (Ed.). *Educational Measurement* (3rd edition) (pp.13–103). Washington, DC: American Council on Education.

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

For a more succinct and/or up-to-date account of the work of these three validity giants, it would be worth consulting:

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50 (1), 1–73.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18 (2), 5–11.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (Ed.). *Intelligence: Measurement, Theory and Public Policy* (pp.147–171). Urbana, IL: University of Illinois Press.

Surprisingly few books have been published specifically on the topic of validity or validation, although this is beginning to change, for instance:

Wainer, H. and Braun, H.I. (1988) (Eds.). *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum.

Weir, C.J. (2005). *Language Testing and Validation: An evidence-based approach*. Hampshire, England: Palgrave Macmillan.

Lissitz, R.W. (2009) (Ed.). *The Concept of Validity: Revisions, new directions, and applications*. Charlotte, NC: Information Age Publishing.

Taylor, C.S. (2013). *Validity and Validation*. New York: Oxford University Press.

Chatterji, M. (2013) (Ed.). *Validity and Test Use: An international dialogue on educational assessment, accountability an equity*. Bingley: Emerald Group.

Markus, K.A. and Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, causation and meaning*. Hove, East Sussex: Psychology Press.

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

Having said that, quite a few special issues of academic journals have been devoted to validity issues, including these two which capture the longstanding debate over the role of consequences in validation research:

Crocker, L. (1997). Editorial: The great validity debate. *Educational Measurement: Issues and Practice*, 16 (2), 4–4.

Newton, P.E. and Baird, J. (2016). Editorial: The great validity debate. *Assessment in Education: Principles, Policy & Practice*, 23 (2), 173–177.

For slightly different approaches to validation argument, see:

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

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

Mislevy, R.J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk*, 2 (4), 237–258.

Bachman, L.F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2 (1), 1–34.

Bachman, L.F. and Palmer, A. (2010). *Language Assessment in Practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Finally, Ofqual has commissioned a number of useful reports on validation over the past few years, including:

AlphaPlus Consultancy Ltd (2014). *Validation of Vocational Qualifications*. Ofqual/14/5373. Coventry: Office of Qualifications and Examinations Regulation.

Curcin, M., Boyle, A., May, T. and Rahman, Z. (2014). *A Validation Framework for Work-Based Observational Assessment in Vocational Qualifications*. Ofqual/14/5374. Coventry: Office of Qualifications and Examinations Regulation.

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346