CHAPTER 1

VALIDITY AND VALIDATION

Validity is the hallmark of quality as far as testing is concerned, being the 'single most important criterion' for evaluating a test (Koretz, 2008: 215). At this level of generality, the many sub-communities of scientists and practitioners that comprise the field of educational and psychological measurement are probably in agreement. However, when interrogated further, this relatively bland consensus can be shown to conceal many different perspectives on the meaning of validity, reflecting claims at a variety of levels, for example, that:

1 it is possible to measure an attribute accurately using test scores (validity as a measurement concept); or
2 it is possible to make accurate and useful decisions on the basis of test scores (validity as a measurement and decision-making concept); or
3 it is acceptable to implement a testing policy (validity as a concept spanning measurement, decision-making and broader impacts and side-effects).

The first of these alternatives is essentially a technical or scientific claim concerning test score meaning: it was how validity was originally defined in the 1920s. The third is ultimately a social or ethical claim concerning test score use. Plenty of debate in recent years has focused on whether or not ethical concerns have anything to do with validity. The second alternative lies somewhere in-between these extremes, having dimensions that are essentially scientific (accuracy) and essentially ethical (usefulness). We begin this book with these three alternatives simply to highlight the fact that the very nature of validity is heavily contested within the field of educational and psychological measurement, despite nearly 100 years of debate on the topic. Part of our intention is to explain why it remains so contested. Not only is the concept heavily contested, but it also can be very confusing to understand for both novice and expert alike. There are many reasons for this:

- There is a very large, disparate literature on validity within educational and psychological measurement, spanning the best part of a century.
- There are large, disparate literatures on validity within other disciplines that are related only tangentially to the testing concept.
- The 'official' meaning of validity among measurement professionals has evolved over time.
- The concept has gained stature over the decades and expanded correspondingly.
- Some of the most important accounts are extremely hard to read (e.g. Messick, 1989a).
- The term 'validity' is employed in so many different ways, in so many different contexts, that often it is entirely unclear what the speaker intends to convey.

The final point is crucial. It is often unclear whether or not the term is being used in a technical sense; and, if it is being used in this way, whether this is the technical sense of educational and psychological measurement, or of another discipline entirely. So this chapter begins by exploring a range of everyday and technical meanings to help set the scene for the account that follows.

What do *we* mean by validity?

When we (the authors) refer to validity, we are referring to a technical term of educational and psychological measurement which relates to the idea of validation. Validity and validation are two sides of the same coin. Validation is an investigation into validity, so validity is the property that is to be investigated; and validation is the process by which it is investigated. *Validity theory* provides a conceptual framework to guide *validation practice*. The main chapters of this book will demonstrate how even this technical meaning of validity has proved to be highly controversial and continually resisted precise definition.

In everyday life, validity is the quality or state of being valid, where this can mean anything from being true, to being cogent to being legally acceptable. It derives from the Latin word *validus*, meaning 'strong', 'healthy' or 'worthy' (Wiktionary, 2013). The technical concept of validity, as it has evolved within educational and psychological measurement, has been associated with all of these different meanings at one point or another, as we shall see as the chapters unfold. First, we consider how validity has evolved within other contexts and disciplines.

Validity across disciplines

Validity has a well-established and very specific meaning in philosophy in the subdiscipline of logic, where it applies to deductive arguments. Fogelin and Sinnott-Armstrong (2001), for instance, defined validity as follows: 'An argument is *valid* if and only if it is not possible for all of its premises to be true when its conclusion is false' (p. 36, italics in original). This is an example of a valid deductive argument:

all men are mortal, Samuel Messick was a man, therefore Samuel Messick was mortal. The following argument would be equally valid: all men are immortal, Samuel Messick was a man, therefore Samuel Messick was immortal. In the terminology of deductive argument, the second of these examples is a valid argument, given its logical structure, but it is not a sound argument because one of its premises is untrue.

Over the past couple of decades, an argument-based approach to validation has become popular (see Chapter 5). From this perspective, validation is seen as the process of developing and appraising the strength of an argument concerning the interpretation and use of test scores. The formal, deductive arguments of philosophy are not actually that relevant to validation which, instead, tends to involve informal, largely inductive arguments. Many informal logicians explicitly avoid using the term 'validity' to describe inductive arguments, since it implies too high a standard. Instead of judging them according to the standard of validity, they are judged according to the standard of *strength* (see Fogelin and Sinnott-Armstrong, 2001). Therefore, we might legitimately refer to the strength of a validity argument, but not to the validity of a validity argument. This is a good thing, because having to refer to the validity of validity arguments is potentially very confusing.

Just as validity has a distinctive meaning in philosophy, it also has distinctive meanings across other academic disciplines (Newton and Shaw, 2013). The concept appears in fields as disparate as law (e.g. Austin, 1995[1832]; Waluchow, 2009), economics (e.g. MacPhail, 1998) and pattern recognition (e.g. Halkidi et al., 2002). It appears in disciplines that are quite distant from educational and psychological measurement and, more confusingly, in disciplines that lie right on its borders, such as genetic testing (e.g. Holtzman and Watson, 1997) and management (e.g. Markus and Robey, 1980).

Validity for research

Perhaps most confusing of all, validity appears as a concept within education and psychology, but without connotations specific to measurement. This is to draw a distinction between *validity for research* and *validity for measurement*. Validity for research is relevant whenever conclusions are to be drawn on the basis of research evidence. Validity for measurement is relevant only for conclusions that relate specifically to measurement.

A particularly important contribution to the literature on validity for research was made by Campbell (1957), who drew a distinction between internal validity and external validity in relation to the design of social science experiments. On the one hand, internal validity concerned the degree of confidence that could be placed in the conclusion that an observed effect was genuine for the experimental group. Confidence in this conclusion could be increased by designs which enabled alternative explanations to be ruled out: for example, the addition of a 'no treatment' control group to a pre-test/post-test design, which enabled the researcher to rule out explanations such as the effect simply being one of maturation between pre-test and

post-test. On the other hand, external validity concerned the degree of confidence that could be placed in the conclusion that the observed effect would generalize from the experimental group to the population from which the group was drawn, or to other populations, settings or variables. Confidence in this conclusion would decrease with biased sampling procedures, or if evidence accrued that the very process of experimentation or observation seemed to be causing the effect.

These ideas were developed by Bracht and Glass (1968), who subdivided external validity into two sub-categories: population validity and ecological validity. Population validity concerned confidence in the generalization of conclusions across populations. Ecological validity concerned confidence in the generalization of conclusions across conditions: for example, settings, treatments, researchers and dependent variables. Over the years, the concept of ecological validity was subdivided further: outcome validity concerned generalization across dependent variables; temporal validity concerned generalization across time; and treatment variation validity concerned generalization across treatment variations.

Campbell also developed his thesis (Campbell and Stanley, 1966[1963]; Cook and Campbell, 1979). Cook and Campbell (1979) divided internal validity into statistical conclusion validity plus internal validity, and divided external validity into construct validity plus external validity, thus deriving a four-way classification which corresponded to the four major decisions facing researchers.

Concepts of validity for research have been developed in different directions by researchers within qualitative traditions, including Lather (1986, 1993), Maxwell (1992), Kvale (1995) and others. In these contexts, validity often has been interpreted to mean confidence in the trustworthiness or credibility of description and interpretation: that is, the legitimacy of the knowledge produced. It also has been extended to embrace the social consequences of qualitative research. Lather (1986) proposed three main conceptions of validity:

1 face validity, which elevated the importance of member-checking – i.e. playing researcher accounts back to participants;
2 construct validity, which stressed systematized reflexivity, exploring how the researcher's theory had changed in response to the data; and
3 catalytic validity, which referred to the degree to which the research process facilitated the transformation of reality by participants, reorienting, focusing and energizing them.

Although it is important not to confuse validity for measurement with validity for research, it is very easy to do so for all sorts of reasons. First, some of the key contributors theorized validity in both contexts – most notably, Campbell. Having said that, even he kept his discussion of validity for measurement (Campbell and Fiske, 1959; Campbell, 1960) quite distinct from his discussion of validity for research (Campbell, 1957; Campbell and Stanley, 1966[1963]; Cook and Campbell, 1979). Second, many of those who theorized validity for research borrowed ideas from the existing literature on validity for measurement. For example, both Cook and Campbell

(1979) and Lather (1986) proposed versions of construct validity which had been originally theorized for measurement by Cronbach and Meehl (1955). Third, quite a few terms are common to both the research and measurement literatures, sometimes with quite different meanings attached. For example, the distinction between internal validity and external validity drawn by Campbell (1957) in the context of research was quite different from the distinction drawn seven years earlier, in the context of measurement, by Guttman (1950), using exactly the same terminology. Finally, it is easy to confuse the two because validation inevitably involves research, which means that validity for research is relevant whenever validity for measurement is being investigated; whereas the converse is not true.

In short, if you want to understand validity for measurement, first you need to appreciate that it is not the same as everyday validity, it is not the same as academic validity across different disciplines, and it is not the same as validity for research. Validity for measurement may resonate with many of these alternative ideas, occasionally share their nomenclature and even have inspired them, but ultimately it is different. Bear this in mind when you are searching for 'validity' on the internet.

Validity for measurement: attributes and decisions

When we (the authors) refer to 'validity', we mean validity for educational and psychological measurement. At the risk of over-generalizing, we might be persuaded to extend this conception to social measurement more broadly, but educational and psychological measurement is our central concern. This is a very broad discipline; in fact, perhaps it is better described as a group of disciplines, some more educational and some more psychological. As we will discuss in Chapter 2, the field coalesced around the turn of the 20th century, as measurement movements united interests as disparate as those of academic psychologists who wished to tap the essence of the human mind, and practising educationalists who wished to hold schools to account for the quality of education delivered.

Educational and psychological measurement, like engineering, is a group of disciplines with a very practical side to it. Generally speaking, individuals are measured in order to make decisions. Thus, measurement and decision-making are intimately linked on the assumption that the more accurate the measurement, the better the decision that ultimately can be taken. The various professions can be characterized very roughly in terms of the most significant decisions that they need to make, and the different attributes that they need to measure in order to make those decisions. Table 1.1 illustrates a range of decisions and attributes for a variety of measurement professionals. It should be read as follows: a clinical psychologist might be interested in diagnosing a personality disorder, for example manic depression, to help them to decide on a treatment for a client. Although the table makes a useful point, the dividing lines between professions are not actually that easy to draw, since many different types of decision are made by professionals working in all of these fields.

Sometimes test scores are aggregated across individuals to provide composite measures of higher-level attributes. This is particularly common in education, when

Table 1.1 Examples of attributes measured to make decisions

Measurement professional	Attribute (example)	Decision
Clinical psychologist	Manic depression	What treatment to prescribe
Vocational counsellor	Compassion	Vocational guidance for a student
Educational psychologist	Dyslexia	What intervention to offer a family
Teacher	Mathematical attainment	The set or stream in which to place a student
Personnel psychologist	Aptitude for clerical work	Whether to select a candidate for a job

results are aggregated across students to provide composite measures of educational effectiveness for teachers, schools, regions or even the nation. For example, test results aggregated to the class level might be used to judge whether one teacher has taught more effectively than another. (For an extended discussion of how test scores are used to make all sorts of decisions within educational settings, see Newton, 2007 or Newton, 2010.)

Validity for measurement is intimately concerned with the accuracy with which it is possible to measure an attribute by using a test; however, historically it has equally concerned the potential to make correct decisions based on those measurements. In this sense, the ultimate purpose of measurement is to improve the ratio of correct-to-incorrect decisions. If we measure in order to make decisions, then the first thing that we need to consider, both when designing and when evaluating tests, is the use to which those test scores will be put: that is, the decisions that will be made on their basis. Ultimately, so the argument goes, it is the decision that determines which attribute needs to be measured. If, for example, we needed to decide which children were ready to start formal schooling, then we would need to find a way of measuring an attribute such as 'readiness for formal schooling'. If, instead, we needed to decide which of 20 candidates to appoint to a clerical post, we would need to find a way of measuring an attribute such as 'aptitude for clerical work'. In the professions of educational and psychological measurement, measurement is the handmaiden of decision-making. This is why many, if not the largest majority of, measurement professionals would insist that validity for measurement is inseparable from – if not tantamount to – validity for decision-making. This is a highly contested area, and one that we shall return to throughout the book.

There is one peculiarity of validity for measurement related to its role in decision-making that is worth emphasizing from the outset. It is of particular importance in educational settings, where often results from a single test are used for multiple purposes. The idea that results from a single test might be interpreted in terms of a variety of different attributes has been part of the received wisdom of the measurement professions since the early days of the 20th century (Table 1.2).

The principle implicit in Table 1.2 is that results from a single arithmetic test might be used to measure, in this case, four qualitatively distinct attributes: achievement, aptitude, dyscalculia and general intelligence. Although the test might have been originally designed to assess nothing more than achievement in arithmetic, it has the

Table 1.2 Validation of a single arithmetic test for different purposes

Testing purpose (decision to be made)	Kind of attribute that might be implicated by the decision
Is the student ready to begin the next instructional sequence?	Achievement in arithmetic, i.e. mastery of the programme that has been taught so far
Should the student be enrolled in a higher-level mathematics class?	Aptitude for higher-level mathematics, i.e. potential to master a higher-level programme
Should certain low-performing students be segregated from others, requiring special intervention?	Specific learning difficulty, e.g. dyscalculia
Should certain high-performing students be placed in a stream for the exceptionally talented?	General intelligence

Adapted from Anastasi (1976), Table 13

potential to be used for measuring other attributes as well. Therefore, the question for the evaluator is whether or not it is possible to defend the use of results from the arithmetic test as measures of each of the attributes in turn.

For some academically-minded psychologists, questions such as this stretch the concept of validity too far. For them, validity is a concept more like truth, so the truth about what a valid arithmetic test measures is simply 'achievement in arithmetic'. However, for many practitioners, especially those in educational settings, validity is less like truth and more like plausibility, and they are willing to embrace the principle that a single test might be used to measure a range of attributes. Again, the issue for the practitioner is whether each attribute can be measured with *sufficient accuracy* to support a claim to validity. The arithmetic test might support an accurate assessment of achievement in arithmetic, but how accurate an assessment of general intelligence does it provide? This is certainly a slightly quirky extension of the basic validity question, but it has been fundamental in shaping thinking on validity and validation since the very early years of the 20th century.

Kinds of validity for measurement

If you want to understand validity for measurement, then trawl the internet at your peril! Not only will you be overwhelmed by a plethora of validities that are not fundamental to measurement, you also may be overwhelmed by the volume of validities that have been theorized specifically for it. It has been observed already that validity for research has been subdivided into many different 'kinds' over the years: internal, external, statistical conclusion, outcome, and so on. Over the decades, validity for measurement has been subdivided into many more. Table 1.3 presents a list of 151 such validities that we have found in the literature while researching this book. Somewhere in the region of 28 are mere synonyms of others in the list. Although quite a few of the terms have been given different meanings by different authors, there is actually quite a lot of overlap in the meanings associated with most of the terms.

People began to divide validity into 'types' or 'kinds' back in the 1940s, generally to classify different approaches to establishing validity for measurement. The most fundamental distinction was captured by Cronbach (1949), in his distinction between:

- *logical validity* – a category for grouping approaches based on logical analysis, typically of test content; and
- *empirical validity* – a category for grouping approaches based on empirical evidence, typically involving the correlation of test scores between tests.

Table 1.3 Kinds of validity that have been proposed over the decades

Abstract	Content sampling	Divergent	Incremental	Nomological	Scoring
Administrative	Context	Domain	Indirect	Occupational	Self-defining
Aetiological	Contextual	Domain-selection	Inferential	Operational	Semantic
Artifactual	Convergent	Edumetric	Instructional	Particular	Single-group
Behavior domain	Correlational	Elaborative	Internal	Performance	Site
Cash	Criteria	Elemental	Internal test	Postdictive	Situational
Circumstantial	Criterion	Empirical	Interpretative	Practical	Specific
Cluster domain	Criterion-oriented	Empirical-judgemental	Interpretive	Predictive	Statistical
Cognitive	Criterion-related	Essential	Intervention	Predictive criterion	Status
Common sense	Criterion-relevant	Etiological	Intrinsic	Predictor	Structural
Communication	Cross-age	External	Intrinsic content	Prima Facie	Substantive
Concept	Cross-cultural	External test	Intrinsic correlational	Procedural	Summative
Conceptual	Cross-sectional	Extratest	Intrinsic rational	Prospective	Symptom
Concrete	Cultural	Face	Item	Psychological & logical	Synthetic
Concurrent	Curricular	Factorial	Job analytic	Psychometric	System
Concurrent Criterion	Decision	Faith	Job component	Quantitative face	Systemic
Concurrent Criterion-related	Definitional	Fiat	Judgemental	Rational	Theoretical
Concurrent true	Derived	Forecast true	Known-groups	Raw	Theory-based
Congruent	Descriptive	Formative	Linguistic	Relational	Trait
Consensual	Design	Functional	Local	Relevant	Translation
Consequential	Diagnostic	General	Logical	Representational	Translational
Construct	Differential	Generalized	Longitudinal	Response	Treatment
Constructor	Direct	Generic	Lower-order	Retrospective	True
Construct-related	Discriminant	Higher-order	Manifest	Sampling	User
Content	Discriminative	In situ	Natural	Scientific	Washback
Content-related					

Over the decades, the validity 'cake' has been sliced in many ways, by many scholars. Some degree of consensus over how best to subdivide the concept has been achieved through official statements issued by joint committees of the North American educational and psychological measurement professions (e.g. American Psychological Association (APA) et al., 1954, 1966, 1974; American Educational Research Association (AERA) et al., 1985, 1999).

Largely influenced by these statements, terms that seemed originally to have been used to describe different *approaches to investigating validity* came to be seen increasingly as terms for describing different *kinds of validity*. Thus, from the 1950s to the 1970s, the concept of validity split along the following lines:

1 *content validity* – a narrowed derivative of logical validity;
2 *criterion validity* – a narrowed derivative of empirical validity; and
3 *construct validity* – a more 'scientific' conception originally theorized as though it were the validity of last resort, when neither content validity nor criterion validity could be relied on.

Oddly, although these three kinds of validity were widely accepted to be conceptually fundamental, all sorts of other kinds of validity continued to be proposed in the literature (e.g. convergent, discriminant, trait, nomological and many more). New kinds of validity continue to be proposed even to the present day (e.g. representational, elaborative, formative, summative and many more). This is all the more bizarre because, since the mid-1980s, the educational and psychological measurement professions have recognized officially that validity is a *unitary concept*. This position was recommended to the professions most forcefully by Samuel Messick, who insisted that it was simply not appropriate to divide validity into different kinds because there was really only one *kind* of validity: *construct validity* (Messick, 1980).

Because the term 'validity' is used in so many different ways, and because so many different 'kinds' of validity have been proposed, even within the field of educational and psychological measurement, it is very easy to become very confused when trying to get to grips with the concept. The following chapters will help to explain how the idea of kinds of validity came into fashion and then went out again. For this purpose, the only three 'kinds' that you really need to worry about are content, criterion and construct (with criterion validity sometimes subdivided into predictive validity and concurrent validity). You will see others referred to at various points throughout the book, and no doubt you already will be familiar with yet others. However, content, criterion and construct validity are the ones that really matter to the historical narrative.

Finally, whenever you read and think about validity, be sure that you are reflecting on validity for *measurement*, rather than validity for *research*. Validation is based on research, of course, so considerations of validity for research always need to be borne in mind. However, they are secondary considerations. They concern the validity of validation research conclusions, as opposed to the validity of measurement conclusions, based on validation research. (Hereafter we shall no longer speak of 'validity for measurement', and simply refer to 'validity'.)

Conventions used in the book

If you are happy to go with the flow of the terminology that we will routinely employ throughout this book, then feel free to skip this section. However, if for example you have felt increasingly disenfranchised with each new mention of tests, scores or measurement, then please read on.

Educational and psychological measurement

We describe our discipline as *educational and psychological* to emphasize, first and foremost, that it spans both education and psychology. This is more important than it initially might appear, due to the disparity of interests and professions within this broad group. Despite substantial differences of perspective, the measurement professions within both education and psychology have strived to maintain a common understanding of validity. This need not have been the case and, had it not been so, our understanding of validity might well have been all the poorer for it.

We use the term *measurement* in a generic sense, aware that it carries both technical and emotional baggage. We certainly do not intend to ostracize those within education and psychology who prefer to call their enterprise 'assessment', 'performance evaluation' or even 'diagnosis' – our use of measurement is intended to embrace each one of these possibilities. In addition, when we refer to measurement, we are not invoking a continuum which has 'real' measurement at one end, and 'faux' measurement at the other end. We are not trying to invoke the idea of measurement scales (cf. Stevens, 1951); neither are we taking a position on the possibility of 'genuine' measurement in the social sciences (cf. Michell, 1999). We use the term in its loosest sense, to embrace even 'nominal' measurement: for example, the binary diagnosis of disordered versus not disordered. Similarly, when we speak of measurement professionals, we are not meaning to imply a group of particularly hardcore quantitative statisticians. We use the term to embrace anyone with a professional remit for any aspect of educational or psychological measurement, assessment, performance evaluation or diagnosis.

In exactly the same way, when we refer to *tests* and *scores*, we in no way mean to ostracize those who prefer to describe their tools in terms of examinations, tasks, performances, questionnaires, marks, judgements, outcomes or suchlike. We use the term 'test' in a generic sense to embrace any structured assessment of behaviour: that is, any measurement *procedure*. This is standard practice in the academic literature on validity. Common to all such procedures will be a set of operations in which behaviour is elicited, evaluated and interpreted; often, the outcome also will be summarized as a score, result or profile report, which is intended to characterize the individual (or higher-level entity, e.g. school) in terms of the attribute being measured.

Traditionally, the concept of validity has referred to the quality or potential of a measurement procedure. For this reason, it is quite hard to discuss validity in the

absence of a procedure: that is, in the absence of something that can be replicated. The claim that a measurement procedure is valid is often tantamount to giving it the 'thumbs up' or a 'green light' or a 'stamp of approval'. In other words, to claim that it is valid is to sanction its use for measuring a certain kind of attribute and, therefore, for making a certain kind of decision. From this perspective, we are actually talking *hypothetically* about a *generic procedure* and its *potential* to support good measurement and decision-making in the future. The claim to validity might well have been based on evidence of how it actually had resulted in good or bad measurement or decision-making in the past. However, *in declaring the measurement procedure valid*, we would be making a claim about its potential to support good measurement and decision-making in the future.

On occasion throughout this book, we may use a range of different terms from 'measurement' to 'assessment', 'test' to 'questionnaire' and 'score' to 'outcome'. However, typically we will not be highlighting significant distinctions between connotations of alternative terms. Thus, for consistency and ease of communication alone, we will tend to default to 'measurement', 'test' and 'score'. The important point to note is that the concept of validity is not specific to any one branch of measurement, assessment, performance evaluation or whatever. It is generic, and applies across the full range of contexts.

Finally, there is a sense in which this book, despite having been written by two British people, may have a whiff of North America about it. This is an inevitable consequence of the majority of the academic literature on validity having been published in the USA. Despite this, we are convinced that lessons learned from the North American literature – situated as they are against a backdrop of large-scale, commercially-produced, standardized tests – are universally relevant. It may well be that hard-fought battles to seek consensus across the very disparate professions of educational and psychological measurement within the USA has helped to ensure an international currency for the concepts of validity and validation.

In theory, just about anything written about quality in educational or psychological measurement could be taken as saying something important about validity. However, to do justice to the explicit literature on validity and validation is hard enough; to try to embrace the implicit literature as well would be impossible for us. This is why we have restricted our discussion to the mainstream literature and, again, why the book has a slightly North American lilt to it.

Attributes or constructs?

There is substantial confusion in the literature over how to refer to the characteristics that we are trying to measure. In particular, it is just not clear what name we ought to give to the category that includes 'things' such as achievement, attainment, aptitude, disorder, learning difficulty, attitude, proficiency, competence, etc. There are a number of obvious contenders, perhaps the most obvious of which is the word 'characteristic' itself. Other contenders include 'trait' or 'disposition', although both

of these tend to suggest human characteristics, whereas we sometimes wish to measure different sorts of 'things' such as school effectiveness. Currently, the two main contenders within the literature are *construct* and *attribute*. It is probably fair to say that for now, 'construct' tends to be used more frequently than 'attribute'.

Unfortunately, plenty of philosophical baggage is associated with both of these terms. Having said that, there is probably more baggage associated with construct than with attribute. There is also a large amount of baggage associated with the word 'construct' that is specific to educational and psychological measurement. The term 'construct validity' came into widespread use during the mid-1950s. Since then its meaning has evolved substantially, transforming from just one kind of validity alongside others to the whole of validity (see Chapters 3 to 5). As it happens, if all of validity is supposed now to be construct validity, then the modifier label, 'construct', is actually redundant. More recently still, various scholars have concluded that the label 'construct' is not simply redundant but misleading, since it implies a philosophical tradition that is no longer considered credible in the 21st century (e.g. Borsboom et al., 2004, 2009).

Although there could be no straightforward resolution of this terminological conundrum, the convention that we will adopt is to use 'attribute' rather than 'construct'. Sometimes we will revert to using the term 'construct' if it is important to follow the exact terminology of the position that we are discussing (e.g. when discussing the genesis of construct validity, or the distinction between 'theoretical constructs' and 'observable attributes'). However, in the main we will refer to 'attributes'. Although there are various reasons why we might choose to do so, our principal reason is to minimize the risks associated with excess baggage; our secondary reason is to be able to reserve the term 'construct' for talk about construct validity.

Particular kinds of attribute

We have identified already certain kinds of attribute of particular significance to educational and psychological measurement, including achievement, aptitude and intelligence. Attribute names fall both in and out of fashion, and rightly too: the everyday meanings of terms such as these can, and do, change over time. Alternatively, the everyday connotations may stay the same while scientific understandings change; or alternative names may be chosen in order to foreground particular implications.

The history of the USA College Board SAT illustrates this case very well. The SAT is designed to assess academic readiness for university (College Board, 2013a). When it was first introduced in 1926, it was known as the 'Scholastic Aptitude Test' and abbreviated to SAT (College Board, 2013b). Particularly then, but even now, the term 'aptitude' had connotations of an innate and largely fixed ability akin to intelligence. Following a long period of academic debate over the nature of intelligence and aptitude, and growing concern that what was measured by the SAT was not well described by a word that had connotations of innateness, the name was changed to 'Scholastic Assessment Test' in 1990. However, just a few years later the name was changed once again, and only the letters 'SAT' were retained: it was no longer an acronym but a name in its own right (Wikipedia, 2013). That which is measured by curriculum-based, end-of-course tests or examinations is also known by a variety of names, including 'achievement', 'attainment', 'proficiency', 'competence', etc. Each of these may have slightly different connotations: for example, 'achievement' may be considered to have evaluative overtones, suggesting an accomplishment, especially after having followed a course of instruction. In fact, some prefer the more neutral term 'attainment' for the simple reason that from a value perspective, mastering the same learning objectives may be more or less of an achievement for one learner in comparison with another. Even attainment is not entirely neutral, since it tends to imply that an attempt has been made to master a particular set of learning outcomes, some of which will have been attained, others not. When the intention is simply to certify the capacity to do x, y or z – regardless of whether or not the test-taker has followed a particular course of learning or instruction – the term 'competence' or 'proficiency' is often preferred (see Messick, 1989a: 68).

The fact that different names (for essentially the same, or very similar, attributes) are associated with different contexts and periods in time, presents a challenge for any historical account. This book tends to default to the names that were most prevalent during the early years of validity theory and validation practice: in particular, *achievement*, *aptitude* and *intelligence*. Of course, it is important how the meanings of these words have changed over time. Even today there is no universal agreement on their precise connotations. However, for the sake of a coherent narrative, it is important to have some consistency across chapters. Incidentally, although evaluative overtones would seem to attach to the term 'achievement', the technical usage has never really carried this implication, so a more neutral reading would be appropriate. Equally, while connotations of innateness often were associated with the term 'aptitude' during the early years, the term should be read more agnostically when it appears in the following chapters.

Reliability and validity

Finally, it is worth saying just a few words about reliability and its relationship with validity. This relationship always has been awkward to express. As we shall see, the earliest definitions of validity were framed in terms of the degree to which a test measures what it is supposed to measure. Reliability was, and always has been, defined more specifically in terms of consistency of outcome.

It is useful to contrast consistency with accuracy. If scores from a test were very inconsistent, then we could stake no claim to accurate measurement at all. For example, a ruler that expands and contracts substantially throughout the course of a day returns very inconsistent readings of the same object from one hour to the next. Although it might happen to record an accurate measurement every so often, we certainly could not rely on it to do so. If we cannot rely upon it to reproduce any particular measurement, then we certainly cannot rely upon it to reproduce the correct measurement. When we need to be able to rely upon

each and every measurement that we make, we need to be able to demonstrate a high level of consistency.[1] In the absence of evidence of consistency, any claim to be able to measure accurately (using a particular measurement procedure) would be indefensible.

However, consistency is not enough. For example, two watches can be consistent with each other, yet both be six hours slow. If a ruler has been calibrated incorrectly it will be consistent but inaccurate: that is, it will always return the same reading for objects of the same length, but those readings will never be accurate. In exactly the same way, a test might lead us to rate certain people intelligent and others unintelligent and consistently so, but we might still be consistently wrong in those ratings. Thus, consistency – that is, reliability – is a necessary condition for ensuring high measurement quality, but it is not a sufficient one. Consequently, if validity is to be understood in terms of measurement quality, then reliability is just one facet of it, and a rather abstract, technical facet at that. The remainder of this book will have relatively little to say about reliability specifically.

An outline of the history of validity

The historical account that follows is our attempt to describe and explain how conceptions of validity have evolved within the field of educational and psychological measurement. It explores answers that have been provided down the ages in response to two fundamental questions: the first concerning validity theory, and the second concerning validation practice.

1 What does it mean to claim validity?
2 How can a validity claim be substantiated?

The narrative that we present may seem occasionally to focus more heavily on the first question than the second. To the extent that this is true, it is because answers to the first question need to be provided before the second can be addressed; yet, as we will see, answers to the first are still hotly debated. Having said that, as we discuss each new contribution to validity theory, their implications for validation practice should become fairly clear. Indeed, a key driver of change in validity theory over the years has been the desire to improve validation practice.

We note that there have been few, if any, comprehensive, coherent and clear accounts of validity theory. The literature is characterized better as a compendium of piecemeal insights, concepts and arguments in search of the holy grail: consensus over a generally accepted theory. The measurement community seemed to be getting close to this holy grail through the pioneering scholarship of Samuel Messick, from the 1970s to the 1990s. Yet while his theory of validity was undeniably comprehensive, it tended to lack clarity. Not only was his magnum opus (Messick, 1989a) long and dense, attempting to draw together a vast body of work on validity and validation, it was also philosophically challenging: one eminent

reviewer described it as 'viscous' (Cronbach, 1989c: 24). In addition, we believe that his position on certain key issues changed gradually over the years, without this being acknowledged explicitly (as will be discussed in Chapter 4). The lack of clarity in his presentation makes it much harder to pass judgement on the coherence of his theory.

Over the years, documents intended to encapsulate professional standards – official statements from the professions – have functioned as a surrogate for a generally accepted theory of validity. Without a shadow of doubt, paramount among these have been successive editions of documents prepared since the mid-1950s by committees of measurement professionals from North America. These were originally presented as *Technical recommendations* (e.g. APA et al., 1954; AERA and National Council on Measurements Used in Education (NCMUE), 1955), but are now described as *Standards* (e.g. AERA et al., 1999), with the sixth edition scheduled for release during 2013, at the time of writing. Throughout this book they will be referred to as successive editions of the *Standards*.

Each edition of the *Standards* has provided succinct guidance on validation, preceded by similarly succinct consensus statements on the meaning of validity. These consensus statements are the nearest that we have come to generally accepted definitions of validity, yet their very succinctness reveals that they are not well-developed theoretical accounts. Moreover, as we shall see, neither are they free from ambiguity, if not occasional contradiction, which further emphasizes that they probably are best viewed as heuristic principles, born as a product of compromise rather than universal satisfaction. We will refer to the conception of validity sketched by successive editions of the *Standards* as the *consensus definition* of the measurement professions (Newton, 2012a). Although these definitions were the product of North American committees, attuned to North American concerns, the principles outlined in the *Standards* have been appropriated internationally. Thus, terms such as 'content validity', 'criterion validity' and 'construct validity' have come to provide a lingua franca for measurement professionals the world over. Successive versions of the consensus definition indubitably have provided an international reference point for understanding validity and validation within the field of educational and psychological measurement.

The fact that successive editions of the *Standards* were never intended as definitive accounts of validity theory may help to explain why their descriptions have sometimes resulted in confusion, over-simplification and poor validation practice (Dunnette and Borman, 1979; Guion, 1980). Indeed, widespread evidence of poor validation practice is the principal reason why the consensus definition has needed to be revised significantly from time to time (see Chapter 4 in particular). This has established an increasingly explicit challenge for generations of validity theorists: to move beyond the disparate heuristic principles of the *Standards* towards a comprehensive, coherent and clear account of validity theory. However, few have been prepared to rise to this challenge. In retrospect, it is clear that successive editions of the textbook that many refer to as the 'holy book' of its field, *Educational Measurement*, have repeatedly stimulated the most definitive

accounts of validity and validation of their generations. It is interesting to note that only two of these accounts were entitled 'Validity' (Cureton, 1951; Messick, 1989a), while the other two were entitled 'Test validation' (Cronbach, 1971) and 'Validation' (Kane, 2006), respectively.

The remainder of this chapter explains the structure of the book, which is divided into four main chapters concerning five key phases in the history of validity theory. These are followed by a concluding chapter. The partition into five key phases is broadly consistent with previous accounts which tend to carve the history of validity into three phases: 'pre-trinitarian', 'trinitarian' and 'unitarian'. This terminology comes from Robert Guion (1980), who described the concept of validity, from the mid-1950s to the 1970s, as a 'holy trinity' comprising content validity, criterion validity and construct validity. We make more than previous accounts of advances in recent years, which we describe as a brand new phase, and we discuss the early years in more detail than is common and with a different emphasis. We believe that the history of validity theory can be usefully mapped to the periods between:

1 the mid-1800s and 1920: a gestational period
2 1921 and 1951: a period of crystallization
3 1952 and 1974: a period of fragmentation
4 1975 and 1999: a period of (re)unification
5 2000 and 2012: a period of deconstruction.

Of course, there are no sharp dividing lines between these phases. They represent a crude attempt to add some structure to a far more rambling, diverse and complex evolutionary course. However, they do seem to capture something important about differences in the zeitgeist between eras, and it is no coincidence that many of the transition points correspond to the publication of revised editions of the *Standards*.[2]

Before introducing each of these five phases, it is worth emphasizing that our historical account focuses primarily on changes in conceptions of validity rather than advances in techniques for validation. In fact many, if not most, of the techniques that we rely upon today, both logical and empirical, were developed in at least a primitive form before the 1950s (as we explain in Chapter 2). Over the decades, debate has focused primarily on how and when to employ these techniques. Therefore, different phases can be characterized by different kinds of answer to a central question: how much of what kind of logical analysis and empirical evidence is required in order to substantiate a claim to validity?

The genesis of validity (mid-1800s–1951)

Chapter 2 covers the first two phases: a gestational period from the mid-1800s to 1920, and a period of crystallization from 1921–1951. It is heavily skewed towards

the crystallization period, during which the concept of validity developed an explicit identity – or perhaps more correctly, a range of different identities.

A gestational period (mid-1800s–1920)

Although the examination was not a product of the 19th century, many countries became increasingly reliant on structured assessments during this period as a basis for making complex decisions about individuals and institutions, especially in Europe and North America. The middle of the century witnessed the introduction of the written examination for schools in the USA (by Horace Mann) and the launch of Local Examinations for schools in England (by the universities of Oxford and Cambridge). Structure, it was assumed, was the key to increased accuracy of assessment, and therefore to better decision-making, facilitating outcomes that were fairer for individuals and more useful for society.

By the end of the 19th century, belief in the potential of structured assessment was high, and results from written examinations were used for all sorts of different purposes, from selecting individuals for jobs in the Civil Service to holding schools to account for the quality of their education. However, the prominence of the written examination was not unquestioned, and many were beginning to wonder whether results from examinations were quite as accurate as they often were assumed to be, particularly given the inevitable element of subjectivity involved in judging them. Attention turned to the development of instruments with more structure and less subjectivity: true/false tests, multiple choice tests, sentence completion tests and the like. The idea of the standardized test was born.

While the more practically-minded professionals were busy developing fairer and more effective techniques for structured assessment, the more academically-minded scientists were busy developing ways of probing the very nature of the human mind. Advances in statistical methodology greatly accelerated this work: in particular, the invention of the correlation coefficient, which enabled the statistical patterns inherent in results from examinations and tests to be quantified and interpreted. The fact that certain individuals tended to score high across a range of assessments, while others tended to score low, was treated as indicative of important differences between them. Instruments designed to investigate such differences increasingly came to be seen as tests of mental capacities. The idea of the mental test was born.

These practical and scientific concerns coalesced within what became known as the measurement movement, which flourished during the early years of the 20th century, particularly in North America, but also in Europe and other parts of the world. A huge industry emerged, which churned out tests of all sorts of attributes (general intelligence, specific aptitudes, educational achievement, etc.) in all sorts of formats. This industry burgeoned in the USA following the success of testing for placement and selection during the First World War, but how was this new industry to be controlled or regulated, if at all, and how were consumers to judge the quality of these new tests?

A period of crystallization (1921–1951)

Education provided a home for many of the uses to which tests increasingly were being put. Tests of educational achievement were used to judge students as well as their schools; tests of general intelligence were used to diagnose 'backwardness' and 'excellence'; and tests of specific aptitudes were used as the basis for vocational guidance. Therefore, it is not surprising that issues of quality and control were hotly debated among members of the educational research community. In 1921, the North American National Association of Directors of Educational Research publicized its intention to seek consensus on the meaning of the terms and procedures that were becoming the stock-in-trade of the measurement movement. At the end of a list of key terms, validity was defined as the degree to which a test measures what it is supposed to measure (cf. reliability, defined in terms of consistency). This provided the foundation for all subsequent thinking on validity.

If validity was to be defined as the degree to which a test measured what it was supposed to measure, then how might a claim to validity be established? Two basic approaches to answering this question were honed during the early years: one based primarily on the logical analysis of test content; and the other based primarily on empirical evidence of correlation. Many believed that the best approach to establishing validity was to gather empirical evidence of correlation between the test and what it was supposed to measure. Evidence of high correlation would support the use of test results for measuring the attribute in question. According to proponents of this approach, the correlation coefficient – or the 'validity coefficient' as it came to be known – provided definitive evidence of validity, even when the content of a test appeared to be quite far removed from the attribute that was supposedly being measured. This was not uncommon for the new-style standardized tests, owing to their brevity and structured formats, making the empirical approach seem all the more attractive.

The key question, from this perspective, was what the test results ought to be correlated against: what should be used as a *criterion* against which to judge the accuracy of results? Often, the judgement of expert practitioners – for example, teachers – was used for this purpose. For example, a teacher might be asked to rank their class in terms of intelligence – thereby constructing a *criterion measure* of intelligence – against which the results from a novel test of intelligence would be correlated. Empirical evidence of high correlation would validate the standardized test as a measure of intelligence. However, there were many other ways of constructing a criterion. For example, one way to validate a (short) standardized test of achievement, which sampled the domain only selectively, was to correlate it against a (long) comprehensive assessment of achievement: i.e. a battery of assessments that sampled the domain in full. While the short test might not actually cover the full range of learning outcomes associated with the long assessment, evidence of high correlation with the long assessment would validate the test as a measure of the full domain.

Although it was quite possible to validate educational achievement tests using empirical evidence, some believed that this was not at all the best approach. Their

argument went that if a test of educational achievement had been adequately designed, then it ought to be its own best criterion. That is, it ought to be obvious, from logical analysis of the test content alone, whether it measured what it was supposed to measure. If a group of expert practitioners scrutinized the content of the test and judged that it matched the content of its curriculum, then the test was valid, period.

As time went by, different communities moulded the classic definition in different ways. Ultimately, many whose interests lay mainly in measuring aptitudes (e.g. certain personnel psychologists) tended to prioritize empirical evidence of correlation. Ultimately, many whose interests lay mainly in measuring achievement (e.g. certain educators) tended to prioritize logical analysis of content. These dividing lines were not absolute by any means, but they certainly were becoming more pronounced. Naturally, the different ways of interpreting validity fed back into different priorities for test development. Aptitude testers developed tests to optimize correlation against criterion measures. Achievement testers developed tests to optimize the sampling of criterion content. Yet aptitude testers could not ignore issues of content, especially the content of the criterion measure; neither could achievement testers ignore issues of correlation, especially the part–whole correlation of question-to-test. Tensions within both camps were substantial and uncomfortable.

The fragmentation of validity (1952–1974)

At the beginning of the 1950s, a committee of the APA, chaired by Lee Cronbach, was given a remit to develop professional standards to govern the information that test producers ought to provide on their tests, to allow users of those tests to judge their quality. An early draft of the very first *Standards* document was published for discussion in 1952 (APA, 1952).

The 1952 draft had a section on validity, which followed the lead of a number of earlier textbooks by classifying validity into 'types' or 'aspects', according to the approach employed to establish it. The most fundamental of distinctions had been captured by Cronbach (1949) in his contrast between logical validity and empirical validity. A few years earlier, Greene et al. (1943) had drawn a similar distinction between curricular validity and statistical validity; although they had added a third type, which they called 'psychological and logical validity'. The 1952 draft went a step further by distinguishing four types of validity: content, predictive, status and congruent. By the time of the final publication in 1954, these types of validity were known as content, predictive, concurrent (previously status) and construct (previously congruent). Greene et al. (1943) had introduced psychological and logical validity to deal with school subjects for which neither curricular methods nor statistical methods could be applied. They described this approach to validation as 'a sort of arm-chair psychological dissection of the total process' (1943: 60). In proposing construct validity, the APA committee's intention was similar: that is, to describe an

approach to validation which could be employed when neither logical analysis nor empirical evidence were deemed to be sufficient.

The idea of construct validity had been first proposed by a subcommittee comprising Paul Meehl and Robert Challman. It was later modified and clarified by the entire committee. Two of the committee members, Cronbach and Meehl, subsequently elaborated its principles within what was to become a landmark paper, 'Construct validity in psychological tests' (Cronbach and Meehl, 1955). They framed their problem as follows. Certain kinds of test (e.g. achievement tests) are evaluated in relation to a universe of content (hence, content validity). Other kinds of test (e.g. aptitude tests) are evaluated in relation to a criterion measure (hence, criterion-related validity, which subsumed predictive and concurrent validity). However, for a substantial number of tests there was no such yardstick, and validation needed to proceed differently. For tests such as these, including many personality tests, it needed to be determined which psychological 'construct' accounted for test performance. By 'construct', they meant the 'postulated attribute' which was presumed to be manifest in test performance. According to Cronbach and Meehl, this new approach subsumed both logical analysis and empirical evidence. Indeed, it embraced any form of evidence or analysis which could be brought to bear on the psychological meaning of test scores. Therefore, construct validation was neither quintessentially logical, nor quintessentially empirical, but quintessentially scientific. It rested on a theory from which predictions were generated and tested out.

The *Standards* was revised in 1966, which resulted in the four types of validity being collapsed into three: content, criterion-related and construct. However, the discussion of validity and validation remained essentially the same, and continued as such into the third edition, which was published in 1974. Although all three editions of the *Standards* from 1954 to 1974 insisted that the three types of validity should not be considered mutually exclusive, the way in which they were presented seemed to suggest otherwise. Thus, content validity tended to be seen as the specialized form of validity for achievement tests, criterion validity for aptitude tests, and construct validity for personality tests. Validity theory and validation practice became fragmented along these lines. Although the nomenclature of validity 'types' had been originally introduced to mark alternative approaches to validation, it increasingly came to be seen as marking alternative conceptions of validity.

This fragmentation was especially pronounced for criterion validity. The classic definition could still be reconciled with content validity and construct validity: achievement tests purported to measure achievement, as defined by curriculum content; while personality tests purported to measure personality, as defined by a theory of the construct in question. However, it was far harder to reconcile the classic definition with criterion validity. Predictor tests were viewed typically as 'black boxes', and whatever they might measure, if anything, was considered to be largely irrelevant – just as long as they were able to predict a criterion measure with some degree of accuracy.

The (re)unification of validity (1974–1999)

Cronbach and Meehl (1955) opened the final paragraph of their treatise on construct validity by insisting that in no way did they advocate it as preferable to either content validity or criterion validity. However, even as early as their landmark paper there were indications that at the very least, construct validity might be considered first among equals. For example, they had stated explicitly that construct validation was important at times for every sort of test, including both aptitude tests and achievement tests. Cronbach later came to emphasize the relevance of constructs to all educational and psychological testing (Cronbach, 1971), which helped to pave the way to the next phase in the history of validity theory.

We describe the period between 1974 and 1999 as the 'Messick years', because his conception on validity had come to dominate the world of educational and psychological measurement by the latter part of it. Developing ideas from Harold Gulliksen and Jane Loevinger, and with the support of allies including Robert Guion, he appeared to bring the majority of measurement professionals of his generation around to the viewpoint that all validity ought to be understood as construct validity.

Although it is not obvious from his complex presentation of validity theory, perhaps the most significant contribution that Messick made during this period was to reclaim measurement as the focus of validation in all contexts. You will recall that the classic definition of validity was framed explicitly in terms of measurement: that is, the degree to which a test measured what it was supposed to measure. However, as time went by, it became commonplace to distinguish between validity for measurement and validity for prediction. Messick demolished this distinction by insisting, along with Guion, that in the absence of a clear understanding of what was being measured, prediction was indefensible. As far as aptitude tests were concerned, there were now three fundamental imperatives for validation:

1 to establish that the criterion measure measured what it was supposed to measure;
2 to establish that the aptitude test measured what it was supposed to measure; and
3 to establish a theoretical rationale for why the aptitude test should predict the criterion measure, in addition to presenting evidence that it actually did.

This was nothing like the 'blind empiricism' that had characterized practice within personnel settings (and elsewhere) for decades prior to the 1970s; neither was this any longer simply criterion validation, but construct validation writ large.

In exactly the same way, Messick upped the ante for the validation of achievement tests. No longer was it sufficient for subject matter experts to claim validity purely on the basis of a logical analysis of test content. Fundamentally, it was not test *content* that needed to be representative of curriculum content, but test *performances* that needed to be representative of the full set of learning outcomes defined by the curriculum. Therefore, validation needed to demonstrate that the proficiency that each question was presumed to tap was actually tapped: that is, that performances were neither inflated nor deflated by factors irrelevant to what

the test was supposed to measure. In exactly the same manner, it needed to be established that the scores awarded by those who marked the test performances were neither inflated nor deflated by construct-irrelevant factors. In short, the evaluator needed to be confident that the variance observed in a set of test scores (i.e. the fact that certain test-takers scored high, while others scored low) was attributable to construct-relevant factors, and not to construct-irrelevant ones. Thus, Messick emphasized the importance of discounting plausible threats to the valid interpretation of test scores: specifically, the twin threats of construct under-representation and construct-irrelevant variance.

In presenting these arguments, Messick promoted a new science of validity to be understood along the lines of construct validity. Validation involved the integration of logical analysis and empirical evidence to substantiate the claim that the test measured what it was supposed to measure, and therefore that it could be used for its intended purpose(s). Messick recast all validation as laborious scientific enquiry into score meaning, encouraging evaluators to accumulate as much evidence and analysis as they could lay their hands on, rather than assuming that they could stake a claim to validity on the basis of a single analytical or empirical study in isolation. This reunification of validity was Messick's triumph.

In addition to providing a comprehensive account of the science of validity, Messick undertook an even bigger challenge: to locate ethics at the heart of validity theory. Ultimately, this proved to be Messick's tribulation. His thesis was very confusing, if not ultimately confused. For example, although he located the consideration of values in the ethical row of his progressive matrix, his discussion seemed to overemphasize the scientific evaluation of values (e.g. investigating consistency between trait implications and the evaluative implications of construct labels), and downplay genuinely ethical evaluation (e.g. representing conflicting theories or ideologies within an overall evaluative argument). Similarly, he stressed the importance of investigating consequences from testing. Yet once again, he ultimately came to treat the evaluation of consequences far more scientifically than ethically. That is, as time went by, he became far more focused on how consequences informed the evaluation of test score meaning, and far less focused on evaluating, for its own sake, the positive or negative social value that attached to consequences from testing, including test misuse.

Messick provided important insights into how values and consequences ought to be included in the science of validity. However, ultimately he failed to provide a persuasive synthesis of science and ethics within validity theory. In doing so, it is fair to say that he confused many within the measurement professions, and left a legacy of a rift between those who continued to exclude ethical analysis, and those who made a point of including it.

The deconstruction of validity (2000–2012)

During the 1990s, work on validity and validation was influenced heavily by Messick. Indeed, while the fourth edition of the *Standards*, published in 1985,

showed clear evidence of his impact, the fifth edition published in 1999 was essentially a consensus interpretation of his position. These were truly the Messick years. Yet with the turn of the millennium, things began to change.

For some time, Michael Kane had been developing a methodology to support validation practice grounded in argumentation (e.g. Kane, 1992). This provided a framework or scaffold for constructing and defending validity claims. Thus, while Messick defined the claim to validity in terms of an overall evaluative judgement, Kane explained how that claim to validity could be constructed and defended. Similarly, whereas Messick and the *Standards* directed evaluators towards the sources of evidence which ought to be used when staking a claim to validity, Kane indicated how those sources ought be integrated within an overall validity argument. Argument provided evaluators with a methodology for subdividing the big question of validity into manageable chunks. It clarified where to begin (with the intended interpretation and use of test scores), how to proceed (by making explicit the claims that would support that interpretation and use in the form of an argument, and by testing their assumptions), and when to stop (when the argument was judged to be coherent and complete, and when its inferences and assumptions were judged to be plausible). The argument-based approach better equipped evaluators to identify the full range of issues that need to be addressed, as well as the issues that required most attention: the weakest links in the argument chain.

With the turn of the millennium, Kane began to commit more of his energy to the development of validity theory, which involved an increasingly trenchant rejection of Messick's very stringent requirements for staking a claim to validity. He was concerned that the new version of construct validity theory bought too heavily into the philosophical baggage associated with Cronbach and Meehl (1955). Indeed, the perspective that Cronbach and Meehl had advocated was immersed in a particular philosophical tradition which, at the dawn of the 21st century, might well be considered at least a little outdated. It was based on the principle that the meaning of a theoretical construct was given by its association with other theoretical constructs within a grand scientific theory. If validation was tantamount to scientific enquiry into score meaning, and the meaning of a score was dependent on the development of a theory relating one theoretical construct to others within a large network of theoretical constructs, then validation would become a truly epic, never-ending quest. It is fair to say that both Cronbach and Messick tended to characterize validation in these 'never-ending' terms. Yet according to Kane, this was unduly complex as a general account of validity, and it was unnecessary to portray validation as quite such a laborious and interminable undertaking.

Whereas Messick gave the impression that validation always ought to require the integration of all sorts of empirical evidence and logical analysis, Kane stated clearly that validation requirements were entirely dependent on the particular interpretation and use of results that the test user had in mind. If the interpretation of results was not especially 'ambitious' – for example, if it involved a fairly simple attribute – then only a small amount of evidence and analysis would be required to substantiate it. In particular, Kane drew a distinction between observable attributes and theoretical

constructs. He implied that interpretations drawn in terms of theoretical constructs might well require the kind of laborious scientific enquiry into score meaning that is characteristic of traditional construct validation; whereas he argued that interpretations drawn in terms of observable attributes did not. Observable attributes – literacy or proficiency in algebra, vocabulary knowledge, achievement in arithmetic, skill in solving quadratic equations, etc. – were far easier to validate than theoretical constructs. Not only did Kane's methodology help to simplify the process of validation, his new theory of validity appeared to reduce the evidential burden, at least for many of the attributes at the heart of educational and psychological measurement. Incidentally, this was not simply a deconstruction of validity in the sense of decomposition and simplification; but also a deconstruction in the sense of downplaying the significance of theoretical constructs.

Kane is not the only theorist of recent years to propose a deconstruction of construct validity theory, since challenges have been mounted on various fronts. In particular, Messick has been repeatedly challenged for attempting to provide an integration of ethics and science within validity theory. Particularly prominent in recent years has been the critique mounted by Gregory Cizek, who has insisted that there can be no integration of scientific and ethical analysis, since the two are mutually incompatible arguments. According to Cizek, it is no coincidence that there is a now a significant disjunction between the theory of validity and the practice of validation: the kind of validation envisaged by Messick – that genuinely integrates ethical and scientific analysis – is simply not feasible.

Others have taken the critique of construct validity theory even further: notably, the research partnerships led by Denny Borsboom, which have argued – contrary to a principle that has been fundamental to educational and psychological measurement for decades – that validity is not a property of the interpretation of test scores after all, but a property of tests. In making this claim, they wished to reinstate the classic definition of validity with no additional baggage, and to take it very literally indeed. New and challenging perspectives on validity also have been championed by Joel Michell, Pamela Moss and Susan Embretson, to name but a few.

Since this chapter is less an account of the history of validity, and more an account of validity-in-the-making, we engage more critically with the literature of this phase.

Twenty-first-century evaluation

In Chapter 6 we bring together insights from the preceding chapters to help identify a full range of issues that need to be taken into account when aspects of testing are to be evaluated, including measurement objectives, decision-making objectives and broader policy objectives. The framework at the heart of Messick's account has been the source of much confusion over the years, and much maligned. However, we think that its underlying logic was basically right, so we end the book by proposing a new framework for the evaluation of testing policy, which is our reinterpretation of the original progressive matrix.

Notes

1 When measurement professionals within education and psychology make decisions, they often rely on single measurements: for example, results from a single achievement test. In fact, it is good practice to triangulate evidence across multiple measurements, but even this is likely to involve just a few measurements from related tests.

2 Work on the first *Standards* began in 1952, the third edition was published in 1974 and the fifth edition was published in 1999. Of course, there is no significance in the final date, 2012, other than it marks the year during which the majority of this book was written.