12

Context and Inference in Language Testing

Glenn Fulcher

Introduction

It is arguably the case that 'The purpose of language testing is always to render information to aid in making intelligent decisions about possible courses of action' (Carroll, 1961, p. 314). This holds true whether the decisions are primarily pedagogic, or affect the future education or employment of the test taker. If fair and useful decisions are to be made, three conditions must hold. Firstly, valid inferences must be made about the meaning of test scores. Secondly, score meaning must be relevant and generalisable to a real-world domain. Thirdly, score meaning should be (at least partially) predictive of post-decision performance. If any of these conditions are not met, the process of assessment and decision-making may be questioned not only in theory but also in the courts (Fulcher, 2014a). It is therefore not surprising that historically, testing practice has rested on the assumption that language competence, however defined, is a relatively stable cognitive trait. This is expressed clearly in classic statements of the role of measurement in the 'human sciences', such as this by father of American psychology James McKeen Cattell (1890, p. 380):

> One of the most important objects of measurement...is to obtain a general knowledge of the capacities of a man by sinking shafts, as it were, at a few critical points. In order to ascertain the best points for the purpose, the sets of measures should be compared with an independent estimate of the man's powers. We thus may learn which of the measures are the most instructive.

The purely cognitive conception of language proficiency (and all human ability) is endemic to most branches of psychology and

psychometrics. This strong brand of realism assumes that variation in test scores is a direct causal effect of the variation of the trait within an individual (see the extensive discussion of validity theory in Fulcher, 2014b). This view of the world entails that any contextual feature that causes variation is a contaminant that pollutes the score. This is referred to as 'construct-irrelevant variance' (Messick, 1989, pp. 38–39). The standardisation of testing processes, from presentation to administration and scoring, is designed to minimise the impact of context on scores.

> In some ways, a good test is like an experiment, in the sense that it must eliminate or at least keep constant all extraneous sources of variation. We want our tests to reflect only the particular kind of variation in knowledge or skill that we are interested in at the moment. (Carroll, 1961, p. 319)

There are also ethical and legal imperatives that encourage this approach to language testing and assessment. If the outcomes of a test are high stakes, it is incumbent upon the test provider to ensure that every test taker has an equal chance of achieving the same test score if they are of identical ability. Score variation due to construct-irrelevant factors is termed 'bias'. If any test taker is disadvantaged by variation in the context of testing, and particularly if this is true of an identifiable sub-group of the test-taking population, litigation is likely. Language tests are therefore necessarily abstractions from real life. The degree of removal may be substantial, as in the case of a multiple-choice test, or less distant, in the case of a performance-based simulation. However, tests never reproduce the variability that is present in the real world. One analogy that illustrates the problem of context is that of tests for lifeguards. Fulcher (2010, pp. 97–100) demonstrates the impossibility of reproducing in a test all the conditions under which a lifeguard may have to operate – weather conditions, swell, currents, tides, distance from shore, victim condition and physical build. The list is potentially endless. Furthermore, health and safety regulations would preclude replicating many of the extremes that could occur within each facet. The solution is to list constructs that are theoretically related to real-world performance, such as stamina, endurance, courage, and so on. The test of stamina (passive drowning victim rear rescue and extraction from a swimming pool, using an average weight/size model) is assumed to be generalisable to many different conditions, and to predict the ability of the test taker to successfully conduct rescues in non-pool domains. The strength of the relationship between the test and real-world performance is an empirical matter.

Recognising the impact of context on test performance may initially look like a serious challenge to the testing enterprise, as score meaning must thereafter be constructed from much more than individual ability. McNamara (1995) referred to this as 'opening Pandora's box', allowing all the plagues of the real world to infect the purity of the link between a score and the mind of the person from whom it was derived. While this may be true in the more radical constructivist treatments of context in language testing, I believe that validity theory is capable of taking complex context into account while maintaining score generalisability for practical decision-making purposes.

In the remainder of this chapter, I first consider three stances towards context: atomism, neobehaviourism and interactionism. This classification is familiar from other fields of applied linguistics, but in language testing each has distinctive implications. Each is described and then discussed under two sub-headings of generalisability and provlepsis. Generalisability is concerned with the breadth or scope of score meaning beyond the immediate context of the test. The latter term is taken from the Greek Προβλέψεις, which I use to refer to the precision with which a score user may use the outcome of a test to look into the future and make predictions about the likely performance of the test taker. Is the most appropriate analogy for the test a barometer, or a crystal ball? I conclude by considering how it is possible to take context seriously within a field that by necessity must decontextualise to remain ethical and legal.

Atomism

An atomistic approach to language testing is reductionist in the sense that it attempts to account for language use in terms of elements, or smaller particles. When the term 'reductionism' is used by applied linguists it normally carries a pejorative connotation, especially within discussions of complexity theory that stress the near impossibility of listing, let alone quantifying, contextual variables (Dörnyei, 2009). Reductionism is, nevertheless, essential to the scientific method. In 1662, Robert Boyle discovered that the pressure and volume of gas are inversely proportionate when temperature is held constant. By using mercury to compress air in a J tube under experimental conditions he established one of the best-known constants in science. Away from the laboratory, the law has informed the invention of the aerosol, tyres, aqualungs, fridges, drinks cans, and syringes, to name just a few examples. Studying

gas in its natural environment would not have produced a generalisable understanding of how gas works. Once a theory had been developed based on a small number of variables, predictions could be made about how gas would behave in other settings. These predictions can be tested, which makes the theory scientific.

Atomism implies a dualist view of the world. Gas may exist within contexts, but it has a separate and essentially unrelated identity and behaviour. When I purchase a new fridge, I expect it to keep my food fresh whether I put it in the garage or my kitchen, whether I have a detached house or a flat, whether I live in a town or the countryside. If the context affected performance, I would suspect the science upon which fridges were constructed.

A test that operates in a context-free way must target a single construct, such that the items selected for inclusion are homogenous indicators of its presence and strength (Loevinger, 1957, p. 645). Indeed, many reliability coefficients such as Cronbach's alpha are measures of homogeneity. Multiple-choice is the preferred item type, written to rules that avoid the possibility of arriving at the correct answer for different reasons. The technologies of item and distractor analysis have evolved over the last 100 years to make multiple-choice items the most researched and understood of all task types.

Generalisability

The atomistic approach has been referred to as 'discrete point testing' because of the focus on single elements of the language in isolation from contexts of use. The strategy was deliberate, because 'The situations in which language is the medium of communication are potentially almost infinite', and as these cannot be replicated in a test 'a situation approach that does not specifically test language elements is not effective' (Lado, 1961, pp. 26–27). It is therefore argued that sampling language elements is practically more efficient, and theoretically more valid, than observing a limited contextualised performance. The validation issue is one of the extent to which one can generalise from the responses to decontextualised test items to language knowledge and use.

> Particularized validation is not only devoid of proper scientific interest but deceptive in its promise of practical economy... Its absurdity is most cogently argued by the demands of practical economy and efficiency alone; for a specific test for every occupation and life situation is its logical and impossible conclusion. (Cattell, 1946, pp. 549–550)

In short, Cattell is arguing that good theory leads to reductionism, and reductionist solutions are more economic and efficient because they are generalisable beyond the immediate context. To understand how the clock works you have to take the back off and look at what makes it tick. One of the enduring empirical findings from language testing research is that discrete tests of grammar are the best predictors of performance on all other task types. Davies' (1978) summary of the evidence is as true today as it was when he wrote:

> What remains a convincing argument in favour of linguistic competence tests (both discrete point and integrative) is that grammar is at the core of language learning... Grammar is far more powerful in terms of generalizability than any other language feature. Therefore, grammar may still be the most salient feature to test. (p. 151)

Provlepsis

The question remains whether testing individual language elements can lead to a large class of real-world decisions. Van Moere (2012) argues that what he calls 'facilitators' (for example, repetition of syntactically acceptable sentences, sentence building from jumbled words, response latencies) can predict scores on more contextualised performance measures. Correlational evidence is presented to support this 'psycholinguistic' approach. If the argument is accepted, atomism provides both generalisability and infinite provlepsis to any context of use. However, this is an example of the correlational fallacy. Evidence suggests that even if high correlations can be established between tests of language elements and scores awarded on performance tests, decisions about future performance are insecure (Kaulfers, 1944). The use of scores to make criterion-referenced decisions requires the specification of domains of use because language is configured by the communicative purpose it serves (Fulcher & Svalberg, 2013). While analysis of the minutiae of language elements is a valuable scientific enterprise that generalises to language competence, there is no guarantee that it predicts language use in a given domain. To put it rather bluntly, I would not wish to certify an air traffic controller as communicatively safe to practice on the basis of her ability to manipulate verb morphology. This is most eloquently expressed in Glaser and Klaus' (1962, p. 435) articulation of criterion-referenced testing.

> The adequacy of a proficiency test depends upon the extent to which it satisfactorily samples the universe of behaviors which constitute criterion performance. In this sense, a test instrument is said to have

content validity; the greater the degree to which the test requires performance representative of the defined universe, the greater is its content validity.

Neobehaviourism

The attack on atomism from the communicative language testing movement was particularly fierce. Lado's use of discrete items to test language elements was ruthlessly savaged; and in the 'revolution' that followed, 'authenticity' became the principal criterion by which tests were judged (Morrow, 1979). Authenticity implied that a test taker must communicate a message for a recognised purpose, that input and prompts should be drawn from beyond the world of language teaching and testing, and context should be provided (Morrow, 2012). The definition of context included specification of physical environment, participant status and levels of formality. The most complex attempt to define authenticity is that of Bachman and Palmer (1996), which lists features of the target language use (TLU) domain, which they claimed could be used to replicate real-life context in test tasks.

The most extreme manifestation of neobehaviourism is in the theory of variable competence, which states that speakers do not possess heterogeneous language ability but only a capability to use language that is differentially realised by every facet of every context of use (Fulcher, 1995). Neobehaviourism implies a monist view of the world. There is no distinction between context and construct, and each context is unique because it is generated by a unreproducible matrix of facets. All meaning is rendered local, and the most that can be expected is that we describe each context in as much detail as possible and collect the descriptions into a compendium that reflects the unimaginably complex variation that exists in the world. When combined with radical constructionism, human identity is sucked into the matrix along with language. Any sense of self, or durable language competence whose performance is an expression of that self, is replaced by a variably context-bound constructed identity that is 'a dynamic, discursive construction and a site of struggle' (Norton, 2013, p. 310). Each act of testing is therefore a singularity which constructs the test taker, their language, the construct and the score meaning.

> We assume in language testing the existence of prior constructs such as language proficiency or language ability. It is the task of the language tester to allow them to be expressed, to be displayed, in test performance. But what if the direction of the action is the reverse, so that the act of testing itself constructs the notion of language proficiency? (McNamara, 2001, p. 339)

Generalisability

Uniqueness of this species brings a plethora of problems. Tarone (1998, p. 95) understates the matter when she asserts that 'the most obvious point here is that testers will need to interpret their test results as applying to very restricted TLU situations and testing conditions... the nature of the L2 oral construct itself may be viewed as context specific, shifting from task to task'. But the rabbit hole runs much deeper.

> If there is no underlying competence, merely variable competence or capability, then all we see is all we see, and there is nothing else to see... We would never be able to generalize from one task to another, from one test to a situation external to the test, or from one classroom exercise to another. (Fulcher, 1995, p. 29)

If no context can be replicated precisely, and there is nothing with duration that is manifest across contexts, we are not able to make any meaningful inferences about the participants or their ability to communicate beyond stating 'that is what they did in that unique act of language use.' By the admission of the neobehaviourists themselves, generalisability is impossible.

Provlepsis

Apart from the uncontrolled mass of unanalysable and uncomparable data collected in the compendium, we are unable to generate any 'knowledge' that we can use to see into the future. The irony is that by abolishing everything but context, context is done away with as well. It is a Nietzschean paradox par excellence.

> If the researcher adopts a monist ontology, then the notion of context becomes largely irrelevant, since the observer, the entity, or construct to be described, induced or explained, and the context are all part of the same phenomenon or reality, and it makes little sense to try to delineate where one ends and the other begins. (Bachman, 2006, pp. 188–189)

The only value for language testing in neobehaviourism more generally, and constructionism in particular, is the attention to detail in the

qualitative analysis of performance data. The real agenda is ideological, rather than analytical or practical. The attack on individual identity as well as durable constructs reveals the poststructuralist political programme. As McNamara (2006, pp. 37–38) argues, 'There is also a growing realization that many language test constructs are explicitly political in character and hence not amenable to influences which are not political.' We are no longer individuals capable of communicating our needs, desires and innermost thoughts with others of our language-using species; rather, we are constructs of the matrix, which is designed by oppressive institutions for their purposes. As Norton (2013) pointed out, it is a 'site of struggle' that binds us to the present, and robs us not only of the ability to make decisions about the likely future communicative success of test takers, but also of the test takers themselves.

Interactionism

If an awareness of context simply means that we say different things to different people in different situations, then it is a truism that we have always known, and it does not need problematising in the neobehaviourist manner (Chapelle, 1998). Research into the contextual variables that impact on test-taker behaviour and score variation in speaking tests has identified systematic effects of concern (for example, Brown, 2003; Galaczi, 2008; Ross & Berwick, 1992), and quasi-experimental research has manipulated individual pragmatic variables such as power, imposition and social status to monitor individual changes in discourse by L1 background (Fulcher & Márquez-Reiter, 2003). However, the purpose of all these studies is either (1) to identify variation that can be defined as part of the construct and therefore can be incorporated into test design specifications and scoring models, or (2) classified as construct-irrelevant variance and controlled. An example of the former is ongoing research to operationalise the new construct of interactive competence (Fulcher, 2010, pp. 108–113); and an instance of the latter is research into the use of interlocutor frames to standardise the level of scaffolding provided in speaking tests (Lazaraton, 1996). The research is therefore an investigation into how things actually are in the real world of individuals communicating with other individuals in a variety of social contexts. Like atomism, interactional approaches to context are dualistic. Individuals are assumed to have abilities that are real in the sense that they have duration and are independent of the thoughts or theories of the language testers (Fulcher, 2014b, p. 1433). How language competence is realised in use is nevertheless influenced by the context of use.

It is in this sense that there is interaction between individuals, their enduring (but evolving) competence, and observable performances.

Acknowledging variability in an interactionist paradigm – even within specified domains – is not inherently problematic for language testing. It is achieved in many other assessment fields with even larger scales and variation, and less clarity in the criteria for assessment. In 1978, Robert Parker decided to start rating wines on a 50- to 100-point scale. The practice has endured and supported the growth of a multi-billion dollar wine trade industry. More recently, wine websites have asked oenophiles to taste a specific vintage in their own homes or other environment of choice, specifically to introduce contextual variation. Their ratings are uploaded, and the wine receives an aggregated rating (CellarTracker, n.d.). The sceptic can amass a range of evidence against the usefulness of this activity, including variation in the taste of individual imbibers:

> taste perception depends on a complex interaction between taste and olfactory stimulation ... the sense of taste comes from five types of receptor cells on the tongue: salty, sweet, sour, bitter, and umami. The last responds to certain amino acid compounds (prevalent, for example, in soy sauce). (Mlodinow, 2008, pp. 131–132)

Expectations also impact on ratings, including price information or the environment in which the wine is served. Similarly, research shows that when context is removed completely in double-blind taste experiments, expert tasters can give very different ratings. Although the readers of *Wine Enthusiast* and *Wine Spectator* know all this, the ratings are still taken seriously as a guide to quality. The ratings affect pricing and purchasing volume.

One reason for the success of wine ratings is that people are generally happy with the idea that a number can act as an index – however inaccurate – of what something is worth. This is why we are asked to rate everything from our shopping experience on the Internet to the service in the hotel we stayed at last weekend. We live in metroscape with rules that everyone implicitly understands (Crease, 2011). The rules are based on Laplace's central limit theorem, which states that the probability that the sum of a large number of independent random factors will take a particular value is normally distributed. For wine tasting, the more ratings are accumulated the more random contextual variance is taken into account in arriving at the final score. This is how we gain control over unimaginably complex variation: we introduce yet more random variation into the process. This critical insight into the value and use

of randomness was first formulated by Peirce and Jastrow (1885), who randomised weights and the sequence in which different weights were given to participants in order to discover if they were able to detect very small differences. Without this insight we would not today have the effective drug trials or other critical research that continually improves our lives by delivering systematic outcomes irrespective of context. Your fridge works as promised wherever you put it; chances are you will like a wine with a rating of 95 whether you drink it in the garden or the dining room; and if you hire an international health professional with a grade A or B on the Occupational English Test, they are highly likely to be able to communicate with patients.

In language testing as wine tasting, the certainty with which we interpret score meaning increases when we maximise tasks (tastings/observations) and scores (tasters/raters). In pilot studies it is possible to vary the number of tasks and raters to discover the optimal mix for an operational language test. But in live testing we cannot accumulate the same number of ratings as the CellarTracker website. We therefore have to be much more circumspect in data collection.

Generalisability

If we treat the test as a data collection tool, the technology that is used to design the tool for a particular purpose is the test specification. The specification first sets out how observational data is converted to a summary score. It then provides an explicit statement of the expected generalisability of the score. It describes in as much detail as possible the range of tasks that might appear in any form of a test and how these tasks are permitted to vary along construct-relevant facets. For example, if I wish to make an inference to a test taker's ability to recognise speaker status and then change level of formality accordingly, the variable 'speaker status' must be allowed to vary over specified parameters. In some tests of academic English this is done through tasks involving peer discussions, tutorials with academic staff and service encounters in the accommodation office or medical centre. Similarly, the specification controls those facets that are not relevant to the intended inferences. For example, any requests that are made in these contexts are restricted to low-level impositions, on the grounds that we have discovered differential performance by L1 background in high-level, but not low-level, imposition tasks (Fulcher & Márquez-Reiter, 2003). I have previously described this role of test specifications as the control and freedom mechanism that manages the complexity of contextual effect on score variation (Fulcher, 2003, pp. 135–136).

Provlepsis

The extent to which test scores provide information about a possible future depends on how well the test specifications articulate the tasks and constructs of interest, and manipulate the control and freedom possible, to represent the language use expected in the domain to which inferences are being drawn. In the case of academic English, this may be achieved through a thorough description of the contexts and nature of language use in higher education, such as that undertaken by Biber (2006). The long tradition of domain-specific analysis to inform test design is incorporated into criterion-referenced assessment, which in turn goes back to the very earliest tests for job-specific selection (Fulcher, 2012).

In traditional approaches to language testing and educational measurement more generally, a 'criterion' has been interpreted as either another test of the same construct (Cureton, 1951) or an external standards document to which cut-scores on a test are 'mapped' (Cizek, 2012; Martyniuk, 2010). These are technical definitions that subvert meaningful validation processes in which the 'criterion' of criterion-referenced assessment is the language use domain of interest (Fulcher & Svalberg, 2013). There is an implicit requirement that the domain informs test design and development, rather than the creation of an instrument for which validation is little more than a post hoc activity to support the permissive inferences that a prospective user may wish to make.

Conclusion

There are two extreme responses to the complexity of context. Lado's answer was to remove context from the language test completely. This was not because Lado was really an atomist, as Morrow and colleagues claimed. Much of Lado's writing is concerned with the expression of culture through language, the primacy of speech, and the use of language to achieve 'higher values' such as tolerance towards minorities and cross-cultural understanding. But the overwhelming sense that meaning-transmission in language is so complex led to the reductionist solution that was a step too far for practitioners just a few decades later.

> By admitting that the ability to speak a foreign language is a complex matter and deliberately attempting to attack it through its linguistic elements we can hope to achieve better coverage of the language itself and more objective scoring of the student's responses. (Lado, 1961, p. 241)

The second extreme response is evident in some uses of complexity theory in modern applied linguistics and social science research, which reject our ability to identify and control key variables that cause contextual variation. For many, dealing with context can be like looking out into the vastness of space. It is boundless and represents infinite potentiality. There is simply no escape from the expanse that faces us, and any choices regarding direction seem meaningless. And as we know, in space no one can hear you scream. Following J. L. Austin, I hereby name this *the Keverberg principle*.

The primary problem with an unfettered interest in a myriad of possible factors that interact to produce a unique outcome is that we are unable to generate generalisable knowledge that leads to provlepsis. The result is that we are unable to make any predictions at all about future success from current performance. In the early days of census taking in Europe, it was argued that the ratio of population to births could be arrived at by segmenting a country into sampling frames, taking a sample from each frame, and dividing the number of births by sample total. The idea was akin to that of modern cost-effective random sampling in polls or surveys to produce a result that is independent of local circumstances – or context. Keverberg objected to this method, arguing that the only solution was to conduct a complete census of every person in a country on a particular day. He applied the same argument to births as to deaths:

> The law regulating mortality is composed of a large number of elements: it is different for towns and for flatlands, for large opulent cities and for smaller and less rich villages, and depending on whether the locality is dense or sparsely populated. This law depends on the terrain (raised or depressed), on the soil (dry or marshy), on the distance to the sea (near or far), on the comfort or distress of the people, on their diet, dress, and general manner of life, and on a multitude of local circumstances that would elude any a priori enumeration.... There would seem to be infinite variety in the nature, the number, the degree of intensity, and the relative proportion of these elements. (Keverberg, 1827, pp. 176–177)

Applied sciences like meteorology, on the other hand, model complex systems to improve provlepsis (for example, Mihailović, Mimić & Arsenić, 2014). In a sense, the purpose of understanding complexity is to simplify it to the extent that we are able to use the information to solve problems. As meteorology has successfully created complex models,

short-term forecasting has become more accurate and the limitations on longer-term forecasting more evident. As Klein (1974, pp. 23–24) states, 'Accurate, repeatable numerical measurement or quantification is a constant objective, even in areas where it is not yet fully achieved.' Wine tasting and language testing aspire to this objective in the full knowledge that what is being measured does not have the same status as natural phenomena. Any such presumption would be to commit the same error of reification as that embodied in Quetelet's *l'homme moyen* (the 'average man'). Nor can wine tasting or language testing pretend that score distributions represent error in any meaningful sense. Social sciences are contingent in far too many ways for this kind of certainty. But just like meteorology we are able to create useful models – abstractions that help us do useful things.

There is a place for both atomism and interactionism in language testing and research, depending upon the kind of inferences that we wish to make about a test taker. In research it is also important to study systems in action to see how the components of the system contribute to its usefulness. This improves generalisation and provlepsis. At the University of Salford in the United Kingdom research is being undertaken to how energy is expended in standard three-bedroom properties, depending upon a range of contextual factors. It would be impossible to do this using the kind of 'natural' research currently in vogue with applied linguists. In naturalistic research, houses might be visited across the country and measurements taken of outside weather conditions, the state of the boiler, inside temperature, level of insulation, air pressure, humidity, and so on. These could be correlated with heat loss. One could go further and add to the model construction materials, type of glass in windows, paint colour (which may reflect heat), number of occupants (body heat), thickness and construction of foundations. The list is potentially endless, and very soon the n-sizes required for what is essentially a correlational fishing trip become mind-bogglingly large and economically unfeasible. A natural response to the problem would be to throw our hands up and say that it is all far too complex. The solution, however, is to create a model that simplifies the problem to reasonable proportions. The engineering team at Salford have created an 'Energy House' inside an experimental building. Each contextual variable that might have an effect on heat loss is given the freedom to vary within specified ranges, while those that are not theoretically presumed to be relevant are controlled. The interaction of the contextual factors may also be studied to produce a complex model of the causes of heat loss. The model allows generalisation to other houses and extrapolation to the real world, where particular

configurations are predicted to be effective in the environment where the house is to be built. Recall Carroll's (1961) view that a language test is like an experiment. In an interactionist view, the action of designing and piloting a language test finds a very close analogy in the energy house project. Both are far from 'naturalistic', but they provide generalisable knowledge that is valuable and make the critical task of provlepsis possible. The only way to understand and use context is to simplify and control it. When Robert Boyle wished to discover what it was about air that was relevant to our understanding of the (newly discovered) circulation of blood, he created a vacuum to find out what would happen to candles and mice when air was not present.

Bachman (2006) was correct when he wrote, '"Context" is a comfortable, albeit slippery term in applied linguistics; we all use the term, and know what it means, yet we don't all agree on precisely what, where or even when it is' (p. 188).

What we are agreed upon is that it is important. In language testing it is essential to establish which aspects of context impact upon learner performance and score variation. The effects must be classified as construct relevant or irrelevant. If they are relevant, they are free to vary within certain parameters in the specifications, and if they are irrelevant they must be controlled. Not all contextual factors are relevant. Theory

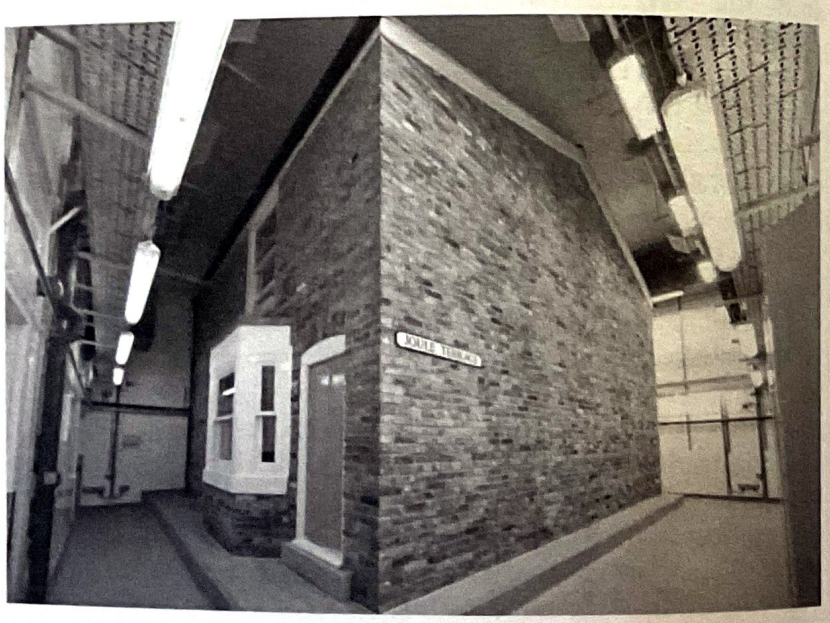

Image 12.1 The Energy House (Reproduced with kind permission of Salford University – http://www.salford.ac.uk/energy/energy-house)

suggests which should be the focus of research and which may safely (for the present) be left to one side. If the problem is not manageable, our research is useless. Messick (1989, pp. 14–15) provides good advice:

> On the one hand, the impact of contextualization can be compelling because the very nature of a task might be altered by the operation of constraining or facilitating factors in the specific situation. On the other hand, we note that the use of multiple task formats along with the standardization of testing conditions contributes to convergent interpretations and comparability of scores across respondents and settings. Rather than opt for one position or the other, in this chapter we urge that the role of context in test interpretation and test use be repeatedly investigated or monitored as a recurrent issue. Thus, the extent to which a measure displays the same properties and patterns of relationships in different population groups and under different ecological conditions becomes a pervasive and perennial empirical question.

My argument in this chapter is that rather than opt for one position or the other, we opt for both. Every test has a purpose that requires context to be modelled anew, and operationalised for assessment in an environment that is remote from the complexity of real-world contexts. This is both natural *and desirable*. Our task is to create instruments to collect data and generate scores sufficient to support inferences about future performance. When contextual features are construct relevant they require modelling, but we must avoid tests in which 'The special circumstances under which they are placed, tell, as we have seen, to the advantage of some, as compared with others' (Latham, 1877, p. 204) for construct-irrelevant reasons. For each new purpose and domain of inference, empirical research is needed to justify the assessment systems that will change the lives of all those who take the test.

References

Bachman, L. F. (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville (Ed.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 165–207). Amsterdam: John Benjamins.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press, Oxford.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing 20*(1), 1–25.
Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In A. Campbell (Ed.), *Teaching English as a second language. A book of readings* (pp. 311–321) (2nd ed., 1965). New York: McGraw-Hill.
Cattell, J. M. (1890). Mental tests and measurements. *Mind, 59*(3), 373–381.
Cattell, R. B. (1946). *Description and measurement of personality*. Yonkers on Hudson: World Books.
CellarTracker (n.d.) Wine review website. Retrieved from http://www.cellartracker.com/
Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge: Cambridge University Press.
Cizek, G. J. (2012). *Setting performance standards: Foundations, methods and innovations* (2nd ed.). New York: Routledge.
Crease, R. P. (2011). *World in the balance: The historic quest for an absolute system of measurement*. New York & London: WW Norton and Co.
Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.
Davies, A. (1978). Language testing. Reprinted from *Language Teaching and Linguistics Abstracts*. In V. Kinsella (Ed.), (1982) *Surveys 1: Eight state-of-the-art articles on key areas in language teaching* (pp. 127–159). Cambridge: Cambridge University Press.
Dörnyei, Z. (2009). Individual differences: Interplay of learner characteristics and learning environment. *Language Learning 59*(Suppl. 1), 230–248.
Fulcher, G. (1995). Variable competence and second language acquisition: A problem for research methodology. *System 25*(1), 25–33.
Fulcher, G. (2003). *Testing second language speaking*. Harlow: Longman.
Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.
Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 378–392). London: Routledge.
Fulcher, G. (2014a). Language testing in the dock. In A. Kunnan (Ed.), *Companion to language assessment* (pp. 1553–1570). London: Wiley-Blackwell.
Fulcher, G. (2014b). Philosophy and language testing. In A. Kunnan (Ed.), *Companion to language assessment* (pp. 1431–1451). London: Wiley-Blackwell.
Fulcher, G., & Márquez-Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing 20*(3), 321–344.
Fulcher, G., & Svalberg, A. M-L. (2013). Limited aspects of reality: Frames of reference in language assessment. *International Journal of English Studies, 13*(2), 1–19.
Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89–119.
Glaser, R., & Klaus, D. (1962). Proficiency measurement: Assessing human performance. In R. M. Gagné (Ed.), *Psychological principles in system development* (pp. 421–427). New York: Holt, Rinehart and Winston.
Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *Modern Language Journal, 28*(2), 136–150.
Keverberg, Baron de. (1827). Notes. Appended to A. Quetelet (1827). Recherches sur la population, les naissances, les décès, les prisons, les dépôts de mendicité, etc., dans le royaume des Pays-Bas (pp. 175–192). Nouveaux mémoirs de l'Académie royale des sciences et belles-lettres de Bruxelles, 4, 117–192.
Klein, H. A. (1974). *Masterpieces, mysteries and muddles of metrology: The world of measurement*. London: George Allen and Unwin Ltd.
Lado, R., (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.
Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge: Dighton, Bell and Company.
Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing, 13*(2), 151–172.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports* 3, *Monograph Supplement, 9*, 635–694.
Martyniuk, W. (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual*. Cambridge: Cambridge University Press.
McNamara, T. (1995). Modelling performance: Opening Pandora's box. *Applied Linguistics, 16*(2), 159–179.
McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing, 18*(4), 333–349.
McNamara, T. (2006). Validity and values: Inferences and generalizability in language testing. In M. Chalhoub-Deville (Ed.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 27–45). Amsterdam: John Benjamins.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: American Council on Education/Macmillan.
Mihailović, D. T., Mimić, G., & Arsenić, I. (2014). Climate predictions: The chaos and complexity in climate models. *Advances in Meteorology*. Retrieved from http://dx.doi.org/10.1155/2014/878249
Mlodinow, L. (2008). *The drunkard's walk: The story of randomness and its role in our lives*. New York: Pantheon.
Morrow, K. (1979). Communicative language testing: revolution of evolution? In C. K. Brumfit & K. Johnson (Eds.), *The communicative approach to language teaching* (pp. 143–159). Oxford: Oxford University Press.
Morrow, K. (2012). Communicative language testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 140–146). Cambridge: Cambridge University Press.
Norton, J. (2013). Performing identities in speaking tests: Co-construction revisited. *Language Assessment Quarterly, 10*(3), 309–330.
Peirce, C. S., & Jastrow, J. (1885). On small differences of sensation. *Memoirs of the National Academy of Sciences 1884* (pp. 73–83). Washington, DC: National Academy of Sciences.
Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*(1), 159–176.
Tarone, E. (1998). Research on interlanguage variation: Implications for language testing. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 71–89). Cambridge: Cambridge University Press.
Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(2), 325–344.