# Criteria for Evaluating Language Quality

Glenn Fulcher

**Abstract**

The assessment of language quality in the modern period can be traced directly to the work of George Fisher in the early nineteenth century. The establishment of a scale with benchmark samples and tasks has been replicated through Thorndike (1912) and into the present day. The tension between assessing observable attributes in performance and underlying constructs that makes performance possible is as real today as in the past. The debate impacts upon the way scales and descriptors are produced, and the criteria selected to make judgments about what constitutes a quality performance, whether in speech or writing. The tensions work themselves through the history of practice, and today we find ourselves in a pluralistic philosophical environment in which consensus has largely broken down. We therefore face a challenging environment in which to address the pressing questions of evaluating language quality.

**Keywords**

Language quality • Performance assessment • Rating scales • Descriptors • Rating criteria

G. Fulcher (✉)
English Department, School of Arts, University of Leicester, Leicester, UK
e-mail: gf39@le.ac.uk



E. Shohamy et al. (eds.), *Language Testing and Assessment*, Encyclopedia of Language and Education, DOI 10.1007/978-3-319-02261-1_13

## Contents



## Introduction

The Oxford English dictionary defines "quality" as "the standard of something as measured against other things of a similar kind; the degree of excellence of something." In language testing the "something" is a language product, which may be a sample of talk or writing. This is "measured" against similar products that have been independently assessed as being appropriate for a particular communicative purpose. The quality of the language sample is the window into the ability of its producer. Or as Latham puts it:

> . . .we cannot lay bare the intellectual mechanism and judge of it by inspection, we can only infer the excellence of the internal apparatus and the perfection of its workmanship from the quality of the work turned out. (Latham 1877, p. 155)

The first attempt to measure language quality by comparison with other samples is found in Fisher's Scale Book (Fulcher 2015a). Between 1834 and 1836, while headmaster of the Royal Hospital School in London, Fisher developed his scale book, in which language performance was classified into five major levels, each with quarter intervals. This produced a 20-level scale. Each level was characterized by writing samples that represented what a pupil was expected to achieve at that level. For spelling there were word lists, and for speaking there were lists of prompts/tasks that should be undertaken successfully. The Scale Book has not survived, but it is clear that Fisher had invented a method for the measurement of quality that is still in use today. There is clear evidence that Thorndike had seen, or was aware of, Fisher's methods (Fulcher 2015b, pp. 84–88). With reference to the assessment of French and German, he suggested attaching performance samples to

levels, together with a brief description of what could be achieved at each level (Thorndike 1912).

It is not clear what criteria were used by Fisher or Thorndike for the selection of samples to characterize each level, other than the professional judgment of experts familiar with the context of the use of the scale. For Fisher, this was a school context in which boys were being educated in preparation for a life in the navy. Thorndike also had a US high school context in mind, but his focus was psychometric and methodological, rather than practical hands-on assessment. But what is clear in both cases is that new language samples collected in an assessment are being evaluated in comparison with criterion samples. Although the term would not be invented for many decades, Fisher was the first to employ criterion-referenced assessment in an educational context.

## Early Developments

The use of criteria external to the assessment context has been central to the evaluation of language quality from the start. It is important to remember that the "criteria" of "criterion-referenced" assessment are not abstract levels that today are frequently referred to as "standards." Rather, the term "criterion" and "standard" were used interchangeably to refer to real-world behaviors that a test taker would be expected to achieve in a non-test environment (Glaser 1963; Fulcher and Svalberg 2013). In the development of the first large-scale language test during the First World War, it was therefore considered essential to reflect such real-world behavior in test content (Fulcher 2012). A group and an individual test of English as a second language were developed to identify soldiers who should be sent to language development batteries rather than deployed to active service. Yerkes (1921, p. 335) reports that the individual test was to be preferred because it was possible to make the content reflect military language more than the group test. Of course, the tasks were still a considerable abstraction from real life, but the criterion was nevertheless the kind of language that was contained in "the drill" (see Fulcher 2015b, pp. 135–140). The score on the test items was interpreted by matching it to a level descriptor from A to E that provided score meaning in absolute criterion terms:

> Men can be tested for English-speaking ability and rated on a scale of A, B, C, D, E. In language the rating E means inability to obey the very simplest commands unless they are repeated and accompanied by gestures, or to answer the simplest questions about name, work, and home unless the questions are repeated and varied. Rating D means an ability to obey very simple comments (e.g., "Sit down," "Put your hat on the table"), or to reply to very simple questions without the aid of gesture or the need of repetition. Rating C is the level required for simple explanation of drill; rating B is the level of understanding of most of the phrases in the Infantry Drill Regulations; rating A is a very superior level. Men rating D or E in language ability should be classified as non-English. (Yerkes 1921, p. 357)

From the First World War, assessing the quality of language performances had two critical components. First was the explicit criterion-referenced relationship

between the content of the test and the domain to which prediction was sought. Second is the level descriptor that summarized what a test taker at a particular level could do with the language in the non-test domain. These two components of performance tests allowed numerical scores to be invested with real-world meaning.

The interwar period was marked by the massive expansion of state provided education in many Western countries. Assessment became critical for accountability, and accountability required controlling the costs of assessment in large systems. There was therefore a focus on the "new-type" multiple choice tests at the expense of performance (Wood 1928). When a new need to assess language performance reemerged in the Second World War, it was as if everything that had been learned during the First World War needed to be reinvented. Thus it was that Kaulfers and others working in the Army Specialized Training Program (ASTP) had to develop new performance tests and descriptors:

> The nature of the individual test items should be such as to provide specific, recognisable evidence of the examinee's readiness to perform in a life-situation, where lack of ability to understand and speak extemporaneously might be a serious handicap to safety and comfort, or to the effective execution of military responsibilities. (Kaulfers 1944, p. 137)

It is the criterion-referenced nature of the decisions being made that requires the quality of language to be assessed through performance. The touchstone was learning to speak a colloquial form of a second language, rather than learning *about* the language (Agard and Dunkel 1948; Velleman 2008). Unlike the individual test created by Yerkes in 1917, the tasks were not domain specific to the military, but covered the functions of securing services and asking for and giving information. This was all that could be achieved in the 5 mins allocated to an individual test. Kaulfers reports that language quality was assessed according to the two criteria of *scope* and *quality* of speech:

**Scope of Oral Performance**

(a) Can make known only a few essential wants in set of phrases of sentences.
(b) Can give and secure the routine information required in independent travel abroad.
(c) Can discuss the common topics and interests of daily life extemporaneously.
(d) Can converse extemporaneously on any topic within the range of his knowledge or experience.

**Quality of Oral Performance**

(0) Unintelligible or no response. A literate native would not understand what the speaker is saying, or would be confused or mislead.
(1) Partially intelligible. A literate native might be able to guess what the speaker is trying to say. The response is either incomplete, or exceedingly hard to understand because of poor pronunciation or usage.
(2) Intelligible but labored. A literate native would understand what the speaker is saying, but would be conscious of his efforts in speaking the language. The delivery is hesitating, or regressive, but does not contain amusing or misleading errors in pronunciation or usage.
(3) Readily intelligible. A literate native would readily understand what the speaker is saying, and would not be able to identify the speaker's particular foreign nationality. (Kaulfers 1944, p. 144)

Under "quality" we can see the emergence of two themes that remain issues of research and controversy to this day. The first is the nature of "intelligibility" and its relation to "comprehensibility," given the constant reference to pronunciation (see Browne and Fulcher 2017). Second is the reference to a "literate" (later to be termed "educated") native speaker as the intended interlocutor.

The ASTP program scored language quality at three levels, under the four headings of fluency, vocabulary, pronunciation and enunciation, and grammatical correctness. The scale for fluency shows that the metaphorical nature of the construct as "flowing" like a river (Kaponen and Riggenbach 2000) emerged very early in performance assessment:

> Fluency
> (2) Speaks smoothly, phrasing naturally according to his thoughts.
> (1) Occasionally hesitates in order to search for the right word or to correct an error.
> (0) Speaks so haltingly that it is difficult to understand the thought he is conveying.
> (Agard and Dunkel 1948, p. 58).

Qualitative level descriptors that closely resemble these early examples have been used ever since, even if they have frequently been disassociated with their original criterion-referenced meaning. They are normally placed in a *rating scale* consisting of two or more levels. Language samples or tasks that are claimed to typify a particular level may be used, following the early practices of Fisher and Thorndike. The rating scale is normally used to match a performance with the most relevant description to generate a score.

## Major Contributions

It should not be surprising that some of the most important contributions have been made within the military context. After the Second World War, the Foreign Service Institute (FSI) was established in the United States to forward the wartime assessment agenda. Although it is still frequently claimed that what emerged from the US military as the Foreign Service Institute (FSI) rating scale was decontextualized (devoid of context, content, or performance conditions) (Hudson 2005, p. 209), as early as 1958 descriptors were attached to the FSI scale. The following example illustrates the level of contextualization that was present:

> FSI Level 2: Limited Working Proficiency.
> Able to satisfy routine social demands and limited work requirements.
> Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on non-technical subjects (i.e., topics which require no specialized knowledge) and has a speaking vocabulary sufficient to express himself simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle elementary constructions quite

> accurately but does not have thorough or confident control of the grammar. (reproduced in Fulcher 2003, p. 226)

The level of contextualization is problematic. The wording suggests "tasks" that a speaker might successfully undertake and the quality of language that might be produced. Yet, it is not specific to its intended military purpose. This is an issue which still exercises language testers today. On the one hand is the argument that all descriptors and scales should refer to constructs only and avoid any reference to context (Bachman and Savignon 1986). The primary purpose of non-contextualization is to achieve greater generalizability of scores across test tasks and real-world contexts. What the language tester is interested in is the underlying constructs or abilities that make communication possible. On the other hand is the argument that by limiting score meaning to specified domains, validation becomes an achievable goal.

The halfway house of the FSI has survived to the present day, despite debates for and against domain specificity. During the 1960s the language and format of the FSI descriptors became standard throughout the military and security agencies in the United States, resulting in a description of language performance known as the Interagency Language Roundtable (ILR), which is still in use today (Lowe 1987). The ILR has also formed the basis of the North Atlantic Treaty Organization (NATO) approach to scoring language quality (Vadász 2012), articulated in Standardization Agreement 6001 (STANAG 6001). The language and structure of the descriptors follows the ILR closely, although additional references to topics and functions have been added in its various revisions (NATO 2010).

The assessment of language quality in the military soon spread to the educational sector. In the early 1980s the American Council on the Teaching of Foreign Languages (ACTFL) and Educational Testing Service (ETS) received US federal grants to adapt the FSI and ILR to create a description of language performance for wider use. The ACTFL *Guidelines* were published in 1986 and revised in 1999 and 2012 (http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012). They have become the de facto framework for describing language performance in the United States in both education and the workplace (Swender 2003).

These descriptions combine linguistic and nonlinguistic criteria and are assumed to be relevant to all languages. The sequence of descriptors on the scale represents an intuitive understanding of the order of second language acquisition and the increasing complexity of real-world tasks that learners can perform, but for which there is little empirical research evidence (Brindley 1998; Chalhoub-Deville and Fulcher 2003).

The FSI descriptors and their subsequent use both in the military and educational sectors have had a profound impact upon the structure and wording of all subsequent scales used for evaluating language quality. The theoretical assumptions, and even the wording, can be traced in all extant scales.

While ACTFL is the dominant system in the United States, the Canadian Language Benchmarks (CLB) is used in Canada, and the Common European Framework of Reference (CEFR) has been developed for use in Europe. These are institutionalized systems and therefore have had a very wide impact on practice (Liskin-Gasparro 2003). However, while both the ACTFL and CLB were designed for operational use in rating language performance, the CEFR bears the hallmarks of a set of abstract standards that cannot be simply taken and used in real assessment contexts (Jones and Saville 2009).

The CLB was developed to assess the English of adult immigrants to Canada and is "a descriptive scale of communicative proficiency in English as a Second Language (ESL) expressed as 12 benchmarks or reference points" (Pawlikowska-Smith 2000, p. 7). Pawlikowska-Smith (2002) argues that the CLB is based on a model of communicative proficiency, drawing specifically on notions of linguistic, textual, functional, sociocultural, and strategic competence, adapted from Bachman and Palmer (1996) and Celce-Murcia et al. (1995). There are three general levels (basic, intermediate, and advanced), each with four subdivisions, for each of the four skill competencies (speaking, listening, reading, and writing).

The CEFR aims to be a pan-European framework for teaching and testing languages (Council of Europe 2001). Like the CLB it has three general levels of basic, independent, and proficient, each subdivided into two levels, providing a six-level system. The system comprises two parts. The first is a qualitative description of each level. For speaking and writing it is elaborated in productive, receptive, and interactive modes. This is "horizontal" in that it does not attempt to help distinguish between levels; it is a taxonomy of the things that language learning is about. The second part is a quantitative description of the levels in terms of "can-do" statements. This is "vertical" in that the levels are defined in terms of hierarchical descriptors.

## Work in Progress

### Rating Scale Development

The major contributions are all institutional systems that perform a policy role within high-stakes testing systems. They are all intuitively developed scales, with the exception of the CEFR, which is a patchwork quilt of descriptors taken from other scales, constructed using a measurement model based on teacher perceptions of descriptor difficulty (Fulcher 2003, pp. 88–113). Dissatisfaction with linear scales that are unlikely to reflect either processes in SLA or performance in specific domains has led to research in scale development that is "data driven." One approach has been through the application of binary choices to separate writing or speaking samples using critical criteria (Upshur and Turner 1995), which has subsequently been applied to TOEFL iBT (Poonpon 2010). The other main approach is the description of performance data to populate descriptors, whether this be taken from test taker performance on tasks (Fulcher 1996) or

expert performance in real-world contexts through performance decision trees (Fulcher et al. 2011). The goal of the latter enterprise is to create a "thick description" of domain-specific performance, thus establishing a true criterion-referenced description against which to match test-generated performances. Data-driven approaches are also being used in prototype writing scales (Knoch 2011). The selection of scale type for particular assessment contexts is a key issue for current research.

### Construct Definition and Validation

While our understanding of what constitutes reasonable performance in specific domains has increased immensely in recent years, the definition and assessment of particular constructs or abilities that enable such performance has been more problematic. Ongoing research into "interactive competence" is particularly important because of the potential to assess individuals in relation to how their own performance and competence is impacted by others (Chalhoub-Deville 2003). Recent work on interactive patterns (Galaczi 2008) and communicative strategies (e.g., May 2011) represents the ongoing attempt to produce operational assessments with richer interactive construct definitions.

### Test Taker Characteristics

Closely related to how participants interact is the question of how individual characteristics affect interaction. The practical implications of this research may impact on how test takers are selected for pair or group speaking tests. Berry (2007) summarizes her extensive research into the impact of personality type, showing that levels of introversion and extroversion can influence speaking scores. Ockey (2009) also found that assertive test takers score higher in group tests, but that less assertive students were not impacted by the pairing. The differences in findings may suggest that the results are conditioned by cultural factors that require further investigation. Nakatsuhara (2011) has also shown that there is variation by proficiency level, personality, and group size. There is clearly much more work to be done here to identify significant variables and their impact on performance.

## Problems and Difficulties

### Generalizability Versus Specificity

Resistance to the use of data-driven or domain-specific scales in large-scale testing is related to restrictions on score meaning. The underlying issue is what constitutes a "criterion" in criterion-referenced testing. For those who argue that domain-specific

inferencing is paramount, the criterion is the language used in real-world applications, which echoes the "job description" tradition of validation (Fulcher and Svalberg 2013). The claim to generalizability of scores to multiple domains and purposes reverts to a criterion-related validation claim based on correlation with an external measure or comparison group (Fulcher 2015b, pp. 100–102). The tension is between substantive language-based interpretations and psychometric expediency. The latter is sometimes used to advocate a robust financial model of "off-the-peg" test use by testing agencies without the need to provide additional validation evidence for changes in test purpose (Fulcher and Davidson 2009). The interplay between the meaning of "criterion" and the economics of global language test use in policy provides plenty of opportunity for conflict.

## Construct Definition

The new interest in content validation (Lissitz and Samuelsen 2007) combined with a lack of interest in construct language within argument-based approaches to validity (Kane 2012, p. 67) has had an impact on validation practices. Chapelle et al. (2010, pp. 3–4) apply this to the scoring of language samples collected in the TOEFL iBT, which moves directly from observation to score, without the requirement for any intervening construct. The point of debate has therefore moved away from construct definition to whether simple content comparison between test tasks and the domain constitutes validation evidence. Kane (2009, pp. 52–58) argues that validation activity remains with the interpretation of scores, and so while the focus may shift to observable attributes in specific performances, there remains a requirement to demonstrate generalizability to all possible test tasks and extrapolation to a domain that cannot be fully represented. But in the simple content validity stance, and the more complex argument-based stance, the room for generalizability of score meaning is considerably reduced. To what extent should construct language be retained? And just how generalizable are the claims that we can reasonably "validate"?

## Rating Scales Versus Standards

The critique of "frameworks" or "standards" documents as tools for policy implementation (Fulcher 2004) has resulted in a recognition that institutional "scales" cannot be used directly to evaluate language quality (Weir 2005; Jones and Saville 2009; Harsch and Guido 2012). But the power of such documents for the control of educational systems has increased the tendency for misuse (Read 2014). The confusion between "standards" and "assessments" is part of the subversion of validity that has been a by-product of the use of scales to create the equivalent of standardized weights and measures in education, similar to those in commerce (Fulcher 2016). This inevitably draws language testers into the field of political action, even if they take the view that they are merely "technicians" producing tools for decision-making processes.

## Future Directions

What has been achieved in the last decade is quite substantial. When the TOEFL Speaking Framework (Butler et al. 2000) is replaced, the new volume will reflect the very significant progress that has been made in assessing the quality of spoken language. We now have considerably more options for scoring models than the simple "more than. . ..less then. . .." descriptors that characterized rating scales in use since the Second World War. These are likely to be richer because of the advances in domain description and referencing. Our deeper understanding of interaction now also informs task design not only for pair and group assessment but also for simulated conversation in a computer-mediated test environment. These developments will inform critical research in the coming years.

## Domains of Inference

The issue of what is "specific" to a domain has come back to the fore in language testing (Krekeler 2006) through the renewed interest in content and the instrumentalism of argument-based approaches to validation. The emphasis must now be on the understanding of what constitutes successful language use in specific domains. Work in the academic domain to support task design in the TOEFL iBT is noteworthy (Biber 2006), as is work on service encounter interactions (Fulcher et al. 2011). There is a long tradition of job-related domain analysis in applied linguistics (e.g., Bhatia 1993), and language testing practice needs to formulate theory and practices for the inclusion of such research into test design.

## Research on Scoring Instruments

Directly related to the previous issue is research into different types of scoring instruments. The efficacy of task-dependent and task-independent rating scales depending on test purpose requires further investigation (Chalhoub-Deville 1995; Hudson 2005; Jacoby and McNamara 1999; Fulcher et al. 2011). As we have found it more difficult to apply general scales to specific instances of language use, it becomes more pressing to show that descriptors adequately characterize the performances actually encountered and can be used reliably by raters (Deygers and Van Gorp 2015).

## Policy Analysis

The most influential approaches to describing language quality are those with the support of governments or cross-border institutions, where there is great pressure for systems to become institutionalized. The dangers associated with this have been

outlined (Fulcher 2004; McNamara 2011), but the motivations for the institutionalization of "frameworks" need further investigation at level of policy and social impact. Of particular concern is the need of bureaucrats to create or defend regional identities or language economies. This leads to the danger that the language testing industry makes claims for tests that cannot be defended and may be particularly dangerous to individual freedoms. As Figueras et al. (2005, pp. 276–277) note, "linkage to the CEFR may in some contexts be required and thus deemed to have taken place. . .."

## Living with Plurality

The immediate post-Messick consensus in educational assessment and language testing has broken down (Newton and Shaw 2014). Fulcher (2015b, pp. 104–124) discusses four clearly identifiable approaches to validity and validation that have emerged in language testing, some of which are mutually incommensurable. At one end of the cline is the emergence of strong realist claims for constructs resident in the individual test taker, and at the other is an approach to co-constructionism that argues for the creation and dissolution of "constructs" during the act of assessing. This clash of philosophies is not new in language testing, but it is more acute today than it has been in the past. The debate over philosophical stance is probably one of the most important to be had over the coming decade, as it will determine the future epistemologies that we bring to bear on understanding the quality of language samples.

## Cross-References


- ▶ Critical Language Testing
- ▶ History of Language Testing
- ▶ Methods of Test Validation
- ▶ Qualitative Methods of Validation
- ▶ The Common European Framework of Reference (CEFR)


## Related Articles in the Encyclopedia of Language and Education


Olga Kagan, Kathleen Dillon: Issues in Heritage Language Learning in the United States. In Volume: Second and Foreign Language Education

Sandra Lee McKay: Sociolinguistics and Language Education. In Volume: Second and Foreign Language Education

Amy Ohta: Sociocultural Theory and Second/Foreign Language Education. In Volume: Second and Foreign Language Education

## References


Agard, F., & Dunkel, H. (1948). *An investigation of second language teaching*. Chicago: Ginn and Company.

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

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal, 70*(4), 380–390.

Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt: Peter.

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Harlow: Longman.

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

Brindley, G. (1998). Describing language development? Rating scales and second language acquisition. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112–140). Cambridge: Cambridge University Press.

Browne, K., & Fulcher, G. (2017). Pronunciation and intelligibility in assessing spoken fluency. In P. Trofimovich & T. Isaacs (Eds.), *Interfaces in second language pronunciation assessment: Interdisciplinary perspectives* (pp. 37–53). London: Multilingual Matters. https://zenodo.org/record/165465#.WDltUbTfWhD.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper*. Princeton: Educational Testing Service.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6*(2), 5–35.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16–33.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Chalhoub-Deville, M., & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals, 36*(4), 409–506.

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

Council of Europe. (2001). *Common European Framework of reference for language learning and teaching*. Cambridge: Cambridge University Press.

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing, Online First* April 7, 1–21. doi:10.1177/0265532215575626.

Figueras, N., North, B., Takala, S., Van Avermaet, P., & Verhelst, N. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing, 22*(3), 261–279.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238.

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1*(4), 253–266.

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 378–392). London/New York: Routledge.

Fulcher, G. (2015a). Assessing second language speaking. *Language Teaching, 48*(2), 198–216.

Fulcher, G. (2015b). *Re-examining language testing: A philosophical and social inquiry*. London/New York: Routledge.

Fulcher, G. (2016). Standards and frameworks. In J. Banerjee & D. Tsagari (Eds.), *Handbook of second language assessment*. Berlin: DeGruyter Mouton.

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), 123–144.

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.
Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29.
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. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.
Harsch, C., & Guido, M. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing, 17*(4), 228–250.
Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics, 25*, 205–227.
Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213–241.
Jones, N., & Saville, N. (2009). European language policy: Assessment, learning and the CEFR. *Annual Review of Applied Linguistics, 29*, 51–63.
Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity* (pp. 39–64). Charlotte: Information Age Publishing.
Kane, M. (2012). All validity is construct validity. Or is it? *Measurement: Interdisciplinary Research and Perspectives, 10*(1–2), 66–70.
Kaponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5–24). Ann Arbor: University of Michigan Press.
Kaulfers, W. V. (1944). War-time developments in modern language achievement tests. *Modern Language Journal, 70*(4), 366–372.
Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81–96.
Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: The effect of background knowledge revisited. *Language Testing, 23*(1), 99–130.
Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge: Dighton, Bell and Company.
Liskin-Gasparro, J. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals, 36*(3), 483–490.
Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437–448.
Lowe, P. (1987). Interagency language roundtable proficiency interview. In J. C. Alderson, K. J. Krahnke, & C. W. Stansfield (Eds.), *Reviews of English language proficiency tests* (pp. 43–46). Washington, DC: TESOL Publications.
May, L. A. (2010). Developing speaking assessment tasks to reflect the 'social turn' in language testing. *University of Sydney Papers in TESOL, 5*, 1–30.
May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127–145.
McNamara, T. F. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333–349.
McNamara, T. (2011). Managing learning: Authority and language assessment. *Language Teaching, 44*(04), 500–515.
Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483–508.
NATO. (2010). *STANAG 6001 NTG language proficiency levels* (4th ed.). Brussels: NATO Standardization Agency.
Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. London: Sage.

Ockey, G. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing, 26*(2), 161–186.

Pawlikowska-Smith, G. (2000). *Canadian language benchmarks 2000: English as a second language – For adults*. Toronto: Centre for Canadian Language Benchmarks.

Pawlikowska-Smith, G. (2002). *Canadian language benchmarks 2000: Theoretical framework*. Toronto: Centre for Canadian Language Benchmarks.

Poonpon, K. (2010). Expanding a second language speaking rating scale for instructional assessment purposes. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 8*, 69–94.

Read, J. (2014). The influence of the Common European Framework of reference (CEFR) in the Asia-Pacific Region. *LEARN Journal: Language Education and Acquisition Research Network,* 33–39.

Swender, E. (2003). Oral proficiency testing in the real world: Answers to frequently asked questions. *Foreign Language Annals, 36*(4), 520–526.

Thorndike, E. L. (1912). The measurement of educational products. *The School Review, 20*(5), 289–299.

Upshur, J., & Turner, C. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*(1), 3–12.

Vadász, I. (2012). What's behind the test? *Academic and Applied Research in Military Science, 10*(2), 287–292.

Velleman, B. L. (2008). The "Scientific Linguist" goes to war. The United States A.S.T. program in foreign languages. *Historiographia Linguistica, 35*(3), 385–416.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281–300.

Wood, B. (1928). *New York experiments with new-type language tests*. New York: Macmillan.

Yerkes, R. M. (1921). *Psychological examining in the United States army* (Memoirs of the National Academy of Sciences, Vol. XV). Washington, DC: GPO.