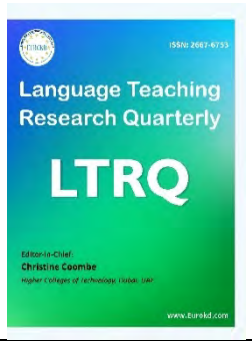

Language Teaching Research Quarterly

2022, Vol. 29, 20–56

Glenn Fulcher's Thirty-Five Years of Contribution to Language Testing and Assessment: A Systematic Review

Glenn Fulcher[1]*, Ali Panahi[2], Hassan Mohebbi[3]

[1]Emeritus Professor of Education and Language Assessment, University of Leicester, UK

[2]Iranian English Language Institute, Ardebil, Iran

[3]European Knowledge Development Institute, Turkey

Received *01 February 2022* ***Accepted*** *23 April 2022*

Abstract

The present systematic review examines Glenn Fulcher's contributions, works, philosophy, and research in language testing and assessment. The data includes his published articles, book chapters, books and interviews (except the one in this special issue) relevant to language testing and assessment from 1987 to March 2022. This study is conducted in two stages: From the sources, Ali Panahi and Hassan Mohebbi derived 127 commonly used main themes, 43 statistical and instrumental concepts, and 14 domains to create a framework for the analysis. We discovered that his research interests were wide-ranging. However, there was a focus on assessing speaking, rating scale design, validity, language assessment literacy and pedagogy, and the broader understanding of the role of assessment from a philosophical and societal perspective. Our analysis provides an overall understanding of the main themes, key concepts and major implications of Glenn Fulcher's work. In the second stage of the study, Glenn presents his personal discussion and reflection of this systematic review.

Keywords: *Glenn Fulcher, Contributions, Analysis, Testing, Assessment, Systematic Review*

Introduction

Language testing and assessment, and educational assessment more generally, has a long and rich history. Many great minds have engaged with assessment practices and research across the centuries, and it is clear that in his appreciation of what has gone before, Fulcher draws heavily on insights from Spolsky (1976; 1995). Both Spolsky and Fulcher are acutely aware that we build upon the work of those who precede us and regret that much "new" research does not credit or

glenn.fulcher@languagetesting.info

doi:10.32038/ltrq.2022.29.03

draw upon lessons already learned, and so treats research questions already addressed as novel (Fulcher, 1999c, 2018b). Taking this as a starting point, we have tried to place Fulcher's work in testing and assessment within a context, to show its relationship to what has gone before, and in other parts of this journey, how it may have impacted the research and practice of others. However, before we present the analysis, we offer some thoughts on language testing and assessment more generally, and Fulcher's contribution to the field.

One of the most obvious major contributions which has lasted throughout his career is the interest in speaking tests and rating scale design (Fulcher, 1987, 1993). It is argued that in examining speaking scales, validity can be enhanced through analyzing what learners actually say in response to tasks, and understanding the language used in target performance domains (Fulcher, 2003b). In 1996, he published a summary of his data-based approach for operationalizing the construct of fluency (Fulcher, 1996a) and explored the generalizability of the fluency scores in a proficiency test (Fulcher, 1996b). He has compared score validity on data-based scales with validity claims for the American Council on the Teaching of Foreign Languages (ACTFL) rating scales (Fulcher, 1996c), among others, and also researched the lack of empirical and theoretical foundations for the Common European Framework of Reference for languages (CEFR) (Fulcher, 2004a, 2010a). Combining the data-based approach with the EBB of Upshur and Turner (1995), Fulcher et al. (2011) used a qualitative approach and elaborated on rating scale design and development for domain-specific inferencing. The data-based approach to rating scale development for performance tests has impacted how we assess both speaking and writing, and there are few testing agencies today that would not claim to use a data-based approach derived from Fulcher's work (Knoch et al., 2021).

It is therefore not surprising that he has engaged with the concept of validity throughout his writing, drawing inspiration primarily from the work of Messick (1989). For example, in 1997, he explored the validity and reliability of a placement test (Fulcher, 1997a) for use in his own institution, the validity of Widdowson's discourse model of communicative competence and performance (Fulcher, 1998a), the reliability and validity of a computer-based test (Fulcher, 1999b), the reliability of two versions of the Vocabulary Levels Test (Xing & Fulcher, 2007) and prototyping a concordance-based cloze test (Kongsuwannakul et al., 2015), as a valid measure of an intended construct. But even when investigating the theoretical aspects of validity and its practical consequences, he has also reflected on the role of validity in society. This is particularly true with regard to legal protections for test-takers, particularly with regard to bias, discrimination, and unfairness (Fulcher & Bamford, 1996; Fulcher, 2013a).

This is intricately linked with his approach to designing tests, which he has compared with the design and architecture of the buildings, as both require specifications - the blueprints and plans from which actual buildings or test forms are created (Fulcher, 2006; Fulcher, 2013b; Davidson & Fulcher, 2012; Fulcher & Davidson, 2009). Also, Fulcher (2009a, 2009b) and Fulcher and Davidson (2007) argue that to avoid validity chaos, it is essential to consider testing as a holistic activity that encompasses consideration of test purpose, impact, utility, consequences, the political philosophy behind the test, and social and legal frameworks. This is likely to lead to ethical assessment, which Fulcher and Davidson have termed "effect-driven testing. This concept enriches

and supports the explicit articulation of consequential validity, value implications, valid inferences and interpretations, social and individual impact of the test, and finally, the decision-making process.

More recently, his contribution to the conceptualization of assessment literacy, classroom-based and learning-oriented assessment (Fulcher, 2020, 2021b) and his focus on score meaning as an inference based on validation and validity evidence (Fulcher, 2013a, 2015c) have impacted on how we conceptualize and teach assessment literacy for teachers. Looking at assessment in pedagogy, Fulcher and Davidson (2008) recommend that valid classroom assessment is based on diagnostics, formative assessment, setting suitable goals, and selecting useful materials and methods. Linking his interest in validation to language assessment literacy and assessment for learning, he has also proposed the validation criterion of "change" as more relevant to learning-oriented assessment contexts. Having outlined some of the enduring themes in his work, we now turn to the analysis of Fulcher's publications from 1987 to the present.

The Analysis

The analysis is divided into four parts: articles, books, book chapters and interviews. For the purpose of this study, the book and software reviews were excluded from the analysis. The categories for the analysis have been developed based on their commonality, frequency, key role and pervasiveness in Glenn Fulcher's works. As Fulcher (2015d) points out, the selection of the themes can be subjective. Added to this, annotations, implications, main themes, statistics, instruments and domain were compiled for the articles (Table 1) and book chapters (Table 2). However, for the books (Table 3) and interviews (Table 4), only annotations and implications were provided. The analysis is hence embedded in four separate tables: Table 1: Analysis of Articles; Table 2: Analysis of Book Chapters; Table 3: Analysis of Books and Table 4: Analysis of Interviews. The publications are listed in chronological order. Before presenting the analysis, we list the analytical categories below.

Main Themes

1. Validity and validity argument
2. Reliability
3. Oral interview scale
4. Content validity
5. Construct and construct validity
6. Communicative oral test
7. English language testing system (ELTS)
8. Communicative testing theory
9. Interagency language roundtable (ILR) oral interview
10. Fairness and ethics
11. Face validity
12. Criterion (or concurrent) validity
13. Validation procedure
14. Operational testing model
15. Rating scale or marking (analytic/holistic/ impression)
16. Performance descriptors, score descriptors or rubrics
17. Multi trait-multimethod matrix (MTMM)
18. Test method
19. Learning and teaching
20. Input and output
21. Cohesive devices and coherence
22. Extralinguistic knowledge and schemata
23. Conditionals and text types
24. Discriminant analysis
25. Fluency and accuracy rating scale
26. Database rating scale for speaking
27. Divergent validity
28. Variable competence model
29. IELTS and TOEFL iBT
30. University of Cambridge Local Examinations Syndicate First Certificate (FCE)
31. Certificate of Proficiency (CPE) examinations
32. Propositional uncertainty or complexity
33. Grammatical and lexical repair
34. Group oral test
35. One-to-one interviews
36. Score generalizability
37. Task validity
38. Task type or test type or item type
39. Task-related anxiety
40. Task difficulty or item difficulty
41. Development of the Texas Oral Proficiency Test (TOPT)
42. ACTFL
43. Assessment of oral proficiency or speaking
44. Foreign Services Institute (FSI) rating scale
45. Interagency Language Roundtable (ILR) rating scales discourse
46. Communicative and interactive strategies
47. Test construction
48. Trait facets (ability continuum)
49. Standards and frameworks
50. Legal framework, politics and economics
51. Placement test
52. Consequential validity
53. Value implications
54. Essay type task
55. Language type task
56. Reading type task
57. Cut score analysis
58. Text difficulty and accessibility
59. Communicative EAP test
60. Evidential basis, evidence-centered design and validity argument
61. ETS
62. TEEP
63. Construct contaminants
64. Computer-based test

65. Pencil-and-paper format
66. Multiple choice tests
67. Distance Learner's Information Service (DiLIS)
68. Document delivery service (DDS)
69. Authenticity
70. Language for academic or specific purposes
71. Item prototypes
72. The oral proficiency interview (OPI)
73. Common European Framework of Reference (CEFR)
74. Pragmatics
75. Large scale (high-stakes) tests
76. Computer adaptive testing
77. Test purpose
78. Test design / test delivery
79. Test model
80. Test framework
81. Test specifications
82. Test retrofit
83. Item facility values
84. Gain score
85. Test architecture
86. Task/ item specifications
87. English as a lingua franca (ELF)
88. Formative assessment, or assessment of/ for learning
89. Diagnostics
90. Practicality, utility, interpretation and inferences
91. Democratic assessment
92. Measurement driven approach
93. Data-driven approach
94. Performance Decision Tree
95. Norm-referenced (NRT)
96. Criterion-referenced testing (CRT)
97. Classroom assessment
98. Performance-based assessment
99. Dynamic assessment
100. Rater Training and Cognition
101. Washback
102. Response validity
103. Concordance-based cloze test
104. Internal and external validity
105. Item banking
106. Predictive validity
107. Cheating
108. Discrimination and bias
109. Score meaning and inference
110. Performance-based data
111. Intelligibility
112. Reverse engineering
113. Measurement-driven instruction
114. Teaching to the test and test taking strategies
115. Scoring rubrics
116. Rater accent familiarity
117. PALS scales
118. Language assessment literacy
119. Apprentice model
120. Learning oriented assessment
121. Teaching English to Speakers of other languages (TESOL)
122. Commercialization of language teaching and testing
123. Canadian Language Benchmarks
124. Assessing writing
125. Continuing professional development (CPD)
126. Teacher assessment/ portfolio assessment
127. Effect-driven testing

Statistics and Instruments

1. Qualitative research
2. Interpretation-based oral interview evidence
3. Descriptive type approach
4. Review paper
5. Quantitative approach
6. Chi-Square
7. Questionnaires
8. Retrospective reports
9. G-study
10. Rash validity scales
11. Correlational method
12. Multitrait-Multi method study
13. (Confirmatory) Factor analysis
14. Rasch model
15. Video recordings or CCTV
16. Audio recordings
17. Transcription analysis
18. Grounded theory methodology
19. Iterative principal axis factor analysis
20. Inter-rater reliability using naive judges
21. Principal Component Analysis
22. Inter-rater and intra rater reliability
23. Equating test forms using anchor items
24. Equating test forms using logistic model (using multiple parallel forms)
25. RASCAL
26. Use of Flesh formula (Flesh reading index and Flesh rating)
27. Expert and Inter-judge agreement
28. Exploratory study
29. Z-test
30. Cronbach's alpha
31. ANCOVA
32. Univariate analysis and univariate general linear model
33. Role play
34. Interviews
35. Exploratory Empirical study
36. Self-report
37. Longitudinal study
38. T-test
39. Simulated recalls
40. Verbalized strategy use
41. Triangulation
42. Analysis of variance
43. Likert-type scale

Domains

A. Papers on validity, reliability, rating scales, scoring and performance tests
B. Papers on language testing and technology
C. Papers on test design and development
D. Papers on language testing and assessment, teaching, learning, pedagogy and applied linguistics
E. Papers on ethics, politics, and law
F. Papers on English for academic and specific Purposes
G. Papers on writing
H. Papers on speaking
I. Papers on listening
J. Papers on reading
K. Papers on vocabulary
L. Papers on grammar
M. Papers on pronunciation
N. Papers on discourse and pragmatic

Table 1
Analysis of Articles

Articles	Annotations	Implications	Main Themes	Stat. Inst.	Domain
Fulcher (1987)	This is the first paper that outlines the notion of "data-based rating scales", derived from an analysis of the mismatch between the ELTS speaking descriptors and discourse recorded from real interactions.	The analysis led to a change in methodology for rating scale design and descriptor construction in Fulcher's Ph.D. thesis and 1996a.	1,2,3,4, 5,6,7,8	1, 2	A, H
Fulcher (1988a)	The ILR's concept of vocabulary is too unclear to be practical in an operational testing model, and data-based discourse analysis techniques for test construction can be used to overcome the scale's shortcomings.	A rating scale can be developed through data-based approach so that the bands will be representative of varying levels or performance.	5,9,10 11,12,13 14,15,16 17,18	3	A, H, K
Fulcher (1988b)	That the classroom is used as a context for research is not a novel idea. Since the 1950s, educators and researchers have used local classroom-based research to inform improved learning and teaching	Issues concerning input and output have not yet been resolved, which has implications for teachers, researchers and applied linguists.	19, 20	4	D
Fulcher (1989)	This paper reviews the role of cohesion in reading theory, arguing that both are important, as reading is simultaneously data-driven and concept-driven.	Teachers can introduce learners to both coherence and cohesion, and researchers should also research both.	21,22	4	D, J
Fulcher (1991a)	The study examines a huge database of written text including academic, narrative, magazine materials and news stories and a simple statistical technique and examines the range of conditional and other if forms.	The implication is that there is a link between the learners' purposes and the need to learn and apply specific kinds of conditional forms.	19, 20, 23	5,6	D, L

Fulcher (1993)	The research investigates the principles, validity and reliability of two data-based oral rating scales (accuracy and fluency) in comparison to an a-priori rating scale (ELTS). The Fluency rating scale evidenced both coherence and continuum validity in three bands.	Data-based approaches to rating scale design provide improved reliability and validity, and so may replace existing design methods.	7,15,24,25,26,27,93	5,7,8,9 10,11,12, 13,14,15	A, H
Fulcher (1995)	This review paper deals with the variable competence approach to Second Language Acquisition. It argues that removing the construct of language competence makes generalizable language research, including score meaning from tests, impossible.	While there is language variation by task and context, individuals bring their own language competence to each performance.	28	4	D
Fulcher (1996a)	The study employs a data-based qualitative and quantitative approach for the description of language use based on Fulcher (1993), and articulated the difference between data-based and other approaches to rating scale design.	Data-based rating scales operationalize language constructs (competence) within instances of performance that improve score validity and rater agreement.	1,2,6,7, 8,12,24, 25,29, 30, 31, 32, 33	1,5,9,14,16, 17,18	A, H
Fulcher (1996b)	This article deals with the use of three tasks in oral tests, with particular reference to the group discussion. The study used Questionnaire techniques and retrospective reports to collect data.	The group oral examination was considered preferable to the traditional one-to-one interview oral tests by test-takers, who said it allowed them to perform at their best.	7,11,34,35, 36,37,38, 39,40,41, 42	7, 9, 5, 8, 14,19	A,H
Fulcher (1996c)	This paper analyses the weaknesses implicit in the American Council on the Teaching of Foreign Languages (ACTFL) rating scales. The scale is used to	Data-based rating scales build validity into scale design and construction, rather than purely a post-hoc activity.	1, 5,13, ,15,18,42 ,	1,5,12, 13 14, 20	A, H

	illustrate problems with a-priori scales in comparison with data-based rating scales.		43,44,45, 46, 48		
Fulcher & Bamford (1996)	This article examines the standards, reliability and validity of EFL tests in the context of the legal framework of the USA and the UK. The review revealed that examination boards might be in danger of legal action unless certain quality issues are addressed.	Language testing does not exist outside the legal and political frameworks of society.	1,2,49, 50	4	A, E
Fulcher (1997a)	The reliability and validity of the placement test used at the University of Surrey were examined in order to place the individuals' inappropriate language support courses.	There is always a call for pretesting all test items before tests become operational and before the decisions are made.	4,5, 12, 51,52, 53, 54, 55, 56, 57	7, 11,14, 21, 22, 23,24, 25	A, E, D
Fulcher (1997b)	A corpus of texts was analyzed to examine text difficulty and accessibility and the results revealed that conceptual structure and poor linguistic structure make the text difficult and less accessible.	The study is useful for teachers, test developers, syllabus designers and materials developers to prepare appropriate educational materials and readings for teaching and assessment purposes.	32, 38, 40, 58	1,5,26, 27	D, J
Fulcher (1998a)	It explores the basic structure and validity of Widdowson's discourse model of communicative competence and performance as the basis for designing and developing reading tests.	The validity of models as basis for the development of reading test must be further evaluated, as they seem efficient.	1,5,27, 28	11, 28,29, 30	A, D, J
Fulcher (1999a)	Traditionally, English for Academic Purposes (EAP) contexts have been assessed with reference to learners' needs analysis and course content analysis. This study assesses the validation and development of EAP tests	Content validity must be used in addition to constructing validity to achieve fair and reliable score interpretations.	1,4, 5,52, 53, 59,60, 61, 62,63	4	A, D, F

	using content specificity and Messick's (1989) theoretical model.				
Fulcher (1999b)	The study reports on the performances on the paper-and-pencil and computer-based tests. The results revealed that the CBT supplied more information than the pencil-and-paper test in placing students into one of two groups.	Since the CBT is reliable enough for its purpose, it can be used by teachers, researchers and testers for placement purposes.	1,2, 64, 65,66	1,2, 30, 31,	A, B
Fulcher (1999d)	The present study reviews the issues of test design and development from the perspectives of test fairness, validity, reliability, washback, stakeholders, learning, teaching and testing. Consequently, a fair approach should take an account of reliability, validity, test writing and scoring. Moreover, the study points out that testing is a support service to teaching and learning.	The implication of the study is that the teachers should assess all the time and use the information derived from the test scores for decision making purpose.	1, 2, 4, 5, 10, 19, 36, 37, 47, 50, 52, 53, 57, 101, 127	1, 3, 4	E
Fulcher & Locke (1999)	The study deals with the ways in which the role of the library in distance learning programs is changing globally to cope with the challenges of the future.	Since the needs of the individuals vary from each other, to meet these needs requires a continuous range of support structures.	67, 68	4	B, D
Fulcher (2000a)	This article deals with communicative language testing as a reaction against multiple-choice tests. First, the history of language testing categorized by scholars (e.g., Morrow, 1979) is briefly explained. Then, it is argued that the jargon of the communicative testing has affected the ways in which language testers approach language teaching problems today, not always for the better.	It is important to understand the history of language testing so that we do not reject or ignore the research or practices of previous generations on the grounds of ideological shifts.	1,2,8,11, 13,19,20, 40 59, 66, 69, 70	4	C, D

Fulcher (2003a)	The article examines a three-phase process model for computer-based test interface design, drawing on good practice from the software industry. It stresses the significance of usability testing and argues that a principled approach to interface design can avoid the threat of interface-related construct-irrelevant variance in test scores	By implication, the findings bring about a mix of validity evidence for the use of CBTs and attempt to avoid construct-irrelevant variances.	1,2, 5, 15, 64, 71	4, 22	B, C
Fulcher & Rosina (2003)	This article deals with the approaches to speaking-related task difficulty. The results revealed that using p-values in a univariate analysis produces a significant three-way interaction between the degree of imposition, language background, and social power.	Designers of language tests for specific purposes can potentially factor cultural elements and pragmatic categories into developing task types and rating scales.	2, 15, 32, 36, 40, 55, 56, 58, 70, 74	5,7,11, 32	C, H
Chalhoub-Deville & Fulcher (2003)	The oral proficiency interview (OPI) resulted from the urgency of the practical needs during World War II when the U.S. military staff needed to fulfill significant foreign language communicative tasks and activities. The article focuses on the OPI and argues that the American Council on the Teaching of Foreign Languages (ACTFL) still needs to develop a coherent mixture of empirical evidence to back up its OPI practice and interpretations.	Language Testing and Assessment agencies need to put in place systematic research agendas to address the validity claims they wish to make for their tests.	15,16, 42,43, 70, 72,	4	B,D,E, F, H
Fulcher (2004a)	This article presents the results of the critical and historical reviews of assessment, teaching and learning as key components in the Common European Framework of Reference (CEFR). It argues that the function and role of CEFR can be associated with socio-political agencies and issues in Europe, some of	The main implication is that CEFR is just one of many models, but one which is being used to achieve the policy goals of European bureaucrats.	49,50,52, 53,73	4	A, C, D, E

	the related functions and implementations and conceptualizations of which can more likely be beyond the language assessors' and testers' control.				
Fulcher (2004b)	The study reviews the Common European Framework of Reference and its dangers. It argues that there seems to be no theoretical basis to the CEFR and many of tests linked to the CEFR do not themselves have a theoretical basis. It recommends that we must be cognizant of the political agenda in standardizing the language of assessment across Europe.	CEFR should be used cautiously, and the social consequences of CEFR use should be considered.	31,42,50, 52,53, 73	4	C, D, E
Marquez Reiter et al. (2005)	The article focuses on the similarities and differences between Britons and Spaniards with regard to the speaker's assumed expectations of compliance. It is revealed that speakers' levels of expectation of compliance are realized in the linguistic elements for conventional and indirect requests.	Since there are differences in social meaning related to conventional indirectness in Spanish and English, analyzing pragmatic categories can inform how speaking performances are evaluated.	8,19,20, 21,22, 33, 74	3, 7,15, 32, 33, 34, 35, 36,	D, N
Fulcher (2005)	This article indicates that TOEFL iBT displays a fundamental change in the way Educational Testing Service (ETS) affects language assessment from test design to test use.	The implication is that TOEFL iBT can be a useful test for assessing English for academic purposes.	1,5,15,13 16,29,38, 49, 52,54, 56, 59,60,61, 64,65,66 70,75,76, 121	1,4	B, D, F, G, H

Fulcher (2006)	The study reviews the purpose and design of tests and compares the process with the design and architecture of buildings, as both require blueprints and plans to develop the actual buildings or test forms.	Close consideration of test purpose will prevent validity chaos and will enrich the consequential validity, value implications and social and individual impact of the test	18, 47,77, 78, 79, 80, 81, 82	4	A, C
Xing & Fulcher (2007)	This article examines the reliability of two versions of the Vocabulary Levels Test at the 5000-word level through a longitudinal study of vocabulary acquisition with use of Version A and Version B of the Vocabulary Levels Test. The results revealed that Version A and Version B were highly reliable and correlated, although the facility values of Version B showed a number of more difficult items.	There are some problems with the 5000-word level tests and those researching vocabulary are warned to take care in their use, especially in the context of longitudinal or gain scores studies.	40,83, 84	2, 5, 11,37, 38	A, D, K
Davidson & Fulcher (2007)	The article argues that the language test development will be more efficient if test impact is considered throughout the test development process. The authors discuss the language of the Common European Framework of Reference for Languages (CEFR) and investigate the utility of such language to pave the way for test development.	Despite the political uses of the CEFR, if it is not reified and used with care, it can be used to inform test content.	73,85, 86,	4	C
Fulcher (2007a)	This article argues that the demand for English on campuses increases at a staggering rate so that private companies see an opportunity for profit in providing English and foundation programs. Evidence suggests the quality of these programmes is often questionable.	The implication is that English language education needs to be re-professionalized and mainstream academia should be revitalized.	19, 70, 87, 122, 125	4	D, F
Fulcher & Davidson (2008)	The article imagines a dialogue between J. S. Mill and Foucault, who hold seemingly different views of the role	Tests affect the individuals, society and all stakeholders	1,10,13 18,47,	4	A, D

	of assessment in education and society. It investigates the social role of assessment and its place in schools.	involved. Therefore, the fate of the individuals and the general benefits of the society should be born in mind.	77, 49, 52, 53, 88, 89		
Fulcher (2009a)	This article deals with the fact that test use is a manifestation of the much broader political philosophy that underpins a society. Political philosophy deals with the collectivist and individualist approach and highlights the way how tests might be used under each condition.	The study recommends we consider the consequential validity of tests and argues that democratic uses of tests should be pursued. .	50, 52,53, 73, 90, 91	4	E
Fulcher & Davidson (2009)	The paper explores architecture as a metaphor for language test development. To this end, the function of test purpose, test use, and its (un)intended effect are examined, as tests are developed for specific purposes and uses. The paper introduced the concept of "test retrofit" as a pre-requisite for making a validity claim for a test used for a purpose different from that contained in the design specifications.	The implication is that a lack of clarity with retrofit restricts validity claims and mystifies the intended impact of tests upon all stakeholders.	1,2, 4, 5, 50, 52, 60, 77, 78, 81, 82, 85, 86, 90	4	C, F
Fulcher (2010a)	This article argues that the use of the CEFR in language education policy and standards-based educational systems serves the political purpose of harmonizing language teaching, learning and assessment. This is achieved through "reification" – or treating the language of CEFR descriptors and their levels – as "true" representations of reality rather than statistically determined constructs.	The language and structure of measurement-driven models like the CEFR should be treated with caution, and not used as the basis for claims about language acquisition or the "true" level of students, without reference to context and performance data.	15, 19, 49, 50, 73, 90,	4	D, E
Fulcher et al. (2011)	This paper uses discourse and domain-specific expert analysis to design a rating scale that combines	Observable action and performance, and the way people	3, 6,8, 15, 72,	1	A, C

	Fulcher's data-driven approach with the empirically-derived boundary-definition method. The new evidence model is termed a "performance decision tree".	use language in real communicative settings to interact with each other, can be used to create improved evidence models for scoring performance tests.	92,93, 94, 98, 110, 113, 115		
Fulcher (2012)	This article presents a research project in which a survey instrument was designed, developed, piloted, and delivered on the Internet to investigate the assessment training needs of the teachers. The data were analyzed both qualitatively and quantitatively. The results were used to construct a theoretical model of language assessment literacy, which was then used to structure learning materials (e.g., Fulcher, 2010b).	The findings can contribute to content and textbook development and can raise the awareness of teachers to issues relevant to assessment literacy.	19, 64	1,3,5,7, 13	A, C, D
Fulcher & Svalberg (2013)	The study indicates that what now passes as CRT is in reality, not criterion-referenced; it is rather the distortion of the original meaning of "criterion" as domain-specific performance. The authors indicate that, unlike NRT, CRT originates in work-based assessment and is a more suitable model in classroom assessment.	CRT is used for formative assessment purposes, assessing learning, achievement tests and providing diagnostic learning	4,5, 38, 49,77, 78, 81, 88, 89, 95,96, 97, 98, 99	4	A, D, H
Fulcher (2015a)	The study examines a wide range of factors involved in second language assessment, framed within an expanded model of speaking test performance. It revealed that the impetus for the growth in speaking assessment came from the educational and military domain to make decisions for recruitment, international mobility, and entrance to higher education.	The historical review can aid teachers and testers in researching and operationalizing particular areas of the language assessment.	3,5,6,8, 13, 15, 18, 36, 70, 100, 101	4	A, D, H

Fulcher (2015d)	The study reviews oral examination used in content-based educational assessment and indicates the novelty of second language speaking assessment in performance tests. The study introduces a wide range of factors and themes in speaking assessment research.	Assessing speaking can help to improve achievement, and meet communicative purposes outside the class context.	3,5,6,8,10,15,18,19, 25,36, 43,52,53, 69,77,78, 79,80,81, 86,89,93, 98,115	4	A, C, D, H
Kongsuwannakul et al. (2015)	The present paper presents the results of prototyping a concordance-based cloze test (ConCloze). The results of the study indicate that ConCloze seemed to be a valid measure of an underlying discrete construct.	ConCloze has some application to ASEAN English language classrooms in terms of English language pedagogy and word-knowledge profiling and evaluation.	1,48,58, 71, 78,102, 103	5	A, D
Yi & Fulcher (2018)	The paper argues that strategy use is one of the assessed constructs in the TOEFL iBT. The findings evaluate the validity claims made by iBT test designers. The study revealed that 84% of strategy types were used similarly in academic tasks and test tasks.	In iBT preparation courses, teaching strategy use can be effective in improving the test takers' performance on the actual test.	1, 2,5, 8, 19, 29, 46, 52, 64, 75, 104,114	39,40, 41	A, D, G, H, I, J

Table 2
Analysis of Book Chapters

Book Chapters	Annotation	Implications	Main Themes	Stat. Inst.	Domain
Fulcher (1991b)	This research paper reports two exploratory studies dealing with issues of reliability and validity of teacher assessment compared with external examinations. The results revealed that teacher assessment is different from other modes of assessment and provides information not supplied by traditional written examinations or aural/oral tests in language examinations.	Teacher assessment within the school setting and classroom can provide valid information, in that it taps those aspects of students' abilities to which formal examinations are not sensitive.	1,2,4, 18, 19, 29,89, 126	4,21,30, 35, 42, 43	A, D
Fulcher (1997c)	The study reviewed issues in assessing writing, including task design, subject matter selection, test method, direct and indirect testing of writing, developing writing tests, various kinds of marking (holistic, analytic and impression marking), alternative assessment, portfolio assessment, classroom research, affective factors in assessing writing, and test developing skills for teachers.	The review can potentially assist teachers to assess learner writing.	11,15, 18, 19, 37, 38, 47, 54, 77, 115, 124, 126	4	A, C, D, G
Fulcher (1998b)	This paper reports on key issues and concepts in computer-based test delivery with use of representative examples. The advantages and disadvantages of using the Internet for test delivery are argued and some aspects of internet delivery are discussed in terms of a technical or measurement	The implication is that test designers and teachers and testers should consider the effectiveness of Internet	2, 64, 78	5,31	B

	perspective. It was predicted that there would soon be a revolution in internet-based testing.	and computer-based tests.			
Fulcher (2000b)	This review reports on the role of the computer in language testing and assessment and argues some of the complex principles and issues to be addressed in the 21st century. It argues that as the technological software has become more developed, a wide range of ethical, research-based and practical issues have arisen.	Since scoring is one of the fundamental factors in computer-based tests, more research needs to be conducted in order to facilitate and automate the process.	5, 10, 64,47, 64, 105	4	B, E
Fulcher (2008)	The present study reviews the criteria for the assessment of language quality in performance tests which dates back to the Second World War. To perform under real life conditions was the main aim of performance testing in both military and academic context. The study also elaborates on the ACTFL Guidelines, the Canadian Language Benchmarks (CLB) and the Common European Framework of Reference as the European system.	Assessment models are often created to serve a military, economic, or political goal. As such, it is important to study their history and underpinning principles.	1,3,6,9,16, 19,29, 36, 42, 49,50, 57, 73, 96, 122, 123	4	A, C, D, E, F
Fulcher (2009b)	The chapter reports the results of a UK-wide survey of universities regarding the organization and outsourcing of English language provision for international students. The results show widespread commercialization and de-professionalisation of EAP staff. The chapter argues that there is evidence that Universities are sacrificing quality for increased income.	TESOL/EAP units increasingly play a role of income generation through teaching large numbers of international students. This is a threat to the professional identity of the field.	19,29,59, 70, 121, 122	4, 7, 41	D, F

Davidson & Fulcher (2012)	The present book chapter reports on a fundamental tool for test development, i.e., test specifications, or specs, also named test blueprints. As a generative document, test specs can be a significant source for the production of multiple equivalent test items or tasks, and maintaining item quality.	As the main test development tool, specifications should be open to critique, debate, and revision.	38, 47,77, 78, 81, 86	4	C
Fulcher (2013a)	This chapter reviews the results of the relationship between language testing and the law. The study reports some exemplar cases drawn from the USA and Europe due to issues such as discrimination, bias, race, and providing equal opportunities for the test takers considering their disability. As a result, the test takers can question whether the evaluations and assessments are fair or just. The chapter updates Fulcher & Bamford (1996).	The implications of the study are that the test takers can question the test result and they also have the right to follow litigation in the court if the test was an invalid measure of their performance.	1,2,4,5,10,13, 49, 50, 77, 78, 82, 106, 107, 108	4	A, E
Fulcher (2013b)	This study examines the test framework and architectural design activity in which the test purpose and use, score inference and interpretation, test retrofit, design documents, rationales, and the intended test-taking population, and the precise nature of the decisions are elaborated. The results indicate that there is a link between test design, test purpose, and validation.	Tests are like buildings in that they often change after they are constructed. However, changes should be planned, audited, purposeful and open to expert or public scrutiny.	13, 77, 78, 81, 82	4	C
Fulcher (2013c)	The chapter is a historical survey of scoring performance tests and related rating scale construction for speaking assessment.	Rating scales and score meaning can serve political objectives; they can create barriers to	1,2, 3, 10, 13, 15, 16, 42, 43, 45, 50, 52, 66, 72,73, 77,	4	A, D

		employment and mobility and can also control the stakeholders and the educational systems.	92, 94, 98, 109, 110, 113		
Fulcher (2014)	This chapter is an attempt to review the significant philosophical principles and issues the language-testing profession faces. It explores the beliefs about the world with use of philosophical argument, support and evidence. What we believe is used to elaborate on the nature of validity theory, interpretive argument, instrumentalism, realism, reductionism, inference and the social facet of language assessment. The main idea is that language assessment and testing practices are considered in both social and philosophical context. This chapter was an exploratory work in preparation for the book-length treatment of similar issues (Fulcher, 2015b).	In examining the language assessment (even teaching and learning) tasks and issues, we need to consider the epistemological and philosophical perspectives in order to understand and interpret practices.	1, 2,19, 20, 36, 52, 53, 69	4	A, D
Fulcher (2015c)	The book chapter mainly deals with the impact of context on the individuals' performance and hence reviews the results of three positions towards the context in language testing, including atomism (discrete-point tests), neo-behaviorism (communicative testing), and interactionism. The study concludes that fair decision-making meets three conditions, including valid inferences, relevance and generalizable score meaning, and prediction to future performance.	In language testing, context should be neither completely neglected nor it should be considered the sole significant element. It should be tapped and considered with reference to the needs	5, 8, 10, 36, 52,66, 77, 109	E	A, D

		and purposes of the tests and individuals.			
Fulcher (2016a)	The study starts with test purpose and elaborates on the next components in the cycle. Moreover, the study reviews the results of repurposing, effect-driven testing, validation process, and substantive validation. The chapter concludes with a discussion of the ecological sensitivity of assessment to local communities and the role of test design in teachers' continuous professional development.	Since test use, inference and interpretation exert micro and macro impact on the individuals and societies, respectively, the process of test design should be discovery-based.	5, 13, 47, 52, 71, 75, 77, 78, 79, 80, 81, 125, 127	4	A, C
Fulcher (2016b)	The study reviews the distinction between the terms "standards" and "frameworks". Standards-based assessment is progressively and universally employed by governmental agencies and can also serve as the expression of power and means of control. On the other hand, they can be used to guide test development and learning. The study describes the three most influential frameworks, including the American Council on the Teaching of Foreign Languages (ACTFL), Canadian Language Benchmarks (CLB), and Common European Framework of Reference for Languages (CEFR).	One of the implications is that standards can facilitate and guide the process of language teaching and learning if used cautiously. But the economic, political and commercial aspects of language testing and assessment need to be rethought and reconsidered.	13, 42, 73, 122, 123,	4	A, C, D, E
Browne & Fulcher (2017)	This chapter argues that the intelligibility of speech is a matter of both perception and performance, so the construct must contain the listener's perception as well as the actual performance of the speaker. The study reveals that both intelligibility and pronunciation test	Pronunciation plays a leading role in the ways listeners perceive intelligibility. This has implications for the	1,2,5, 19,49, 111	5,14	A, D, H, I, M

	scores vary as a result of listener familiarity with accent of the speaker.	selection and training of raters for speaking tests.			
Fulcher & Owen (2016)	This chapter presents the main terminological concepts and constructs required for understanding assessment and standardized language testing. It, therefore, introduces key topics including validity, reliability, norm-referencing, criterion-referencing, test purpose, fairness, politics, assessment for learning, standards, teachers' perceptions, preparing learners for examinations, washback, and social consequences.	An accessible list of key elements of assessment literacy for language teachers.	1, 2, 5, 10, 19, 49, 50, 52, 5375, 77, 88, 95, 96, 97, 101, 112, 113, 114	4	A, C, D, E,
Fulcher (2018a)	The study examines assessment in the speaking classroom with a focus on providing feedback to learners on their performance. Feedback can be supplied within the framework of assessment for learning, so that it helps the learners to be aware of their current level of performance and the target at which they are aiming, identifying the gap between the two.	Assessing speaking and providing quality feedback to learners should be incorporated into continuous professional development.	3, 9, 15, 16, 25, 42, 88, 117	4	A, C, D, H
Fulcher (2019)	Language testing and assessment is presented as a very practical activity (Fulcher, 2010c), but it is supported by theoretical justifications and evidential basis.	There can be a common understanding and interpretation regarding test design, development, constructs, tasks and assessment practices. So, the activities need to be placed on the continuum of CPD.	5, 18, 38, 47, 52, 53, 59, 60, 77, 78, 79, 80, 81, 82, 86, 90, 118, 125	4	A, C, D, E, F

Fulcher (2020)	This chapter explores language assessment literacy, the apprentice model of teaching language testing, the characteristics of apprenticeship tasks, and the theory of pedagogy highlighting teaching and learning for LAL in one specific context. The approach is elaborated within a Pragmatic theory of learning with the use of the metaphor of the apprentice. The study extends the definition of language assessment literacy to the practice of learning and teaching.	The pedagogic implementation of LAL models requires materials and methods that combine the acquisition of both theory and practice.	19, 38, 78, 118, 119	4	A, D, N
Green & Fulcher (2021)	The chapter Introduces SLA researchers and language testers to the language test design cycle. It views the test design cycle as a set of systematically interconnected and interrelated actions and activities that manage assessment instruments and procedures in a way to consistently achieve design goals.	As tests are used to collect evidence in SLA studies, SLA researchers need to be aware of how to construct valid tests.	1, 2, 5, 13, 18, 71, 77, 78, 79, 80, 81, 109	4	A, C , D
Fulcher (2021a)	This chapter reports on validity in an LOA context. The study pinpoints the key differences between a high-stakes and an LOA assessment paradigm. The chapter argues that *change* is the most significant validity criterion for LOA.	The implication of the study is that LAL for LOA is what teachers need to know in order to put assessment in the service of change.	1, 49, 75, 88, 118, 120	4	A, D
Fulcher (2021b)	The present outlines critical research questions in language testing research, such as evidence for supporting test score, use and interpretations, fairness, test purpose and decision making, and formative assessment.	Novice scholars, graduates, post graduates and language teachers, may find useful guidance on topics for local and personal research projects.	1, 10, 15, 52, 75, 77, 78, 88, 90, 118, 120	4	A, C, D

Table 3
Analysis of Books

Books	Annotation	Implications
Fulcher (2003b)	This book provides a comprehensive discussion of testing speaking in a second language in eight chapters. First, it elaborates on the history of testing second language speaking. Then, it defines the constructs, tasks, rating scales and test specs. In chapter 6, it deals with the raters, trainers, and administration and finally, in the last two chapters, it considers the evaluation and research of second language speaking tests.	The kinds of questions the teachers would more probably ask and the related answers can be found in the book. The book will be useful for those who would like to develop speaking tests in their own institutions. Also, the book can be effective for the test designers, applied linguists and course designers.
Fulcher & Davidson (2007)	The book contains major themes and key terms, models, concepts and practical considerations in language assessment and testing through bringing together influential articles and discussing their contribution to the field. Moreover, it presents reflective tasks which enable and engage the readers. Therefore, the book provides a thorough review of test development, item and task development, ethical practice, pragmatism in assessment, data analysis, washback, scoring performance tests, validity, validity argument, test validation, evidence-based design, analysis of test results, study of test revision or change, design of arguments for test validation and effect-driven testing.	By implication, the book is effective in both theoretical and practical terms, operationalizing and conceptualizing the ins and outs of testing and assessment.
Fulcher (2010b)	In 10 chapters, illustrated with real tests and assessments, language assessment issues are discussed with reference to both qualitative and quantitative research methods. Fulcher initiates the readers, testers, teachers and non-testers into the purpose of testing, large-scale standardized testing, classroom assessment, the process of test design, creating test specifications, test architecture, evaluation of the test specifications and items, scoring the multiple choice and performance tests,	The key purpose of the book is to equip the readers, testers and teachers with what is required to put theory into practice and observe the impact of assessment on learning and

	automated scoring systems, establishing cut scores, absolute standards, statistical tools, the practicalities of test administration, and the effect of tests on learning and teaching.	teaching, prepare learners to take tests and help the teachers to assess their learners formatively.
Fulcher & Davidson (2012)	This handbook is an indispensable reference which covers some of the most significant key issues in language testing, including validity argument, classroom assessment and washback, assessing younger learners, assessment for immigration and citizenship, the social and ethical uses of tests, test specifications, evidence-centered design, test-taking strategies, research methods and techniques in the validation of a language test, writing items and tasks, prototyping and field tests, test administration and training, measurement theory, reliability, scoring, ethics, and language policy.	A comprehensive reference for language teachers, postgraduate students, scholars, testers and all those working in the field of language assessment and applied linguistics.
Fulcher (2015b)	Re-examining Language Testing examines the evolution of language assessment within the framework of philosophical, social, historical and cultural beliefs and perspectives. The book elaborates on more fundamental topics such as validity, validity argument, validity claims, consequential validity, content validity, construct validity, interpretive argument, scoring models, test design and specifications, models of language competence and performance, measurement and psychometrics, meritocracy and language testing, ethics and fairness, and socio-political issues and values.	Language Testing and educational assessment more widely exists and evolves to serve purposes within society. It is therefore impacted by the philosophy and values of its users. An awareness of the wider context of assessment practice is important for ethical practice in any era.
Fulcher & Harding (2022)	This handbook including 35 authoritative articles written by 51 leading specialists, divided into ten sections, provides an overall view of the key concepts and issues in language testing and assessment, such as validity, test use, classroom assessment and washback, assessing the language skills, test design and administration, writing	An authoritative portrait of the field today, with predictions for the future that may guide

	test items and tasks, prototyping and field test, measurement theory, technology on language testing, and ethics, fairness and policy. In the end, they provide an epilogue to provide an opportunity for further rethinking and reconsideration of language testing and assessment into the future.	research efforts in the coming decade.

Table 4

Analysis of Interviews

Interview	Annotation	Implications
Fulcher (2007b)	Fulcher elaborates on his interest in and familiarity with language testing and then indicates that language testing community is gradually growing in ways that gives the stakeholders confidence. Talking about professionalism, he discusses the evolution of codes in the International Language Testing Association (ILTA). He argues for the importance of effect-driven testing, evidence-centered design, and the philosophical basis for test purpose, intended test use, the end test users and consequences. Finally, he provides advice for classroom teachers and suggests that teachers should use their skills and creative talents to produce proficient learners.	The interview communicates some effective messages regarding the role of language testing and its usefulness in real classroom contexts.
Fulcher (2010c)	Learners may be considered as "consumers" partly due to the fact that they use the language to work, study, learn or socially integrate. What he recommends is that the testing products should be well made and useful for the intended purposes. Then he argues that in order for the politicians not to lose their position in the global economic market, they try to control the educational system to generate the kind of society they wish to create. As learners will prepare for tests, he argues that teaching to the test should not focus on practicing test items, but rather should develop communicative skills and abilities which will in turn boost the score for construct relevant reasons. Finally, he believed that technology provides communicative materials and opportunities and computer-based tests should be also noticed, as computer scoring seems efficient and reliable.	Globalized market and needs make the learners consumers; teaching to the test should be appropriately treated and the role of technology and automated scoring should be considered.

Fulcher (2018b)	Fulcher first credits and acknowledges the work of others conducted in the past, and deals with the ethical issues, social consequences, philosophical foundations and more significantly elaborates on effect-driven testing. Then, he argues that the testing agencies and the policy makers are directly responsible for the unintended consequences of the tests. Considering test purpose, test retrofit and test consequences, he also states that it is natural for high-stakes tests, such as IELTS and TOEFL iBT, to be challenged. He concludes the with a consideration of ethics and fairness,	The history testing and assessment, along with the study of changing values and ethical systems, is important to understanding fairness in the present.

Discussion and Reflection (Glenn Fulcher)

I was somewhat taken aback when asked to read the analysis of my publications by Hassan and Ali, as it had never occurred to me that anyone would think the outcome of such a review might be either informative, or indeed of interest, to a wider audience. But without their commitment and thoroughness, I would not have had such a tangible focus for the questions that everyone from time immemorial asks of themselves as an active career draws to a close: in the words of Marcus Aurelius, "In this river, then, where there can be no foothold, what should anyone prize of all that races past him?" (Meditations, Book 4, v. 36). I therefore register my thanks to them.

My research and publications have emerged from the attempts of a teacher to satisfy his own curiosity about assessment practices, set alongside a realization that social systems like assessment are both contingent and interactive. We can do things in different ways and, if we judge what we do to be substandard, we can improve them. Teachers can make a difference; research-informed choices can improve the world for our students. The social critique of test impact by scholars like Shohamy and McNamara can help focus our attention on where change is needed, or extreme care exercised. My own analysis of the originally unintended use of the CEFR through a process of reification and institutionalization is part of the same enterprise. But unlike many constructivists, I have never believed that what we now refer to as "constructs" are merely socially convenient artefacts. Not everything is socially constructed. It may be that the names for constructs are abstract nouns that we cannot observe, but the maxim of the logical positivists for testing reality is not inevitable. C. S. Peirce asked of a construct, "What does its reality consist of?" His answer was: "Why it consists in something being true of something else that has a more primary mode of substantiality. Here we have, I believe, the materials for a good definition of an abstraction." The meaning of "hardness" lies in our observation of which materials can scratch others, and what cannot be scratched. And so, the meaning of "fluency" lies in our observation of…? I will return to this below, because observations of language are much more complex than "hardness" – despite the attempts of the "new realism" movement to argue otherwise. But these initial comments are designed to show that a life spent in language testing is definitely NOT just about language testing. If we are to do it well, we must concern ourselves with many of the fundamental questions of philosophy: What is knowledge? How do we arrive at knowledge? What is ethical practice? What do we mean by a just society, and what is the role of testing (and education) in creating it? It is no coincidence that the first work of politics by Plato was also a treatise on education. And with the increasing use of artificial intelligence (AI) in automated assessment, what is the nature of mind? Is language a behaviour, or part of consciousness? And what is the role of value systems in making assessment choices?

Social science research cannot be separated from these questions, or from history. Hassan and Ali perceptively recognize this in the very first paragraph of their summary. It is not coincidental that great minds in our field like Bernard Spolsky and Alan Davies have grounded their work in both philosophy and an understanding of the past. And Lado's (1961) work on language testing was conducted to enhance intercultural understanding and communication on "…a basic assumption of and belief in the unity of all mankind. All races have the same origin and are capable of the same emotions and the same needs encompassing the whole range of human experience from hunger and the craving for food to theological inquiry and the seeking of God" (ibid., p. 276). The misrepresentation of Lado in the British communicative language testing movement was as much a motivation for my 2000a paper, as the growing awareness

from discourse studies that their depiction of "real" speech was largely inaccurate. But by the same token, we must also credit the communicative movement with providing an impetus for increased research into the representation of more complex constructs, and the role of context. The test called *The Communicative Use of English as a Foreign Language* produced jointly by Cambridge and the Royal Society of Arts (RSA) in 1988 was truly revolutionary in ways that we would now describe as "integrated" – but probably too avantgarde at the time to survive a single administration. Without an understanding of our history, we are not able to build upon what others have achieved, extend and deepen our knowledge, or identify and fill gaps. Fields of endeavor that are successful know their history and learn from it.

A reflection of this nature permits an anecdote. In a recent seminar for postgraduate students approaching assessment time, programme tutors were invited to field questions. The inevitable happened very early in the proceedings when a student asked: "How many references do I need?" And sadly, the conventional absentminded response materialized: "Somewhere between 15 and 20, and only reference work from the last 5 years to show your awareness is current." My response to this "how long is a piece of string" question would have been very different. In his summary of my book chapter on context in language testing (2015c), the editor writes: "In an engaging chapter which sees the author unafraid to draw upon the still-pertinent ideas of some slightly dusty Victorian scholars, Fulcher employs a series of analogies, which include such disparate pursuits as life-saving, purchasing a new fridge, and wine tasting…." (King, 2015, p. 9). I take the explicit reference to Victorian scholars – who are far from "dusty" and very pertinent – as a great compliment. We ignore the journey of our profession at our peril. The history of ideas is also important when teaching students how to use statistical tools for test analysis. I even draw examples of distributions from the tables in the English version of Quetelet's (1842) *A Treatise on Man*, from which we also derive our modern Body Mass Index. Why? Because, as Hacking (1990, pp. 108-109) puts it so wonderfully:

> "Given a lot of measurement of heights, are these the measurements of the same individual? Or are they the measurements of different individuals? If and only if they are sufficiently like the distribution of figures derived from measurements on a single individual....at this exact point there occurred one of the fundamental transitions in thought, that was to determine the entire future of statistics....Here we pass from a real physical unknown, the height of one person, to a postulated reality, an objective property of a population....This postulated truth unknown value of the mean was thought of not as an arithmetical abstract of real heights, but as itself a number that objectively describes the population."

In social science and psychological research, thinking about deviation from a population mean as error terminated with Galton; but in testing, the standard error of measurement for an individual is still derived from the distribution of the population. It is a double-transition, in Hacking's terms. A move from individuals to a population "reality", and from the new postulated reality back to a specific individual. This is but one fascinating example. The general point is that only teaching students about "descriptive" statistics (are they ever descriptive?) and running lab-based classes showing them how to push buttons in SPSS or FACETS, masks the philosophical, historical and social complexity of assessment practice and its assumptions.

At worst, it can also lead to statistical determinism at the expense of understanding people. The title of Hacking's book is highly appropriate: *The Taming of Chance*; as is the festschrift for Alan Davies: *Experimenting with Uncertainty*. The theoretical aspect of construct definition, practical test design, and creating explicit evidence models, help us understand the uncertainty attached to score-based inferences, and the risks associated with subsequent unsound decisions. And at a much more mundane level, ignoring history has caused me as a journal editor to return many papers with the advice to read research that is more than five years old. Reinventing the wheel can be avoided by extensive reading.

I was also surprised to see just how much I have written about assessing speaking, although on reflection it is probably not unexpected. As Lado (1961, p. 239) says, "The ability to speak a foreign language is without doubt the most highly prized language skill, and rightly so…." Without an ability to speak, intercultural communication is severely restricted; and so too is our ability to understand our fellow human beings. As a young teacher, I began by asking why I could not predict the (external) examination grades of my own students. When I studied discourse analysis and experienced the birth of corpus linguistics in 1980s Birmingham, I thought I had found a way to address the conundrum. I hypothesized that scores arrived at through an a-priori evidence model did not reflect my students' speaking ability, because the descriptors on the scale did not "describe" the speech elicited by the tasks. I first explored the idea in an assignment for my MA testing class (Fulcher, 1987), which evolved into doctoral research. My proposed solution was to create "data-based" rating scales arrived at through the analysis of speech generated by test tasks, and later speech in target use domains. The research was messy, as most research is. But I received expert guidance from Charles Alderson, and much welcome critique and support from Caroline Clapham and Dianne Wall.

I discovered (at least) two important things. Firstly, there are constructs which are horrendously difficult to operationalize. One of these was "grammatical accuracy". This was broadly in line with existing evidence from "world Englishes" research (e.g., Lowenberg, 1993); but also with the growing realization that "nativeness" was enormously complex (Davies, 2011). The demise of the criterion of "educated native speaker" as a hook upon which to hang lower-level descriptors became inevitable. Secondly, uninterpreted observational elements could not be scaled because they were not linear, and counting them did not correlate with speaking proficiency. This is why our constructs in language testing are just not as uncomplicated as Peirce's "hardness", which can be defined by a set of simple observations. The only strong validity evidence to emerge from my research was for a fluency rating scale (Fulcher, 1993; 1996a) that was constructed from high-inference categories. For example, the low-inference "counting" of number and length of pauses did not predict speaking proficiency, but the interpretation of *why* the pauses occurred (e.g., speech repair, turn-taking, content planning, humour) did. Eight interpretive categories were found to account for the data in the speech corpus, and descriptors were generated using discriminant analysis. The robustness of these descriptors in my research was confirmed during scaling for the Common European Framework of Reference. Using my fluency descriptors, North (2007, p. 657) reported that "…the fluency descriptors proved to have a rock-solid interpretation of difficulty across sectors, regions, and languages, and so…they were used as anchor items in the project…." The data-based approach has taken a number of different turns in subsequent years, but the

fundamental principles have been widely embedded into the practice of rating scale design. What was novel in the 1980s and 1990s is now mainstream (Knoch et al., 2021).

But I believe the research is important beyond the immediate practical application to assessing speaking. The ability to comprehend language and meaning – the heart of Lado's goal of intercultural communication and understanding – is a fundamental human ability. It is part of what we are; it defines our humanity. It IS a high-inference activity. And yet, all approaches to the automated assessment of speech rely entirely on low-inference categories. That is, algorithms count and quantify what machines can readily identify. What I have called "the folly of low-inference categories" (Fulcher, 2015b, pp. 72-77) rests upon an assumption that observable phenomena are direct realizations of language processing capacity; in circular argumentation, it is also claimed they can be used as indicators of their cause for scoring. This is a "software solution" to the theory of mind, which has most famously been challenged by Searle (1980) in his Chinese room analogy. Searle (2002, p. 15) puts the problem like this: "Instead of recognizing that consciousness is essentially a subjective, qualitative phenomenon, many people mistakenly suppose that its essence is that of a control mechanism or a certain kind of set of dispositions to behavior or a computer program". The solution to the problem of language and mind for Searle is that "…all meaning and understanding goes on against a background which is not itself meant or understood, but which forms the boundary conditions on meaning and understanding, whether in conversations or in isolated utterances" (ibid., p. 202). In the science fiction of Star Trek, this is what Data cannot achieve in his striving to be "more human". In Peircean terms, the relationship between language and mind may be characterized as one of "evolutionary love" (Peirce, 1882/1998). All of which is echoed in theology: "if there were any need of proof of how utterly man is rooted in mankind, one only need pause at the fact of language" (Ebeling, 1993, p. 92). Indeed, without language there could be no "courage to be", for "In every encounter with reality the structures of self and world are independently present. The most fundamental expression of this fact is the language which gives man the power to abstract from the concretely given and, after having abstracted from it, return to it, to interpret and transform it" (Tillich, 1952, p. 82).

And so, I conclude that the attempt to create a model of language and mind through latent trait modelling or correlational data is futile. Cattell and Galton are worth reading still, but the days of "mental mining" are well and truly in the past. The automated scoring of speaking may serve a useful function if, and only if, there is a clear link between low-inference categories and processing ability, which occurs most frequently in the early stages of language learning. But the evidential link soon evaporates along with validity, and our regard for humanity.

The last sentence was intentionally provocative. It shows that we exist in an endless state of tension between values/beliefs, empirical evidence, theory building, social policy and commercial viability. But we should not be afraid of this. Peirce (1863; 1958, p. 11) helps explain why: "Human learning must fail somewhere. Materialism fails on the side of incompleteness. Idealism always presents a systematic totality, but it must always have some vagueness and thus lead to error. Materialism is destitute of philosophy. Thus, it is necessarily one-sided…. But if materialism without idealism is blind, idealism without materialism is void." I would argue that the evidence gained from rating scale research supports a particular set of values and beliefs about language and mind, and what it means to be human. This may, of course, be challenged. But the value implications of the alternative, as well as the evidential

basis and policy implications, should be made explicit (see Fulcher & Davidson, 2008, for an historical example). In language testing, this is part of what Fred Davidson and I refer to as "effect-driven testing".

My career-long interests are also directly reflected in the model of Language Assessment Literacy (LAL) presented in Fulcher (2012), and particularly the three-tier model (ibid., 126) of contexts, principles and practices. I have enjoyed teaching immensely – both language and language testing. And so, it was inevitable that how I teach would be increasingly influenced by the research. Fred Davidson and I were very proud that Alan Davies had written of our book (Fulcher & Davidson, 2007) that it "…does seem to provide the most complete coverage of skills, knowledge and principles" (Davies, 2008, p. 341). Fulcher (2010b) attempted to expand this coverage based on the 2012 model (the research having been conducted in 2009), and go further in developing tasks and activities that improved on what Annie Brown (2011) described as a "deliberate pedagogy". Read (2011) seems to agree that this was achieved. Although it was perhaps too difficult for the intended audience, which was supposed to be pitched somewhere between Douglas (2010) and Fulcher & Davidson (2012) (all texts by this stage residing with Routledge). Fulcher (2015b) was the subsequent attempt to address the "contexts" part of the LAL model in a single volume, although it was certainly not written as a pedagogic text. Along with the website (http://languagetesting.info), which had existed since 1995, but updated in 2009, I had what I thought to be a complete set of pedagogic resources for teaching language testing. The one piece of the jigsaw that was missing was an account of how I used the resources in my own teaching. I first articulated this at a conference on LAL organized at Lancaster University, although I don't recall the date; it was expanded for a paper delivered for the TALE project at the University of Cyprus in 2018, and published as Fulcher (2020). This chapter articulates the model of the language tester as a pre-Aristotelian craftsman, using an understanding of theory and the world to fashion an artefact that either enables meritocratic societies to function (high-stakes proficiency), or supports Deweyan-style learning and personal growth (low-stakes formative). Students of language testing are apprentices who learn by doing: using theory to design, research to create, values to assess. And through the subsequent practice of our craft, we make a small (often unseen) contribution to improving people's lives. That's what being a language tester is for me, at least. Idealistic? Yes. Optimistic? For sure. But I'm not that keen on the alternatives. Oh yes – and it's been *fun*.

References

Brown, A. (2011). Book review: Glenn Fulcher and Fred Davidson; *Language testing and assessment: An advanced resource book. Routledge*: London and New York, 2007, xix + 367 pp.: 9780415339469 (hbk), 9780415339476 (pbk). *Language Testing, 28*(1), 145-148. https://doi.org/10.1177/0265532210386932

Browne, K., & Fulcher, G. (2017). Pronunciation and Intelligibility in Assessing Spoken Fluency. In T. Isaacs & P. Trofimovich (Eds.), *Second Language Pronunciation Assessment: Interdisciplinary perspectives Bristol: Multilingual Matters*. https://doi.org/10.21832/ISAACS6848

Chalhoub-Deville, M., & Fulcher, G. (2003). The Oral Proficiency Interview: A Research Agenda. *Foreign Language Annals, 36*(4), 498-506. https://doi.org/10.1111/j.1944-9720.2003.tb02139.x

Davidson, F., & Fulcher, G. (2007). The Common European Framework of Reference (CEFR) and the design of language tests: A matter of effect. *Language Teaching, 40*(3), 231-241. https://doi.org/10.1017/S0261444807004351

Davidson, G., & Fulcher, G. (2012). Developing test specifications for language assessment. In Coombe, C., & Stoynoff, S., Osullivan, B. & Davidson, P. (Eds.), *The Cambridge guide to second language assessment* (pp. 59-65). Cambridge: Cambridge University Press.

Davies, A. (2008). Textbook trends in teaching language testing. *Language testing*, *25*(3), 327-347. https://doi.org/10.1177/0265532208090156

Davies, A. (2011). Does language testing need the native speaker? *Language Assessment Quarterly*, *8*(3), 291-308. https://doi.org/10.1080/15434303.2011.570827

Douglas, D. (2010). *Understanding Language Testing*. Abbingdon, Oxon: Hodder Education. https://doi.org/10.1177/0265532210373604

Ebeling, G. (1993). *Introduction to a Theological Theory of Language*. London: Collins.

Fulcher, G. (1987). Tests of oral performance: the need for data-based criteria. *ELT Journal, 41(4),* 287-291. https://doi.org/10.1093/elt/41.4.287

Fulcher, G. (1988a). *Lexis and Reality in Oral Evaluation*. Revised and expanded version of a paper presented at the Annual Meeting of the International Association of Teachers of English as a Foreign Language (22nd, Edinburgh, Scotland, April 11-14, 1988).

Fulcher, G. (1988b). The EFL classroom as a place. *Education Today: Journal of the College of Preceptors*, *38*, 107.

Fulcher, G. (1989). Cohesion and coherence in theory and reading research. *Journal of Research in Reading, 12(2),* 146-163. https://doi.org/10.1111/j.1467-9817.1989.tb00163.x

Fulcher, G. (1991a). Conditionals revisited. *ELT Journal, 45(2),* 164-168. https://doi.org/10.1093/elt/45.2.164

Fulcher, G. (1991b). The role of assessment by teachers in schools. In T. Caudery (Ed.), *New Thinking in TEFL*. (The Dolphin Series, No. 21), Denmark, Aarhus University Press, 138-158.

Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language* [Unpublished PhD Dissertation]. Lancaster: University of Lancaster

Fulcher, G. (1995). Variable competence in second language acquisition: A problem for research methodology? *System*, *23*(1), 25-33. https://doi.org/10.1016/0346-251X(94)00055-B

Fulcher, G. (1996a). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13(2),* 208-238. https://doi.org/10.1177/026553229601300205

Fulcher, G. (1996b). Testing tasks: issues in task design and the group oral. *Language Testing, 13*(1), 23-51. https://doi.org/10.1177/026553229601300103

Fulcher, G. (1996c). Invalidating validity claims for the ACTFL oral rating scale. *System*, *24*(2), 163-172. https://doi.org/10.1016/0346-251x(96)00001-2

Fulcher, G., & Bamford, R. (1996). I didn't get the grade I need. Where's my solicitor? *System, 24(4),* 437-448. https://doi.org/10.1016/s0346-251x(96)00040-1

Fulcher, G. (1997a). An English language placement test: issues in reliability and validity. *Language Testing, 14(2),* 113-139. https://doi.org/10.1177/026553229701400201

Fulcher, G. (1997b). Text difficulty and accessibility: Reading formulae and expert judgement. *System, 25(4),* 497-513. https://doi.org/10.1016/s0346-251x(97)00048-1

Fulcher, G. (1997c). *Writing in the English Language Classroom*. Macmillan.

Fulcher, G. (1998a). Widdowson's model of communicative competence and the testing of reading: an exploratory study. *System, 26(3)* ,281-302. https://doi.org/10.1016/s0346-251x(98)00020-7

Fulcher, G. (1998b). Computer-based language testing: The call of the Internet. In C. A. Coombe (Ed.), *Current Trends in English language Testing* (pp. 1-14). UAE: Al Ain University Press.

Fulcher, G (1999a). Assessment in English for academic purposes: putting content validity in its place. *Applied Linguistics, 20(2),* 221-236. https://doi.org/10.1093/applin/20.2.221

Fulcher, G. (1999b). Computerizing an English language placement test. *ELT Journal, 53(4),* 289-299. https://doi.org/10.1093/elt/53.4.289

Fulcher, G. (1999c). Book Review: A history of foreign language testing in the United States: from its beginnings to the present. *Language Testing, 16(3),* 389-394. https://doi.org/10.1177/026553229901600307

Fulcher, G. (1999d). Ethics in language testing. *TAE SIG Newsletter, 1(1),* 1-4.

Fulcher, G., & Locke, D. (1999). Distance education: The future of library and information services requirements. *Distance Education, 20(2),* 313-329. https://doi.org/10.1080/0158791990200209

Fulcher, G. (2000a). The 'communicative' legacy in language testing. *System, 28(4),* 483-497. https://doi.org/10.1016/s0346-251x(00)00033-6

Fulcher, G. (2000b). Computers in language testing. In Brett, P. & Motteram, G. (Eds.) *A Special Interest in Computers: Learning and teaching with information and communications technologies*. Manchester: IATEFL publications, 93-107.

Fulcher, G. (2003a). Interface design in computer-based language testing. *Language Testing, 20(4),* 384-408. https://doi.org/10.1191/0265532203lt265oa

Fulcher, G. (2003b). *Testing Second Language Speaking*. London and New York: Routledge. https://doi.org/10.4324/9781315837376

Fulcher, G., & Rosina, M. R. (2003). Task difficulty in speaking tests. *Language Testing, 20(3),* 321-344. https://doi.org/10.1191/0265532203lt259oa

Fulcher, G. (2004a). Deluded by artifices? the common European framework and harmonization. *Language Assessment Quarterly,1(4),* 253-266. https://doi.org/10.1207/s15434311laq0104_4
Fulcher, G. (2004b). Are Europe's tests being built on an unsafe framework? *Guardian Weekly, 18th March.* Availablehttp://education.guardian.co.uk/tefl/story/0,,1170569,00.
Fulcher, G. (2005). *Better Communications Test will Silence the Critics*. Guardian Education. Retrieved on 18 November from https://www.theguardian.com/education/2005/nov/18/tefl3
Fulcher, G. (2006). Test architecture. *Foreign Language Education Research, 9,* 1-22.
Fulcher, G. (2007a). Universities undermine their own foundations. *The Guardian Weekly TEFL Supplement, 13,* 74-96.
Fulcher, G. (2007b). *An Interview with Glenn Fulcher. Shiken, Japan*, JALT.
Fulcher, G. (2008). Criteria for Evaluating language quality. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education (2nd Ed.) Language Testing and Assessment* (pp. 157-176). New York, NY: Springer.
Fulcher, G. (2009a). Test Use and Political philosophy. *Annual Review of Applied Linguistics, 29,* 3-20. https://doi.org/10.1017/S0267190509090023
Fulcher, G. (2009b). The commercialization of language provision at university. In J. C. Alderson (Ed.), *The Politics of Language Education: People and Institutions* (pp. 125-146). London: Multilingual Matters.
Fulcher, G. (2010a). The reification of the Common European Framework of Reference (CEFR) and effect-driven testing. In A. Psaltou-Joycey & M. Matthaioudakis (Eds.), *Advances in Research on Language Acquisition and Teaching* (pp. 15-26). Thessaloniki: GALA.
Fulcher, G. (2010b). *Practical Language Testing.* London: Hodder Education/Routledge. https://doi.org/10.4324/980203767399
Fulcher, G. (2010c). *Glenn Fulcher Talks to ELT News*. Athens, Greece.
Fulcher, G. (2012). Assessment Literacy for the Language Classroom. *Language Assessment Quarterly, 9(2),* 113-132. https://doi.org/10.1080/15434303.2011.642041
Fulcher, G. (2013a). Language testing in the dock. In Kunnan, A. J. (Ed.), *The Companion to Language Testing* (pp. 1553-1570). London: Wiley-Blackwell.
Fulcher, G. (2013b). Test Design and Retrofit. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 5809 - 5817). Malden MA: Wiley Blackwell. https://doi.org/10.1002/9781405198431.wbeal1199
Fulcher, G. (2013c). Scoring performance tests. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 392-406). Routledge. https://doi.org/10.4324/9780203181287
Fulcher, G. (2014). Philosophy and Language Testing. In A. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1431-1451). London: Wiley-Blackwell. https://doi.org/10.1002/9781118411360.
Fulcher, G. (2015a). Assessing second language speaking. *Language Teaching*, 48(2), 198-216. https://doi.org/10.1017/S0261444814000391
Fulcher, G. (2015b). *Re-examining language testing: A philosophical and Social Inquiry.* London and New York: Routledge. https://doi.org/10.4324/9781315695518
Fulcher, G. (2015c). Context and inference in language testing. In King, J. (Ed.), *Context and the Learner in Second Language Learning* (pp. 225 -241). London: Palgrave Macmillan.
Fulcher, G. (2015d). Assessing second language speaking. *Language Teaching, 48(2),* 198-216. https://doi.org/10.1017/S0261444814000391
Fulcher, G. (2016a). The Practice of Language Assessment. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (pp. 463-475). New York: Routledge. https://doi.org/10.4324/9781315716893
Fulcher, G. (2016b). Standards and Frameworks. In J. Banerjee & D. Tsagari (Eds.), *Handbook of Second Language Assessment* (pp. 29-44). Berlin: De Gruyter. https://doi.org/10.1515/9781614513827-005
Fulcher, G. (2018a). Assessing Spoken Production, in *The TESOL Encyclopedia of English Language Teaching, Vol 8, (PP 4900-4906). Edited by* Liontas, John I. (Project Editor DelliCarpini, Margo; Volume Editor: Christine Coombe), 1-6. https://doi.org/10.1002/9781118784235.eelt0364
Fulcher, G. (2018b). *Glenn Fulcher Talks to Bahram Behin*. JALDA, Iran.
Fulcher, G. (2019). Cultivating language assessment literacy as collaborative CPD. In Gillway, M. (Ed.), *Addressing the State of the Union: Working Together, Learning Together* (pp. 27-35). Garnet.
Fulcher, G. (2020). Operationalizing Language Assessment Literacy. In Tsagari, D. (Ed.), *Language Assessment Literacy: From Theory to Practice* (pp. 8-28). Cambridge Scholars.
Fulcher, G. (2021a). Language Assessment Literacy in a Learning-Oriented Assessment Framework. In A. Gebril (Ed.), *Learning-oriented assessment: Putting theory into practice* (pp. 254-270). New York: Routledge. https://doi.org/10.4324/9781003014102
Fulcher, G. (2021b). Language Testing. In Mohebbi, H., & Coombe, C. (Eds.), *Research Questions in Language Education and Applied Linguistics* (pp. 349-352). London: Springer.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment: An advance Resource Book.* Routledge: London and New York.

Fulcher, G., & Davidson, F. (2008). Tests in Life and Learning: A deathly dialogue. *Educational Philosophy and Theory, 40(3),* 407-417. https://doi.org/10.1111/j.1469-5812.2007.00358.x

Fulcher, G., & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26(1),* 123-144. https://doi.org/10.1177/0265532208097339

Fulcher, G., & Davidson, F. (2012). *The Routledge Handbook of Language Testing.* Routledge. https://doi.org/10.4324/9780203181287

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28(1),* 5-29. https://doi.org/10.1177/0265532209359514

Fulcher, G., & Harding, L. (2022). *The Routledge Handbook of Language Testing.* Routledge. https://doi.org/10.4324/9781003220756

Fulcher, G., & Owen, N. (2016). Dealing with the Demands of Language Testing and Assessment. In H. Graham (Ed.), *The Routledge Handbook of English Language Teaching. Routledge Handbooks in Applied Linguistics* (pp. 109-120). Oxford: Routledge.

Fuclcher, G., & Svalberg, A. (2013). Limited aspects of reality: Frames of reference in language assessment. *International Journal of English Studies, 13(2),* 1-19. https://doi.org/10.6018/ijes.13.2.184061

Green, A., & Fulcher, G. (2021). *Test Design Cycle*. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second language Acquisition and Language Testing* (pp. 69-77). New York/London: Routledge.

Hacking, I. (1990). *The Taming of Chance*. Cambridge: Cambridge University Press.

King, J. (Ed.) (2015). *The Dynamic Interplay Between Context and the Language Learner*. London: Palgrave Macmillan.

Knoch, U., Deygers, B., & Khamboonruang, A. (2021). Revisiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing, 38*(4), 602-626. https://doi.org/10.1177/0265532221994052

Kongsuwannakul, K., Fulcher, G., & Smith, N. (2015). *Prototyping a concordance cloze test: Preliminary results.* Proceedings of 7th International Conference on Humanities and Social Sciences: ASEAN 2015: Challenges and Opportunities. School of Foreign Languages Institute of Social Technology Suranaree University of Technology.

Lado, R. (1961). *Language Testing.* London: Longman

Lowenberg, P. H. (1993). Issues of validity in tests of English as a world language: Whose standards? *World Englishes, 12*(1), 95-106.

Marquez Reiter, R., Rainey, I., & Fulcher, G. (2005). A Comparative Study of Certainty and Conventional Indirectness: Evidence from British English and Peninsular Spanish. *Applied Linguistics, 26(1),* 1-31. https://doi.org/10.1093/applin/amh018

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (pp. 13-103).* New York: Macmillan/American Council on Education.

Morrow, K. (1979). *Communicative language testing: revolution of evolution?* In Brumfit, C.K. & Johnson, K. (Eds.), The Communicative Approach to Language Teaching. Oxford University Press, Oxford, pp. 143-159.

North, B. (2007). The CEFR Illustrative Descriptor Scales. *The Modern Language Journal, 91*(4), 656-659. **https://doi.org/10.1111/j.1540-4781.2007.00627_3.x**

Peirce, C. S. (1882/1998) Evolutionary Love. In C. E. Moore (Ed.), *The Essential Writings of Charles S. Peirce* (pp. 237-260). New York: Prometheus Books.

Peirce, C. S. (1863/1958) The place of our age in the history of civilization. In P. P. Wiener (Ed.) *Values in a Universe of Chance. Selected Writings of Charles S. Peirce*. (pp. 4 -14). New York: Doubleday Anchor Books.

Quetelet, A. (1842). *A Treatise on Man*. Reprinted 1968. New York: Burt Franklin.

Read, J. (2011). Review of Book review: G. Fulcher (2010). *Practical language testing. London: Hodder Education.* 304 pp. ISBN: 9780340984482. *Language Testing, 28*(2), 302-304. **https://doi.org/10.1177/0265532210394641**

Searle, J. R. (1980). *Minds, Brains, and Programs*. *Behavioral and Brain Sciences, 3*(3), 417-424. http://doi.org/10.1017/S0140525X00005756

Searle, J. R. (2002). *Consciousness and Language*. Cambridge: Cambridge University Press.

Tillich, P. (1952). *The Courage to Be*. New Haven and London: Yale University Press.

Spolsky, B. (1995). *Measured Words*. Oxford: Oxford University Press.

Spolsky, B. (1976). *Language Testing: Art or Science?* [Conference Paper]. The 4th International Congress of Applied Linguistics. Stuttgart, Germany.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12. https://doi.org/10.1093/elt/49.1.3

Xing, P., & Fulcher, G. (2007). Reliability assessment for two versions of Vocabulary Levels Tests. *System, 35(2),* 182-191. https://doi.org/10.1016/j.system.2006.12.009

Yi, J., & Fulcher, G. (2018). Strategy Use in the TOEFL iBT Speaking and Academic Classroom. *Korean Journal of Applied Linguistics, 34(1),* 223-252. https://doi.org/10.17154/kjal.2018.3.34.1.219

Acknowledgments

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

Rights and Permissions

Open Access