Research timeline: Assessing Second Language Speaking

Glenn Fulcher

University of Leicester, United Kingdom

School of Education

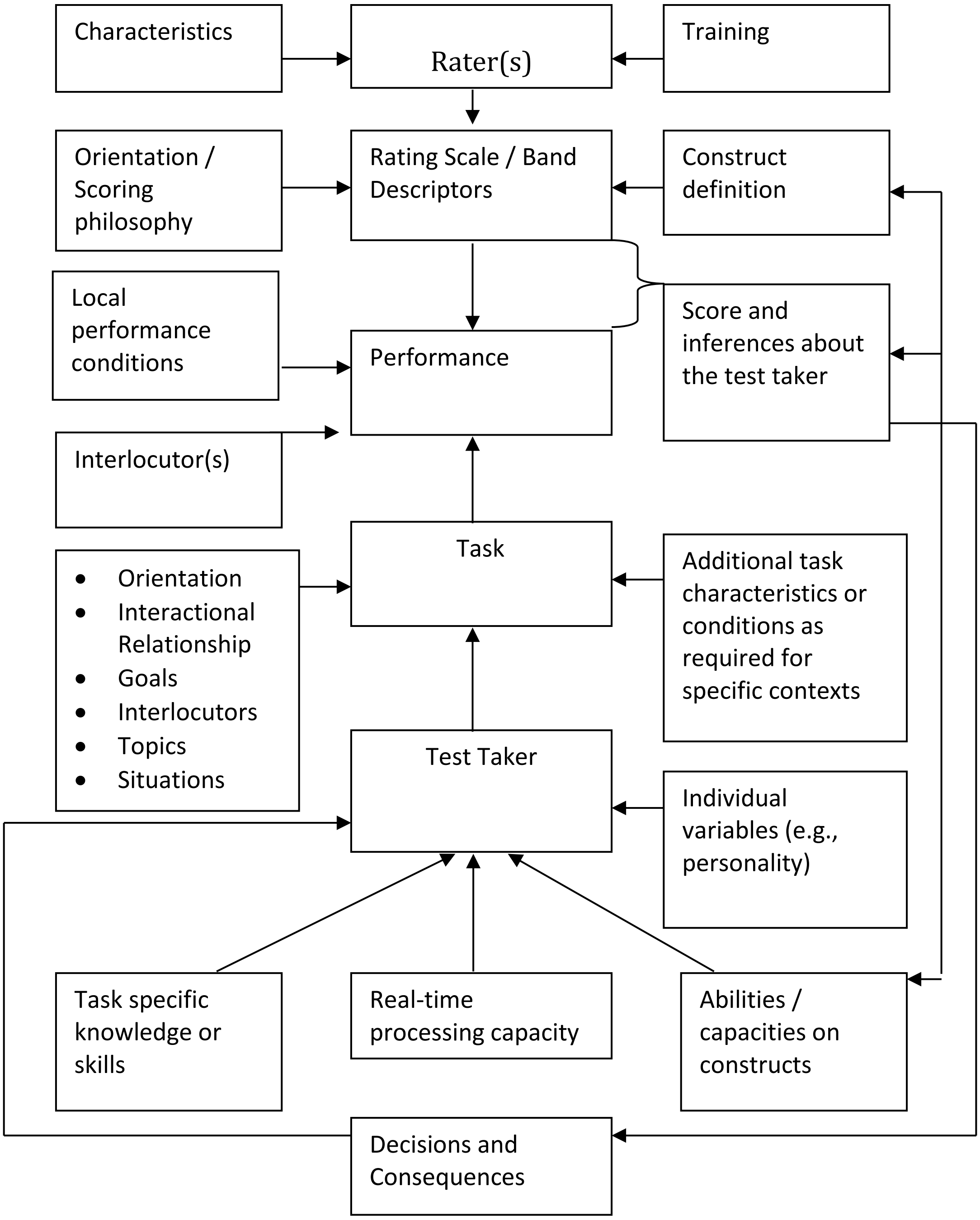
Biodata: Glenn Fulcher is Professor of Education and Language Assessment at the University of Leicester, and Head of the School of Education. He has published widely in the field of language testing, from journals such as *Language Testing*, *Language Assessment Quarterly*, *Applied Linguistics* and *System*, to monographs and edited volumes. His books include *Testing second language speaking* (Longman, 2003), *Language testing and assessment: An advanced resource book* (Routledge 2007), *Practical language testing* (Hodder 2010), and the *Routledge handbook of language testing* (Routledge 2012). He currently co-edits the Sage journal *Language Testing*.

Introduction

While the viva voce (oral) examination has always been used in content-based educational assessment (Latham 1877, p. 132), the assessment of second language speaking in performance tests is relatively recent. The impetus for the growth in testing speaking during the 19th and 20th Centuries is twofold. Firstly, in educational settings the development of rating scales was driven by the need to improve achievement in public schools, and to communicate that improvement to the outside world. Chadwick (1864) implies that the rating scales first devised in the 1830s served two purposes: providing information to the classroom teacher on learner progress for formative use, and generating data for school accountability. From the earliest days, such data was used for parents to select schools for their children in order to ‘maximize the benefit of their investment’ (Chadwick 1858). Secondly, in military settings it was imperative to be able to predict which soldiers were able to undertake tasks in the field without risk to themselves or other personnel (Kaulfers 1944). Many of the key developments in speaking test design and rating scales are linked to military needs.

The speaking assessment project is therefore primarily a practical one. The need for speaking tests has expanded from the educational and military domain to decision making for international mobility, entrance to higher education, and employment. But investigating how we make sound decisions based on inferences from speaking test scores remains the central concern of research. A model of speaking test performance is essential in this context, as it helps focus attention on facets of the testing context under investigation. The first such model developed by Kenyon (1992) was subsequently extended by McNamara (1995), Milanovic & Saville (1996), Skehan (2001), Bachman (2001), and most recently by Fulcher (2003, p. 115), providing a framework within which research might be structured. The latter is reproduced here to indicate the extensive range of factors that have been and continue to be investigated in speaking assessment research, and these are reflected in my selection of themes and associated papers for this timeline.

Figure 1. An expanded model of speaking test performance (Fulcher 2003, p. 115).


Characteristics
Rater(s)
Training
Orientation / Scoring philosophy
Rating Scale / Band Descriptors
Construct definition
Local performance conditions
Score and inferences about the test taker
Performance
Interlocutor(s)
Task
• Orientation
• Interactional Relationship
• Goals
• Interlocutors
• Topics
• Situations
Additional task characteristics or conditions as required for specific contexts
Test Taker
Individual variables (e.g., personality)
Task specific knowledge or skills
Real-time processing capacity
Abilities / capacities on constructs
Decisions and Consequences

Overviews of the issues illustrated in figure 1 are discussed in a number of texts devoted to assessing speaking that I have not included in the timeline (Fulcher 2003; Lazaraton 2002; Luoma 2004; Taylor (ed. 2011). Rather, I have selected publications based on 12 themes that arise from these texts, from figure 1, and from my analysis of the literature.

Themes that pervade the research literature are rating scale development, construct definition, operationalisation, and validation. Scale development and construct definition are inextricably bound together because it is the rating scale descriptors that define the construct. Yet, rating scales are developed in a number of different ways. The data-based approach requires detailed analysis of performance. Others are informed by the views expert judges using performance samples to describe levels. Some scales are a patchwork quilt created by bundling descriptors from other scales together based on scaled teacher judgments. How we define the speaking construct and how we design the rating scale descriptors are therefore interconnected. Design decisions therefore need to be informed by testing purpose and relevant theoretical frameworks.

Underlying design decisions are research issues that are extremely contentious. Perhaps these can be presented in a series of binary alternatives to show stark contrasts, although in reality there are clines at work.

Specific Purposes Tests vs. Generalizability. Should the construct definition and task design be related to specific communicative purposes and domains? Or is it possible to produce test scores that are relevant to any and every type of real-world decision that we may wish to make? This is critical not least because the more generalizable we wish scores to be, the more difficult it becomes to select test content.

Psycholinguistic Criteria vs. Sociolinguistic Criteria. Closely related to the specific purpose issue is the selection of scoring criteria. Usually, the more abstract or psycholinguistic the criteria used, the greater the claims made for generalizability. These criteria or ‘facilities’ are said to be part of the construct of speaking that is not context dependent. These may be the more traditional constructs of ‘fluency’ or ‘accuracy’, or more basic observable variables related to automaticity of language processing, such as response latency or speed of delivery. The latter are required for the automated assessment of speaking. Yet, as the generalizability claim grows, the relationship between score and any specific language use context is eroded. This particular antithesis is not only a research issue, but one that impacts upon the commercial viability of tests; it is therefore not surprising that from time to time the arguments flare up, and research is called into the service of confirmatory defence (Chun 2006; Downey et al. 2008).

Normal Conversation vs. Domain Specific Interaction. It is widely claimed that the ‘gold standard’ of spoken language is ‘normal’ conversation, loosely defined as interactions in which there are no power differentials, so that all participants have equal speaking rights. Other types of interaction are compared to this ‘norm’ and the validity of test formats such as the interview are brought into question (e.g. Johnson 2001). But we must question whether ‘friends chatting’ is indeed the ‘norm’ in most spoken interaction. In higher education, for example, this kind of talk is very rare, and scores from simulated ‘normal’ conversations are unlikely to be relevant to communication with a professor, accommodation staff, or library assistants. Research that describes the language used in specific communicative contexts to support test design is becoming more common, such as that in academic contexts to underpin task design (Biber 2006).

Rater Cognition vs. Performance Analysis. It has become increasingly common to look at ‘what raters pay attention to’. When we discover what is going on in their heads, should it be treated as construct irrelevant if it is at odds with the rating scale descriptors and/or an analysis of performance on test tasks? Or should it be used to define the construct and populate the rating scale descriptors? Do all raters bring the same analysis of performance to the task? Or are we merely incorporating variable degrees of perverseness that dilutes the construct? The most challenging question is perhaps: Are rater perceptions at odds with reality?

Freedom vs. Control. Left to their own devices, raters tend to vary in how they score the same performance. The variability decreases if they are trained; and it decreases over time through the process of social moderation. With repeated practice raters start to interpret performances in the same way as their peers. But when severed from the collective for a period of time, judges begin to reassert their own individuality, and disagreement rises. How do we identify and control this variability? This question now extends to interlocutor behaviour, as we know that interlocutors provide differing levels of scaffolding and support to test takers. This variability may lead to different scores for the same test taker depending on which interlocutor they work with. Much work has been done in the co-construction of speech in test contexts. And here comes the crunch. For some, this variation is part of a richer speaking construct and should therefore be built into the test. For others, the variation removes the principle of equality of experience and opportunity at the moment of testing, and therefore the interlocutors should be controlled in what they say. In face-to-face speaking tests we have seen the growth of the interlocutor frame to control speakers, and proponents of indirect speaking tests claim that the removal of an interlocutor eliminates subjective variation.

Publications selected to illustrate a timeline are inevitably subjective to some degree, and the list cannot be exhaustive. My selection avoids clustering in particular years or decades, and attempts to show how the contrasts and themes identified play out historically. You will notice that themes H and I are different from the others in that they are about particular methodologies. I have included these because of their pervasiveness in speaking assessment research, and may help others to identify key discourse or multi-faceted Rasch measurement studies (MFRM). What I have not been able to cover is the assessment of pronunciation and intonation, or the detailed issues surrounding semi-direct (or simulated) tests of speaking, both of which require separate timelines. Finally, I am very much aware that the assessment of speaking was common in the United Kingdom from the early 20th Century. Yet, there is sparse reference to research outside the United States in the early part of the of the timeline. The reason for this is that apart from Roach (1945, reprinted as an appendix in Weir, Vidaković & Galaczi (2013) (eds.) there is very little published research from Europe (Fulcher 2003, p. 1). The requirement that research is in the public domain for independent inspection and critique was a criterion for selection in this timeline. For a retrospective interpretation of the early period in the United Kingdom with reference to unpublished material and confidential internal examination board reports to which we do not have access, see Weir & Milanovic (2003) and Vidaković & Galaczi (2013).

Themes

A. Rating scale development

B. Construct definition and validation

C. Task design and format

D. Specific purposes testing and generalizability

E. Reliability and rater training

F. The native speaker criterion

G. Washback

H. Discourse analysis

I. Multi-faceted Rasch Measurement (MFRM)

J. Interlocutor behaviour and training

K. Rater cognition

L. Test-taker characteristics

References

Bachman, L. F. (2001). Speaking as a realization of communicative competence. Paper presented at the meeting of the American Association of Applied Linguistics. St. Louis, Missouri, February.

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

Chadwick, E. (1858). On the economical, social, educational, and political influences of competitive examinations, as tests of qualifications for admission to the junior appointments in the public service. *Journal of the Statistical Society of London* 21.1, 18 – 51.

Chadwick, E. (1864). Statistics of educational results. *Museum: A Quarterly Magazine* of *Education, Literature and Science* 3, 479-484.

Chun, C. W. (2006). Commentary: An Analysis of a Language Testing for Employment: The Authenticity of the PhonePass Test. *Language Assessment Quarterly* 3.3, 295 – 306.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M. & Van Moere, A. (2008). Evaluation of the Usefulness of the Versant for English Test: A Response. *Language Assessment Quarterly* 5.2, 160 – 167.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Longman/Pearson Education.

Johnson, M. (2001). *The Art of Non-conversation. A re-examination of the validity of the Oral Proficiency Interview*. New Haven and London: Yale University Press.

Kaulfers, W. V. (1944). War-time developments in modern language achievement tests. *Modern Language Journal* 28, 136 – 150.

Kenyon, D. (1992). Introductory remarks at symposium on development and use of rating scales in language testing. Paper delivered at the 14th Language Testing Research Colloquium, Vancouver, March.

Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge: Dighton, Bell and Company.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests*. Cambridge: Cambridge University Press.

Luoma, S. (2004). *Assessing second language speaking*. Cambridge: Cambridge University Press.

McNamara, T. F. (1995). Modelling performance: Opening Pandora's Box. *Applied Linguistics* 16.2, 159 – 179.

Milanovic, M. & Saville, N. (1996). Introduction. In Milanovic, M. (ed.), *Performance testing, cognition and assessment* (pp. 1 – 17). Cambridge: Cambridge University Press.

Skehan, P. (2001). Tasks and language performance assessment. In Bygate, M., Skehan, P. & Swain, M. (eds.), *Researching pedagogic tasks: Second language learning, teaching and testing*. (pp. 167 – 185). London: Longman.

Taylor, L. (2011). *Examining Speaking. Research and practice in assessing second language speaking*. Cambridge: University of Cambridge Press.

Weir, C. & Milanovic, M. (2003). (eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913 – 2002*. Cambridge: Cambridge University Press.

Weir, C. J., Vidaković, I. & Galaczi, E. D. (2013). (eds.), *Measured constructs. A history of Cambridge English language examinations 1913 – 2012*. Cambridge: Cambridge University Press.

Vidaković, I. & Galaczi, E. D. (2013). The measurement of speaking ability 1913 – 2012. In Weir, C. J., Vidaković, I. & Galaczi, E. D. (eds.), *Measured constructs. A history of Cambridge English language examinations 1913 – 2012*. Cambridge: Cambridge University Press.

Year	References	Annotations	Theme
1864	Chadwick, E. (1864). Statistics of educational results. *Museum: A Quarterly Magazine* of *Education, Literature and Science* 3, 479-484. Also see discussion in: Cadenhead, K. & Robinson, R. (1987). Fisher's 'Scale Book': An Early Attempt at Educational Measurement. *Educational Measurement: Issues and Practice* 6.4, 15 – 18.	The earliest record of an attempt to assess second language speaking dates to the first few years after **Rev. George Fisher** became Headmaster of the Greenwich Royal Hospital School in 1834. In order to improve and record academic achievement, he instituted a 'Scale Book', which recorded performance on a scale of 1 to 5 with quarter intervals. A scale was created for French as a second language, with typical speaking prompts to which boys would be expected to respond at each level. The Scale Book has not survived.	**A**
1912	Thorndike, E. L. (1912). The measurement of educational products. *The School Review* 20.5, 289–299.	Scales of various kinds were developed by social scientists like **Galton** and **Cattell** towards the end of the 19th Century, but it was not until the work of Thorndike in the early 20th Century that the definition of each point on an equal interval scale was revived. With reference to speaking German, he suggested that performance samples should be attached to each level of a scale, along with a descriptor that	**A, B**

		summarizes the ability being tested.	
1920	Yerkes, R. M. (1920). What psychology contributed to the war. In R. M. Yerkes (ed.), *The new world of science:* *Its development during the war*. New York, NY: The Century Co, 364 – 389. Also see discussion in: Fulcher, G. (2012). Scoring performance tests. In Fulcher, G. & Davidson, F. (eds.), *The Routledge handbook of language testing*. London and New York: Routledge, 378 – 392.	**Yerkes** describes the development of the first large-scale speaking test for military purposes in 1917. It was designed to place army recruits into language development battalions. It consisted of a verbal section and a performance section (following instructions), with tasks linked to scale level by difficulty. Although the development of the test is not described, the generic approach is outlined, and involved the identification of typical tasks from the military domain that were piloted in test conditions. It is arguably the case that this was the first English for Specific Purposes test based on domain specific criteria. In addition, there was clearly an element of domain analysis to support Criterion-referenced assessment.	**A, B, C, D**
1944	Kaulfers, W. V. (1944). War-time developments in modern language achievement tests. *Modern Language Journal*, 28, 136 – 150. Also see discussion in:	The interwar years saw a rapid growth in large-scale assessment that relied on the multiple-choice item for efficiency. In the Second World War **Kaulfers** quickly realized that these tests could not adequately predict ability to speak in	**A, B, D**

	Velleman, B. L. (2008). The 'scientific linguist' goes to war: the United States A.S.T. program in foreign languages. *Historiographia Linguistica* 35, 385–416.	potentially life-threatening contexts. Teaching and assessment of speaking was quickly geared towards the military context once again. **Kaulfers** presents scoring criteria according to the scope and quality of performance. However, all descriptors are generic and not domain specific.	
1945	Roach, J. O. (1945). *Some problems of oral examinations in modern languages. An experimental approach based on the Cambridge examinations in English for Foreign Students*. University of Cambridge Examinations Syndicate: Internal report circulated to oral examiners and local representatives for these examinations. (Reprinted as facsimile in Weir et al. 2013)	**Roach** was among the first to investigate rater reliability in speaking tests. He was concerned primarily with maintaining 'standards', by which he meant that examiners would agree on which test takers were awarded a pass, a good pass, and a very good pass, on the Certificate of Proficiency in English. He was the first to recommend what we now call 'social moderation' (see MISLEVY 1992) – familiarization with the system through team work, which results in agreement evolving over time.	**E**
1952/ 1958	Foreign Service Institute. (1952/1958). *FSI Proficiency Ratings*. Washington D.C.: Foreign Service Institute. Also see discussion in:	Little progress was made in testing second language speaking until the outbreak of the Korean War in 1950. The Foreign Service Institute (FSI) was established, and the first widely used semantic-differential rating scale put	**A, B, C, D, F**

	Sollenberger, H. E. (1978) Development and current use of the FSI oral interview test. In Clark, J. L. D. (ed.), *Direct testing of speaking proficiency: Theory and application*. Princeton, NJ: Educational Testing Service, 1–12.	into use in 1952. This operationalized the ‘native speaker’ construct at the top band (level six). With the Vietnam war on the horizon, a decision was taken to register the language skills of US diplomatic and military personnel. Work began to expand the FSI scale by adding verbal descriptors at each of the six levels from zero proficiency to native speaker, and to include multiple holistic traits. This went hand in hand with the creation of the Oral Proficiency Interview (OPI), which was a mix of interview, prepared dialogue, and simulation. The wording of the 1958 FSI scale and the tasks associated with the OPI have been copied into many other testing systems still in use.	
1967	Carroll, J. B. (1967). The foreign language attainments of language majors in the senior year: A survey conducted in US colleges and Universities. *Foreign Language Annals* 1.2, 131 – 151.	Despite little validation evidence the FSI/ILR approach became popular in education because of its face validity, inter-rater reliability through social moderation, and perceived coherence with new communicative teaching methods. Carroll’s study of 1967 showed that the military system was not sensitive to language acquisition in an	**E, G**

		educational context, and hence was demotivating. It would be over a decade before this research had an impact on policy.	
1979	*Strength Through Wisdom: A Critique of U.S. Capability. A Report to the President from the President's Commission on Foreign Language and International Studies*. (1979). Wahington DC: US Government Printing Office.	Further impetus to extend speaking assessment in educational settings came from a report submitted to **President Carter** on shortcomings in the US military because of lack of foreign language skills. It is not coincidental that in the same year attention was drawn to a study published by **Carroll** in 1967. The American Council on the Teaching of Foreign Languages (ACTFL) was given the task of revising the FSI/ILR scales for wider use.	
1979	Adams, M. L. & Frith, J. R. (1979). *Testing kit: French and Spanish*. Washington DC: Department of State and the Foreign Service Institute.	As part of the ACTFL research into new rating scales the first testing kits were developed for training and assessment purposes in US Colleges. The articles and resources in **Adams & Frith** provided a comprehensive guide for raters of the Oral Proficiency Interview for educational purposes.	**A, C, E, G**
1980	Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. In Frith, J. R. (ed.),	**Adams** conducted the first structural validation study designed to investigate which of the five FSI subscales	**B**

	Measuring spoken language proficiency. Washington DC: Georgetown University Press, 1 – 6.	discriminated between learners at each proficiency level. The study was not theoretically motivated, and no patterns could be discerned in the data.	
1980	Reves, T. (1980). The group-oral test: an experiment. *English Teachers Journal* 24, 19 – 21.	**Reves** questioned whether the OPI could generate ‘real-life conversation’ and began experimenting with group tasks to generate richer speaking samples.	**C**
1981	Bachman, L. F. & Palmer, A. S. (1981). The construct validity of the FSI oral interview. *Language Learning* 31.1, 67 – 86.	The first construct validation studies were carried out in the early 1980s, using the multitrait-multmethod technique and confirmatory factor analysis. These demonstrated that the FSI OPI loaded most heavily on the speaking trait, and lowest of all methods on the method trait. These studies concluded that there was significant convergent and divergent evidence for construct validity in the OPI.	**B**
1983	Lowe, P. (1983). The ILR oral interview: origins, applications, pitfalls, and implications. *Die Unterrichtspraxis* 16, 230 – 244.	In the 1960s the FSI approach to assessing speaking was adopted by the Defense Language Institute, the Central Intelligence Agency, and the Peace Corp. In 1968 the various adaptations were standardized as the Interagency Language Roundtable (ILR), which is still the accepted tool for the	**A, C, D**

		certification of second language speaking proficiency throughout the United States military, intelligence and diplomatic services (http://www.govtilr.org/). Via the Peace Corp it spread to academia, and the assessment of speaking proficiency worldwide. It also provides the basis for the current NATO language standards, known as STANAG 6001.	
1984	Liskin-Gasparro, J. E. (1984). The ACTFL Proficiency Guidelines: Gateway to testing and curriculum. *Foreign Language Annals* 17.5, 475 – 489.	Following the publication of *Strength Through Wisdom* and the concerns raised by **Carroll's** 1967 study, the ACTFL Guidelines were developed throughout the 80s, with preliminary publications in 1982, and the final Guidelines issued in 1986 (revised 1999). Levels from 0 to 5 were broken down into subsections, with finer gradations at lower proficiency levels. Level descriptors provided longer prose definitions of what could be done at each level. New constructs were introduced at each level, drawing on new theoretical models of communicative competence of the time, particularly those of Canale and Swain.	**A, B**

		These included discourse competence, interaction, and communicative strategies.	
1985	Lantolf, J. P. & Frawley, W. (1985). Oral proficiency testing: A critical analysis. *Modern Language Journal* 69.4, 337 – 345.	**Lantolf** and **Frawley** were among the first to question the ACTFL approach. They claimed the scales were ‘analytical’ rather than ‘empirical’, depending on their own internal logic of non-contradiction between levels. The claim that the descriptors bear no relationship to how language is acquired or used set off a whole chain of research into scale analysis and development.	**A, B**
1986	Kramsch, C. J. (1986). From language proficiency to interactional competence. *Modern language journal* 70.4, 366 – 372.	**Kramsch’s** research into interactional competence spurred further research into task types that might elicit interaction, and the construction of ‘interaction’ descriptors for rating scales. This research had a particular impact on future discourse related studies by HE & YOUNG (1998).	**B**
1986	Bachman, L. F. and Savignon, S. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview. *Modern Language Journal* 79, 380 – 390.	This very influential paper questioned the use of the native speaker to define the top level of a rating scale, and the notion of zero proficiency at the bottom. Secondly, they questioned reference to context within scales as confounding	**B, D, F**

		constructs with test method facets, unless the test is for a defined ESP setting. This paper therefore set the agenda for debates around score generalizability, which we still wrestle with today.	
1987	Fulcher, G. (1987). Tests of oral performance: the need for data-based criteria. *English Language Teaching Journal* 41.4, 287 - 291	Using discourse analysis of native speaker interaction, this paper provided the first evidence that rating scales did not describe what typically happened in naturally occurring speech, and advocated a data-based approach to writing descriptors and constructing scales. This was the first use of discourse analysis to understand under-specification in rating scale descriptors, and was expanded into a larger research agenda (see FULCHER 1996).	**A, B, H**
1989	Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly* 23.3, 489 – 508.	In another discourse analysis study, **Van Lier** showed that interview language was not like ‘normal conversation’. Although the work of finding formats that encouraged ‘conversation’ had started with REVES (1980) and colleagues in Israel, this paper encouraged wider research in the area.	**B, H**
1991	Linacre, J. M. (1991*). FACETS*	Rater variation had been a concern since	**E, I**

	computer programme for many-faceted Rasch measurement. Chicago, IL: Mesa Press.	the work of Roach during the war, but only with the publication of **Linacre's** FACETS did it become possible to model rater harshness/leniency in relation to task difficulty and learner ability. MFRM remains the standard tool for studying rater behaviour today and test facets today, as in the studies by LUMLEY & MCNAMARA (1995), and BONK & OCKEY (2003).	
1991	Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (eds.), *Language Testing in the 1990s*. London: Modern English Publications and the British Council, 71 – 86.	Based on research driving the IELTS revision project, **Alderson** categorized rating scales as use-oriented, rater-oriented, and constructor-oriented. These categories have been useful in guiding descriptor content with audience in mind.	**A**
1992	Young, R. & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition* 14.4, 403 – 424.	An early and significant use of discourse analysis to characterize the interaction of test takers with interviewers in the First Certificate Test of English. Discourse structure was demonstrated to be related to examiner, task and gender variables.	**B, C, H, L**
1992	Douglas, D. & Selinker, L. (1992). Analyzing Oral Proficiency Test performance in general and specific purpose	Douglas & Selinker show that a discipline specific test (chemistry) is a better predictor of domain specific performance than a general speaking	**A, B, D**

	contexts. *System* 20.3, 317 – 328).	test. In this and a series of publications on ESP testing they show that reducing generalizability by introducing context increases score usefulness. This is the other side of the coin to BACHMAN & SAVIGNON'S (1986) generalizability argument.	
1992	Ross, S. & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14.1, 159 – 176.	Reacting to critiques of the OPI from VAN LIER (1989), LANTOLF & FRAWLEY (1985; 1988), and others, **Ross** & **Berwick** undertook discourse analysis of OPIs to study how interviewers accommodated to the discourse of candidates. They concluded that the OPI had features of both interview and conversation. However, it also raised the question of how interlocutor variation might result in test takers being treated differentially. This sparked a chain of similar research by scholars such as LAZARATON (1996).	**B, C, H, J**
1992	Mislevy, R. J. (1992). *Linking Educational Assessments. Concepts. Issues. Methods and Prospects.* Princeton NJ: Educational Testing Service.	LOWE (1983; 1987) and others had argued that the meaning of descriptors was socially acquired. In this publication the term 'social moderation' was formalized. NORTH (1998) and the Council of Europe have taken this	**E**

		concept and made it central to the project of using the Common European Framework of Reference (CEFR) scales as a European-wide lens for viewing speaking proficiency.	
1995	Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12.1, 16 – 33.	**Chalhoub-Deville** investigated the inter-relationship of diverse tasks and raters using multidimensional scaling to identify components speaking proficiency that were being assessed. She found that these varied by task and rater group, and therefore called for the construct to be defined anew for each task x rater combination. The issue at stake is whether the construct 'exits' separately from those who make judgments and the facets of the test method.	**A, B, E**
1995	Lumley, T. and McNamara, T. (1995). Rater characteristics and rater bias: implications or training. *Language Testing* 12.10, 54 – 71.	Rater variability is studied across time using FACETS, showing that there is considerable variation in harshness irrespective of training. The researchers question the use of single ratings in high-stakes speaking tests, and recommend the use of rater calibrations to provide training feedback or adjust scores.	**E, I**

1995	Upshur, J. & Turner, C. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal* 49.1, 3 – 12.	The paper in which **Upshur** & **Turner** introduce Empirically-derived binary-choice boundary-definition scales (EBB). These address the long-standing concern over a-priori scale development outlined by LANTOLF & FRAWLEY (1985), and start to tie decisions to specific examples of performance as recommended by FULCHER (1987). The scales are task specific rather than generic. The methodology has specific impact on later studies like those of POONPON (2010).	**A, B, C, D, K**
1996	McNamara, T. (1996). *Measuring Second Language Performance*. Harlow: Longman.	The research around the development of the Occupational English Test (OET) for health professionals is described. This is a specific purpose test with a clearly specified audience, and scores from this instrument are shown to be more reliable and valid for decision making than generic English tests.	**A, B, C, D**
1996	Fulcher, G. Testing tasks: issues in task design and the group oral. *Language Testing* 13.1, 23 – 51.	Building on REVES (1980) and others, this study compared a group oral (3 participants) and two interview-type tasks. Discourse was more varied in the group task, and participants reported a preference for working in a group with	**C, G**

		other test-takers.	
1996	Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13.2, 208 - 238.	Based on work conducted since FULCHER (1987), primarily an unpublished PhD project, this paper describes the research underpinning the design of data-based rating scales. The methodology employs discourse analysis of speech samples produce scale descriptors. The use of the resulting scale is compared with generic a-priori scales. Using discriminant analysis the data-based scores are found to be more reliable, and using MFRM rater variation is significantly decreased. The data-based approach therefore solves the problems identified by researchers like LUMLEY & MCNAMARA (1995). The study also generated the Fluency Rating Scale descriptors, which were used as anchor items in the CEFR project.	**A, B, C, D, H**
1996	Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews. The case of CASE. *Language Testing* 13.2, 151 – 172.	In the ROSS & BERWICK (1992) tradition, and inspired by VAN LIER, **Lazaraton** identifies 8 kinds of support provided by a rater/interlocutor in an OPI. She concludes that the variation is problematic, and calls for additional rater training and possibly the use of an	**B, H, J**

		‘interlocutor support scale’ as part of the rating procedure.	
1996	Pollitt, A. & Murray, N. L. (1996). What raters *really* pay attention to. In Milanovic, M. & Saville, N. (eds.), *Performance testing, cognition and assessment. Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*. Studies in Language Testing 3. Cambridge: Cambridge University Press.	The use of Thurstone’s Paired Comparisons, and Kelly’s Repertory Grid Technique, to investigate how raters use rating scales and what they notice in candidate spoken performances. The research showed raters bring their own conceptual baggage to the rating process, but used constructs such as discourse, sociolinguistic, and grammatical competence, as well as fluency and ‘naturalness’.	**B, K**
1997	McNamara, T. (1997). Modelling performance: Opening Pandora’s Box. *Applied Linguistics* 18.4, 446 – 465.	Speaking had generally been characterized in cognitive terms as traits resident in the speaker being assessed. Building on the work of KRAMSCH (1986) and others, **McNamara** showed that interaction implied the co-construction of speech, and argued that in social contexts there was shared responsibility for performance. The question of shared responsibility, the role of the interlocutor, become active areas of research.	**B**
1998	Young, R. & He, A. W. (1998)	An important collection of research	**B, C, H**

	(eds.), *Talking and testing. Discourse approaches to the assessment of oral proficiency*. Amsterdam: John Benjamins.	papers analysing the discourse of test-taker speech in speaking tests. The speaking test is characterized as an 'interactive practice' co-constructed by the participants.	
1998	North, B. & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing* 15.2, 217 – 262.	This paper describes the measurement-driven approach to scale development as embodied in the CEFR. Descriptors from existing speaking scales are extracted from context and scaled using MFRM using teacher judgments as data.	**A, I**
1999	Jacoby, S. & McNamara, T. (1999). Locating competence. *English for Specific Purposes* 18.3, 213 – 241.	In two studies, **Jacoby** & **McNamara** discovered that the linguistic criteria used by applied linguists to rate speaking performance did not capture the kind of communication valued by subject specialists. They recommended studying 'indigenous criteria' to expand what is valued in performances. This work has impacted on domain specific studies, such as Fulcher et al. 2011. It also raises serious questions about psycholinguistic approaches such as those advocated by VAN MOERE (2012).	**B, K**
2002	Young, R. (2002). Discourse approaches to oral language	A careful investigation of the 'layers' of discourse in naturally occurring speech	**B, C, H**

	assessment. *Annual Review of Applied Linguistics* 22, 243 – 262.	and test tasks. This is combined with a review of various approaches to testing speaking, with an indication of which test formats are likely to elicit the most useful speech samples for rating.	
2002	O’Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing* 19.1, 33 – 56.	A methodological study to compare the ‘informational and interactional functions’ produced on speaking test tasks with those the test designer intended to elicit. The instrument provided to be unwieldy and impractical, but the study established the important principle for examination boards that evidence of congruence between intention and reality is an important aspect of construct validation.	**B, H**
2003	Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing* 20.1, 1 – 25.	A much quoted study into variation in the speech of the same test taker with two different interlocutors. **Brown** also demonstrated that scores also varied, although not by as much as one may have expected. Builds on ROSS & BERWICK (1992), LAZARATON (1996) and McNAMARA (1996). Raises the critical issue of whether variation should be allowed because it is part of the construct, or controlled	**B, H, I, J**

		because it leads to inequality of opportunity.	
2003	Fulcher, G. & Marquez-Reiter, R. (2003). Task difficulty in speaking tests. *Language Testing* 20.3, 321 – 344.	An investigation into the effects of task features (social power and level of imposition) and L1 cultural background, on task difficulty and score variation. Like BROWN (2003) it was discovered that although significant variation occurred when extreme conditions were used, effect sizes were not substantial.	**B, C, H**
2003	Bonk, W. J. & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing* 20.1, 89 – 110.	Using FACETS, the researchers investigated variability due to test taker, prompt, rater, and rating categories. Test taker ability was the largest facet. Although there was evidence of rater variability this did not threaten validity, and indicated that raters became more stable in their judgments over time. This adds to the evidence that socialization over time has an impact on rater behaviour.	**B, E, I**
2005	Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2005). *A teacher-verification study of speaking and writing prototype tasks for a new TOEFL Test*. TOEFL Monograph No.	An important prototyping study. Pre-operational tasks were shown to experts who judge whether they represent the kinds of tasks that students would undertake at University. They are also presented with their own student's	**B, C, K**

	MS-26. Princeton, NJ: Educational Testing Service.	responses to the tasks and asked whether these are 'typical' of their work. The study shows that test development is a research-led activity, and not merely a technical task. Design decisions and the evidence for those decisions are part of a validation narrative.	
2007	Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt: Peter Lang.	Based on many years of research into personality and speaking test performance, **Berry** shows that levels of introversion and extroversion impact on contributions to conversation in paired- and group-formats, and results in differential score levels when ability is controlled for.	**B, C, L**
2008	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.	A discourse analytic study of the paired test format. The research identified three interactive patterns in the data: 'collaborative', 'parallel' and 'asymmetric'. Tentative evidence is also presented to suggest that there is a relationship between scores on an 'Interactive Communication' rating scale.	**B, C, H**
2009	Ockey, G. (2009). The effects of group members' personalities on a test taker's L2 group oral	Building on BERRY (2007), **Ockey** investigates the effect of levels of 'assertiveness' on speaking scores in a	**B, C, L**

	discussion test scores. *Language Testing* 26.2, 161 – 186.	group oral test, using MANCOVA analyses. Assertive students are found to have lower scores when placed in all assertive groups, and higher scores when placed with less assertive participants. The scores of non-assertive students did not change depending on group makeup. The results differ from BERRY, indicating that much more research is needed in this area.	
2010	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.	A study that brings together the EBB approach of UPSHUR & TURNER with the data-based approach of FULCHER (1996) to create a rich data-based EBB for use with TOEFL iBT tasks. In the process the nature of the academic speaking construct is further explored and defined.	**A, B, H, K**
2011	Fulcher, G., Davidson, F. & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance Decision Trees. *Language Testing* 28.1, 5 - 29.	Like POONPON (2010), this study brings together UPSHUR & TURNER'S (1995) EBB and FULCHER'S (1996) data-based approach in the context of service encounters. It also incorporates indigenous insights following JACOBY & MCNAMARA (1999). It describes interaction in service encounters through a performance decision tree that focuses	**A, B, H**

		rater attention on observable criteria related to discourse and pragmatic constructs.	
2011	Frost, K., Elder, C. & Wigglesworth, G. (2011). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing* 29(3), 345 – 369.	Integrated task types have become widely used since their incorporation into TOEFL iBT. Yet, little research has been carried out into the use of source material in spoken responses, or how the integrated skill can be described in rating scale descriptors. The 'integration' remains elusive. In this study a discourse approach is adopted following ideas in DOUGLAS & SELINKER (1992) and FULCHER (1996) to define content related aspects of validity in integrated task types. The study provides evidence for the usefulness of integrated tasks in broadening construct definition.	**A, B, C**

2011	May, L. (2011). Interactional Competence in a Paired Speaking Test: Features Salient to Raters. *Language Assessment Quarterly* 8.2, 127 – 145.	Following KRAMSCH (1986), MCNAMARA (1997) and YOUNG (2002), **May** problematizes the notion of the speaking construct in a paired speaking test. However, she attempts to deal with the problem of how to award scores to individuals by looking at how raters focus on features of the speech of individual participants. The three categories of interpretation: understanding interlocutor's message, responding appropriately, and using communicative strategies, are not as important as the attempt to disentangle the individual from the event, while recognizing that discourse is co-constructed.	**B, C, K**
2011	Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing* 28.4, 483 – 508.	Building on BONK & OCKEY (2003) and other research into the group speaking test, **Nakatsuhara** used conversation analysis to investigate group size in relation to proficiency level and personality type. She discovered that more proficient extroverts talked more and initiated topic more when in groups of 4 than in groups of 3. However, proficiency level	**B, H**

		resulted in more variation in groups of 3. With reference to GALACZI (2008), she concludes that groups of 3 are more collaborative.	
2012	Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing* 29.1, 325 – 344.	Very much against the trend, **Van Moere** makes a case for a return to assessing psycholinguistic speech ‘facilitators’, related to processing automaticity. These include response latency, speed of speech, length of pauses, reproduction of syntactically accurate sequences, with appropriate pronunciation intonation and stress. Task types are sentence repetition and sentence building. This approach is driven by an a-priori decision to use an automated scoring engine to rate speech samples, and the validation argument points to the objective nature of the decisions made in comparison with interactive human scored tests, which are claimed to be unreliable and contain too much construct-irrelevant variance. This is an exercise in reductionism par excellence, and is likely to reignite the debate on prediction to domain performance from ‘atomistic’ features	**B, C**

		that last raged in the early communicative language testing era.	
2012	Tan, J. Mak, B, & Zhou, P. (2012). Confidence scoring of speaking performance: How does fuzziness become exact? *Language Testing* 29.1, 43 – 65.	The application of fuzzy logic to our understanding of how raters score performances. This approach takes into account both rater decisions, and the levels of uncertainty in arriving at those decisions.	**E, J**
2014	Nitta, R & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral performance. *Language Testing* 31.2, 147 – 175.	This research investigates providing test-takers with planning time prior to undertaking a paired speaking test. The unexpected findings are that planning time results in stilted prepared output, and reduced interaction between speakers.	**C, H**

Acknowledgements

I would like to thank Dr. Gary Ockey of Educational Testing Service for reviewing my first draft, and providing valuable critical feedback. My thanks are also due to the very constructive criticism of the three reviewers, which has considerably improved the coverage and coherence of the timeline. Finally to the editor of *Language Teaching* for timely guidance and advice.