Research timeline: assessment of second language fluency

Article

Reading's research outputs online

CAMBRIDGE
UNIVERSITY PRESS

RESEARCH TIMELINE

Assessment of second language fluency

Parvaneh Tavakoli ⓘ

University of Reading, Reading, UK
Email: p.tavakoli@reading.ac.uk

Introduction

Investigating speech fluency has, for a long time, been at the core of second language (L2) studies, as fluency is believed to epitomise successful acquisition of L2, characterise effective communication, elucidate the complex process of acquisition, and predict L2 speakers' proficiency. The significance attributed to fluency in these areas explicates the research attention paid to it over the past decades. An important area of development in this regard is L2 assessment in which fluency is recognised as a key underlying construct of spoken language ability by international language tests (e.g., IELTS, TEEP, APTIS) and language benchmarks (e.g., CEFR). Many high-stakes tests of English and other languages include fluency in their rating scales, with the earliest on record tracing back to the 1930s – the College Board's English Competence Examination (1930) in America. Including fluency as a fundamental aspect of speaking ability in the rating scales, rating descriptors, and rater training materials, either as an independent criterion or combined with others (e.g., delivery), has become common practice in language testing over the past decades. What has made assessment of fluency even more appealing to researchers and test providers in recent years is the objectivity and reliability of its measurement and its compatibility with the technological developments in automated assessment of speaking. Fluency is now largely recognised as a construct that can be efficiently and reliably assessed in automated assessment of spoken language ability and used to predict proficiency (de Jong, 2018*; Ginther et al., 2010*; Kang & Johnson, 2021*; Tavakoli et al., 2023).

Fluency is commonly regarded as a complex and multidimensional construct, often reported as difficult to define with a degree of openness to interpretation in assessment contexts (de Jong, 2018*). Lennon (1990*) provided one of the earliest L2 definitions of fluency, considering it as the 'rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language' (Lennon, 1990, p. 26). Since its publication, this definition has been widely cited and has become the basis for other conceptualisations of fluency to emerge. Lennon's (1990*) work was also pioneering as it distinguished between a broad versus narrow sense of fluency with the former referring to the general concept of proficiency and the latter characterising the speakers' fluidity of speech. A decade later, Koponen and Riggenbach (2000, p. 6) offered their own interpretation of fluency as 'flow, continuity, automaticity, or smoothness of speech'. What these definitions have in common is that flow and fluidity are central to fluency and largely embody efficiency of language processing. Segalowitz's (2010*) work was a turning point in the process of developing a better understanding of the construct of fluency as it provided a more overarching and structured approach to its conceptualization and examination. In his triadic model, Segalowitz (2010*) argued that fluency should be understood as a multidimensional construct with at least three distinct but interrelated aspects: *cognitive fluency*

---

*Indicates full reference appears in the subsequent timeline.

Note. Authors' names are shown in small capitals when the study referred to appears in this timeline.

(i.e., efficiency of the operations underlying speech production), *utterance fluency* (i.e., the measurable aspects of speech performance such as speed and silence), and *perceived fluency* (i.e., inferences listeners make based on their perceptions of the speaker's fluidity of speech). The relationship between the three aspects has been examined in the literature and the findings, so far, provide an in-depth understanding of the three aspects and how they interact with one another during the speech production processes.

To complement the predominantly cognitive perspective on fluency, more recently researchers have emphasised the significance of conceptualising fluency from a sociolinguistic and interactional perspective (Segalowitz, 2010*, 2016*). Tavakoli and Wright (2020, p. 3), for example, underlined the social and interactional nature of fluency, in both L1 and L2, and maintained that a more complete understanding of fluency will only be achieved when it is examined in relation to 'the context, purpose and audience'. They argued that a speaker considered fluent in one context addressing a specific audience may not be as fluent in another context, discussing a different topic, or interacting with a different audience. Similarly, speakers may project different fluency behaviour when speaking for different purposes (e.g., giving bad news is usually delivered more slowly than breaking good news). This new perspective is being adopted in recent studies (e.g., Morrison & Tavakoli, 2023), offering emerging evidence about the interactional and sociolinguistic dimensions of fluency.

The earliest L2 studies in this timeline that have aimed to develop a systematic and objective framework for measuring fluency date back to the 1980s (e.g., Raupach, 1980*). In such studies, a native speaker baseline was usually considered and measures such as number of words or pauses per minute were calculated. Considerable developments have emerged since then. Skehan (2003*) and Tavakoli and Skehan (2005*) provided evidence that utterance fluency consists of three distinct factors: speed, breakdown, and repair. Other studies (Suzuki, Kormos & Uchihara, 2020*) have supported the three-factor model and argued that speed, breakdown, and repair constitute the underlying construct, consolidating the validity of the triadic approach to measuring fluency. Adopting this measurement framework, other researchers have complemented it by including more nuanced measures. Skehan (2009*) and Kahng (2014*), for example, have made a distinction between pure (e.g., speed only) and composite (e.g., speed plus pauses) measures; de Jong and Bosker (2013*) have shown that a threshold of 250–300 milliseconds (ms) is optimal for measuring the number of pauses when examining proficiency, and Hunter (2017) has argued that pause should be examined in terms of duration, frequency, location, and character (e.g., filled or unfilled). Tavakoli's (2011*) findings highlighting the significance of pause location in distinguishing L1 from L2 speech were followed by a growing trend in fluency studies that systematically examined the location of pauses (e.g., final vs mid-clause position).

The development of technology has had two main influences on the assessment of fluency. First, the use of digital technology has changed the way fluency is measured. While earlier studies measured fluency more manually (e.g., using a watch or chronometer to measure pauses), often in longer time units (e.g., one second), the introduction of digital technology in the 1990s (e.g., GoldWave Digital Audio Editor, 2024) made it possible to measure the temporal aspects of fluency more objectively, in smaller units, and with more precision. Subsequently, the free availability of technical software specifically developed to analyse speech (e.g., PRAAT, Boersma & Weenink, 2013) further helped spread the use of this technology. Second, the rapid development of the use of Artificial Intelligence (e.g., speech recognition and auto scoring technology) has changed assessment of fluency in automated assessment of speaking ability. Using such technological developments, some key test providers (e.g., Pearson Test of English) have started measuring fluency automatically or using it to predict proficiency for other aspects of speech (e.g., comprehensibility).

Since the 1980s, L2 research has additionally made remarkable progress in understanding a range of factors that affect L2 fluency. Studies in task-based language teaching (TBLT) have been particularly important in delineating the impact of task design, task characteristics, and performance conditions on fluency (Foster & Skehan, 1996; Michel, 2011). As can be seen in the timeline below, the findings of cognitive and psycholinguistic research in fluency have also helped foster a more in-depth

understanding of the relationship between fluency and processing and production demands. Factors such as L1 fluency, personal style, social and pragmatic aspects of fluency, and task type are now recognised as variables with a significant impact on L2 fluency.

The complex nature of fluency and the range of factors affecting it have made its assessment difficult. Tavakoli and Wright (2020) noted the disparity between the rating descriptors, rating criteria, and the common practice in the assessment of fluency in different high-stakes tests. Despite the widespread agreement about including fluency as a key criterion in the assessment of speaking ability, little agreement is observed about how fluency is characterised in the tests' rating scales and descriptors or how it is measured. Tavakoli and Wright (2020, p. 110) reported that the rating descriptors in these tests lack an unequivocal and specific definition to the extent that they can lead to 'a degree of personal interpretation of fluency'. In addition, rating descriptors are often not based on research evidence (Fulcher, 1996*). What makes assessment of fluency even more difficult is the paucity of research evidence about the extent to which fluency is expected to develop in relation to proficiency. So far there is little research in this area to show whether a linear relationship is expected between fluency and proficiency, and whether the different aspects of fluency (speed, breakdown, repair) develop consistently as L2 ability progresses. Strikingly, most language testing fluency descriptors assume fluency develops as a whole construct (i.e., in its different aspects of speed, breakdown, and repair) as proficiency develops, but this assumption is not supported by solid empirical evidence (e.g., see Tavakoli et al., 2020*). Previous research has reported that most rating scales and rating descriptors are not based on empirical evidence (de Jong, 2018*; Fulcher, 1996*) and, as can be seen below, there are few studies that have examined fluency in a language testing context or made a contribution to the development of rating descriptors and rating scales.

The purpose of this article is to provide a timeline of studies (1979–2022) that have aimed at helping develop an accurate, evidence-based, and reliable understanding of fluency and its measurement and assessment. The timeline demonstrates a selection of seminal work and primary studies by both established and emerging researchers that have made an impactful contribution to the development of the current fluency frameworks (i.e., conceptual, measurement, and assessment frameworks) over the past four decades.

The timeline presented below has categorised these studies based on the following themes:

A. Understanding and defining the construct of fluency
1. Understanding the construct of fluency
2. Cognitive perspectives to conceptualising fluency
3. Interactional perspectives to conceptualising fluency
4. Emerging models of fluency

B. Fluency in language assessment
1. Fluency in international tests of L2 proficiency
2. Developing fluency rating scales, rating descriptors, rater training
3. Relationship between fluency and levels of proficiency
4. Automated assessment of fluency in international tests of English

C. Factors affecting L2 fluency
1. Individual speaker factors (e.g., personal styles)
2. Cross-linguistic factors
3. Social and contextual factors (e.g., study abroad)
4. Task related factors
5. Fluency in relation to other aspects of linguistic knowledge (e.g., lexis)

D. Measuring fluency
1. Subjective approaches to measuring fluency and perceived fluency
2. Objective approaches to measuring utterance fluency
3. Technological advancement in assessing L2 fluency

References

Boersma, P., & Weenink, D. (2013). *Praat: Doing phonetics by computer.* http://www.praat.org/

Foster, P., & Skehan, P. (1996). The influence of planning time on performance in task- based learning. *Studies in Second Language Acquisition*, *18*(2), 299–234. doi:10.1017/S0272263100015047

GoldWave Digital Audio Editor. (2024). *About GoldWave Inc.* Retrieved in May 2024 from https://www.goldwave.com/about.php

Hunter, A.-M. (2017). *Fluency development in the ESL classroom: The impact of immediate task repetition and procedural repetition on learners' oral fluency* [Doctoral dissertation]. St Mary's University. https://research.stmarys.ac.uk/id/eprint/1868

Michel, M. (2011). Effects of task complexity and interaction in L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 141–174). John Benjamins Publishing Company. doi:10.1075/tblt.2.12ch6

Morrison, A., & Tavakoli, P. (2023). Task communicative function and oral fluency of L1 and L2 speakers. *Modern Language Journal*, *107*(4), 896–921. doi:10.1111/modl.12883

Tavakoli, P., Kendon, G., Muzhurnaya, S., & Ziomek, A. (2023). Assessing fluency in the Test of English for Educational Purposes. *Language Testing*, *40*(3), 607–629. doi:10.1177/02655322231151384

Tavakoli, P., & Wright, C. (2020). *Second language speech fluency: From research to practice.* Cambridge University Press.

Parvaneh Tavakoli is Professor of Applied Linguistics at the University of Reading. Parvaneh's main research interest lies in the interface of second language acquisition, language teaching, and language testing. She is specifically interested in task-based language teaching, task design, and development of oral fluency in second language acquisition. Parvaneh has led several international research projects investigating second language performance, acquisition, assessment, and policy in different contexts. She has disseminated her research in the form of articles in prestigious peer-reviewed journals (e.g., *Modern Language Journal*, *SSLA*, and *Language Learning*), policy reports (e.g., Report to Welsh Government), and books by key publishers (e.g., Cambridge University Press and TESOL Press). Her latest monograph 'Comprehensibility in Language Testing' has been published Open Access by Equinox.

Year	References	Annotations	Themes
1979	Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler & W. S. Y. Wang (Eds.), *Individual differences in language ability and language behaviour* (pp. 85–102). Academic Press. https://doi.org/10.1016/B978-0-12-255950-1.50012-3	This is an early and influential attempt at conceptualising the construct of fluency and highlighting its multidimensional nature. In this article, **Fillmore** focuses on different dimensions of fluency highlighting qualities such as filling time with talk, having few pauses, being coherent, having appropriate things to say, and using language innovatively. Although the definition was initially proposed for first language (L1) speech, it was soon adopted by many as a baseline to define second language (L2) fluency.	**A1, A2, A3**
1980	Raupach, M. (1980). Temporal variables in first and second language speech production. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honour of Frieda Goldman-Eisler* (pp. 263–270). Mouton. https://doi.org/10.1515/9783110816570.263	This is one of the earliest studies investigating the temporal aspects of fluency in both L1 and L2 speech. **Raupach** examined 10 L2 speakers of English (five L1 French and five L1 German) and analysed their speech in terms of a range of temporal variables including speech rate, articulation rate, silent and filled pauses, and length of run. The findings suggested that L1 speech was more fluent than L2, cross-linguistic differences were observed, and differences in personal styles were clearly seen. Despite the small scale of the study, many of Raupach's measures have been used and similar findings have been achieved in subsequent studies.	**C2, D1, D2**
1983	Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–225). Routledge. https://doi.org/10.4324/9781315836027	This article highlighted the processing and production demands of speaking and emphasised the role of what they termed as 'nativelike fluency', linking fluency of speech with frequent use of formulaic sequences. Highlighting the formulaic nature of language use, **Pawley & Syder** argued that the NATIVELIKE FLUENT use of language relies largely on availability and use of prefabricated chunks that native speakers typically use.	**A1, A4, C5**
1990	Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40(3), 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x	This article advanced our understanding of what constituted fluency, how fluency differences were perceived, and which fluency features were quantifiable. **Lennon** argued that adopting a quantitative analysis would help develop a more reliable and evidence-based approach to assessing fluency and a more useful strategy to detect its development in L2 learning. Drawing on FILLMORE (1979) the article offers a dichotomous classification of fluency (i.e., broad vs narrow), which was widely adopted by others in the coming years.	**A1, A2, A4, D1, D2**
1991	Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, 14(4), 423–441. https://doi.org/10.1080/01638539109544795	Comparing three 'highly fluent' and three 'highly nonfluent' non-native speakers of English, **Riggenbach** examined their speech in terms of measures of filled and unfilled pauses, speed, and repairs. The fluent	**A1, D1, D2,**

Year	References	Annotations	Themes
		speakers spoke faster with fewer pauses than the nonfluent speakers. Objective measures of speech rate and pausing used by Riggenbach were widely adopted by later studies and were shown to be significant predictors of perceived fluency.	
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. https://doi.org/10.1177/026553229601300205	This article makes a significant contribution to the field of language testing as it was the first to take a data-driven approach to developing a fluency rating scale. Using a Grounded Theory approach and analysing samples of speech from 21 test-takers (both qualitatively and quantitatively), **Fulcher** offered a thick description of fluency characterised by eight fluency features. He then ran a discriminant analysis to identify the extent to which these features distinguished speakers at different levels of proficiency. Based on this, he developed the first evidence-based fluency rating scales.	**B1, B2, B3, C5, D1, D2**
2000	Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243–265). University of Michigan Press.	In this article, **Freed** drew attention to the interactive nature of fluency and the role the listener plays in it. This is perhaps one of the earliest studies examining what became known as the relationship between PERCEIVED and UTTERANCE fluency. Asking non-expert raters to rate samples of L2 learners' speech collected in an oral interview task, Freed examined the rating scores and the raters' explanations of the scores to identify objective measures of fluency. The results suggested that in addition to specific fluency measures (e.g., speech rate and pauses), raters considered other linguistic features in students' speech (e.g., vocabulary and grammar) when rating their fluency. She emphasised that LENNON's (1990) classification would be a useful framework to work with.	**A1, A2, A3, D1, D2**
2000	Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5–24). University of Michigan Press.	This introductory chapter to the influential volume *Perspectives on fluency* is a key text in which the authors explained the different views on fluency and opened up a platform for discussing fluency from a range of perspectives, from language teaching to discourse analysis and clinical use. The researchers emphasised LENNON's (1990) and PAWLEY & SYDERS' (1983) perspectives, and highlighted the multifaceted nature of fluency and the different concepts it represented; for example, FLUENCY AS SMOOTHNESS OF SPEECH, FLUENCY AS AUTOMATICITY, FLUENCY AS PROFICIENCY, and FLUENCY AS OPPOSED TO ACCURACY.	**A1, A3, C2, C3**
2001	Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition*, *23*(3), 497–526. https://doi.org/10.1017/S027226310100403X	This is one of the earliest studies examining the relationship between L1 and L2 fluency in relation to crosslinguistic influences. Comparing data elicited from advanced and intermediate Russian learners of L2	**A1, B3, C2**

		English in both L1 Russian and L2 English, **Riazantseva** examined the speakers' pausing patterns by measuring pauses of longer than 100 ms. The results suggested that the two languages had different pausing patterns, implying the existence of cross-cultural differences. They also demonstrated that while intermediate learners demonstrated less L1-like behaviour than advanced learners, they overcame the L1 effects as their L2 proficiency developed.	
2002	Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, *111*(6), 2862–2873. https://doi.org/10.1121/1.428279	This is a very early attempt at examining the extent to which expert judgments of fluency can be predicted by automatically measured temporal aspects of fluency. Based on the data from 90 L1 and L2 Dutch speakers and subjective ratings of nine expert raters, the results of the study showed that expert ratings of fluency can be predicted by several fluency measures with rate of speech, articulation rate, and number of pauses being the best predictors (varying correlations ranging between 0.57 and 0.92 for different proficiency levels). This implied that fluency measured automatically can represent perceived fluency, a concept used in automated assessment of fluency in the years to come. It should be noted, however, that the results indicated that the predictive power of the measures was stronger for read speech than for spontaneous speech.	**B1, B2, D1, D2, D3**
2003	Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1–14. https://doi.org/10.1017/S026144480200188X	In this article, **Skehan** argued that fluency is a key aspect of performance and in a trade-off relationship with complexity and accuracy. To capture a comprehensive picture of fluency, Skehan argued for an analysis of fluency across three dimensions: speed (e.g., number of words/syllables per minute), breakdown (e.g., frequency and length of pauses), and repair (e.g., reformulation and repetition). He also called for a measure of the degree of automatisation (e.g., length of run) to be included in the measurement of fluency.	**A1, A4, C4, C5**
2004	Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655–680. https://doi.org/10.1111/j.1467-9922.2004.00282.x	In this study, **Derwing et al.** asked 28 untrained and three trained raters to rate speech samples of 20 Mandarin learners of English in three tasks in terms of measures of fluency, comprehensibility, and accentedness. The findings are important for fluency studies for two reasons. First, the results indicated that fluency was affected by task type (learners were less fluent in narrative tasks), although raters' judgements were not affected by task type. Second, a high correlation was observed between perceptions of fluency and comprehensibility, implying fluent speech was usually associated with ease of understanding.	**A1, C4, C5**

(*Continued*)

Year	References	Annotations	Themes
2004	Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition, 26*(2), 275–301. https://doi.org/10.1017/S0272263104262064	In this article, **Freed et al.** investigated the effects of context of learning on fluency. By comparing speech samples from students studying French at home, in an intensive immersion program, and on a study abroad program, the authors claimed that the immersion programme students showed more fluency gains than others (in number of words spoken, length of the longest turn, rate of speech, and speech fluidity). The study abroad group did better (in terms of fluidity) than those studying at home. The findings raised questions about the effects of study abroad on fluency. The results also demonstrated that language use in terms of number of hours spent on using the language (in writing) outside class was linked to gains in fluency.	**A1, C3, D1, D2**
2004	Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001	This is one of the earliest studies investigating teachers' perceptions of fluency. Examining speech samples produced by 16 Hungarian L2 learners of English at two levels of proficiency, the authors asked three experienced native speaking teachers and three non-native speaking teachers to rate the learners' fluency. The results demonstrated that speech rate, mean length of utterance, phonation time ratio, and number of stressed words per minute were the best predictors of fluency scores. The authors concluded that fluency was essentially a temporal and intonational phenomenon.	**A1, D1, D2**
2005	Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks. *ETS Research Report Series, 1*, 1–157. https://doi.org/10.1002/j.2333-8504.2005.tb01982.x	This study stands out as the first to use verbal report methodology to examine expert raters' criteria for assessing fluency and to use the findings as the basis for developing rating scales for both independent and integrated speaking tasks of the Test of English as a Foreign Language (TOEFL). Providing their verbal reports when assessing 198 TOEFL test-takers' speech samples, the expert raters referred to a range of fluency features including speech rate, hesitation, pauses, and repair measures. The list of fluency features, similar to FULCHER's (1996) study, was to be adopted by TOEFL for the rating scales and rater training purposes. The second part of **Brown et al.**'s study shows that measures of speech rate, unfilled pauses, and total amount of pause were the best fluency features to characterise proficiency at each level. The study used a 1-second pause threshold and did not distinguish between clause internal and external pauses. The findings, also obtained by others in subsequent studies, were used as the basis for automatic assessment of fluency.	**B1, B2, B3, C5, D1, D2, D3**

2005	Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–277). John Benjamins. https://doi.org/10.1075/lllt.11.15tav	Adopting SKEHAN's 2003 triadic model of fluency, the authors used a range of ten measures to assess the fluency of speech samples from 80 L2 learners performing four narrative tasks. **Tavakoli & Skehan** chose their measures carefully to reflect the latest developments in second language acquisition (SLA) (e.g., FREED, 2000) by lowering the pause threshold to 400 ms and distinguishing between mid-clause and end-clause pauses. They used factor analysis on performances from each of the four tasks to identify the underlying constructs of fluency. The results showed that repair fluency consistently loaded on a different factor than speed and pausing across all tasks. The results also suggested task structure affected fluency, as performances in more structured tasks were more fluent.	A1, A4, C4, C5, D2
2006	Kormos, J. (2006). *Speech production and second language acquisition.* Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9780203763964	In this monograph, **Kormos** proposed a model of L2 speech production process and discussed the similarities and differences between it and the existing L1 models. Kormos outlined the developmental processes of L2 speech production and highlighted the factors that affected them. She explained that greater fluency is achieved by having access to a larger repertoire of lexical and syntactic knowledge and a more automatised access to that knowledge. She also argued that a limited linguistic repertoire or lack of automaticity in using it would lead to disfluency (e.g., slower speech, breakdown, and repair), a hypothesis that has been tested and supported by others (e.g., KAHNG, 2014).	A1, A2, A4, C5
2008	Skehan, P., & Foster, P. (2008). Complexity, accuracy, fluency and lexis in task-based performance: A synthesis of the Ealing research. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy, and fluency in second language use, learning, and teaching* (pp. 207–226). University of Brussels Press. https://doi.org/10.1075/lllt.32.09fos	In this paper, the authors ran a meta-analysis of a group of studies they had conducted in the 1990s and early 2000s, examining the effects of task characteristics and conditions on aspects of performance, including its fluency. The paper highlighted the importance of using effective and valid measures of fluency, as these would be more sensitive to experimental conditions.	A1, A4, A4, C5
2009	de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385	The authors provided a script written in Praat that allowed users to automatically detect syllables in audio files without needing a transcription. Using sound files, the script developed a TextGrid file with syllable nuclei marked in a point tier system. The authors reported high correlations between human-measured speech rate and automatically measured speech rate using this script.	D1, D2, D3
2009	Derwing, T. M., Munro, M. J., Thomson, R. I. & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*(3), 533–557. https://www.jstor.org/stable/44485884	This article is particularly important as it reported one of the pioneering studies to investigate the relationship between L1 and L2 fluency and cross-linguistic differences in a longitudinal design over a period of two years. Collecting speech samples from L1 Slavic and Mandarin speakers performing the same tasks in L1 and L2 English and analysing the data for a range of temporal measures of fluency,	C1, C2, C3

(*Continued*)

Year	References	Annotations	Themes
		Derwing et al. reported significant correlations between L1 and L2 fluency measures. The correlations were found to be stronger for the Slavic L1 group, implying cross-linguistic differences may play a further role. Despite such clear results, the authors acknowledged that 'a straightforward relationship between fluency in the L1 and the L2 cannot be expected' (p. 554), denoting several factors contribute to the relationship between L1 and L2 fluency.	
2009	Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047	In this article, **Skehan** aimed to present 'a theoretically motivated and empirically grounded account' (p. 512) of measures of complexity, accuracy, and fluency to portray L2 performance. He drew attention to the significant role of fluency and lexis (Pawley & Syder, 1983) in modelling L2 performance and called for a more systematic and sophisticated approach to measuring fluency. Skehan emphasised the importance of measuring automatisation (e.g., through calculating length of run) in addition to the triadic measures of breakdown, speed, and repair (set by Skehan, 2003). To understand the complex nature of fluency, Skehan argued that it was necessary to research pause location (mid- vs end-clause), pause character (e.g., filled vs unfilled), the relationship between L1 and L2 fluency, and task effects.	**C1, C2, C4**
2010	Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–399. https://doi.org/10.1177/0265532210364407	The authors examined 150 test-takers' speech samples when taking the Oral English Proficiency Test (OEPT) and analysed them for speed and breakdown measures. The results suggested strong correlations between OEPT scores and measures of speech rate, articulation rate, and mean length of run. The analysis failed to show any significant correlations between OEPT scores and filled pause ratio or length of filled pauses. The results also suggested raters had assigned higher scores to those producing longer runs, confirming Skehan's (2009) hypothesis about the importance of measures such as length of run.	**B1, B2, B3, D1, D2, D3, D5**
2010	Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge. https://doi.org/10.4324/9780203851357	This is one of the most influential and frequently cited sources in fluency studies. In this book, **Segalowitz** presented an explanatory model for conceptualising fluency, arguing fluency should be recognised in three different but interrelated aspects: cognitive (efficiency of the processes underlying speech production), utterance (observable features of fluency), and perceived fluency (listeners' perceptions). Importantly, Segalowitz reinforced the interactional nature of fluency by highlighting the listener's role in the process.	**A1, A2, A3, A4**

2011	Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal*, *65*(1), 71–79. https://doi.org/10.1093/elt/ccq020	This study is the first to highlight the importance of mid-clause pausing when measuring fluency. Comparing samples of speech collected from 40 native speakers of English and 40 L2 learners of English performing four different tasks, **Tavakoli**'s results suggested that the key difference between the two groups' fluency behaviour was in the frequency and length of pauses produced in mid-clause positions. The results indicated many mid-clause pauses occurred when speakers were involved in repair (e.g., repeating, replacing or reformulating a word or an utterance), implying mid-clause pauses were associated with other cognitive and linguistic processes. No pauses were observed within formulaic sequences, confirming Pawley & Syder's (1983) concept of nativelike selection and nativelike fluency.	C1, C2, D2, D5
2013	Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177/0265532212455394	This study aimed at investigating the contribution of the three aspects of utterance fluency (speed, pauses, and repair) to perceived fluency. Using Segalowitz's (2010) model as the theoretical framework and Skehan's (2009) model as the measurement framework, the authors employed 80 untrained raters to judge fluency of L1 and L2 Dutch speakers. The authors selected their measures carefully to ensure they were largely independent. **Bosker et al.** found that measures of speed and pause made a significant contribution to the judgements of fluency, while the contribution of repair measures was negligible. They also found that listeners were sensitive to pause features of speech more than they were to speed or repair. It is important to note that measures of mid-clause pause and composite measure of speech rate were not used in this study.	A1, B1, B3, C1, C2, D1, D2
2013	de Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of the 6th workshop on disfluency in spontaneous speech (DiSS)* (pp. 17–20). Royal Institute of Technology (KTH).	This article is particularly important as it empirically investigated the pause threshold that can be optimally used in L2 studies. While in prior studies fluency researchers had used a range of different pause thresholds (ranging from 100 ms to 3 s), this study provided evidence that setting the cut-off point at 250–300 ms yields the highest correlation between the frequency of silent pauses and the L2 proficiency measure (vocabulary knowledge). It is worth mentioning that proficiency was only measured through the vocabulary knowledge test. Following this publication, many studies used this threshold for measuring silent pauses.	D1, D2, D3
2013	Foster, P. (2013). Fluency. In C. A. Chapelle (Ed.), *The international encyclopaedia of applied linguistics.* (pp. 2124–2130). Wiley-Blackwell.	In this article, **Foster** discussed fluency from a broader perspective, considering it in relation to L1 speech production models, reinforcing the significance of developing an L2 model of speech production. Analysing issues around patterns of pausing and speed in L1 and L2 speech production, Foster highlighted the importance of using	A1, A4, C5

(*Continued*)

Year	References	Annotations	Themes
		formulaic sequences and their relationship to fluency (PAWLEY & SYDER, 1983).	
2014	Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, *64*(4), 809–854. https://doi.org/10.1111/lang.12084	**Kahng**'s study is important as it not only examined L1 and L2 speakers' fluency in terms of their differences, but it employed stimulated recall to validate the examination of objective measurement of fluency in a qualitative manner. Her findings indicated the differences between L1 and L2 fluency behaviour, emphasising the role of mid-clause pausing that potentially underlines difficulties in speech production (KORMOS, 2006; PAWLEY & SYDER, 1983).	A1, C2, D1, D2
2014	Pinget, A.-F., Bosker, H. R., Quené, H., & de Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*, *31*(3), 349–365. https://doi.org/10.1177/0265532214526177	The authors examined raters', both trained and novice, judgements of fluency and accentedness by exploring the relationship between objective measures of speech and raters' subjective ratings. The results suggested that acoustic measures of fluency (e.g., frequency of pauses) were reliable predictors of fluency ratings, while segmental and suprasegmental measures of accent predicted variance in accent ratings. The results were important as they supported the validity argument of the fluency construct by indicating that perceived fluency and perceived accent were separate constructs.	A1, B2, C5, D1, D2
2015	de Jong, N. H. (2015). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(1), 113–132. https://doi.org/10.1515/IRAL-2016-9993	Examining the distribution of silent and filled pauses in L1 and L2 speech, **de Jong** investigated speech samples from 70 L1 and L2 Dutch speakers in terms of utterance boundary and word frequency. Achieving similar findings as TAVAKOLI's (2011), de Jong reported that L2 speakers paused more frequently than L1 speakers within utterances. These pauses were also longer than those produced by L1 speakers. Interestingly, both groups of speakers paused more before low frequency words, suggesting the cognitive demands were high for such words, highlighting the importance of examining pause location.	A1, C2, C5,
2015	de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behaviour. *Applied Psycholinguistics*, *36*(2), 223–243. https://doi.org/10.1017/S0142716413000210	In this study, **de Jong et al.** questioned the validity of L2 fluency measures in representing the L2 fluency construct. The authors investigated the extent to which non-linguistic characteristics of the speaker (e.g., personality or speaking style) may affect their fluency. By collecting L1 data from L2 learners of Dutch while assessing their proficiency, the authors concluded that some corrected L2 measures predicted proficiency better. Corrected measures were calculated by partialing out the L1 variance from the L2 measures. They concluded that L2 studies will benefit from using corrected measures of L2	C1, C2, B1, B3

		fluency, and therefore both L2 and L1 speech samples should be examined to develop an in-depth understanding of the construct.	
2016	Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*(6), 828–848. https://doi.org/10.1093/applin/amu069	Aiming to identify what linguistic features contribute to the construct of communicative adequacy (i.e., the ability to communicate successfully in the real world), the authors examined data collected from 80 L2 learners performing four different tasks. The data were then rated by 20 native speakers for adequacy and coded for a range of fluency, accuracy, and complexity measures. Interestingly, filled pause frequency was the strongest predictor of communicative adequacy. Silent pause frequency and speech rate were other reliable predictors of adequacy. All in all, **Révész et al.** concluded that 'fluency emerged as a critical determinant of communicative adequacy' (p. 844).	**B3, C4, C5**
2016	Tavakoli, P. (2016). Speech fluency in monologic and dialogic task performance. *International Review of Applied Linguistics (IRAL): Special Issue on Fluency*, *54*(2), 131–150. https://doi.org/10.1515/iral-2016-9994	In this article, **Tavakoli** drew attention to the limited progress the field had made in measuring fluency in dialogic task performance. Arguing that the existing measures were only suitable for monologic data, she maintained that a different set of measures were needed when analysing fluency in a dialogic format. Underlining the challenges researchers would experience when working with dialogic data, Tavakoli offered an exploratory approach to the measurement of the interactive aspects of dialogic performance (e.g., between turn pauses, overlaps, and interruptions).	**A2, A3, C4, D1, D2**
2016	Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 79–95. https://doi.org/10.1515/iral-2016-9991	Building on his previous work, in this article, **Segalowitz** argued that providing a description of patterns of fluency in different learning conditions is not sufficient if a full understanding of the construct is expected. He proposed an exploratory framework, based on a dynamical systems perspective, for understanding both the cognitive dimension of fluency and the social context that contributed to attainment of fluency. He argued that without considering the pragmatic and sociolinguistic nature of the context of communication, it is very difficult to understand the relationship between cognitive and utterance fluency.	
2017	Huensch, A., & Tracy-Ventura, N. (2017a). L2 utterance fluency development before, during, and after residence abroad: A multidimensional investigation. *Modern Language Journal*, *101*(2), 275–293. https://doi.org/10.1111/modl.12395	This study stands out as it examined the development of different aspects of fluency across a two-year period before, during, and after completing a year of study abroad. The authors collected data from L1 English learners of Spanish at six different time points while retelling an oral narrative task. The results suggested that while speed fluency developed quickly and was usually maintained in time, breakdown fluency took a long time to improve and was more sensitive to attrition.	**A1, C1, C2, D1, D2**

(*Continued*)

(*Continued*)

Year	References	Annotations	Themes
2017	Huensch, A., & Tracy-Ventura, N. (2017b). Understanding second language fluency behaviour: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics*, 38(4), 755–785. https://doi.org/10.1017/S0142716416000424	This study belongs to the group of studies investigating the relationship between L1 and L2 fluency, making an important contribution to the understanding of this relationship. By examining speech performances of L1 English speakers learning French or Spanish during a study abroad period, the authors reported that L1 fluency, cross-linguistic differences, and proficiency level all contributed to L2 fluency, although to different extents. The authors concluded that L1 fluency behaviour cannot solely explain variations in L2 fluency.	**A1, B3, C1, C2, C3, D1, D2**
2018	de Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, 15(3), 237–254. https://doi.org/10.1080/15434303.2018.1477780	This article is particularly important as **de Jong** argued that assessment of fluency in practice is not based on the findings of applied linguistics. Critically examining assessment of fluency in four different tests, the article raised questions and concerns about the way fluency was conceptualised and operationalised in high-stakes language tests. de Jong encouraged language test designers and providers to ensure the fluency measures they employed in their rating scales and rating descriptors actually reflected 'the ability to talk fluently and efficiently' (p. 237).	**B1, B2, B3, B4**
2018	Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics*, 39(3), 593–617. https://doi.org/10.1017/S0142716417000571	**Saito et al.** investigated the relationship between perceived, utterance, and cognitive fluency (Segalowitz, 2010) by asking 10 untrained English native speaking raters to judge fluency levels of 90 Japanese English L2 speakers and 10 English native speakers. The results suggested that specific measures of utterance fluency characterised different fluency profiles: frequency of final-clause pauses distinguished low- and mid-level fluency performance; frequency of mid-clause pauses differentiated mid- and high-level performance; and articulation rate distinguished high-level and nativelike performance. Interestingly, the repair measures neither related to the listeners' ratings nor distinguished between the proficiency groups. The authors linked the different aspects of the fluency profiles to different stages of the speech production process.	**A4, B3, D1, D2**
2018	Tavakoli, P., & Hunter, A-M. (2018). Is fluency being 'neglected' in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research*, 22(3), 330–349. https://doi.org/10.1177/1362168817708462	In reaction to Lennon's (1990) dichotomy of broad versus narrow senses of fluency, the authors of this article aimed to develop a better understanding of how L2 teachers conceptualised and defined fluency. Collecting data from 84 teachers in England, **Tavakoli & Hunter** (2018) argued that fluency was considered at four levels of very broad, broad, narrow, and very narrow. The very broad category reflected fluency as overall proficiency of the L2; the broad category represented fluency as	**A1, A4, D1, D2**

		L2 speaking ability; the NARROW classification indicated flow and continuity of speech; and the VERY NARROW category referred to fluency in relation to objective and measurable aspects of speech. The authors argue L2 teachers could improve their understanding and practice by adopting a NARROW or VERY NARROW perspective.	
2018	Peltonen, P. (2018). Exploring connections between first and second language fluency: A mixed methods approach. *Modern Language Journal*, *102*(4), 676–692. https://doi.org/10.1111/modl.12516	While investigating the effects of L1 fluency on L2 fluency, this study differed from previous work in this area as it adopted a mixed-methods approach to examine the relationship. In line with HUENSCH AND TRACY-VENTURA (2017a), the results showed that most temporal L2 measures were predicted from L1 fluency measures, although the predictive power was different across the measures. Speakers' idiosyncratic patterns of fluency were also observed in both languages, implying the speakers use the same patterns (e.g., repeating words) in both their L1 and L2 speech.	**A1, C1, C2, D1, D2**
2020	Duran-Karaoz, Z., & Tavakoli, P. (2020). Predicting L2 fluency from L1 fluency behaviour: The case of L1 Turkish and L2 English speakers. *Studies in Second Language Acquisition*, *42*(4), 671–695. https://doi.org/10.1017/S0272263119000755	To respond to HUENSCH & TRACY-VENTURA's, (2017a) call for examining the role of proficiency in the relationship between L1 and L2 fluency, the authors collected data from Turkish L1 speakers performing narrative tasks in both L1 Turkish and L2 English. The authors also examined their proficiency carefully by using two different tests tapping into the participants' declarative and procedural knowledge. The results showed significant correlations between L1 and L2 fluency measures for breakdown and repair; the relationships were not moderated by proficiency level, implying L1 and L2 fluency behaviour correlate, to a certain extent, regardless of the speakers' proficiency.	**A1B3 C1 C2 C4**
2020	Foster P. (2020). Oral fluency in a second language: A research agenda for the next ten years. *Language Teaching*, *53*(4), 446–461. https://doi.org/10.1017/S026144482000018X	This *Thinking Aloud* article in *Language Teaching* offered a comprehensive review of how the construct of fluency has been defined and examined over the past years. In this article, **Foster** provided an inclusive overview of research and set the agenda for future research in this area. Of particular interest to this research is Foster's emphasis on the need to identify the relationship between perceived fluency and idiomaticity of use (SKEHAN, 2009 and KAHNG, 2014), and the relationship between vocabulary size and fluency measures (PAWLEY & SYDER, 1983).	**A1, A2, A3, C3**
2020	Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*(2), 506–547. https://doi.org/10.1111/lang.12384	By examining the relationship between fluency and multiword sequences (MWSs) across assessed levels of proficiency, the authors showed that a linear relationship existed between proficiency and many of the MWS measures. Their results also suggested that fluency correlated with measures of MWSs. For example, high frequency MWSs were positively related to articulation rate, and the proportion of MWSs was negatively associated with mid-clause pauses. The results taken	**B1, B2, B3, B4, C4, C5**

(*Continued*)

Year	References	Annotations	Themes
		together provided robust empirical evidence to support Pawley & Syder's (1983) concept of nativelike fluency.	
2020	Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *Modern Language Journal, 104*(1), 169–191. https://doi.org/10.1111/modl.12620	The study aimed at investigating fluency in a detailed micro-analytic framework across assessed levels of proficiency. The study is particularly important as it adopted a detailed framework to examine utterance fluency and compare it with raters' perceived fluency at a range of different proficiency levels in a standardised test. The results indicated a few important findings, including the fact that speed distinguished between different levels of proficiency, but with a ceiling effect when the speaker reached a B2 level. The results also indicated that neither a clear pattern nor a linear progression was observed for repair across different proficiency levels.	B1, B2, B3, C4
2021	de Jong, N. H., Pacilly, J., & Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy and Practice, 28*(4), 456–476. https://doi.org/10.1080/0969594X.2021.1951162	The study aimed at developing a less time-consuming approach to measuring certain aspects of fluency while ensuring accuracy and consistency. They revised the existing script from de Jong and Wempe (2009) to measure a range of features, including filled pauses, automatically. Using primary and secondary corpora in L2 English and L2 Dutch, they used annotators to identify the optimal algorithm to detect filled pauses. Correlations between the manual and automatic measures of filled pauses were 0.53 to 0.78, with the Dutch data showing lower coefficients.	B1, B2, B3, B4, D1, D2, D3
2021	Kang, O., & Johnson, D. (2021). Linguistic features and automatic scoring of Aptis speaking performances. *The Assessment Research Awards and Grants Publications.* https://www.britishcouncil.org/linguistic-features-and-automatic-scoring-aptis-speaking-performances https://www.britishcouncil.org/sites/default/files/kang_and_johnson_layout.pdf	The authors used a computer model to score the proficiency of Aptis test-takers automatically for a range of linguistic measures. The correlation coefficients between the computer-generated scores and the raters' scores showed high levels of agreement of 0.90 and 0.76, figures that were higher than those reported in previous studies. As for fluency, the results suggested that proficient candidates produced more syllables per second and fewer pauses, replicating Ginther et al. (2010) and Kormos and Dénes's (2004) findings while using an automatic scoring machine.	B1, B2, B3, B4, C5, D1, D2
2021	Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *Modern Language Journal, 105*(2), 435–463. https://doi.org/10.1111/modl.12706	In this meta-analysis, the authors analysed primary studies based on correlation coefficients to examine the relationship between temporal features of utterance fluency and judgements of perceived fluency. Analysing 263 effect sizes from 22 studies, the authors reported that perceived fluency was strongly associated with speed, composite measure, and pause frequency; moderately with pause length; and weakly with repair measures. The strongest effect size was observed	A1, A4, C5, D1, D2

		between the composite measure of speech rate and subjective judgements of fluency. The results confirmed Tᴀᴠᴀᴋᴏʟɪ et al.'s (2020) findings that speed and pauses distinguish speakers at different levels of proficiency.	
2022	Suzuki, S., & Kormos, J. (2022). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, *45*(1), 38–64. https://doi.org/10.1017/S0272263121000899	This study examined the construct of fluency by investigating the relationship between utterance and cognitive fluency across four different task types. **Suzuki & Kormos**'s results highlighted the two dimensions of 'linguistic resources' and 'processing speed' as the core of the construct and confirmed the robustness of Tᴀᴠᴀᴋᴏʟɪ and Sᴋᴇʜᴀɴ's (2005) three-dimensional model (e.g., speed, breakdown, and repair fluency) across different tasks.	A1, C4, C5, D1, D2