The Multidimensionality of Second Language Oral Fluency:

The Interface Between Cognitive, Utterance, and Perceived Fluency

Shungo Suzuki

This thesis is submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy in Linguistics

April 2021

Lancaster University

Department of Linguistics and English Language

Declaration

I declare that this thesis has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted in this thesis is my own, except where work which has formed part of a co-authored publication has been included. My contribution and those the other authors to the work have been explicitly indicated below. I also confirm that appropriate credit has been provided throughout the thesis where reference has been made to the work of others. The work presented in Chapters 4 and 5 was previously published in the following articles:

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency : A meta-analysis of correlational studies. *The Modern Language Journal*, *105*(2).

This study was conceived by all of the authors. I carried out the review of literature, the data collection and analysis including statistical analyses, and drafting the paper, while I discussed various issues with the other authors throughout the research project.

Date 26/04/2021

Name: Shungo Suzuki

Signature: 

Acknowledgements

First and foremost, my deepest gratitude goes to my supervisor, Prof. Judit Kormos. Without her insightful advice, patient encouragement and thoughtful support, this thesis would never have been completed. Our discussion on theoretical contributions to the field, methodological rigour, publications, and philosophy as a researcher has always inspired me tremendously and have been an essential core of myself as a researcher. Since I met her book—*Speech Production and Second Language Acquisition*—during my undergraduate study in Japan, I have dreamed of becoming a researcher like her. She is still my goal and role model as a researcher, but I have eventually been able to be her co-author in some publications, which I could never think of when I was an undergraduate student. I am absolutely certain that I have walked the best path of academia. I am sincerely honoured to stand at the starting point to the next phase of academic career as her student.

I would also like to express my gratitude to my MA supervisor in Waseda University, Prof. Tetsuo Harada, for supporting my data collection in Japan, encouraging me to become a researcher, and being a referee of my Japanese scholarship. I am also grateful to Prof. Yasuyo Sawaki for helping me to attain the expertise in advanced statistics used in the thesis and also for being another referee of the scholarship.

I gratefully acknowledge that the materials in this thesis were developed with the assistance of Dr. Yuichi Suzuki and Prof. Alastair Graham-Marr. I appreciate that Dr. Yuichi Suzuki kindly shared the material of his Maze task and provided insightful feedback on the analysis of the data of the Maze task. I also truly thank Prof. Alastair Graham-Marr for his professional commitments for creating audio-recorded stimuli for read-aloud assistance.

I'm also grateful to two of my senior colleagues. First, I sincerely thank Dr. Michael Ratajczak, who is my mentor and best friend in Lancaster, for his kind and thorough assistance for my understanding of mixed-effects modelling. Second, many thanks go to Dr. Takumi Uchihara, who is my senior colleague and peer role model since my undergraduate study, for his support for meta-analytic procedures.

Finally, I would like to thank the scholarships from JASSO (Japan Student Services Organization) and FASS (the Faculty of Arts and Social Sciences, Lancaster University) for their financial support for my PhD study. Without these scholarships, I could not have fully committed to the PhD research project as well as publications.

Abstract

In the context of the learning, teaching, and assessment of second language (L2) speaking skills, L2 fluency has been regarded as one of the important constructs. However, L2 fluency research has witnessed a long debate over the definition and measurements of L2 oral fluency; scholars have interchangeably used the term, "fluency", with different connotations, such as speakers' ability, speech features, and listeners' perception. In order to distinguish different conceptualizations of fluency, Segalowitz (2010) proposed three subconstructs of fluency: *utterance fluency* (i.e., observable temporal features of speech), *cognitive fluency* (i.e., speaker's ability to manipulate L2 knowledge efficiently), and *perceived fluency* (i.e., listener's subjective judgements of fluency). However, it is still unclear how these three subconstructs of L2 fluency are interrelated with each other.

The overarching goal of the thesis is to examine the construct of L2 oral fluency, particularly focusing on the interrelationship between cognitive, utterance, and perceived fluency. To this end, this thesis consists of four separate studies. Study 1 took a meta-analytic approach to synthesizing previous findings on the relationship between perceived and utterance fluency. Study 2 compared utterance fluency performance across speaking tasks which were designed to differ in the quality of speech processing demands, operationalized by task design features (i.e., task effects). Study 3 examined the contribution of cognitive fluency to utterance fluency, taking a structural equation modelling (SEM) approach. The study also analysed the stability of the factor structure of utterance fluency (Tavakoli & Skehan, 2005)—speed, breakdown, and repair fluency—and cognitive fluency across speaking tasks. Finally, Study 4 investigated the extent to which L2 utterance fluency can be predicted from L1 utterance fluency with regard to the moderator effects of L2 proficiency on the L1-L2 utterance fluency link.

Study 1 collected 263 effect sizes from 22 studies reporting the correlation coefficients between listener-based judgements of fluency and objective measures of temporal features (N = 335–746). Among the pooled utterance fluency measures, Study 1 selected the common measures from four different categories: speed (articulation rate), breakdown (silent pause frequency, silent pause duration), repair (disfluency rate), and composite fluency (mean length of run, speech rate). Methodological moderator variables were selected with respect to the major phases of research into the utterance-perceived fluency link: Speech stimulus preparation (e.g., task type, target L2), Rater background (e.g., L1 vs. L2 listeners), Perceived fluency rating procedure (Definition of fluency, the number of point scales), and Utterance fluency measure calculation (length of pauses, manual vs. automated annotation).

Studies 2–4 were conducted using the same dataset which included a set of cognitive and utterance fluency measures from Japanese-speaking learners of English (N = 128). Using a range of psycholinguistic tests, cognitive fluency was assessed in terms of linguistic resources and processing speed at different linguistic levels: vocabulary (vocabulary size, lexical retrieval speed), grammar (sentence construction speed and accuracy, grammaticality judgement speed and accuracy), and pronunciation (articulatory speed). In order to measure utterance fluency, speech data were elicited via four speaking tasks which differed in the quality of speech processing demands: argumentative task, picture narrative task, and text retelling tasks with/without read-aloud assistance. The speech data were analysed in terms of three subconstructs of utterance fluency (speed, breakdown, and repair fluency). The participants' L1 fluency was also assessed, using another L1 argumentative speech task. Their proficiency scores were operationalized as two factor scores of cognitive fluency (linguistic resources and processing speed) in Study 3.

Study 1 demonstrated that perceived fluency was strongly associated with speed and pause frequency (r = |.59–.62|), moderately with pause duration (r = |.46|), and weakly with repair fluency (r = |.20|), while composite measures showed the strongest effect sizes (r = |.72–.76|). A series of moderator analyses also revealed that the utterance-perceived fluency link may be influenced by methodological variables particularly related to speech stimulus preparation (target L2, task type, length of stimuli) and perceived fluency rating procedure (the definition of fluency presented to raters).

Study 2 compared utterance fluency across four speaking tasks, using Generalized Linear Mixed-effect modelling (GLMM) with the tasks as a categorical fixed-effects predictor. The results showed that conceptualizing demands (content generation) increased the frequency of filled pauses, while the demands on formulation (activation of linguistic and phonological representations) had an impact on articulation rate, mid-clause pause ratio, and mid-clause pause duration.

In Study 3, prior to an SEM analysis, a set of confirmatory factor analyses (CFA) demonstrated that utterance fluency has a three-factor structure (speed, breakdown, and repair fluency) and that cognitive fluency has a two-factor structure (linguistic resource and processing speed). An SEM analysis, based on these factor structures of cognitive and utterance fluency, showed that speed fluency was primarily associated with processing speed, while both linguistic resource and processing speed equally contributed to breakdown fluency. Repair fluency was significantly linked to linguistic resource, only when the content of speech was predefined (picture narrative and text summary tasks). Meanwhile, repair fluency was found to be independent of processing speed in all the speaking tasks.

Study 4 examined the L1-L2 utterance fluency link using a set of GLMMs. The results suggested that all the L2 utterance fluency measures were predicted from their L1 counterparts. In addition, significant moderator effects of L2 proficiency on the L1-L2 fluency link were found only in speed fluency measures. The L1-L2 fluency link was weakened as a function of L2 linguistic resource but was strengthened as a function of L2 processing speed.

The results of Study 1–4 confirmed that the relative importance of three subdimensions of utterance fluency—speed, breakdown, and repair fluency—can vary, depending on the perspective of assessment (perceived vs. cognitive fluency). These findings provide several practical implications for language assessment, such as the development of assessment tools and guidance for examiner training, as well as for L2 fluency learning and teaching.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Importance of Oral Fluency in L2 Communication and Assessment

In the context of the learning and teaching of second language (L2) speaking skills, oral fluency is commonly regarded as one of the major learning goals, due to its important role in real-world communication. A certain level of fluency is necessary to maintain the interlocuter's attention in oral communication and to be able to save speakers' own face (Lennon, 2000). Similarly, oral fluency is an essential component of speech that determines listener-perceived comprehensibility (S. Suzuki & Kormos, 2020). In the context of L2 assessment, a variety of high-stakes oral proficiency tests, such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS), have adopted oral fluency as one of the main constructs of L2 proficiency in line with the research finding that L2 oral fluency is a robust indicator of L2 proficiency (Baker-Smemoe et al., 2014; Tavakoli et al., 2020). Therefore, it is essential to better understand L2 oral fluency as a construct.

1.2 Theoretical Background to the Thesis

Although L2 oral fluency has been commonly regarded as one of the essential aspects of L2 speaking performance, L2 fluency research has witnessed a long debate over the definition of oral fluency. As a pioneer work, Fillmore (1979) conceptualized fluency as a holistic construct equivalent to overall oral proficiency, covering (a) temporal smoothness, (b) linguistic repertoires and accuracy, (c) sociolinguistic appropriateness, and (d) content sophistication. Building on Fillmore's definition, Lennon (1990, 2000) proposed two different scopes for oral fluency. While acknowledging that the term, fluency, can be used as overall command of language (i.e., higher-order fluency), Lennon (1990, 2000) also narrowly defined oral fluency as the temporal characteristics of speech (i.e., lower-order fluency). Note

that Lennon (1990) also emphasized that the notion of fluency should closely align with listeners' perception about the speaker's processing efficiency based on the given speech. Taken together, fluency has been traditionally conceptualized from the perspectives of speakers' ability, speech features, and listeners' perceptions.

In response to the fact that the term, fluency, has been interchangeably used with different connotations, Segalowitz (2010, 2016) proposed the three subconstructs of oral fluency: cognitive fluency (CF), utterance fluency (UF), and perceived fluency (PF). CF refers to "the efficiency of the speaker's underlying processes responsible for fluency-relevant features of utterance" (Segalowitz, 2010, p. 50). Segalowitz (2010) explicitly mentions that the construct of CF is strongly linked to L2 speech production mechanisms (e.g., de Bot, 1992; Kormos, 2006; Levelt, 1999). UF is concerned with "the oral features of utterances that reflect the operation of underlying cognitive processes" (Segalowitz, 2010. p. 50), including the speed of delivery and hesitations. PF is defined as "the inferences that listeners make about a speaker's cognitive fluency [CF] based on perception of the utterance fluency [UF] features of the speaker's speech output" (Segalowitz, 2010, p. 50). In other words, PF is conceptualized as the listeners' intuitive judgements of CF, while UF can reflect CF in the form of observable temporal features of speech. Segalowitz (2010) clarifies that PF should be distinguished from UF, meaning that PF is not simply equated with subjective judgements of UF features. Listeners' judgements and perceptions are normally shaped by ignoring the acceptable disfluency phenomena which they believe are irrelevant to the efficiency of L2 system (i.e., CF; Segalowitz, 2010). Therefore, PF can be regarded as the subjective judgements of speakers' CF. The interrelationship between CF, UF, and PF is visualized in Figure 1 (for more details, see Chapter 3).

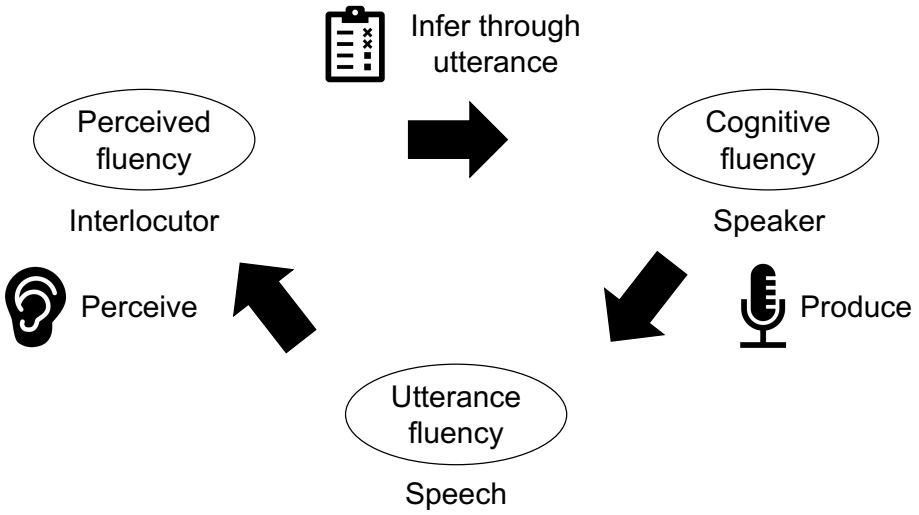


Figure 1. The visualization of the interrelationship between cognitive, utterance, and perceived fluency.

Segalowitz's (2010, 2016) triad model of fluency has assisted researchers to define, operationalize, and measure L2 fluency in a theoretically valid manner. Among Segalowitz's (2010) three subconstructs of fluency, previous studies had extensively examined the relationship between UF and PF, even before Segalowitz (2010). However, L2 fluency research has faced several methodological challenges. Regarding UF, a variety of measures have been developed. Accordingly, different UF measures have been employed to predict PF scores, lowering the comparability of findings across studies. Another challenge lies in the methodological procedures for PF judgements. Depending on the research focus, previous studies have adopted different methodologies and have used different rating procedures and speech elicitation methods. However, the extent to which these methodological variables affect the relationship between PF and UF has not yet been systematically examined.

CF has been relatively underresearched among the three subconstructs of fluency (cf. De Jong et al., 2013; Kahng, 2020). According to Segalowitz (2010, 2016), the conceptualization of CF theoretically corresponds to L2 speech production mechanisms (e.g., de Bot, 1992; Kormos, 2006; Levelt, 1999). Due to the particular focus on CF in the thesis, the

psycholinguistic models of L2 speech production thus serve as a theoretical framework of the thesis (see Chapter 2). Moreover, another important characteristic of CF is L2 specificity (Segalowitz, 2016). From a theoretical perspective, speech production entails both language-general and language-specific processes (Kormos, 2006; Levelt, 1989; Segalowitz, 2010). In the case of L2 speech production, speech is generated through cognitive processes shared across L1 and L2 as well as the manipulation of L2-specific linguistic knowledge. Building on the assumption that CF should reflect L2-specific competence related to fluent speech production (Segalowitz, 2016), it is also essential to understand what components of speech production are assumed to be language-general and language-specific for a better understanding of CF as a construct.

1.3 Aims and Research Designs of the Thesis

For a better understanding of L2 oral fluency as a construct, it is essential to clarify how three subconstructs of fluency (Segalowitz, 2010, 2016) are similar to and distinct from each other. The overarching goal of the thesis is thus to understand the construct of fluency from three different perspectives, that is, PF (listeners' perception), CF (speakers' competence), and UF (speech characteristics), with a particular focus on the relationship between UF and PF and between CF and UF. To this end, this thesis consists of four separate studies.

To understand how listeners selectively pay attention to different temporal features of speech, Study 1 focuses on the relationship between UF and PF (hereafter, UF-PF link) and queries how each dimension of UF is associated with PF judgements. Assuming that PF is equivalent to listeners' subjective judgements of CF (Segalowitz, 2010), the investigation into the contribution of UF to PF would give some insights into what temporal features listeners believe reflect the speakers' CF. Although the UF-PF link has been relatively extensively

examined (Bosker et al., 2013; Derwing et al., 2004; Kormos & Dénes, 2004), prior work has provided some inconsistent findings. Therefore, Study 1 adopts meta-analytic techniques to synthesize previous findings on the UF-PF link. Study 1 also aims to identify which methodological factors can moderate the strengths of the UF-PF link. The identification of significant moderator factors would give insights into how listeners establish their perceptions of fluency.[1]

Motivated by the lack of studies on the relationship between CF and UF (hereafter, CF-UF link), the remaining three studies aim to examine the CF-UF link, using different methodological approaches. Operationalizing CF as a set of speech processing components (e.g., lexical retrieval, syntactic phrase construction), Study 2 aims to identify which components of speech production are related to UF measures. Using an experimental approach, UF performance is compared across four speaking tasks which are designed to differ in the quality of speech processing demands. To this end, task design features are manipulated to create different speech processing demands, such as the degree of content generation and the availability of relevant linguistic items. Assuming that UF measures are reflective of CF (Segalowitz, 2010), the manipulation of task design features is expected to affect cognitive demands of the target speech processing components and subsequently influence the relevant temporal features of speech.

Study 3 investigates the contribution of CF to UF and its variability across speaking tasks at the level of constructs, taking a structural equation modelling (SEM) approach. Study 3 measures CF as speaker's linguistic resources and processing skills, using a set of

[1] Some part of Study 1 was accepted for publication in *The Modern Language Journal* as Suzuki, Kormos, and Uchihara (in press, 2021).

psycholinguistic tests, such as a picture naming task and a grammaticality judgement test. Study 3 also delves into the dimensionality of CF and UF with respect to the stability of the factor structures. Accordingly, the SEM models predicting the latent variables of UF from those of CF are constructed separately for different speaking tasks, and the stability and variability of the CF-UF link across tasks are also discussed.

For a better understanding of the validity of UF measurements, Study 4 examines the extent to which L2 UF measures are associated with the corresponding L1 UF measures. Assuming that L2 speech production entails both L2-specific and language-general processes (Kormos, 2006; Segalowitz, 2010), some UF measures may be reflective of language-general processes or idiosyncratic factors, as opposed to L2-specific CF. Operationalizing the covariance between L1 and L2 UF measures as the contributions of language-general factors to L2 UF measures, Study 4 predicts L2 UF measures from their L1 counterparts. Prior research has reported the moderator effects of L2 proficiency on the strength of L1-L2 UF link. Therefore, Study 4 also investigates the extent to which the predictive power of L1 UF for L2 UF can be moderated by L2 proficiency. Using the factor scores of CF from Study 3 as a proxy for L2 proficiency, Study 4 tests the interaction effects by L1 UF measures and the proficiency scores on the corresponding L2 UF measures.

1.4 Thesis Structure

This thesis consists of 10 chapters including the current chapter of Introduction. Chapter 2 reviews the models of L2 speech production mechanisms and considers their importance in defining and operationalizing the three subconstructs of L2 oral fluency. In Chapter 3, the literature of L2 oral fluency is introduced with regard to the definition and valid operationalization of UF, PF, and CF. The chapter also provides the synthesis of research into

the UF-PF link, the effects of speech processing demands on UF performance, the CF-UF link, and the L1-L2 UF link. Chapter 4 outlines the methodology of Study 1, while Chapter 5 reports and discusses the findings of Study 1. Since Studies 2–4 are conducted based on the same dataset, Chapter 6 describes the methodologies of these three studies. Chapters 7, 8, and 9 present and discuss the findings of Studies 2, 3, and 4, respectively. Finally, Chapter 10 summarises the findings from four separate studies, followed by the discussion of theoretical and methodological contributions of these findings to L2 fluency research. The chapter concludes the whole thesis, by summarizing major findings, reporting methodological limitations, and suggesting future directions for L2 fluency research.

Chapter 2: Literature Review of Second Language Speech Production Mechanism

2.1 Introduction

This chapter provides a theoretical overview of L2 speech production mechanisms. I briefly review two major approaches to speech production mechanisms, namely, spreading activation and modular theories (Section 2.2). For a better understanding of differences between L1 and L2 speech production mechanisms, I also address the following theoretical issues specific to L2 speech processing: partially automatized L2 knowledge, the limited range of L2 linguistic resources, and simultaneous activation of L1 system (Section 2.3). I then introduce Kormos' (2006) model of L2 speech production to explain different phases and processes involved in L2 speech production (Sections 2.4–2.7), followed by its explicit connection to L2 oral fluency (Section 2.8). Finally, the chapter summarizes the similarities and differences of major phases in L2 speech production (Section 2.9).

2.2 Theoretical Assumptions of Speech Production Mechanisms

In this chapter, I mainly discuss Kormos' (2006) L2 speech production model which is based on modular theories of L1 speech production (e.g., Levelt, 1989,1999). Modular theories of speech production (e.g., Levelt, 1989, 1999) assume that the L2 speech processing system consists of several *modules* which are specialized in converting one particular type of input into a particular type of output, but not vice versa (i.e., one-way flow of activation). As this thesis adopts Kormos' (2006) model as a theoretical framework, I introduce one fundamental principle of modular theories, that is, the *competition-based mechanism.* This principle is applied to all the encoding processes in L2 speech production. As mentioned earlier, modular theories assume that the L2 system consists of different modules which produce a particular type of output in response to the input. However, the input simultaneously activates not only

the intended information but also other related information. The competition-based mechanisms postulate that the appropriate information is selected among the activated information according to the level of activation. The potential items for the output (e.g., concepts, words, phonemes) are automatically activated by the input and compete for the selection. The efficiency and accuracy of selecting the intended item can be facilitated by the higher resting activation level of the intended item. The level of activation of candidate items is determined by the level of correspondence with the input. Taking the example of lexical encoding, where lexical entries are selected according to the conceptual specifications of the preverbal message, the activation level of candidate lexical entries is dependent on the correspondence between the conceptual specifications of the preverbal message and the semantic representations of the candidate lexical entries stored in the mental lexicon. If the preverbal message includes the concept of DOG, semantically and thematically related lexical entries, such as hypernym (e.g., *animal*, *Golden Retriever*), similar objects (e.g., *cat*), and prototypical movements (e.g., *bark*, *walk*), are activated with varying activation levels. In successful lexical encoding, the lexical entry, *dog,* is selected and processed further in subsequent stages of speech production.

2.3 Issues Specific to L2 Speech Production

Historically speaking, L2 speech production models have been developed based on the theories and findings from L1 speech production research (e.g., Dell, 1986; Levelt, 1989, 1999; for a review, see Kormos, 2006). Since existing L1 speech production models generally focus on monolingual L1 speakers, the direct application of L1 speech production models might fail to explain some aspects of L2 speech production processes. When adapting L1 speech production models to the context of L2 speech, it should be considered how L2 learners and bilingual speakers (hereafter, L2 speakers) differ from monolingual L1 speakers.

Existing L2 models (e.g., de Bot, 1992; Segalowitz, 2010), including Kormos (2006), have specified the following essential characteristics of L2 system: (a) the controlled nature of L2 linguistic processing, (b) the limited range of L2 linguistic resources, and (c) simultaneous activation of L1 system. First, L2 learners are likely to rely, at least to some extent, on controlled processing. Particularly for those who have started L2 learning after the age of puberty (i.e., so-called the Critical Period Hypothesis; DeKeyser, 2000), it is rarely possible to learn the target language without analytic explanation of linguistic forms (e.g., word order, inflectional forms). In other words, such late learners have to rely on explicit learning to some extent. Due to the analytic nature of explicit learning, the outcome of their L2 learning tend to be declarative knowledge, which refers to factual and consciously accessible information and is thus often describable (DeKeyser, 2017). In contrast, L1 speakers acquire their L1 primarily by implicit learning without the awareness of learning objects. Accordingly, their L1 knowledge is considered to largely consist of procedural knowledge, which is responsible for implementing cognitive activities including language production (DeKeyser, 2017), and thus their speech production processes are mostly automatic (Kormos, 2006). Although declarative and procedural knowledge can be acquired for the same linguistic rules, the status of these two types of knowledge is different. From the perspective of cognitive psychology, declarative knowledge requires much more attentional resources to be retrieved than procedural knowledge. Accordingly, more attentional resources are required for late learners to produce L2 speech, compared to L1 monolingual speakers or early bilingual speakers who have a large amount of exposure to the target language in naturalistic and/or immersion contexts (DeKeyser, 2015; Ullman, 2015). More recently, DeKeyser and his colleagues have distinguished the knowledge acquired through implicit learning (i.e., implicit knowledge) from automatized explicit knowledge (Maie & Dekeyser, 2020; Y. Suzuki & DeKeyser, 2017). Although both types of knowledge are characterized by rapid

access, they differ in the extent to which awareness of linguistic forms is involved. Considering the importance of the speed of access to linguistic knowledge in speech production, this thesis follows the distinction of declarative/procedural knowledge.

The difference in the cost of attentional resources between declarative and procedural knowledge is assumed to be derived from the nature of cognitive processing of these two knowledge types. From the neurolinguistic perspective, declarative knowledge is assumed to induce controlled processing, while procedural knowledge is processed automatically (Paradis, 2009). Controlled processing requires conscious control and deliberate decisions, whereas automatic processing is characterized by effortlessness and is ballistic in nature. Controlled processing requires a great amount of attentional resources, and hence it is unlikely to be executed in parallel with other processing mechanisms. Meanwhile, automatic processing does not rely on attentional resources as much as controlled processing. It is thus assumed that automatic processing permits parallel execution of cognitive mechanisms and results in fluent performance. This distinction of knowledge and processing types is important in the context of L2 speech production, because only automatic processing can be executed in parallel with other processes. Therefore, one can assume that fluent speaking performance should be underpinned by automatic processing rather than controlled processing. Similarly, it can be proposed that less fluent L2 speakers may rely on declarative knowledge and controlled processing, whereas fluent speakers are likely to utilize procedural knowledge and automatic processing.

In addition to the degree of automaticity in speech processing, L2 speakers considerably vary in terms of their range of L2 linguistic resources. Prior research has showed that L2 speakers with lower proficiency tend to have a limited repertoire of linguistic resources, while those

with higher proficiency use relatively diverse linguistic resources (e.g., De Clercq & Housen, 2017; Lindqvist, Bardel, & Gudmundson, 2011; Zareva, 2007). These studies suggest that L2 linguistic resources can be enhanced as a function of L2 proficiency. Subsequently, it is possible that less competent L2 speakers have not acquired some linguistic knowledge which is needed to express their intended message. In such situations, they may compensate for the lack of particular linguistic knowledge by using communicative strategies or by substituting it with L1 linguistic knowledge (i.e., L1 transfer). Although some of those communicative strategies are common in L1 speech production (e.g., paraphrasing), the conscious or subconscious transfer of L1 knowledge is more characteristic of L2 speech production. Alternatively, the lack of linguistic knowledge corresponding to the intended message may result in disruptions of speech processing. For instance, when learners cannot retrieve some lexical item essential to communicate their intended message, the ongoing linguistic process (here, lexical retrieval) has to stop and thus cannot move forward to the subsequent processing stage. Accordingly, such disruptions of speech processing can be observed as breakdowns or silent pauses in the form of speaking performance.

Finally, as L2 speakers have a fully developed L1 linguistic knowledge and processing mechanisms, L1 linguistic knowledge, such as lemmas and phonological categories, are activated even in the course of L2 speech production (e.g., Kroll, Sumutka, & Schwartz, 2005). The level of activation of L1 linguistic knowledge during L2 speech processing may vary, depending on the degree of consolidation of L2 knowledge (cf. French & Jacquet, 2004; Pavlenko, 2009). For instance, beginning level L2 speakers may learn L2 vocabulary items by associating them with their L1 translation equivalents. As a result, they can retrieve L2 vocabulary items only through consciously activating L1 corresponding vocabulary items. Therefore, even during L2 speech production, their relevant L1 vocabulary items are highly activated. In contrast, advanced L2 speakers can directly retrieve L2 vocabulary items based

on their conceptual information. This is achieved by the gradual establishment of a direct memory trace between the concept and the corresponding L2 vocabulary item through extensive exposure and retrieval opportunities. Accordingly, such dissociation of L1 and L2 lexical knowledge is assumed to be established as a function of proficiency (e.g., the Revised Hierarchical Model of bilingual mental lexicon; Kroll & Stewart, 1994). However, both L1 and L2 speech production begin with the conceptual aspects of the intended message (i.e., conceptualization; see Section 2.5). Accordingly, even though one intends to speak in their L2, due to the language-general nature of concepts, their L1 linguistic knowledge is automatically activated to some extent during L2 speech production, even for those at the higher levels of L2 proficiency (Hoshino & Thierry, 2011; Kroll et al., 2006).

2.4 Knowledge Stores in L2 Speech Production Model

Considering the preceding characteristics specific to L2 speech production, Kormos (2006) proposed a L2 speech production model by integrating the modular models of L1 speech production (Levelt, 1989, 1999) with the developmental perspective of L2 competence. Her model assumes that L2 speech production proceeds by retrieving necessary information from four distinct knowledge stores in long-term memory: the episodic memory, the mental lexicon, the syllabary, and the declarative knowledge of L2 rules. The *episodic memory* contains the knowledge about discourse models and sociolinguistic competence, such as genre awareness and politeness, as well as the encyclopaedic knowledge including common knowledge/facts in the real-world and one's own memory of experience (so-called *knowledge of external and internal world*; Levelt, 1999). The episodic memory also includes the situational information including the interlocutors (e.g., where they are, who they are).

The second knowledge store—the *mental lexicon*—is the storage of L1 and L2 lexical entries with three hierarchical and interconnected levels: concept, lemma, and lexeme. A *concept* here is substantively equivalent to the semantic meaning of the given lexical entry and is closely related to the episodic memory. Among L1 and L2 lexical entries, some concepts can be identical, while other concepts are partly shared or distinct between languages (Pavlenko, 2009). For instance, *silla* in Spanish refers to a larger domain than *chair* in English, because the Spanish word, *silla*, can include the objects of *stools* in English (Graham & Belnap, 1986). It is also possible that some concepts are specific to one language and do not exist in another language. Meanwhile, a *lemma* represents the morphosyntactic properties of the lexical entry, such as obligatory/optional complements and inflectional features (e.g., gender). Finally, a *lexeme* contains the phonological (and orthographical) forms of the lexical entry. Notably, the mental lexicon stores not only single-word lemmas, but also multiword sequences (Pawley & Syder, 1983; Wolter & Yamashita, 2017; Yamashita & Jiang, 2010). Although there are different routes to establish multiword sequences as a lexical entry in the mental lexicon (e.g., chunking, association between words; see Durrant & Schmitt, 2010), the stored multiword sequences are assumed to have the same three-dimensional representation as single lemmas: semantic/pragmatic meaning, morphosyntactic properties, and phonological information (i.e., superlemma; cf. Sprenger, Levelt, & Kempen, 2006).

In addition to L2 knowledge related to the mental lexicon, L2 speakers also need to retrieve the articulatory motor gestures from a third knowledge store, that is, the *syllabary.* The syllabary stores a repertoire of syllable gestures for L1 and L2 altogether. Due to the shared storage of L1 and L2 syllable gestures, when one has not established the appropriate syllable gestures for L2, the pronunciation of the syllables tends to be substituted by the corresponding L1 gestures (see Section 2.6).

The final knowledge store, which is unique to Kormos's (2006) model, is the *declarative knowledge of L2 rules* (mainly, syntactic and phonological rules). This knowledge store is not proposed in L1 speech production (e.g., Levelt, 1989, 1999) possibly due to the fact that monolingual L1 speakers predominantly rely on procedural knowledge. In contrast, the extent to which L2 speakers' linguistic knowledge is proceduralized varies considerably, depending on the level of proficiency and/or the amount of L2 learning experience. L2 speakers may thus use declarative knowledge to carry out some linguistic encoding processes, particularly when partially proceduralized L2 knowledge is required to produce the intended message. The rationale for proposing L2 declarative knowledge as the distinct linguistic knowledge store from proceduralized rule-based mechanisms is supported by the findings from neuroimaging research.[2] For instance, comparing L2 speakers at different levels of proficiency and L1 speakers, Ullman (2015) showed that different brain regions were responsible for declarative knowledge and procedural knowledge of grammatical rules. More recently, researchers have claimed that highly automatized procedural knowledge is specific to the input of the encoding process, whereas declarative knowledge is, by nature, relatively flexible and abstract (DeKeyser, 2015; Ullman, 2015). More specifically, automatized procedural knowledge is accessible for the linguistic processing that the speaker has experienced before (e.g., prototypical usages), but not for the infrequent or unfamiliar linguistic context that the speaker has rarely experienced. Considering the fact that most L2 learners experience less linguistic exposure to the target language than L1 speakers do, it is likely that L2 learners' intended message includes some unfamiliar linguistic processing.

[2] The proceduralized knowledge for each rule-based linguistic procedure is presented as a separate module in the speaker's L2 speech production system.

Therefore, for L2 speakers to communicate their message, declarative knowledge should be accessed particularly in rule-based creative production.[3]

2.5 Conceptualization

L2 speech production proceeds in response to a certain communicative intention or demand. Accordingly, the first step of L2 speech production is to plan what message to convey to satisfy the given communicative intention, that is, *conceptualization*. L2 speech production models commonly assume that conceptualization is language-general, meaning that L1 monolingual speakers and L2 speakers are supposed to carry out conceptualization in substantively the same manner (de Bot, 1992; Kormos, 2006; Segalowitz, 2010). Thus, existing L2 speech production models follow Levelt's (1989, 1999) model for the description of conceptualization processes. This section introduces major processes of conceptualization based on Levelt's (1989, 1999) model.

Conceptualization consists of two sequenced processes: macroplanning and microplanning. During *macroplanning*, the speaker consciously or subconsciously decides (a) what speech acts and informational content to express and (b) the order of presenting such information to appropriately guide the interlocutor's attention so that the intended message will be understood. In order to perform speech acts appropriately, the speaker draws on the knowledge of discourse management from the episodic memory (e.g., sociolinguistic norms). In addition, for the sake of coherence throughout the speaker's utterances as well as the turns between the interlocutors, the episodic memory cooperates with the episodic buffer to temporarily store the information of the present communicative situation as well as the

[3] Of note, declarative and procedural knowledge were considered as continuum while they have been recently assumed as simply distinct memory systems.

ongoing discourse (Baddeley, 2003; Baddeley & Hitch, 2018). In accordance with the speaker's communicative intention, macroplanning computes an ordered sequences of speech acts or information as an input to the subsequent process, that is, *microplanning*.

During microplanning, the speaker further specifies the outcome of macroplanning (i.e., the conceptual structure of the message) by adding the informational perspective, such as semantic representations of the message and the *given* versus *new* status of information. The ultimate function of microplanning is to convert the conceptual structure from macroplanning into a linguistically expressible form by providing all necessary conceptual specifications for the subsequent linguistic encoding processes. More specifically, the speaker needs to translate the conceptual structure into the propositional form, in which the semantic entities and relations are embedded. For the message to take the propositional form, the message needs to contain the following major information (see Levelt, 1989, 1999): (a) the specification of referents, (b) the function/argument structure (i.e., predication about the referents), (c) the referents' thematic roles in the predication, (d) the head/modifier structure (e.g., quantification), and (e) the mood of message (e.g., interrogative or imperative for requesting).

Moreover, for each concept in the message, the speaker is assumed to assign language cues for the subsequent lemma selection as one of the conceptual specifications at the phase of microplanning (de Bot & Schreuder, 1993; Kormos, 2006). This means that the selection of the intended language is largely determined by the present communicative situation during conceptualization. Taken together, the product of microplanning involves the informational content in the propositional form and all the necessary conceptual specifications with the language cues being assigned to each lexical concept. This form of the conceptual message is

called the *preverbal message.* As suggested by its name, the content of the preverbal message is not yet linguistically encoded but is accessible for the subsequent linguistic encoding processes.

It is noteworthy that macroplanning is, in general, universal regardless of the language selected for speech, whereas microplanning is assumed to be language-specific (see also, de Bot, 1992; Segalowitz, 2010). One of the rationales for the language-specific nature of microplanning is the differential range of lexical concepts across languages. Another reason for the language-specific view of microplanning is the fact that each language requires a differential range of conceptual information that is obligatorily expressed in the surface structure of language, which is "an ordered string of lemmas grouped into phrases and subphrases of various kinds" (Levelt, 1989, p. 11). For instance, tense information is necessarily expressed in the form of verb conjugations in tense-marking languages such as English. However, some languages, such as Mandarin Chinese, Thai, and Japanese, do not have the system of verb tense. Such tenseless languages do not require the temporal property of tense, whereas they can optionally express timeframes using either lexical items (e.g., *kinou* [yesterday], *ashita* [tomorrow] in Japanese) or particles of aspects (e.g., *cengjing* for a past event, *jiangyao* for a future event in Mandarin;Qiu & Zhou, 2012).

2.6 Formulation and Articulation

The preverbal message generated through macroplanning and microplanning is sent to the *formulation* stage where the conceptual specifications in the preverbal message are converted into the corresponding linguistic forms. Formulation entails lexical encoding, morphosyntactic encoding, phonological encoding, and phonetic encoding, and proceeds in this order. Subsequently, the linguistically encoded message is then pronounced using speech

organs, that is, *articulation*. The following sections explain how the linguistic form of the message is specified through different encoding processes of formulation and is articulated.

2.6.1 Lexical encoding

As mentioned above, the preverbal message contains all necessary specifications for linguistic encoding processes. In lexical encoding processes, the speaker can retrieve the intended lexical entries from the mental lexicon with respect to the semantic representations and language cues of each concept (La Heji, 2005). Lexical encoding consists of two major processes: the activation of related lexical entries and the selection of the appropriate ones corresponding to the intended concept. First, the conceptual specifications of the preverbal message send activation to both L1 and L2 semantically relevant lexical entries (Hermans et al., 1998). Although the intended language of the concept is specified by the language cue, both L1 and L2 lexical entries are assumed to receive activation, because the mental lexicon is composed of various connections among L1 and L2 lexical entries (Hermans et al., 1998). Subsequently, all the L1 and L2 related lexical entries are activated according to the conceptual specifications and then compete for selection (for the alternative process of langauge selection, see Kroll, Bobb, Misra, & Guo, 2008).

Second, following the abovementioned competition-based mechanism, the lexical item which received the highest activation among the related items is selected (La Heji, 2005). In general, the level of activation of the target lexical item can be enhanced by the correspondence to the conceptual specification of the lexical concept. Moreover, research into slips of the tongue suggests that the relative levels of activation between the intended item and unintended items may affect the success of lexical retrieval (Poulisse, 1999; Poulisse & Bongaerts, 1994). When the activation level of the intended item is comparable to the other related items, the unintended items can also be selected. However, in the context of

L2 speech production, the success and efficiency of lexical selection is also affected by the extent to which the knowledge of the target lexical entry is mastered. L2 learners' mental lexicon varies in terms of the depth of knowledge of lexical entries as well as the number of lexical entries (for the recent review, see also Dóczi & Kormos, 2016). Therefore, in order to achieve efficient lexical encoding, L2 speakers need to establish or strengthen the memory traces between the lexical concept and its corresponding lexical entry. The outcome of lexical encoding processes is the set of lexical entries corresponding to the conceptual specifications of the preverbal message (see Figure 2). This lexicalized message is then sent to the morphosyntactic encoding stage.

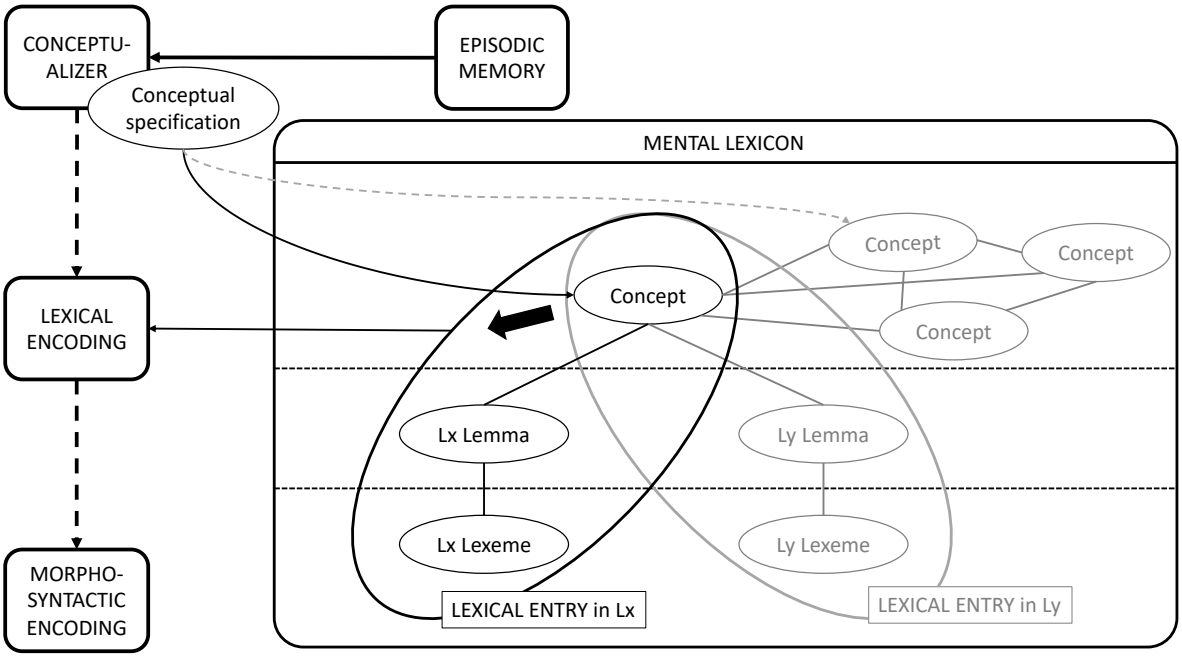


Figure 2. The visualization of lexical encoding process from conceptual specification of the preverbal message to lexical entry in the mental lexicon on the basis of Kormos (2006).

2.6.2 Morphosyntactic encoding

Kormos (2006) claims that the general mechanisms of syntactic encoding do not substantively differ between L1 monolinguals and L2 speakers. Accordingly, her model adopts Kempen and Hoenkamp's (1987) Incremental Procedural Grammar (IGP), which is

also adopted in Levelt's (1989) L1 speech production model. The IPG assumes that syntactic encoding processing comprises a collection of grammatical and functional procedures. Notably, each of these grammatical and functional procedures only produces their own special kind of the output when they are called upon by a particular type of input. This presumption of the IGP can be considered theoretically compatible with the principle of modular theories which Kormos' (2006) speech production model is based on. Another important assumption of the IPG is the existence of a storage buffer for the intermediary products of syntactic procedures. The storage buffer plays a role when the already activated information is necessary for some syntactic processing at later phases of syntactic encoding (for the details, see the explanation below).

According to the IPG, syntactic encoding consists of two major processes: the activation of morphosyntactic properties of selected lexical entries (i.e., lemmas) and the construction of phrase/clause structures based on the syntactic rules of the language. In L1 speech production model, it is assumed that L1 knowledge about lemmas is fundamentally declarative in nature, whereas the building procedures of phrases/clauses are realized with the help of procedural knowledge (Kormos, 2006; Levelt, 1989). In contrast, as mentioned previously, L2 speakers vary in the degree of proceduralization/automatization of L2 rule-based knowledge. Depending on the level of proficiency, L2 speakers may thus rely on their declarative knowledge of L2 syntactic rules (e.g., word order) even in the second phase of syntactic processing. Therefore, Kormos' (2006) model assumes that the declarative knowledge of L2 syntactic rules contributes to L2 speakers' speech production to a large extent, especially for late learners, compared to L1 speech production.

Syntactic processing starts in response to the outcome of lexical processing. In other words, the input of syntactic processing is a set of lexical entries embedded in the propositional form. As the propositional form includes all the necessary information for linguistic encoding processes, the lexicalized message from lexical encoding conceptually specifies the hierarchical relationship between lexical entries, such as head/modifier structures (see Section 2.5). Regarding the first process of syntactic encoding, the speaker activates the selected lexical entries to retrieve their morphosyntactic properties (i.e., lemma), such as gender and obligatory/optional complements. In both L1 and L2 speech production, even a fragment of the input can trigger the target encoding processing mechanism, which is described as the incremental nature of speech production (Kormos, 2006; Levelt, 1989; see also Section 2.8). In the case of L2 syntactic encoding, once the lexical entry is selected based on the conceptual specifications in the preverbal message (i.e., lexical encoding), the lemma is retrieved for syntactic encoding. Therefore, there must be an orderly fashion in the activation of lemmas; the order of specifications of concepts and lexical items may determine in what order the morphosyntactic properties of lemmas are retrieved.

The second process of syntactic encoding consists of multiple morphosyntactic procedures to construct phrases and clauses. First, the syntactic category of the lemma (e.g., noun vs. verb) initiates a *categorical procedure*, which establishes the phrasal category to which the selected lemma belongs (e.g., noun phrase vs. verb phrase). For instance, the word–*cat*–has a syntactic category of a noun, indicating that this word can be the head of a noun phrase (NP). Subsequently, the building procedure specific to NPs will proceed.

Second, the categorical procedure also identifies the other lexical entries conceptually attached to the lemma in the message, which can fill the lemma's obligatory/optional

complements and specifiers (e.g., modifiers in the head/modifier structure). Moreover, the categorical procedure sets the diacritic parameters for the lemma, such as accessibility status and singularity status. For instance, when the message refers to two cats which are identifiable to the interlocutor, the information of quantification (here, *two*) and the status of [+ accessible] will be attached to the head of the NP, *cat*. The outcome will be *the two cats* with the status of [+accessible, -singular]. In other words, although most of the conceptual specifications are processed during the phase of lexical encoding, some informational perspective of the preverbal message is processed at the syntactic encoding phase.

Third, once the phrasal structure is constructed, the categorical procedure selects the grammatical function (e.g., subject, object) of the processed phrase structure. The grammatical functions of the phrase structures are specified based on the thematic roles assigned in the preverbal message (see Section 2.5; Levelt, 1989, 1999). For instance, when a NP is processed, the categorical procedure identifies whether the NP will be the subject of the given clause (or sentence [S]) or the object of the main verb in response to the assigned thematic role (e.g., agent vs. recipient). Subsequently, the identification of the grammatical function assigns the diacritic feature of case (e.g., nominative case for NP_{SUBJ}) which specifies the word form in the surface structure at the later stage. Afterwards, the categorical procedure further decides whether the phrase structure will be a head or a complement of a higher-order categorical procedure, if applicable. These processes of categorical procedure are depicted in the upper part of Figure 3.

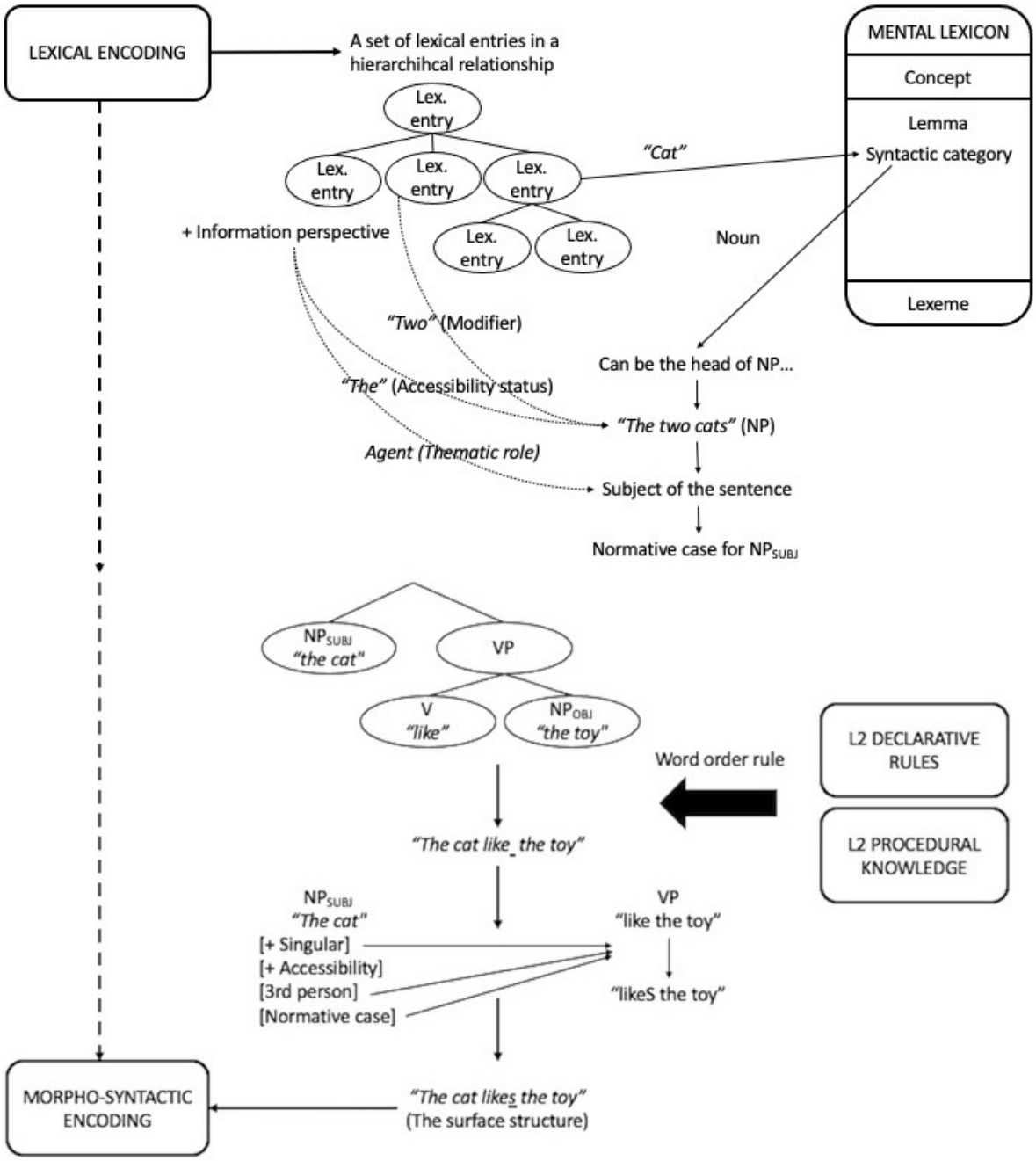


Figure 3. The visualization of lexical encoding process from conceptual specification to lexical entry on the basis of Kormos (2006).

The fourth step of syntactic encoding is the activation of the word order rule of the intended language and to specify the positions of the words in the surface structure of the sentence. More specifically, according to the word order rule of the language, the hierarchical relationship between the processed words is converted into the ordered sequence of words. For instance, the NP with the grammatical function of subject will be located in the initial

position of the clause in English. At this phase, the higher-order categorical procedure mentioned in the fourth step can also be activated. If the clause is a constituent of a higher-order sentence/clause, the subordinate clause is constructed and attached to the phrase which belongs to a super-ordinate clause (S).

After the order and grammatical functions of lexical entries are specified, some diacritic parameters can be specified at the final phase of the IPG. This is because these diacritic parameters are dependent on the information outside the phrase to which the target lexical word belongs (i.e., so-called *immediate maximal projection*; Myers-Scotton & Jake, 2000). One of the representative examples of this type of linguistic phenomena is a subject-verb agreement. For instance, when the grammatical subject is a NP, *the cat*, the head of this NP (*cat*) has the parameters of third person, nominative case, and singular. If the tense of the main verb which belongs to the NP's predicate is the present tense, the third person singular-*s* will be attached to the verb (e.g., *like the toy* → *likeS the toy*). These fourth and fifth phases of morphosyntactic encoding are visualized in Figure 3. The eventual output of these syntactic encoding procedures is the surface structure which is the ordered sequence of lemmas grouped into phrases and subphrases. The surface structure is then delivered to phonological encoding processes.

2.6.3 Phonological encoding

After the surface structure of the message is created through morphosyntactic encoding, the speaker converts the surface structure into an audible stream of sounds (i.e., overt speech). This conversion is achieved through three sequenced processes: phonological encoding, phonetic encoding, and articulation. In order to understand the differential roles of these three processes, a distinction should be made between two levels of pronunciation: *phonological*

and *phonetic*. Generally speaking, the phonological level refers to the abstract and representative level of sounds, whereas the phonetic level is concerned with the actual sound (Ladefoged, 2015). For instance, the word-initial and word-final sounds of /t/ in English (e.g., *tie* vs. *cat*) are phonologically identical but phonetically different (e.g., /taɪ/–[t^h]; aspirated] vs. /kæt/–[t]; unaspirated]). These phonologically identical but phonetically distinct sounds (phones) are called *allophones*. These phonetic differences can be found not only within languages but also between languages. For example, English word-initial /t/ is aspirated, while Japanese word-initial /t/ is unaspirated. It is thus possible that some L1 and L2 phonological representations can be identical but phonetically different, while other phonological representations, such as vowels, tend to differ across languages. In the former case, the L1 influence on L2 pronunciation is often observed at the phonetic level. Taking the example of the aspiration at the word-initial position, the length of the voice onset time (i.e., the time interval between the burst release of the plosive and the onset of voicing) is commonly used as an acoustic cue to detect the degree of L1 influence in L2 speakers (Stoehr et al., 2017).[4] However, when it comes to the latter case (i.e., distinctive phonological categories between languages), it is substantively difficult to identify whether L2 pronunciation is affected by L1 at either phonological or phonetic levels.

The distinction between phonological and phonetic levels is useful to understand how phonological and phonetic encoding processes are different. Phonological encoding converts the surface structure into its phonological representations. In the course of phonological encoding, three subprocesses establish the phonological representations of the message. First, the speaker activates the phonological information of lexical items and inflectional

[4] Stoehr et al. (2017) mentioned that the attainment of L2 phonetics can cause the attrition of L1 phonetics in late bilingual speakers. The influence of phonetic systems across languages can be reciprocal, suggesting some possibility that L2 acquisition may modify the speaker's L1 phonetic system.

morphemes embedded in the surface structure. In other words, phonological encoding mechanisms send activation to the mental lexicon and retrieve the phonological information about what phonemes are contained in the lexical entry (i.e., lexeme; see Figure 2) in a serial fashion (i.e., from the word-initial to the word-final). One of the theoretical issues related to the retrieval of phonological forms of lexical entries is the unit of stored phonological representations, such as features, segments and syllables. In general, speech production models based on modular theories (e.g., Kormos, 2006; Levelt, 1989, 1999) follow the segmental view of phonological representations (e.g., Roelofs, 1999). The assumption of the segmental view is that each phonological segment has its own abstract representation in phonological encoding system as a chunk of features. For instance, the phonological representation of /b/ is stored as a minimum unit consisting of features such as [+ voiced] and [+Labial]. In other words, the same feature is also stored in other phonological segments (e.g., the [+voiced] feature in /b/, /d/, /g/, etc.).

After the activation of phonological segments for each word in the message, the syllabification process begins. Due to the serial nature of the activation of phonological segments in the lexeme, the phonological encoding process identifies the order and position of each phoneme within the word. The syllabification process groups phonemes into syllables, following the language-specific syllabification rule (i.e., phonotactics). More specifically, a vowel or a diphthong is assigned to a different syllable node as a nucleus. As depicted in Figure 4, a syllable node generally consists of the onset (i.e., consonant(s) before the nucleus; e.g., /sp/ in *speech*), the nucleus (e.g., /iː/ in *speech*), and the coda (i.e., consonant(s) after the nucleus; e.g., /tʃ/ in *speech*). In L2 speech production, speakers can apply the L1 syllabification rule if the L2 syllabification rule is not sufficiently established. For instance, Japanese speakers of L2 English are likely to apply the rule of Japanese syllable

structure to L2 English syllabification processes (Japanese–one consonant in the onset vs. English–at maximum, three consonants in the onset). As a result, they may insert an extra vowel after each consonant to keep the consonant–vowel (CV) syllable structures when producing English words (e.g., English word *sky* /skaɪ/ [one syllable] as /su･ka･i/ [three syllables]; Vance, 2008).

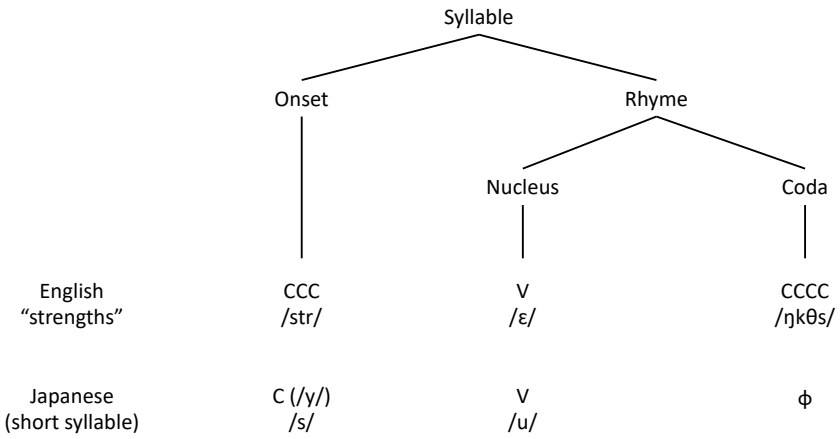


Figure 4. A visual representation of syllable structures of English and Japanese short syllable.

This syllabification process occurs not only within words but also between words. When a word-initial syllable begins with a vowel (i.e., no consonants in its onset), the vowel can be attached to the coda of the word-final syllable of the previous word. Taking an example of "*select us*" from Levelt (1999), the phonological forms of each word in this phrase can be described as /sə･lekt/ (two syllables) and /ʌs/ (one syllable). The coda of "*select*" (/kt/) has two consonants of /k/ and /t/, and the nucleus of the second syllable (/e/) requires the former consonant (/k/) as the coda of the syllable to keep the default structure of syllables in English (i.e., CVC). Although English syllables also tend to have a syllable-initial consonant (i.e., maximization of onset; Levelt, 1999), /kt/ in "*select*" cannot be an onset of syllables due to phonotactic rules in English (i.e., */ktʌs/). Accordingly, while the former consonant (/k/) remains in the position of coda of the previous syllable (/le/→/lek/), the vowel of the

following word (/ʌ/ in "*us*" /ʌs/) takes the latter consonant (/t/) as its onset in order to follow the default structure of English syllables. Then, the consonant /t/ of *"select"* is combined with the rhyme of the subsequent word "*us*" (/ʌs/). Finally, the whole phrase consisting of two words (*select us*) is produced like a single word as "*se-lec-tus*" (i.e., cliticization).

In addition to the syllabification process, phonological encoding specifies the metrical features for the message, such as pitch and stress. It is noteworthy that some metrical features including the location of the prominences in intonational units are determined by microplanning at the phase of conceptualization. This is because some intonational features are closely related to the informational aspects of the message, such as the accessibility status of lexical entries (*given* vs. *new* information to the interlocutor). Accordingly, as with syntactic encoding, some informational perspective of the message can be conveyed directly from conceptualization. The eventual outcome of phonological encoding is the syllabified phonological representations of the message with the specifications of metrical features. This form of the message is called the *phonological score* which is ready to activate the corresponding articulatory gestures.

2.6.4 Phonetic encoding and articulation

The pronunciation of speech is assumed to be established through phonological and phonetic encoding processes. However, as described previously (see Section 2.6.3), these encoding processes are distinct in terms of underlying linguistic processing. Phonological encoding is concerned with the transformation of the surface structure of the message into the phonological representations. The outcome of phonological encoding is a string of syllabified phonemes and metrical features (i.e., phonological score). Using this phonological score as the input resource, phonetic encoding converts the phonological score into a plan of

articulatory movements/gestures corresponding to the phonological score, that is, *articulatory score*).[5] In addition, phonological and phonetic encoding mechanisms have access to different knowledge stores. Phonological encoding retrieves the lexemes from the mental lexicon, whereas phonetic encoding retrieves articulatory gestures from the syllabary, which is one of the knowledge stores in the speaker's long-term memory (Levelt, 1989; see Section 2.4).

As mentioned in Section 2.4, the syllabary stores both L1 and L2 motor programs for syllables. Moreover, both L1 and L2 can have some identical phonological segments which differ at the phonetic level (i.e., allophones). Accordingly, beginning-level L2 learners are likely to substitute L2 articulatory gestures with their L1 counterparts, because they might have not established target-like L2 articulatory gestures. Taking the example of the /t/ sound produced by Japanese speakers of English, the /t/ sound at the beginning of the word is generally unaspirated in Japanese, while the word-initial /t/ sound is aspirated in English (Harada, 2007). English speakers distinguish the word-initial /t/ from /d/ by relying more on whether the sound is aspirated or not (i.e., [+/- aspirated]) than on whether the sound is voiced or not (i.e., [+ voiced]). On the other hand, Japanese speakers usually distinguish /t/ and /d/ in Japanese in terms of [+/- voiced] status, and thus the distinction based on aspiration is difficult for Japanese learners of English to attain. In addition, since the aspirated /t/ sound does not exist in (standard) Japanese, Japanese learners of English perceive both [t^h] and [t] as the same phonetic segment of /t/. Therefore, they tend to produce the word-initial /t/ sound in English words as [t] (i.e., L1 phonetic substitution) unless they establish L2-specific articulatory gestural scores in their syllabary (Harada, 2007).

[5] Of note, the outcome of phonetic encoding can be called either the articulatory score or the internal speech. The articulatory score is the name as the prerequisite information for the subsequent process–articulation, while the internal speech is the name as the speech before/without articulating to the possible interlocutor (i.e., speech to oneself).

The final stage of speech production is *articulation*, where the articulatory score from the phonetic encoding is executed to produce speech sounds. In order to produce intelligible sounds to the interlocutor, the speaker utilizes various speech organs such as tongue and tooth. The outcome of the articulation is called the *overt speech*. Articulation is theoretically distinct from the formulation mechanisms, because the execution of articulation involves the use of motor movements rather than linguistic representations and encoding processes.

2.7 Self-monitoring

In addition to three major stages of speech production (Conceptualization, Formulation, and Articulation), self-monitoring is also essential for successful speech production. Kormos' (2006) model follows Levelt's (1983, 1989) *perceptual loop theory* (PLT), assuming that the self-monitoring processes should work similarly both in L1 and L2 speech production. In this section, I describe the PLT which proposes three different loops in the course of speech production: (a) the conceptual loop, (b) the prearticulatory loop, and (c) the external loop. These three monitoring loops inspect the intermediary or eventual outcome of different encoding processes, meaning that they are different in the target encoding modules and scope for the inspection. The first monitoring loop, the *conceptual loop*, inspects the outcome of the conceptualization (i.e., the preverbal message) in terms of the extent to which the preverbal message is in line with the speaker's original communicative intention. As the preverbal plan is not linguistically converted, the conceptual loop exclusively examines the communicative appropriacy of the message which the speaker intends to express. The second loop is the *prearticulatory loop* which checks the processed message in terms of linguistic accuracy before articulating the message. More precisely, the prearticulatory loop inspects the outcome of formulation (i.e., the internal speech). Meanwhile, the *external loop*, which is the final loop of self-monitoring, examines the articulated speech, once it is parsed by the speakers

themselves. The scope of the external loop of monitoring can include both communicative appropriacy and linguistic accuracy. In actual speech, the external loop of self-monitoring can be observed as self-repair behaviors, such as false starts and reformulations (Kormos, 2000; Williams & Korko, 2019). Self-repair behaviors represent the resolution of the problem which is detected by self-monitoring processes. Self-monitoring based on the external loop is called *overt monitoring,* as the monitoring processes can be evidenced by self-repair behaviors. In contrast, monitoring via the prearticulatory loop is called *covert monitoring,* as its problem-solving mechanisms are only observable as pauses or silence. However, pauses in speech can reflect the time required either for encoding processes (e.g., lexical retrieval) or covert monitoring (see also Kormos, 2006; Postma & Kolk, 1993). Pauses can thus derive from either breakdown in encoding processes or operations of self-monitoring, and therefore listeners are not always aware of the ongoing process of covert monitoring in their interlocutors' speech.

Regardless of the loop of self-monitoring, it is assumed that self-monitoring is based on the speaker's comprehension system (Levelt, 1983, 1989; for an alternative view, see Nooteboom, 1980). According to Levelt (1989), the rationale for locating the self-monitoring function as part of comprehension system rather than production system is that if self-monitoring function is distributed in the production modules, it requires the speakers to have a replicated linguistic knowledge store for encoding and monitoring (Levelt, 1989). This view is against the principle of ecology and simplicity in human cognition. Another piece of indirect evidence for this assumption is that research findings show that the relative speed of executing different types of self-repairs is similar to that of detecting content- and form-related errors when processing the interlocutor's speech (i.e., speech comprehension; see

Kormos, 2000b). Accordingly, speakers' self-monitoring processes can be similar to the way they monitor the interlocutors' speech.

In the literature of self-monitoring, scholars have debated "what kind of representation or code" the prearticulatory loop of self-monitoring processes can access when internal speech is monitored (Levelt et al., 1999, p. 6). Although the prearticulatory loop can detect a wide range of errors in speech, including lexical, syntactic, morphological, and pronunciation errors, Levelt's (1989) PLT supposed that the speakers can only inspect the phonetic representation (i.e., the outcome of phonetic encoding; immediately before articulation) for linguistic accuracy, probably because Levelt (1983, 1989) assumed that the self-monitoring function is rooted in the speakers' comprehension system. In order for the comprehension system to monitor the speakers' own internal or overt speech, the phonetic representation of the message needs to be activated at least internally (i.e., as a result of phonetic encoding). The activation of phonetic representations triggers the parsing process of speech comprehension, which segments the phonetic representation of speech into meaningful units, such as words and chunks, to retrieve their semantic representations (Levelt, 1993). Note that prior to the parsing process, comprehension of interlocuter's speech carries out speech perception processes (i.e., so-called *acoustic-phonetic processor* or *audition*), which extract and retain a phonetic representation of speech from a stream of sounds (Anderson, 2009). However, when it comes to monitoring the speakers' own internal speech, they do not need to perceive this overt speech signal, because the phonetic representation of the message is already activated and stored in their phonological short-term memory (Jacquemot & Scott, 2006).[6] Therefore, it can be argued that internal speech is inspected by the prearticulatory

[6] Although Jacquemot and Scott (2006) assume that the phonological representation of the message is inspected for self-monitoring the internal speech, their fundamental tenet of temporarily buffering the outcome of encoding processes (either phonological or phonetic levels) was adopted here.

loop only after the phonetic representation of the message is produced. On the other hand, more recent studies, including Levelt's subsequent work (Indefrey & Levelt, 2004; Levelt et al., 1999; Wheeldon & Levelt, 1995), have claimed that the internal speech is the consequence of activating the phonological representation of words rather than their phonetic representations or their articulatory features (for a review, see Oppenheim & Dell, 2010). However, due to the incremental nature of processing from phonological encoding to articulation (Levelt, 1993), it is rare that the representation of the message is purely composed of either phonological or phonetic representations in the context of spontaneous speech production, because the initial syllable of the word can be articulated even before the phonetic encoding of the whole word is completed (Bachoud-Lévi et al., 1998). In sum, it might be feasible to argue that the representation of internal speech can be a combination of phonological and phonetic representations of the message with a varying degree of dominance between these two levels.

Following Levelt's (1983, 1989) PLT, the self-monitoring process is regarded as a conscious activity, meaning that the speaker needs to use attentional resources for self-monitoring. As mentioned early (see Section 2.3), since L2 learners are likely to have partially automatized L2 knowledge, a relatively large amount of attentional resources is assigned for linguistic encoding processes. Subsequently, fewer attentional resources can be available for self-monitoring in L2 speech production, especially for beginning-level L2 speakers. It can thus be assumed that L2 learners may pay attention to one aspect of speech, as opposed to multiple aspects simultaneously (e.g., content vs. well-formedness; Skehan, 2014a). Self-repair behaviours are not supposed to automatically occur according to the encoding flaws that the speakers detected. Depending on the severity of errors and the situational characteristics (e.g., communicative situations, interlocutors), speakers might not correct their

errors even if they detect them. For instance, when L2 learners talk with other learners with the same L1 background, some wrong lexical choice due to L1 substitutions may be ignored, especially when the intended meaning of the words is comprehensible and transparent to the interlocutors. Meanwhile, when they talk to learners from diverse language backgrounds, speakers may be aware that such lexical errors can result in communication breakdowns. Accordingly, once they produce such lexical errors, they may tend to self-repair the errors.

2.8 Connecting L2 Speech Production Mechanisms with L2 Fluency Research

L2 fluency research has adopted L2 speech production models as a theoretical framework to explain how L2 linguistic knowledge or competence are reflected in the form of utterances or speaking performance (Kormos, 2006; Segalowitz, 2010; Skehan, 2014a; Tavakoli & Wright, 2020). As mentioned in Section 1.2, this association between linguistic knowledge and actual speech is equivalent to the relationship between CF (cognitive fluency) and UF (utterance fluency) within Segalowitz's (2010) three subconstructs of fluency. However, one may argue that the way that CF (i.e., components of speech production mechanisms) is manifested in UF (i.e., temporal characteristics of speech) is not always straightforward. In addition to major components of speech production mechanisms, three important characteristics of L2 speech production may explain how effectiveness and efficiency of speech processing are manifested in fluency of speech. The first characteristic is the *incremental* nature of speech production. In speech processing, even a fragment of the input can trigger a particular type of speech processing mechanism. This is specific to speech production models based on modular theories, which assume that each module starts processing if and only if it has received its specific input. For instance, as soon as part of the preverbal message is passed on to the formulator, lexical and phonological representations of each concept in the message can be

activated and even articulated, even while for the remaining part of the preverbal message, appropriate lexical entries are still being selected (i.e., lexical encoding).

Relating to the incremental nature of speech production, the second characteristic is that different speech production processes proceed in *parallel*. In other words, speakers can engage in different speech processing simultaneously. Taking the example above, speakers can articulate part of the intended message and simultaneously retrieve the lexical entries corresponding to the other parts of the message. This characteristic allows for the smooth continuation of speech across utterances. Once the preverbal message is sent to the formulator, the conceptualizer starts again working on the next chunk of message, even while the previous chunk is still being processed in the formulator. Due to such incremental and parallel processing at the between-utterance level, speakers can maintain coherence across utterances without longer breakdowns between them. Similarly, even at the within-utterance level, these characteristics enable speakers to process a chunk of words and produce fluid utterances rather than word-by-word speech production.

These two incremental and parallel characteristics, which are essential for smooth oral communication, are underpinned by the third feature of speech production—*automaticity*. Especially in the case of L1 speakers, their spontaneous speech is generally smooth and efficient, presumably because their linguistic knowledge is fully automatized. Thanks to the automatized linguistic knowledge, they can engage with multiple types of speech processing simultaneously with the small amount of attentional resources. In order to achieve an optimal level of fluency in L2, learners need to carry out speech processing incrementally and in parallel. To this end, they need to attain automatized/procedural L2 knowledge, which allows for smooth speech processing with a small amount of attentional resources. Otherwise,

learners are forced to rely on declarative knowledge and to engage with controlled processing (see Section 2.4). As a result, they may experience a shortage of attentional resources, leading to the failure of operating multiple processing in parallel. Therefore, L2 speakers' variability in oral fluency (i.e., UF) can be assumed to reflect the extent to which their L2 knowledge is automatized (i.e., CF; Segalowitz, 2010, 2016).

2.9 Summary

Motivated by the assumption that the construct of CF is theoretically reflective of speaker's L2 speech production system (Segalowitz, 2010, 2016), this chapter illustrated what and how linguistic knowledge contributes to the production of utterances, based on Kormos' (2006) model of L2 speech production. According to Kormos' (2006) model, L2 speech production mechanisms have four major components: conceptualization, formulation, articulation, and self-monitoring. Conceptualization is responsible for the generation of the preverbal message in response to the given communicative demands, while formulation converts the preverbal message into the corresponding linguistic representations. Articulation produces the overt speech of the linguistic representation by the motoric execution of speech organs. Self-monitoring refers to the function of checking the correctness and appropriateness of the outcome of these three processes in terms of meaning and forms.

The major distinctions of these four components of L2 speech production can be made in terms of the degree of language-specificity and the nature and scope of processing (see Table 1 below). First, the conceptualization processes are largely shared across individual's languages (e.g., L1 and L2) and thus are considered language-general. Accordingly, the conceptualization in L2 speech production is assumed to be relatively independent of L2 proficiency. Meanwhile, formulation and articulation can be regarded as language-specific

processes, as they draw on some form of L2 knowledge. However, in order for conceptualization to produce the preverbal message, speakers are required to create the language-specific propositional form of the message, which is accessible at the subsequent linguistic encoding processes. Therefore, although conceptualization is not supposed to draw on L2-specific linguistic knowledge, conceptualization in L2 oral production is affected by the effectiveness and efficiency of subsequent processes, such as formulation and articulation.

Table 1. *Summary of similarities and differences between conceptualization, formulation, articulation and self-monitoring.*

	Language specificity	Level of representation	Focus of processing
Conceptualization	Language-general	Conceptual/semantic	Meaning
Formulation	L2-specific	Linguistic	Form
Articulation	L2-specific	Motoric/gestural	Form
Self-monitoring	Language-general and L2-specific	All of the above	Meaning and form

Another distinction should be noted between two language-specific components—formulation and articulation. Formulation is composed of linguistic encoding processes at the different linguistic levels, including lexical, morphosyntactic, phonological and phonetic levels. These encoding processes manipulate different types of linguistic representations (e.g., lexical and phonological representations). In contrast, articulation is considered purely motoric, meaning that the execution of articulation involves the use gestural movements rather than information processing. Despite the motoric nature of articulation, there is variability among L2 speakers in the efficiency of the execution of articulation processes (Broos et al., 2018), which confirms the L2-specific nature of articulation.

Finally, the focus of these components of speech production also differs from each other. Due to its preverbal and language-general nature, conceptualization is closely linked with the meaning/content aspects of the message. Formulation and articulation are responsible for the formal conversion of the intended message so that the outcome of speech production (i.e., the overt speech) is comprehensible for the interlocutors. Meanwhile, self-monitoring can be concerned with both the content and form of the message. The focus of self-monitoring can thus be either on content accuracy or on linguistic well-formedness.

In relation to the contribution of CF to UF, the chapter also reviewed three essential features of fluent speech production—incremental, parallel, and automatic speech processing. L2 speech production mechanisms suggest that oral fluency can be achieved by the automatization of L2 knowledge, which allows for efficient speech processing with a small amount of attentional resources. This is in line with the construct definition of CF (Segalowitz, 2010, 2016), confirming the validity of speech production mechanisms as the theoretical framework for L2 fluency research.

Chapter 3: Literature Review of Second Language Oral Fluency[7]

3.1 Introduction

This chapter outlines the theoretical definitions, operationalizations, and common measurements of three subconstructs of L2 oral fluency in Segalowitz's (2010, 2016) framework—UF (utterance fluency; Section 3.2), PF (perceived fluency; Section 3.3), and CF (cognitive fluency; Section 3.6). In addition, empirical studies on the relationship between UF and PF (Sections 3.4–3.5) and between CF and UF (Section 3.7) are also synthesized with regard to the potential moderator effects of methodological factors. Motivated by the essential role of task design in the association between CF and UF, the current chapter also theorizes how task design features affect the processes of L2 speech production, drawing on the framework of speech processing demands (Section 3.8). Finally, in light of the validity of UF measures, the association between L1 and L2 UF is reviewed (Section 3.9).

3.2 Utterance Fluency

Within Segalowitz's (2010, 2016) framework, UF (utterance fluency) refers to the observable temporal features reflecting the speaker's operation of L2 speech production mechanisms (i.e., CF), such as the speed of delivery, pauses, and hesitations. Researchers have commonly divided UF into three subcomponents—speed, breakdown, and repair fluency (Skehan, 2003; Tavakoli & Skehan, 2005). *Speed fluency* is concerned with the density of information or the speed of delivery and thus is typically measured by articulation rate (i.e., the mean number of syllables produced per minute excluding pauses). *Breakdown fluency* refers to pausing behaviours including silent and filled pauses. Assuming that pauses may reflect disruptions in speech processing, scholars have traditionally operationalized breakdown fluency in terms of

[7] Several sections of this chapter were accepted for publication in *The Modern Language Journal* as Suzuki, Kormos, and Uchihara (in press, 2021).

the length and frequency of pauses. There has been a long debate over the minimum length of pauses attributed to breakdowns in speech processing, such as lexical retrieval and syntactic procedures (De Jong et al., 2013; De Jong & Bosker, 2013), because short pauses are less likely to reflect such breakdowns in speech processing (i.e., so-called *micropauses*; Riggenbach, 1991). Thus, scholars have attempted to identify an optimal threshold of silent pause length. Recent studies tend to define a pause as a silence longer than 250 ms, considering its predictive power in PF ratings and L2 proficiency (De Jong, 2016b; De Jong & Bosker, 2013). In addition to the length and frequency of pauses, recent studies have also recognized the importance of pause location in predicting PF (Kahng, 2018; Saito et al., 2018; S. Suzuki & Kormos, 2020). These studies have commonly shown that pauses in the middle of clauses were more strongly associated with PF than pauses at clause boundaries, confirming the findings from small-scale qualitative studies (e.g., Hawkins, 1971; Riggenbach, 1991). From the perspective of L2 speech production mechanisms (Kormos, 2006; Segalowitz, 2010), pauses within clauses are hypothesized to reflect disruptions in L2-specific linguistic processing, while pauses at clausal boundaries are supposed to capture the breakdown in conceptualization-related processes, such as content planning (De Jong, 2016b; Götz, 2013; Skehan et al., 2016; Tavakoli, 2011). Finally, repair fluency is, by definition, a range of disfluency phenomena, including self-corrections, false starts, and verbatim repetitions. It can be argued that repair fluency is in a supplementary relationship with breakdown fluency. From a theoretical perspective, both breakdown and repair fluency are assumed to reflect the operation of self-monitoring processes (i.e., covert and overt repairs, respectively; see Kormos, 2000, 2006) and also are regarded as an opportunity for speakers to buy time to deal with disruptions in speech processing (Bui et al., 2019; De Jong et al., 2015). Accordingly, some studies even examine breakdown and repair fluency as inseparable phenomena (e.g., Williams & Korko, 2019).

L2 fluency research has conventionally measured temporal features of speech, following Tavakoli and Skehan's (2005) triad model of UF (speed, breakdown, and repair fluency). Building on this triad model of UF, scholars have raised methodological concerns about the validity of UF measurements (cf. Housen & Kuiken, 2009; Lambert & Kormos, 2014; Michel, 2017). A variety of UF measures have been developed, and the validity of those measurements has received increasing attention (see Bosker et al., 2013; Lambert & Kormos, 2014). In some studies, the construct validity of UF measurements might be negatively affected if the measurement taps into multiple dimensions of UF (Bosker et al., 2013). For instance, one of the composite measures, mean length of run (MLR), is calculated based on the total number of syllables produced and the number of pauses. Accordingly, MLR can tap simultaneously into both speed and breakdown fluency. The selection of such composite measures can affect the interpretability of results due to the conceptual collinearity among the selected measures (Bosker et al., 2013). For instance, in the research context where the temporal correlates of PF are examined, even if MLR is found as a significant predictor of PF, it is unclear which dimensions of UF—speed or breakdown fluency—contribute to listeners' perceptions of L2 fluency (see also Section 3.4).

According to Segalowitz (2010), UF refers to the observable temporal features of speech which can represent the efficiency of the speaker's L2 system (i.e., CF). Following the conceptualization of CF as part of L2 proficiency, one can assume that UF should indicate some aspects of L2 proficiency (cf. Baker-Smemoe et al., 2014; Iwashita et al., 2008; Tavakoli et al., 2020). Similarly, L2 performance research has also considered the validity of UF measures from the perspective of the extent to which different UF measures can capture

some aspects of L2 proficiency and/or developmental changes (cf. Lambert & Kormos, 2014).

3.3 Perceived Fluency

According to Segalowitz (2010), PF (perceived fluency) is defined as the listener's inference about the speaker's efficiency in speech production (i.e., CF) based on the listener's perception of utterance features of speech (i.e., UF). In other words, PF is conceptualized as the listeners' intuitive judgement of CF. Although the concept of PF was established by Segalowitz in 2010, listener-based judgements of L2 fluency had been extensively examined even before Segalowitz (2010). A body of prior work has measured PF either by instructing listeners to focus on temporal characteristics of speech or by providing no definitions to allow them to intuitively judge speakers' fluency. Depending on the instructions of the researchers, listeners' perception of fluency can thus be established either based on their perception of speech characteristics, including temporal and non-temporal features (e.g., grammatical errors), or their intuitive judgements of the speaker's efficiency in speech production (i.e., CF). However, Segalowitz (2010) claims that in the research context of speech judgement tasks, listeners naturally tend to engage with the latter scenario, that is, they make "subjective judgements of L2 speakers' oral fluency" (Segalowitz, 2016, p. 86). Segalowitz's (2010, 2016) definition of PF explicitly points out the association with CF. More specifically, listeners inherently make an inference about how efficiently the speaker is able to produce their speech in the target language, based on their perception of utterance features. Listeners are assumed to establish their PF judgements by paying attention to particular utterance features that they believe reflect the speaker's CF rather than to all the different kinds of utterance features (Segalowitz, 2010). This assumption has been supported by the differential weights of the predictive power of temporal features for listener-based judgements (see Section 3.4).

Another issue relating to the construct definition of PF lies in variations in the scope of fluency. It has been suggested that people in general tend to regard fluency as overall L2 proficiency, while the research-informed definition of fluency refers exclusively to temporal aspects of speech (Tavakoli & Hunter, 2018). The former and the latter have been respectively termed *higher-order fluency* or fluency in a broad sense; and *lower-order fluency* or fluency in a narrow sense (Lennon, 1990, 2000). Therefore, depending on how the definition of fluency is specified by researchers, the construct of PF can be different in its scope.

3.4 The Utterance-Perceived Fluency Connection

Building on the preceding theoretical background of PF and UF, L2 fluency research has investigated which temporal features of utterances can explain listeners' perception of fluency. Previous studies have commonly shown that PF is primarily associated with speed and breakdown fluency and also, to a lesser degree, with repair fluency (for a similar review, see Saito et al., 2018; S Suzuki & Kormos, 2020). Although previous studies tended to report that a large amount of the variance in PF scores was explained by a set of UF measures, there remains variability in the amount of variance explained across studies (e.g., $R^2 = 0.84$ in Bosker et al., 2013 vs. $R^2 = 0.57$ in Saito et al., 2018). Therefore, it seems plausible to argue that the connection between UF and PF is affected by methodological differences across studies.

In addition to the amount of explained variance of PF scores, there are several inconsistent findings regarding the UF-PF link. First, comparing speed and breakdown fluency measures, some studies showed that speed fluency measures had higher correlation coefficients with PF

scores than breakdown fluency measures (Bosker et al., 2013; Kormos & Dénes, 2004). However, other studies, especially which considered pause location, reported that breakdown fluency measures correlated with PF scores more strongly than speed fluency measures (Cucchiarini et al., 2002; S. Suzuki & Kormos, 2020). These contradictory findings may indicate that mid-clause pause measures tend to have a strong predictive power for PF judgements. In other words, distinguishing mid-clause pauses from end-clause pauses, scholars have developed more valid breakdown fluency measures in terms of the predictive validity for PF. Such an advantage of mid-clause pause measures can also be found in the context of L2 listeners' judgements of PF (Magne et al., 2019). Moreover, relating to pause location, pause type—silent versus filled pauses—may affect the predictive power of breakdown fluency for PF scores. In general, the measures based on silent pauses tended to correlate with PF scores more strongly than those based on filled pauses (Bosker et al., 2013; Cucchiarini et al., 2002; S. Suzuki & Kormos, 2020).

Second, another inconsistent finding in L2 fluency research is the role of repair fluency in PF judgements. Since repair fluency entails a range of disfluency phenomena, the selection of target disfluency phenomena has varied across previous studies. Considering the difficulty of validly categorizing disfluency phenomena into different sub-groups without the speakers' own retrospective accounts (Kormos, 1999b, 1999a), some studies simply counted different disfluency phenomena altogether (e.g., disfluency rate), but tended to find non-significant correlation between the repair fluency measures and PF scores (e.g., Cucchiarini et al., 2002; Kormos & Dénes, 2004; S. Suzuki & Kormos, 2020). Meanwhile, scholars have also used some repair fluency measures with the particular focus on specific disfluency phenomena, such as self-repetitions and self-corrections. These measures were found to significantly correlate with PF scores in some studies (e.g., Bosker et al., 2013), but not in other studies

(Magne et al., 2019; Saito et al., 2018). Although these quantitative investigations have provided mixed findings on the predictive power of repair fluency for PF ratings, qualitative findings suggested that disfluency phenomena affected listeners' perception of L2 fluency (Magne et al., 2019; Préfontaine & Kormos, 2016; Rossiter, 2009).

Third, as mentioned previously, composite measures, such as speech rate and MLR (mean length of run), can capture multiple dimensions of UF and thus tend to strongly correlate with PF scores (Derwing et al., 2009; Kormos & Dénes, 2004; Préfontaine et al., 2016; Rossiter, 2009). Despite such a strong predictive power for PF, it is not always appropriate to select those composite measures especially when research aims to use multiple UF measures to predict PF scores (Bosker et al., 2013). From a statistical perspective, such a problem is regarded as a multicollinearity issue (Plonsky & Oswald, 2016). Even from a theoretical perspective, these composite measures make it difficult to interpret the findings, because it is unclear which temporal features a given composite measure represents (e.g., speed vs. pause frequency for MLR).

3.5 Moderator Variables of the Utterance-Perceived Fluency Link

The previous section suggests that methodological differences across studies may contribute to inconsistent results regarding the relationship between UF and PF. As illustrated in Figure 5, research into the UF-PF link involves five major methodological phases, each of which entails a set of methodological decisions. Since Study 1 focuses on correlation coefficients (see Chapter 4) as a target type of effect sizes, this section introduces potential moderator variables for the first four methodological phases with regard to relevant previous studies.

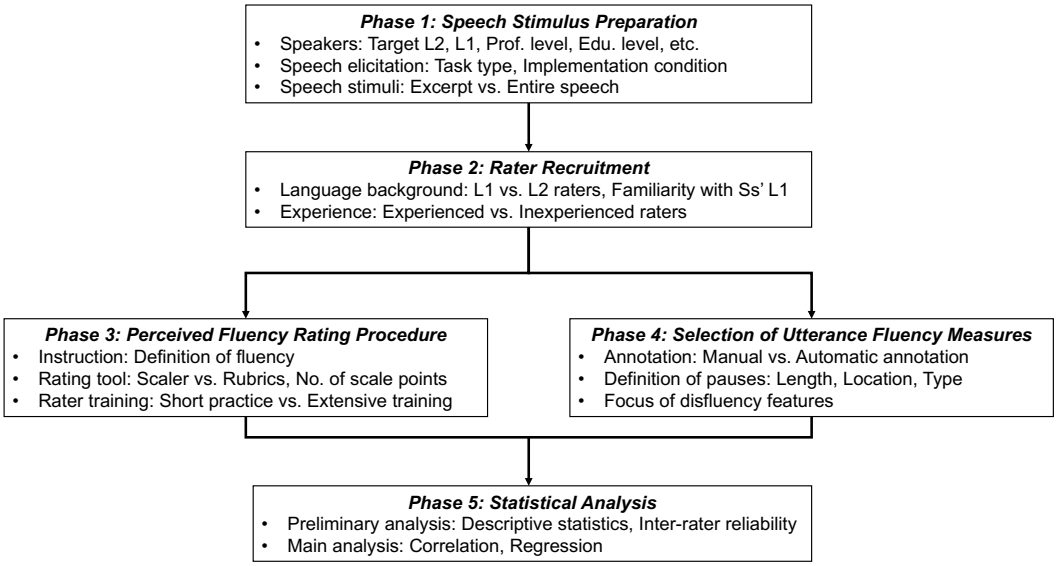


Figure 5. Five major phases in L2 research into utterance-perceived fluency link.

3.5.1 Speech stimulus preparation

The first phase of L2 PF research is the preparation of speech stimuli for PF judgements. Unless researchers use existing dataset or corpora, they need to begin with the collection of speech data. First, researchers specify the target population of speakers in terms of L1, L2, proficiency levels, age, and so on. Second, researchers determine speech elicitation methods, such as speaking task type and condition. Task effects on the UF-PF connection have been rarely examined within single studies. As an exception, Cucchiarini et al. (2002) compared the UF-PF link between controlled and spontaneous speech production (read-aloud task vs. opinion-giving speech), showing that the correlation coefficients between PF scores and various UF measures were overall higher in controlled speech than in spontaneous speech. Similarly, Préfontaine et al. (2016) contrasted fluency measures between three tasks which differed in the extent to which task prompts predefined the content of speech (related and unrelated picture narrative tasks, text retelling task). Using mixed-effects regression modelling, they reported that the relative magnitudes of regression coefficients among UF measures varied across tasks. In addition, L2 fluency research has also been recently extended to dialogic speaking tasks, showing that dialogic fluency is theoretically distinctive

from monologic fluency (Tavakoli, 2016; van Os et al., 2020). Possibly due to the budding phase of dialogic fluency research, prior research has tended to use different methodologies to examine the UF-PF link in dialogic speech, such as the selection of UF measures (Ahmadi & Sadeghi, 2016; Sato, 2014).

After collecting speech data, researchers need to decide whether speech stimuli are presented to raters in the form of the entire speech sample or just as its short excerpt. Some scholars claim that short excerpts of speech (e.g., initial 30 seconds) are sufficient to elicit listener perception data in research contexts (Derwing et al., 2006, 2009), whereas some studies have presented entire speech as speech stimuli, emphasising the ecological validity of findings in language assessment contexts (e.g., Préfontaine et al., 2016; S. Suzuki & Kormos, 2020). However, it has been unclear how the length of speech stimuli affects the connection between utterance features and listeners' perception of L2 fluency.

3.5.2 Rater recruitment

The second phase of L2 UF-PF link research is the recruitment of listeners for PF judgements. One of the relevant listener background factors may be the language background of raters. Although few in number, previous studies have examined the effects of language background within a single study (Rossiter, 2009) or between studies (Magne et al., 2019; Saito et al., 2018), commonly reporting that listeners' perceptions of fluency tend to be largely similar but more or less different between L1 and L2 raters. Moreover, prior research has been aware of the potential effects of raters' experience, such as experience as examiners for high-stakes tests, teaching experiences, and expertise in linguistics, on listener-based judgements of speech (Isaacs & Thomson, 2013). However, the effects of listener

background factors on the UF-PF link has not yet been extensively examined in L2 fluency research (for rare exceptions, see Rossiter, 2009; Saito, Trofimovich, & Isaacs, 2017).

3.5.3 Rating procedure

The third phase of L2 UF-PF link research is the actual implementation of PF ratings. As mentioned previously, the construct of PF can vary, in its range of scope. Thus, depending on the focus of research, previous studies either instructed their listeners to focus narrowly on temporal aspects of speech (i.e., lower-order fluency; e.g., Bosker et al., 2013) or provided no definition to allow for their intuitive judgements of fluency (higher-order fluency; e.g., S. Suzuki & Kormos, 2020). In the former case, most studies presented the definition of fluency based on research findings, while some studies employed existing assessment tools, such as the CEFR assessment scale (Préfontaine et al., 2016) or even created rubrics for their research purposes (Sato, 2014). In the case of ratings without definitions, listeners may regard the concept of oral fluency as an equivalent of overall proficiency beyond temporal performance (Tavakoli & Hunter, 2018). Accordingly, studies providing no definition for listeners tend to employ a range of linguistic predictor measures covering non-temporal features, such as grammatical errors and lexical diversity (e.g., S. Suzuki & Kormos, 2020). Although scholars have theorized that the different definitions of fluency may differentiate raters' judgements of fluency, this issue has scarcely been investigated within a single study (cf. Dressler & O'Brien, 2019).

When it comes to rating scales, there is huge variation in the number of scale points. Isaacs and Thomson (2013) found that five- and nine-point scales did not significantly differ in the severity of ratings, but their Rasch analysis revealed that the distinguishability of adjacent levels on scales was more meaningful on the five-point scale than on the nine-point scale.

Although the advantage of the five-point scale was statistically supported, their qualitative data indicated that the number of levels on the five-point scale could have led to difficulty with judging medium-level performance precisely (e.g., the score of 3 vs. 4 on the five-point scale).

Similarly, researchers also need to decide on the amount of practice that raters will have before the main rating task. Most previous studies have asked their raters to judge several speech samples to familiarize them with the use of rating scales. On the other hand, especially when using rubrics for fluency judgements, careful training (e.g., feedback and discussion among raters) is often provided for their raters to ensure that they assign the same meaning to scores on the scale (e.g., Doe, 2017; Sato, 2014). However, the extent to which such careful training affects the association between utterance features and fluency judgements is still unclear.

3.5.4 Selection of utterance fluency measures

The fourth phase of L2 UF-PF link research is the computation of UF measures by annotating temporal features of speech samples, which entails several methodological decisions. First and foremost, researchers need to select UF measures to predict PF ratings. Although the selection of UF measures is dependent on the focus of research including the scope of PF (e.g., higher- vs. lower-order fluency), researchers are recommended to ensure the construct validity of measures selected (Lambert & Kormos, 2014), the comparability of the measures with previous studies (Michel, 2017), and the intercollinearity among the measures (Bosker et al., 2013). Note that research focusing on higher-order fluency tends to employ linguistic measures in addition to UF measures, such as grammatical errors and lexical diversity (e.g., Kormos & Dénes, 2004; S. Suzuki & Kormos, 2020). Building on the selection of UF

measures, several methodological decisions have to be made. Researchers might also decide to annotate speech samples either manually or automatically. In the case of manual annotation of speech, scholars commonly use some assistance of acoustic analysis software such as Praat (Boersma & Weenink, 2012). On the other hand, scholars can entirely rely on the software for automatic speech annotation, such as scripts in Praat (e.g., De Jong & Wempe, 2009) and the continuous speech recognizer (Strik, Russel, Van Den Heuvel, Cucchiarini, & Boves, 1997).

Second, in order to annotate speech samples either manually or automatically, researchers need to specify the temporal features relevant to selected UF measures. Among various disfluency features, researchers need to carefully specify the definition of silent pauses. L2 fluency research has been engaged in a long debate over the threshold for silent pauses. Although De Jong and Bosker (2013) suggested the optimal minimum length of silent pauses as 250 ms based on their simulation data, a body of research, especially before De Jong and Bosker (2013), employed the different thresholds for silent pauses (e.g., 200 ms, 400 ms). Besides, some studies set the maximum length of pauses (e.g., 3000 ms, Kormos & Dénes, 2004) to avoid breakdowns due to non-linguistic processing.

Moreover, in response to the multidimensional nature of pausing behaviour, scholars have recently distinguished pauses based on their location. This methodological trend is motivated by the theoretical relationship between pause location and underlying cognitive processes (see Section 3.2). However, it has not been directly examined to what extent the predictive power of pause-related measures for PF ratings may vary, depending on whether pauses are distinguished by location. Another methodological issue around breakdown fluency measures is the distinction between silent and filled pauses. Some studies counted silent and filled

pauses separately (e.g., Bosker et al., 2013; S. Suzuki & Kormos, 2020), but others did not make a distinction between them (e.g., Trofimovich et al., 2017). However, it is still unclear the extent to which pause type (silent vs. filled pauses) differentiates the predictive power of pause-related measures for PF judgements.

Regarding repair fluency measures, scholars have focused on different disfluency features, possibly due to various types of disfluency features as well as the substantive difficulty with their reliable categorization (cf. Kormos, 1999b). It is thus unclear how the association between PF scores and repair fluency measures differs, depending on the disfluency features in focus.

3.6 Cognitive Fluency

CF (cognitive fluency) refers to speakers' ability to produce fluent speech in L2 and thus is concerned with how efficiently L2 speakers operate their speech production mechanisms (Segalowitz, 2010). Specifically, CF entails both general cognitive control capacities and L2-specific cognitive systems (Segalowitz, 2016). General cognitive control capacities include a variety of language-general cognitive capacities and processes, and among them, working memory capacity and attention control may play a central role in L2 speech production (Segalowitz, 2010; Skehan, 2014b). These are supposed to be independent of L2-specific competence, as they are shared across different language systems of individuals. Meanwhile, L2-specific cognitive systems consist of a range of linguistic encoding processes at different linguistic levels including lexis, morphosyntax, and pronunciation (Kormos, 2006; Segalowitz, 2010; see also Section 2.6). Therefore, the valid operationalization of CF should correspond to different cognitive/linguistic processes involved in L2 speech production.

L2 speech production models (e.g., de Bot, 1992; Kormos, 2006; Segalowitz, 2010), all of which are based on Levelt (1989, 1999), commonly assume that L2 speech production entails three major phases—*conceptualization*, *formulation*, and *articulation*—which are executed serially in this order. In addition to these major processes of speech production, the *self-monitoring* function examines the interim and eventual outcomes of the preceding processes in terms of content appropriacy and linguistic correctness (see Chapter 2). Among them, conceptualization is theoretically independent of L2-specific proficiency, because conceptualization is responsible for the manipulation of conceptual information prior to linguistic encoding processes. Thus, one can argue that conceptualization should be related to the general cognitive category of CF due to its non-linguistic nature. On the other hand, formulation and articulation are categorized as components of L2-specific CF (Kahng, 2020; Segalowitz, 2016). Regarding formulation, there are four major linguistic levels of encoding modules involved in formulation: lexical, syntactic, morphological, and phonological levels. Articulation should be regarded as another distinct component of CF due to the fact that it only involves motor skill processing (for details, see Kormos, 2006; see also Section 2.6.4). Meanwhile, the self-monitoring function may tap into both general cognitive and L2-specific aspects of CF, because it is driven by either conceptual (e.g., illogical connection of the utterance to the previous utterance) or linguistic problems (e.g., inappropriate lexical choice) identified in the course of speech production. Building on the notion of L2-specific CF (Kahng, 2020; Segalowitz, 2016), CF measures are supposed to tap into formulation and articulation processes and some linguistic aspects of self-monitoring.

Previous studies on CF have used both broad and narrow definitions of L2-specific CF. In a narrow sense, in accordance with Segalowitz's (2016) original conceptualization, CF refers to the speed/efficiency aspects of linguistic encoding processes. In a broad sense, often adopted

in empirical studies (e.g., De Jong et al., 2013; Kahng, 2020), CF may include linguistic knowledge resources as well as the speed of processing. For instance, lexical processing in L2 speech production is related to the range of available lexical resources (i.e., vocabulary size) as well as the speed of lexical retrieval (i.e., lexical fluency) (see Kormos, 2006). In a narrow sense, only lexical fluency is regarded as a lexical component of CF. On the other hand, following the broad definition of CF, both vocabulary size and lexical fluency should be included in the lexical component of CF. From a theoretical perspective, both linguistic resources and processing speed are interconnected. L2 speakers may sometimes need to express their intended message, even though they have not acquired the corresponding linguistic items. It can happen that L2 speakers need to reconceptualize the intended message in such a way that they can convey their thoughts using their own linguistic resources. In this case, there might be disruptions in the flow of speech processing, and the reformulation of the message can be observed as pauses in the utterance. Such pausing behavior is one of the essential aspects of UF (Tavakoli & Skehan, 2005; see also Section 3.2). According to Segalowitz's (2010, 2016) framework, CF is conceptualized as a set of components of L2 speech production mechanisms which can explain observable temporal features of utterances (i.e., UF). Following this conceptualization of CF, the valid operationalization of CF may involve both linguistic resources and processing speed, because pausing behavior in L2 speech may be caused by either the lack of linguistic resources or slow processing speed. In other words, the temporal nature of pausing behavior, such as the frequency and duration of pauses, at least theoretically, can capture both the availability of linguistic resources and the speed of linguistic processing. Therefore, this thesis follows the broad definition of CF and subsequently operationalizes CF as linguistic resources and processing efficiency at the level of vocabulary, grammar and pronunciation (see Chapter 6; for a similar methodological decision, see also De Jong et al., 2013; Kahng, 2020). Although the potential components of

CF have been theoretically specified, to the best of my knowledge, the structure and dimensionality of these CF components have not yet been examined. Therefore, Study 3 in the thesis aims to test different factor structures of L2-speific CF, using a range of CF measures capturing the abovementioned components.

Finally, another theoretical issue around the conceptualization of CF is the distinguishability of CF from UF (Segalowitz, 2010, 2016). Although both subconstructs of fluency are closely related, they are supposed to be operationalized and measured at different levels, that is, the CF at the level of underlying cognitive processes and UF at the level of observable spontaneous speech. This distinction between CF and UF can be supported by the multi-componential nature of L2 speech production (De Jong et al., 2013). Each process of speech production, such as lexical retrieval and syntactic procedure, is assumed to proceed serially; thus, each process is responsible for different aspects of speech production but is simultaneously related to each other (Kormos, 2006). In other words, actual spontaneous speech (i.e., UF) can be regarded as an eventual outcome of various components of speech processing. For the sake of valid assessment of L2 speech production mechanisms (i.e., CF), CF components should be assessed individually while controlling for the effects of other components.

3.7 The Cognitive-Utterance Fluency Connection

According to Segalowitz's (2010, 2016) framework, temporal performance of speech (i.e., UF) is assumed to be achieved by the speaker's ability to mobilize linguistic resources and processing skills (i.e., CF). Although few in number, previous studies have examined what cognitive and linguistic processes underly L2 UF. Even before Segalowitz (2010), Segalowitz and Freed's (2004) pioneering study investigated the role of L2-specific cognitive ability in

L2 oral fluency in the context of English-speaking learners of Spanish (N = 40). Using a semantic classification task and a repeat-and-shift task in both L1 and L2, they computed L2-specific cognitive measures for lexical access and attention control in terms of both general speed and stability of processing speed (i.e., so-called the coefficient of variance [CV] index) by partialing out the corresponding L1 measures. They found that mean length of run without fillers in L2 speech was supportively associated with both speed and stability of lexical access. Meanwhile, their correlational analyses also suggested that L2 speech rate was negatively associated with the measure of the processing stability of attention control, contrary to their expectation that efficient cognitive processing contributes to oral fluency. Despite the narrow range of cognitive processing measures, these findings confirmed the role of cognitive ability in L2 UF.

Building on Segalowitz's (2010) framework of oral fluency, De Jong et al. (2013) employed a range of linguistic knowledge and processing measures to predict different UF measures. Their data were collected from a total of 179 learners of L2 Dutch from various L1 backgrounds. Their CF measures tapped into lexical (vocabulary size, lexical retrieval speed), grammatical (grammatical knowledge, sentence construction speed), and pronunciation (phonetic accuracy, articulatory speed) knowledge. Their UF measures covered speed, breakdown, and repair fluency. A series of correlational analyses showed that the relevant components of CF varied across UF measures. For instance, mean syllable duration (the inversed measure of articulation rate; speed fluency) was correlated with a whole range of CF measures including vocabulary, grammar, and pronunciation. Meanwhile, breakdown fluency measures were related to more specific dimensions of CF; mean duration of pauses was significantly but weakly correlated only with lexical retrieval speed. Moreover, both silent and filled pause ratio measures were mainly correlated with lexicogrammatical knowledge

and processing. In addition, their linear mixed-effects models revealed that speaking task type moderated the strengths of the relationships between CF and UF measures. These findings showed that different aspects of UF may represent different components of CF and also that the CF-UF connection can be strengthened or weakened, depending on speaking task design.

Similarly, Leonard and Shea (2017), as part of their longitudinal study with English-speaking learners of Spanish ($N = 39$), correlated the composite score of UF with vocabulary and grammatical resource and processing measures. Their regression model predicting the fluency score included only lexical and grammatical processing speed measures based on the reaction time (RT) in their picture naming and sentence-picture verification tasks. Their finding shows the strong explanatory power of processing speed measures for UF, compared to linguistic resource measures, in accordance with Segalowitz's (2016) narrow conceptualization of CF.

Assuming that L1 speakers have a more efficient language system than L2 speakers, Kahng (2014) compared various UF measures between L1 speakers of English and Korean-speaking learners of L2 English. The results showed that L1 and L2 speakers differed in speed (mean syllable duration), breakdown (mid- and end-clause pause frequency), and repair fluency (self-correction frequency). Focusing on L2 speakers, she found that all of these L2 UF measures, except for self-correction frequency, were also significantly correlated with a proficiency test score as a proxy for CF. In addition, drawing on the stimulated recall data about L2 speakers' cognitive speech processes, Kahng (2014) found that the primary causes of pauses might differ as a function of L2 proficiency. High proficiency learners were likely to encounter breakdowns due to content and pragmatic aspects, while low proficiency

learners tended to produce pauses due to retrieval problems with vocabulary and grammar. This suggests that, from the perspective of speakers' perceptions, overall proficiency might moderate the relationship between CF and UF.

Finally, Kahng (2020) examined the predictive power of CF measures for UF measures, using a personal narrative task with Chinese learners of English ($N = 44$). Uniquely, Kahng (2020) included the corresponding L1 UF measures as another predictor variable. In her study, CF measures covered vocabulary size for single words and phrases, lexical retrieval speed, grammatical resources and processing speed, and articulatory speed, largely following De Jong et al. (2013). The results of stepwise multiple regression analyses indicated three major findings. First, although mean syllable duration (speed fluency) and mid-clause pause ratio (breakdown fluency) correlated with both lexical and syntactic measures of CF, different CF measures were identified as predictor variables in the regression models. Mean syllable duration was predicted from lexical measures of CF (lexical retrieval speed, phrasal vocabulary size), while mid-clause pause ratio was predicted from the measure of syntactic processing speed. This finding thus indicated that the primary component of CF can be different across dimensions of UF. Second, the regression models of mid-clause pause ratio and self-correction rate did not include the corresponding L1 UF measures as predictor variables. This finding indicates that pauses in the middle of clauses and self-repair are reflective of L2-specific processing. Third, the strongest predictors in the regression models of mean pause duration and filled pause ratio were their corresponding L1 UF measures, suggesting that the length of silent pauses and the frequency of filled pauses are more closely related to speakers' language-general idiosyncratic factors than to L2-specific CF.

Taken together, previous studies suggested two common patterns of the CF-UF link. First, different components of CF can be associated with different dimensions of UF to a varying degree. Therefore, for a better understanding of the CF-UF link, it is essential to consider the dimensionality of CF and UF. Second, the association between CF and UF can vary, depending on speaking task design (De Jong et al., 2013). However, it is still unclear what task design features moderate the CF-UF link, because in their study, De Jong et al. (2013) handled speaking task as a random-effects predictor in their regression models. Finally, previous studies employed similar but different measurements of CF and analyzed the measured scores only at the level of observed variables. It can thus be hypothesized that the findings in previous studies may entail some measurement errors. Accordingly, L2 CF-UF link research may be extended by examining the relationship between CF and UF at the level of constructs (statistically speaking, latent variables) using SEM (structural equation modelling).

3.8 Speech Processing Demands as a Framework for Moderating Task Effects on the Cognitive-Utterance Fluency Link

As mentioned in the previous section, the association between CF and UF can vary across different tasks (De Jong et al., 2013). In this section, I further theorize how task design features affect speech production processes and UF performance, drawing on speech processing demands as a theoretical framework of task effects on oral fluency.

From a theoretical perspective, speech production proceeds by consuming attentional resources, because it entails various types of information processing such as the retrieval of schematic and linguistic knowledge (Baddeley, 2003). In addition, it is assumed that the attentional resources available for spontaneous L2 speech production are limited due to the

partially automatized status of L2 knowledge (de Bot, 1992; Kormos, 2006; Skehan, 2014b). Accordingly, L2 learners need to distribute their limited amount of attentional resources to different speech processing phases, such as conceptualization and formulation (cf. Limited attentional capacity model; Skehan, 2014a). Since each phase of speech production is responsible for different aspects of speech, L2 speaking performance can be affected by how L2 learners distribute their attentional resources (Foster & Tavakoli, 2009; Skehan, 2009, 2014b). For instance, conceptualization is responsible for content planning (Kormos, 2006; Levelt, 1989, 1999). If L2 learners are required to elaborate the content of their speech in detail and to pay close attention to the organization of their speech, they might spend a large amount of attentional resources on conceptualization. However, due to the high demands on conceptualization, they might devote only a limited amount of attentional resources to the subsequent formulation processes. Consequently, their flow of speech processing at the formulation stage is likely to be disrupted, which can be observed as breakdowns in their speech (Foster & Tavakoli, 2009; Skehan, 2009, 2014b).

Although task design may generally affect the amount of attentional resources available for different aspects of speech processing, L2 learners differ in the efficiency in using the given amount of attentional resources for L2-specific processing. The more automatized their L2 processing is, the more processes L2 learners can carry out with the same amount of attentional resources (Skehan, 2009, 2014b). In other words, learners' automaticity in L2 processing might modulate the effects of speech processing demands on their speaking performance. Automaticity in L2 processing is substantively equivalent to Segalowitz's (2010) conception of L2-specific CF. The process of task effects on L2 UF is visualized in Figure 6 below. From the observable level (right side) to the speaker-internal level (left side), a speaking task first poses different processing demands on each stage of speech production

(arrow A). In response to the given task, speakers use their attentional resources at each stage of speech production in a serial manner (allow B). Depending on the demands on conceptualization, the amount of attentional resources available for L2-specific CF (i.e., formulation and articulation) varies (arrow C). Finally, speakers' automaticity as well as the amount of attentional resources available can determine the overall smoothness of speech processing (arrow D), which is observed as UF of the speech produced (arrow E).

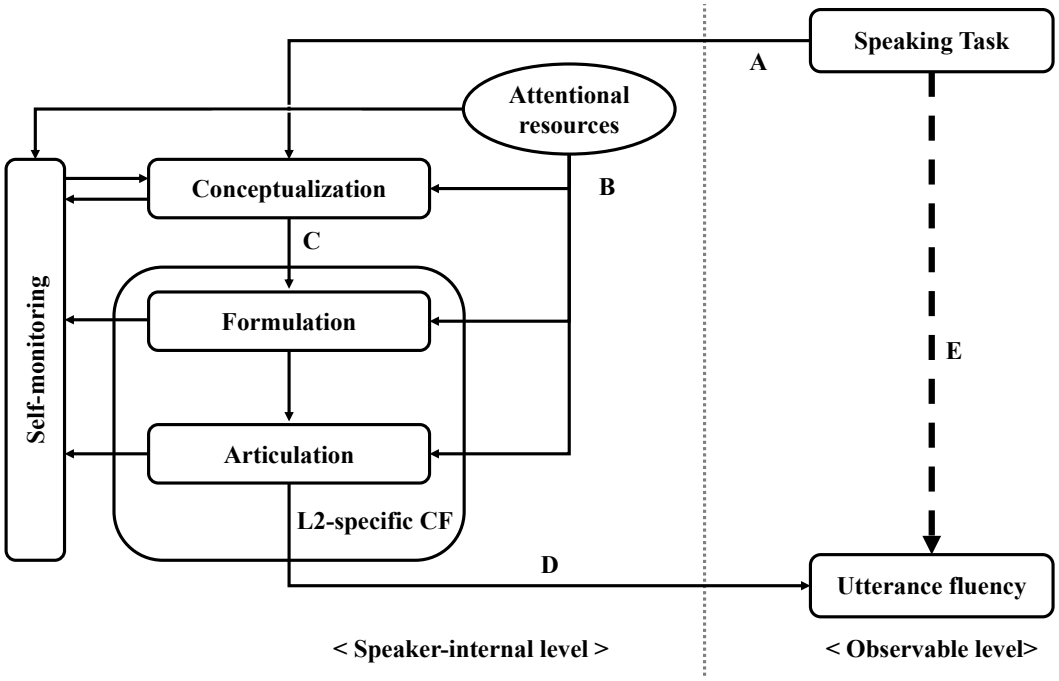


Figure 6. A visual representation of task effects on L2 utterance fluency (on the basis of Kormos, 2006; Segalowitz, 2010; Skehan, 2014).

Building on these theoretical assumptions, scholars have proposed that different speaking tasks might result in different constraints on L2 learners' distribution of attentional resources (Skehan, 2009, 2014b). Task design can thus influence processing demands on different aspects of speech production mechanisms. Therefore, the framework of speech processing demands has attempted to explain the intra-individual variability in speaking performance across tasks (e.g., Préfontaine & Kormos, 2015; Skehan, 2009). In the literature of L2 speech processing demands, previous studies have focused on the effects of manipulating

conceptualizing demands, using either of three major approaches—the tightness of task structure, the requirement of content generation, and the necessity to choose an option from alternative choices.

The first type of the manipulation of conceptualizing demands is the adjustment of the tightness of task structure (Foster & Tavakoli, 2009; Tavakoli & Foster, 2008; Tavakoli & Skehan, 2005). The task structure refers to the extent to which task design provides a clear macrostructure of the speech for speakers to achieve the given task requirements. For instance, in the case of picture narrative tasks, the time sequences across scenes and the conventional storyline development can provide learners with a clear macrostructure of their speaking performance (Tavakoli & Skehan, 2005). Accordingly, with the assistance of the macrostructure of the message, conceptualization processes can be accomplished with a small amount of attentional resources. Tightly structured tasks would save relatively large amount of attentional resources for formulation and articulation processes, subsequently enhancing learners' UF performance (Tavakoli & Skehan, 2005). In accordance with this assumption, previous studies have found that the fluency of performance in structured conditions of picture narrative tasks was enhanced (Tavakoli & Foster, 2008; Tavakoli & Skehan, 2005). Tavakoli and Skehan (2005) found that the speech in their structured picture narrative tasks was characterized by longer MLR (mean length of run) and fewer pauses. The similar pattern of MLR and pause frequency can be mathematically expected, because the number of pauses *plus one* (i.e., the number of runs segmented by pauses) makes up the denominator in the formula for MLR. Accordingly, the similar pattern of these two measures confirmed that the reduced conceptualizing demands may reduce breakdowns in L2 speech production. Similarly, Tavakoli and Foster (2008) reported that EFL learners (Iranian learners of English) produced fewer mid-clause pauses and fewer false starts in the tightly structured version of

picture narrative tasks than in the loosely structured ones. The reduction of conceptualizing demands may thus lead to the reduction of breakdowns specific to L2 formulation processing. Their follow-up study also found that L1 English speakers' fluency did not vary according to the different degree of task structure (Foster &Tavakoli, 2009). Considering the language-general nature of conceptualization processes, the manipulation of conceptualizing demands by making changes in task structure may not be sufficient to affect UF performance of speakers with highly automatized linguistic knowledge (e.g., L1 speakers). However, in the case of L2 learners, structured tasks may free up attentional resources for formulation processes and thus reduce breakdown fluency (Tavakoli & Skehan, 2005).

Another approach to manipulating the demands on conceptualization is the degree of necessity for speakers to plan the content of speech (i.e., content generation; Préfontaine and Kormos, 2015). Préfontaine and Kormos (2015) employed three speaking tasks which differed in the quality of speech processing demands—unrelated and related picture narrative tasks and a text retelling task. Among the three tasks, the effects of conceptualization demands on L2 fluency were examined by comparing the unrelated and related picture narrative tasks. The unrelated picture narrative task required L2 speakers to create a storyline from six unrelated pictures, whereas the related picture narrative task provided a predefined sequence of events with a 11-frame cartoon. The results showed that learners' speech in the unrelated picture narrative task was characterized by lower articulation rate (speed fluency) but shorter silent pauses (breakdown fluency), compared to the related picture narrative task. Their study partially confirmed that the enhanced demands on conceptualization can have a negative impact on speed fluency. However, shorter silent pauses in the unrelated picture narrative task indicate the positive impact of conceptualizing demands on breakdown fluency. Interestingly, the articulation rate in the related picture narrative task was also

significantly higher than in the text retelling task where learners read the source text in their L1 and narrated it in L2. This result suggests that even though learners were commonly provided with the predefined content for the speech through either visual prompts or L1 text, speed fluency was different between two tasks. One possible explanation for this might be that in the text retelling task, L1 lemmas from the source text and L2 lemmas retrieved for speech might have competed for selection, consequently slowing down L2 speech processing. In addition, Préfontaine and Kormos (2015) cautioned that content generation can theoretically enhance UF performance; the higher demand of content generation in open tasks may allow learners to avoid using difficult linguistic items by modifying their intended message. In contrast, closed task, such as related picture narrative tasks, may force learners to express key information to describe the storyline. If learners have not acquired the lexical items corresponding to key information to be conveyed, they may experience breakdown in speech production. This might explain the shorter pause duration in the unrelated picture narrative task than in the related one. To sum up, Préfontaine and Kormos' (2015) results supported that the manipulation of conceptualizing demands by content generation might affect the overall efficiency of speech production as observed in speed fluency of performance. The predefined content of speech may also increase the demands on formulation, leading to longer pauses in the utterances. Accordingly, in a related picture narrative task, its predefined content can reduce the demands on conceptualization and simultaneously increase the demands on formulation. The former can enhance UF performance, while the latter can lead to less fluent speech, indicating that related picture narrative tasks can positively and negatively affect UF performance.

Reviewing two approaches to manipulating the demands on conceptualization, one can argue that the manipulation of conceptualizing demands would entail the confounding influence on

the demands on formulation, due to the serial nature of speech production. However, the final approach—the alternative choice and its online change—can manipulate the demands on conceptualization, while controlling for the demands on formulation, albeit specific to one particular type of task, that is, network description task (Felker et al., 2019). The network description task is one type of picture naming task where participants are presented with a network of objects linked via paths and are asked to describe the route of the paths connecting the objects highlighted (e.g., Felker et al., 2019). It is assumed that the more potential alternative choices (i.e., distractors) participants are presented with, the more attentional resources they would need for the macroplanning stage of conceptualization. Accordingly, the easy and difficult conditions in Felker et al.'s (2019) study were manipulated by the number of distractors (one non-target object in the easy condition vs. two or three objects in the difficult condition) in their first experiment. In their follow-up experiment, the conceptualizing demands were further increased by changing the target picture and its path immediately after participants fixated at some picture stimuli for 500 ms. The assumption behind the enhancement of conceptualizing demands by changing paths is that the online changes in the target paths after eye-fixation started (as an indicator of the onset of speech planning) may force participants to revise their plan of the preverbal message, consuming a certain amount of attentional resources. The results of their first experiment showed that the increasing number of distractors did not affect UF features (e.g., silent and filled pause frequency, speech rate) except for syllable lengthening in both L1 and L2 production. In contrast, the results of the second experiment revealed that both L1 and L2 speech was less fluent in most UF features, including filled and silent pause frequency, syllable lengthening, speech rate, under the changing path condition than under the baseline condition (no online changes). In addition, their results also showed that the effects of conceptualizing demands on speech rate were larger in speakers' L2 than in their L1. These

findings suggest that the online change of paths (i.e., the revision of the preverbal message) may result in more enhanced conceptualizing demands than the increased number of alternative choices (i.e., the selection of information for speech). However, it should be noted that participants in Felker et al. (2019) were familiarized with all the picture names prior to the network description task. In addition, Felker et al. (2019) acknowledged that the output of the task was not syntactically complex, and the same syntactic structures might have been recycled across paths. Accordingly, UF features in the network task may largely reflect the demands on the retrieval of the phonological form and articulation of picture names. Therefore, despite the careful control of factors other than conceptualizing demands, the ecological validity of Felker et al.'s findings to spontaneous L2 speech might be questioned.

Previous studies have manipulated task design features of picture narrative tasks with regard to conceptualization processes. However, the validity of operationalizing conceptualizing demands can be further enhanced in accordance with speech production mechanisms. There are three major components of conceptualization—the specification of communicative intention, macroplanning, and microplanning (see Section 2.5). First, as reviewed earlier, conceptualization starts with specifying what communicative intention the speaker intends to achieve by producing speech. In the context of task-based performance research, the required communicative intention can be identical to task requirements (e.g., information coverage and precision in picture narrative tasks). Accordingly, task-relevant conceptualizing demands are likely to affect the smoothness of conceptualization processes and subsequently have an effect on fluency performance. Second, macroplanning is responsible for the generation and selection of information to be verbalized. In this sense, the presentation of distractors in network description tasks (Felker et al., 2019), content generation in unrelated picture narrative tasks (Préfontaine & Kormos, 2015), and loose narrative task structure (Tavakoli &

Foster, 2008; Tavakoli & Skehan, 2005) are reflective of macroplanning processes, because these task design features require learners to plan the potential pieces of information and to decide which information they would convey. In addition to the generation and selection of speech content, the literature of L1 speech production suggests that the increased number of ideational or discourse units, such as topic shifts and moves, may enhance demands on macroplanning (Greene & Cappella, 1986; Roberts & Kirsner, 2000). This is possibly because macroplanning is also responsible for guiding listeners' attention by signposting the organization of discourse (Levelt, 1989). Accordingly, the number of discourse transitions within speech might be another factor determining conceptualizing demands. Third, microplanning entails a range of information processing mechanisms, which transforms the conceptual specifications of the message into the language-specific form of proposition, including the specification of the referents and the mood of the planned message (for more details, see Section 2.5). Difficulty with microplanning, particularly if it is closely related to task requirements (e.g., the number of referents in picture narrative tasks), may thus lead to the enhancement of conceptualizing demands.

Considering these three major components of conceptualization, the theoretical validity of manipulating conceptualizing demands in picture narrative tasks can be discussed. First, the task requirement of narrative tasks can be regarded as identifying the characters and referring to them consistently, specifying the main events of the story and telling the events in a coherent manner (Luoma, 2004). Due to the support of the visual prompt, the identification of characters can be considered relatively easy, because the same characters usually appear throughout the scenes of the picture prompt (see Tavakoli & Foster, 2008). Even with some modifications in task structure, picture narrative tasks might thus provide relatively low conceptualizing demands. Considering the fact that the online changes of referents is specific

to network description tasks (i.e., low ecological validity to real-world speech production), content generation and the number of alternative choices might be feasible factors to consider when manipulating conceptualizing demands. Regarding content generation, open tasks, as opposed to closed tasks (e.g., picture narratives), require learners to plan a large part of speech content (Pallotti, 2009). Among different types of open tasks, a relatively larger number of discourse transitions can be expected in opinion-giving or argumentative tasks. The task requirement of argumentative tasks involves choosing and expressing an opinion on a given issue, contrasting it with other opinions, and discussing rationales and supporting information for the opinion (Luoma, 2004). Argumentative tasks thus require learners to engage with the selection of information among various alternatives and the organization of discourse so that listeners can understand how different pieces of information are connected (for a similar decision, see Lambert et al., 2017). In other words, the demands on macroplanning in argumentative tasks would be relatively high. Furthermore, the ecological validity of argumentative tasks can be considered relatively high, because due to the open-ended nature of the task, argumentative tasks can create a communicative situation where listeners do not have fully developed expectations about the speech content.

Reviewing the methodologies of manipulating conceptualizing demands, one can argue that research on L2 speech processing demands has focused mainly on conceptualization. Due to the serial nature of speech production, the amount of attention required at the conceptualization phase has an influence on the attentional resources available in subsequent processes, such as formulation (Skehan, 2009). Scholars have been interested in how such a shortage of attentional resources in formulation would affect UF performance, because this line of research would yield information on what temporal features are reflective of formulation processes, which have also been regarded as L2-specific CF in Segalowitz's

(2010, 2016) framework. However, enhanced conceptualizing demands may not always lead to the increased demands on formulation processes. For instance, the network description task with online changes of paths in Felker et al. (2019) imposed high demands on conceptualization, while the output of the task was not syntactically and lexically complex, meaning that the preverbal message should have been rather simple and thus linguistic demands for formulation should have been low. Therefore, in order to better understand the temporal or fluency features responsible for formulation processes, it might be argued that linguistic formulation demands should be manipulated in a direct manner. Drawing on the literature of speech production, one of the fundamental characteristics of formulation is activation spreading. When it comes to the retrieval of linguistic items for formulation, the higher the activation of the item is, the higher the probability that the item is selected is (see Section 2.2). One of the methodologies to operationalize this characteristic is the priming technique, which assumes that activating target items in advance would assist the retrieval of the items in the subsequent production (for a review, see McDonough & Trofimovich, 2008). For instance, in a broad sense, the reading-to-speaking tasks where learners read a source text and retell the content of the text can be regarded as one speaking task type in which the source text activates in-text lexical and grammatical items prior to speech production.

Finally, but not limited to, another methodological issue in the research into speech processing demands is the selection of UF measures. As argued previously, L2 fluency research has employed a different set of UF measures, depending on the research focus. For the purpose of understanding how speech processing demands are related to UF, fine-grained measures, which are supposed to tap into a particular speech processing phase, might be more interpretable, compared to composite measures.

3.9 The Association Between First and Second Language Utterance Fluency

Segalowitz (2010) emphasizes that L2 UF can be more or less explained by individual speakers' idiosyncratic factors, such as general cognitive skills and speaking style. Such idiosyncratic factors are assumed to play a similar role in L1 and L2 speech production (see Peltonen, 2018; Williams & Korko, 2019). Motivated by the potential role of idiosyncratic factors in L2 UF, previous studies have operationalized L2 learners' idiosyncratic pattern by the covariance between L1 and L2 UF and thus have examined the association between L1 and L2 UF performance (e.g., Bradlow et al., 2017; De Jong et al., 2015; De Jong & Mora, 2019; Peltonen, 2018). In terms of the validity of UF measures, UF measures are expected to reflect L2-specific CF (Segalowitz, 2010, 2016). The validity of UF measures can thus be discussed in terms of not only the association with PF and CF, but also the optimal independence from L1 fluency. Therefore, the understanding of the association between L1 and L2 UF measures may give insights into the selection of L2 UF measures for research and assessment purposes.

Despite the limited number of studies on the L1-L2 association in UF performance, scholars have identified two important factors affecting the L1-L2 UF link: cross-linguistic effects and the role of L2 proficiency (for a review, see Huensch & Tracy-Ventura, 2017). Regarding cross-linguistic effects, cross-linguistic similarities and differences are assumed to differentiate the strength of the L1-L2 UF association. Bradlow et al.'s (2017) cross-linguistic study confirmed that in speaking rate measures (speech rate and articulation rate), the predictive power of L1 UF for L2 UF was robust across different L1 backgrounds, while the strengths of the L1-L2 association tended to vary, according to speakers' L1. There are theoretically relevant linguistic features to temporal aspects of speech: the rhythmic pattern (e.g., stress-timed vs. syllable-timed language) and the syllable structure or complexity

(Pellegrino et al., 2011). Closely looking at the literature of the L1-L2 UF link, one can argue that while the most researched L2 is English, researched L1s were mainly either stress-timed languages, such as Slavic (Derwing et al., 2009), Swedish (Peltonen & Lintunen, 2016), or syllable-based languages, such as Finnish (Peltonen, 2018), Mandarin (Derwing et al., 2009), Spanish (De Jong & Mora, 2019), and Turkish (Duran-Karaoz & Tavakoli, 2020). However, to the best of my knowledge, there have been no studies examining the L1-L2 UF link in the context of learners of L2 English with mora-timed L1 language backgrounds. One of the representative mora-timed languages is Japanese (Vance, 2008). According to Pellegrino et al. (2011), the Japanese language has relatively low syllable complexity (indexed by the average number of constituents per syllable; 2.65 in Japanese vs. 3.70 in English and 3.87 in Mandarin) and information density (indexed by the average semantic information per syllable; 0.49 in Japanese vs. 0.91 in English and 0.94 in Mandarin). Therefore, for a better understanding of the cross-linguistic effects on the L1-L2 UF link, the association between L1 and L2 UF behaviours should be examined with Japanese-speaking learners of English.

Although the L1-L2 UF association was found to be cross-linguistically robust in speaking rate measures (Bradlow et al., 2017), the effect sizes in previous studies have not always been consistent, possibly because of the multidimensionality of UF (i.e., speed, breakdown, and repair fluency). The remaining part of this section summarizes previous studies about the L1-L2 UF link for each dimension of UF. First, speed fluency tends to show a relatively stable strength of the L1-L2 UF association. The correlation coefficients between L1 and L2 speed fluency measures are moderate (De Jong et al., 2015) or strong (De Jong & Mora, 2019; Huensch & Tracy-Ventura, 2017; Peltonen, 2018). Although most studies employed picture narrative tasks to elicit learners' speech, the strongest correlation was found in the role-play speaking task (r = .70; De Jong & Mora, 2019). The relatively stronger association between

L1 and L2 UF in role-play tasks than in picture narrative tasks might be expectable, because role-play tasks allow for a flexible storyline or speech content and thus provide more room for learners' speaking style to be reflected. In other words, learners' speaking style can be more apparent in open tasks than in closed tasks. However, some studies reported substantively no correlation between L1 and L2 articulation rate (r = .07; Duran-Karaoz & Tavakoli, 2020). One of the possible reasons for the non-significant corelation in their study might be the use of unpruned transcription to calculate the measure of articulation rate. The articulation rate measure based on unpruned transcription included disfluency words (e.g., repeated or modified words), which are theoretically more characteristic of L2 speech than of L1 speech. As a result, the inclusion of such disfluency phenomena might have obscured the L1-L2 link in Duran-Karaoz and Tavakoli (2020). Furthermore, another important finding regarding the L1-L2 speed fluency association is that longitudinal development may increase the strength of the association (Huensch & Tracy-Ventura, 2017), suggesting the potential moderator effects of L2 proficiency on the L1-L2 UF link. Derwing et al. (2009) also argued that the effects of proficiency on the L1-L2 link might be moderated by the cross-linguistic similarities between learners' L1 and L2. In the group of Mandarin learners of English (cross-linguistically different L1 and L2), L1-L2 correlations were lowered with increased L2 proficiency by reducing the transfer from L1, while in the group of Slavic learners of English (cross-linguistically similar L1 and L2), L1-L2 correlations were enhanced as a function of proficiency levels possibly by increasing beneficial L1 transfer.

As for breakdown fluency, both pause frequency and pause duration measures overall tend to show moderate to strong correlation coefficients between L1 and L2 speech (De Jong et al., 2015; Peltonen, 2018). However, the effects of pause type (silent vs. filled pauses) on the predictive power of L1 measures for L2 counterparts varied across studies. Although the

correlation coefficients between L1 and L2 silent pause frequency tend to be moderate-to-strong with a relatively stable tendency across studies, the strengths of the L1-L2 association of filled pause frequency varies considerably across studies. Some studies reported highly strong predictive power of L1 filled pauses for the L2 counterparts (r = .73 in De Jong et al., 2015). Even with moderate correlation coefficients, the correlation coefficients of filled pauses (r = .50 for mid-clause pauses, r = .30 for end-clause pauses) were higher than those of silent pauses (r = .34 for mid-clause pauses, r = -.05 for end-clause pauses) in Duran-Karaoz and Tavakoli's (2020) study. Meanwhile, the correlation coefficients between L1 and L2 filled pauses were non-significant in Huensch and Tracy-Ventura's (2017) study. Taken together, previous studies suggest that there is variability of the predictive power of L1 filled pause measures for their L2 counterparts. One of the possible reasons might be cross-linguistic and/or cross-cultural differences of filled pauses (Tian et al., 2017). Considering the fact that these studies reporting both filled and silent pauses (De Jong et al., 2015; Duran-Karaoz & Tavakoli, 2020; Huensch & Tracy-Ventura, 2017) commonly employed the combination of stress-timed and syllable-timed languages for speakers' L1 and L2, the varied strength of association between L1 and L2 filled pause frequency may not be explained by cross-linguistic similarities (e.g., rhythm, tempo). An alternative possible reason for the varied L1-L2 link in filled pauses is cross-cultural differences, because phonologically similar languages can have different norms for temporal features. There is a consensus that filled pauses have some communicative functions, such as the signal of complexity of the upcoming utterance and planning problems to interlocutors (see Tian et al., 2017). The relative frequency of such functional use of filled pauses and the correspondence between fillers and communicative functions differed even between American and British English. For instance, Tian et al. (2017) found that the use of *um* signaling a severe planning problem

was observed only in the corpus of British English, but not in the corpus of American English.

With regard to pause duration measures, previous studies suggested that the correlation coefficients between L1 and L2 pause duration measures tend to be moderate to strong (De Jong et al., 2015; De Jong & Mora, 2019; Huensch & Tracy-Ventura, 2017; Peltonen, 2018). It can also be argued that end-clause pause duration may result in a stronger L1-L2 association than mid-clause pause duration (De Jong et al., 2015; Huensch & Tracy-Ventura, 2017; Peltonen, 2018). This might be explained by the assumption that end-clause pauses are reflective of conceptualization processes, which are shared across different language systems within individuals (i.e., language-general processes; De Jong, 2016b; Götz, 2013; Tavakoli, 2011). Accordingly, the covariance of end-clause pause duration between L1 and L2 may derive from the language-general processes underlying end-clause pauses.

Closely looking at previous studies, one may argue that the effect sizes of the L1-L2 UF link in repair fluency measures can considerably vary. One of the most plausible reasons for such varying effect sizes is the incompatibility of the focus of repair fluency measures across studies. Some studies count different kinds of disfluency altogether for the sake of validity of classification (Duran-Karaoz & Tavakoli, 2020; see also Section 4.7.4), while other studies have employed fine-grained measures, such as self-repetition frequency and self-correction frequency (De Jong et al., 2015; Huensch & Tracy-Ventura, 2017). However, even with the same target repair features, different effect sizes have been reported. For instance, in the case of the correlation between L1 and L2 self-repetition frequency, a strong effect size was reported in De Jong et al. (2015; r = .60), while non-significant correlation was found in Peltonen (2018; r_s = .094). These inconsistent findings may indicate that there are some

moderator variables affecting the association between L1 and L2 UF measures. Another important finding regarding the L1-L2 UF link in repair fluency is that the output of English-speaking learners of L2 French in Huensch and Tracy-Ventura's (2017) study showed a moderate-to-strong correlation coefficient in self-repetition frequency (r = .55) only after 5-month residence in the L2-speaking environment (i.e., French-speaking country), despite non-significant correlation before the residence abroad (r = .12). This longitudinal change of the strength of the L1-L2 UF association seemed to be related to the development of L2 proficiency. However, the same pattern was not found in English-speaking learners of Spanish. Therefore, as with speed fluency, the strengths of the L1-L2 UF association may reflect the complex interplay of cross-linguistic effects and learners' L2 proficiency (for a review, see Huensch & Tracy-Ventura, 2017).

3.10 Summary

This chapter reviewed the literature of L2 fluency research with particular focus on three key subconstructs of fluency—CF (cognitive fluency), UF (utterance fluency), and PF (perceived fluency). Moreover, previous studies about the relationship between UF and PF and between CF and UF were synthesized with regard to the potential moderator effects of methodological factors. Regarding the UF-PF link, previous studies showed that listener-based PF judgements are primarily associated with speed and breakdown fluency and secondarily with repair fluency (Saito et al., 2018; S. Suzuki & Kormos, 2020). However, the best predictors of UF measures for PF ratings varied among studies (e.g., speed vs. breakdown fluency), and the role of repair fluency measures in PF judgements has been found inconsistent across studies. A close investigation into previous studies indicated that those contradictory findings might have derived from methodological differences across studies.

As for the CF-UF link, although few in number, previous studies operationalized CF as a set of psycholinguistic tests, such as picture naming tasks and sentence construction tasks, and predicted UF measures from those CF measures (De Jong et al., 2013; Kahng, 2020). These studies suggested that different components of CF can be associated with different dimensions of UF to a varying degree. For a better understanding of the CF-UF link, it is thus essential to consider the dimensionality of CF and UF. Moreover, De Jong et al. (2013) reported the potential moderator effects of speaking task type on the CF-UF link. Further studies are thus needed to examine what task design features moderate the CF-UF link.

In response to the potential moderator effects of task design features on the CF-UF link, the chapter reviewed the framework of speech processing demands which can theorize how task design features affect speech production processes (Préfontaine & Kormos, 2015; Skehan, 2009, 2014b). The review of studies about L2 speech processing demands suggested that the operationalization of conceptualizing demands in prior work might not have been aligned with task requirements and that the direct manipulation of formulation demands has not been tested yet.

Finally, in accordance with Segalowitz's definition of UF as the indicator of speaker's L2-specific automaticity (i.e., CF), the validity of UF measures can be discussed in terms of not only the predictive/concurrent validity with PF and CF, but also the independence from language-general processing skills. Prior research has operationalized such language-general aspects of UF measures as the covariance between L1 and L2 UF measures and examined the association between L1 and L2 measures with regard to cross-linguistic influences and the role of L2 proficiency (Bradlow et al., 2017; Duran-Karaoz & Tavakoli, 2020; Huensch & Tracy-Ventura, 2017). However, the L1-L2 UF link has been researched in the pairs of stress-

timed and syllable-timed languages. Accordingly, it is still unclear how L1 and L2 UF measures are related to each other in the context of learners with mora-timed L1 language backgrounds.

Chapter 4: Methodology for Study 1[8]

4.1 Introduction

Study 1 examines the relationship between PF (perceived fluency) and UF (utterance fluency), that is, the predictive power of UF measures for listener-based PF judgements. As the UF-PF link has been relatively extensively investigated in the literature of L2 oral fluency, Study 1 takes a meta-analytic approach to synthesizing previous findings and computing the overall effect sizes of correlation coefficients between PF and UF measures. The current chapter illustrates RQs of Study 1 (Section 4.2), the procedures of library search and data coding (Sections 4.3–4.6), the methodological trends of collected primary studies (Section 4.7), and the statistical analysis (Section 4.8).

4.2 Research Questions

A body of previous research has extensively examined the predictive role of UF measures in listeners' judgements of PF (Bosker et al., 2013; Derwing et al., 2004; Kormos & Dénes, 2004; Préfontaine et al., 2016; Rossiter, 2009; Saito et al., 2018; S. Suzuki & Kormos, 2020). Although previous studies have commonly showed that speed and breakdown fluency measures are primarily associated with PF, the best predictors for PF varied among studies (see Section 3.4). Moreover, another component of UF—repair fluency—has been found to be inconsistently related to PF across studies. Taken together, these findings in previous studies suggest the possibility that methodological variables, such as rating procedures and listeners' background, may affect the relationship between UF and PF. Therefore, Study 1 meta-analysed the overall correlation coefficients between PF scores and UF measures and also examined the moderator effects of methodological factors on the UF-PF connection. In

[8] Several sections of this chapter were accepted for publication in *The Modern Language Journal* as Suzuki, Kormos, and Uchihara (in press, 2021).

addition, motivated by the multidimensional nature of breakdown fluency, I shed light on methodological considerations in measuring pausing behaviours such as the threshold for silent pauses. I also examined the extent to which methodological differences in pause measurements can moderate the strength of association between pause-related measures and PF scores. Study 1 was thus guided by two RQs:

RQ1-1. What is the overall relationship between perceived fluency and subdimensions of utterance fluency–*speed*, *breakdown*, and *repair fluency*–as well as composite measures?

RQ1-2. To what extent does the relationship between perceived fluency and utterance fluency vary, according to methodological factors in different phases of L2 perceived fluency research—*speech stimuli preparation*, *rater recruitment, rating procedure*, and *selection of utterance fluency measures*?

4.3 Literature Search

As the first and foremost step of meta-analysis methodology (see Plonsky & Brown, 2015), the research domain of the current meta-analysis was specified—the relationship between listener-based judgements of L2 oral fluency (i.e., PF) and objectively measured temporal features of speech (i.e., UF). From a statistical perspective, this target research domain was regarded as correlational coefficients between PF scores and UF measures. In order to obtain a comprehensive pool of previous studies, Study 1 employed three different literature resources: database searching, journal search, and ancestry search from review papers.

Regarding database searching, following the guidelines on literature search for a meta-analysis (In'nami & Koizumi, 2010; Plonsky, 2015; Plonsky & Brown, 2015), five databases

were included: *Linguistics and Language Behaviour Abstract* (LLBA), *the Educational Resources Information Center* (ERIC), *ProQuest Dissertations and Theses*, *PsycINFO*, and *Academic Search Ultimate*. Besides, in order to reduce the effects of publication bias (i.e., the tendency of published studies to report larger or significant effect sizes and, subsequently, the potential suppression of small or non-significant effect sizes in published articles; Pigott & Polanin, 2019), I took an inclusive approach by including dissertations, conference proceedings, and book chapters. Building on the research domain, the following keywords were collected, covering target variables (i.e., PF, UF) and relevant methodologies including statistical analyses: *second language/L2, foreign language, fluency, correlation, speech, oral, speaking, production, performance, spoken, utterance, perception, assessment, rating, judgement, temporal, speed, breakdown, pause, repair, disfluency, pronunciation, acoustic, speech rate, articulation rate, mean length of run, syllable, syllable duration, hesitation, reformulation, false starts, repetition, pause frequency, pause duration, pause ratio, pause length, phonation time ratio, self-repair, self-correction*. I also used a NOT function of Boolean operators with *reading fluency* and *writing fluency* to exclude studies about fluency in written modality.

For the sake of the comprehensive literature search, I also conducted journal search with the same keywords, from the following 23 journals of applied linguistics and speech-related phenomena: *Annual Review of Applied Linguistics, Applied Linguistics, Applied Psycholinguistics, Bilingualism: Language and Cognition, Canadian Modern Language Review, ELT Journal, Foreign Language Annals, International Journal of Applied Linguistics, International Review of Applied Linguistics in Language Teaching, Journal of Speech, Language and Hearing Research, Language and Speech, Language Assessment Quarterly, Language Awareness, Language Learning, Language Teaching, Language*

Teaching Research, Language Testing, Modern Language Journal, RELC Journal, Second Language Research, Studies in Second Language Acquisition, System, TESOL Quarterly. In addition, I conducted ancestry search from recent review papers of L2 fluency (De Jong, 2016a, 2018; Segalowitz, 2016).

The literature search identified 5,061 papers eligible for the meta-analysis. Afterwards, their titles, abstracts, and study descriptors (e.g., keywords, subject categories) were inspected if (a) the study measured any aspects of oral fluency in any forms (either PF or UF) and (b) the speech data were produced by L2 learners. A sample of approximately 2% (k = 100) of the 5,061 studies was independently examined by the trained research assistant. As a result, 93% agreements were reached at this screening, and then the disagreements were solved through discussion. These screening processes identified 318 studies (for the entire process of retrieving studies, see Figure 7).

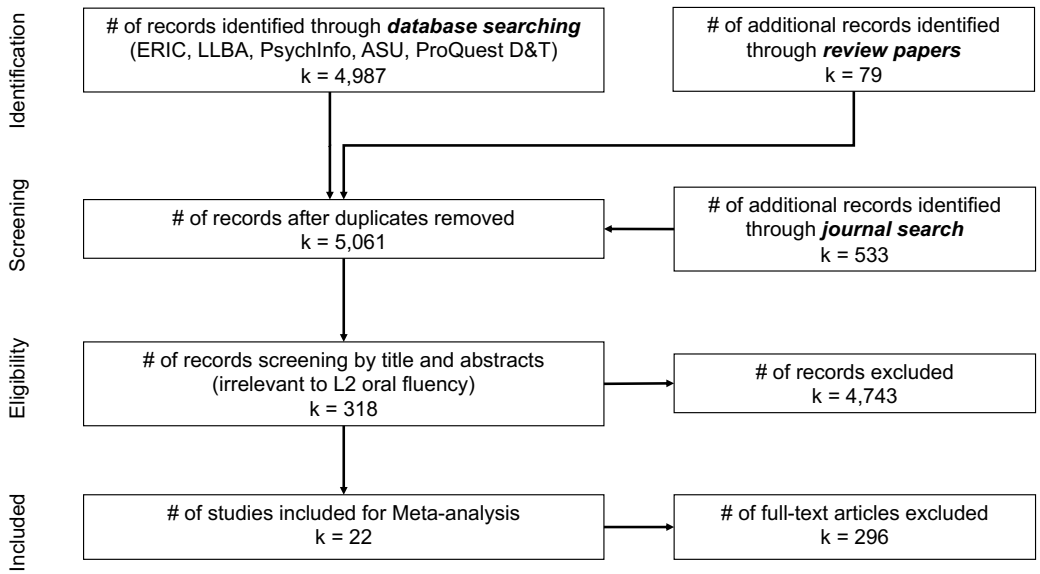


Figure 7. The entire process of retrieving studies.

4.4 Criteria for Eligibility

In order to identify the relevant studies to the current meta-analysis among the retrieved studies, the following nine criteria were set:

1. The study explicitly mentioned that speech samples were collected from L2 learners. I excluded studies that employed speech data for the purpose of clinical assessment (e.g., speech disorder, neuropsychological disease).
2. The study may include speech samples elicited from L1 speakers of the target language. However, L1 speakers' speech samples must be only used as the reference points for PF judgements (cf. Bosker et al., 2013). Based on this criterion, I excluded van Gelderen (1994) where L1 speech samples accounted for the large portion of participants (48 out of 60 speech samples).
3. The study may employ different speaking tasks for speech elicitation because I aim to examine the moderator effect of task types on the explanatory power of UF measures for PF judgements. However, since the target research domain exclusively focuses on L2 speech, I excluded interpreting speech which requires speakers to process two languages simultaneously (e.g., Yu & van Heuven, 2017).
4. The study evaluated L2 oral fluency using listener-based scaler ratings. The study may have either used any type of existing rating tools (e.g., CEFR assessment grid; Préfontaine, Kormos, & Johnson, 2016) or created researchers' own definition of fluency to raters (e.g., Saito et al., 2018). Note that the study may have even provided no predefined definition of fluency to raters if the focus of the research was on listeners' intuitive perception of fluency (e.g., Kormos & Dénes, 2004; S. Suzuki & Kormos, 2020).

5. The study may have employed analytic rating scores of fluency of some oral proficiency tests (e.g., TOEFL iBT) if the fluency scores were determined by listener-based judgements. However, it must have provided sufficient information of how fluency scores were assigned, such as the number of scale points and the definition of the target "fluency" in the target assessment context. I excluded the study if such fluency score mingled with other constructs (e.g., *fluency and coherence* in IELTS band descriptor; Korko & Williams, 2017). Similarly, some of those studies examined the relationship between UF measures and the holistic scores of high-stakes tests, such as ACTFL OPI (Baker-Smemoe et al., 2014; Silvio et al., 2016) and TOEFL variants (Higgins et al., 2011; Kang et al., 2010; Lee & Winke, 2018). Although the descriptors of those holistic scores referred to fluency or temporal features of speech, they were not designed to calculate the analytic scores for fluency. I thus excluded these studies as well.
6. The study employed at least one objective measure of UF (e.g., speech rate, pause frequency).
7. The study must have either reported correlation coefficients between listener-based and objective measures of fluency or provided information needed to calculate correlation coefficients such as raw data (e.g., Kormos & Dénes, 2004).
8. The study may have employed any group of raters for PF judgements, because Study 1 aims to investigate the effects of listeners' background on the relationship between UF measures and PF judgements.
9. The article reporting on the study must have been written in English.

Using randomly selected 30 studies, the reliability of inclusion was established by 96.7% agreements with the trained research assistant. After disagreements were solved through

discussion, I coded the remaining studies and thus identified 28 out of 318 studies that met all nine criteria in the meta-analysis. However, some studies were found to use identical datasets across studies. Therefore, six studies were excluded, and thus 22 studies were included for the current meta-analysis, which provided in total 263 effect sizes. These 22 studies comprised 17 journal articles, one book chapter, one conference proceeding, two PhD theses, and one MA thesis.

4.5 Selection of Utterance Fluency Measures

Due to a large number of different UF measures across studies, I decided to reduce the number of UF measures for the current meta-analysis. In order to select appropriate UF measures for the meta-analysis, I combined an a priori theoretically driven approach with the methodological trends of the pooled previous studies. To this end, the frequency of different UF measures is summarized in Table 2. I decided to set selection criteria with reference to one of the motivations of the current meta-analysis—the informed practice for selecting valid UF measures in future fluency research. Accordingly, the theoretically distinctive subdimensions of UF were first considered—speed, breakdown, and repair fluency (Tavakoli & Skehan, 2005). I then decided to include one or two representative UF measures for each subdimension of UF. Second, I also decided to include some composite measures, considering their prevalent use in research on L2 fluency development (e.g., Baker-Smemoe et al., 2014; Iwashita et al., 2008; Tavakoli et al., 2020). Third, for practical reasons for statistical analyses, I also take into account the number of effect sizes of UF measures in the current pool of effect sizes.

Table 2. *Descriptive summary of utterance fluency measures in the pooled studies through the library search.*

Construct	UF measures	*k*	*n*
Speed	Articulation rate	11	28
Breakdown (Frequency)	Pause frequency	10	38
	Mid-clause pause frequency	7	9
	End-clause pause frequency	5	5
	Filled pause frequency	7	17
Breakdown (Duration)	Pause duration	7	20
	Mid-clause pause duration	3	5
	End-clause pause duration	2	2
Repair	Disfluency rate	5	10
	False start	1	1
	Repetition	4	6
	Self-correction	4	6
Composite	Speech rate	12	44
	Mean length of run	10	30
	Phonation time ratio	4	9

Note. k = number of studies; n = number of effect sizes. The total number of studies = 22.

Regarding speed fluency, there is only one fine-grained measure, namely, articulation rate (Bosker et al., 2013; Tavakoli et al., 2020). Note that some studies used the measure of mean duration of syllable which is the mathematically inversed measure of articulation rate (De Jong, 2016a). However, I coded and counted the measure as articulation rate.

As for breakdown fluency, scholars have confirmed the multidimensional nature of pausing behaviours in terms of frequency, length, and location. I thus first labelled breakdown fluency measures as either frequency or duration measures. Both measures were then further coded in terms of pause location (mid-clause, final-clause or both) so that the labels of pause location could be used for moderator analysis. Considering the independence of observations for the effect size aggregation, multiple effect sizes should not be included for effect size aggregations from the same dataset (e.g., both of mid- and end-clause pause measures for the

effect size aggregation for breakdown fluency). Considering the number of effect sizes available (see Table 2), I focused on silent pause frequency and mean duration of silent pauses for the effect size aggregation (RQ1-1) and also decided to use pause location (mid- vs. end-clause pauses), minimum pause length, and pause type (silent vs. filled pauses) for moderator analyses (RQ1-2). Another problem was that especially recent studies employed only mid- and end-clause pause measures, instead of pause measures counting pauses regardless of pause locations. In such cases, I averaged the effect sizes between mid- and end-clause pause measures for the overall effect size calculations (RQ1-1).

As shown in Table 2, there was variability in the focus of disfluency phenomena among studies. Considering the comprehensive range of features, I selected disfluency rate—the mean number of all types of disfluency features—as a representative measure for repair fluency in the research context of the UF-PF link. However, to obtain a relatively large number of effect sizes for the effect size aggregation, I decided to include repair fluency measures capturing the frequency of any type of disfluency features. For the sake of the independence of observations, the selected effect sizes were averaged across measures for the effect size aggregation if the study reported multiple repair measures (e.g., self-repetition and self-correction; Saito et al., 2018). I labelled the targeted disfluency features for the subsequent moderator analysis (RQ1-2).

As regards composite measures of UF, the number of effect sizes available was prioritised over their theoretical accounts. Accordingly, I only selected speech rate (k = 44) and mean length or run (k = 30) for the current meta-analysis.

4.6 Coding

To establish the reliability for coding effect sizes and relevant moderator variables, the randomly selected 10 out of 22 studies were blind-coded by myself and the trained research assistant. The inter-coder agreements reached 95.8%, and disagreements were solved through discussion. Accordingly, the coding scheme was revised based on the discussion. Then, I coded the remaining studies.

4.7 Moderator Variables

Motivated by methodological differences among previous studies, I initially intended to code a total of 16 moderator variables. However, due to an incomparability of some of the moderator variables across studies (e.g., different criteria for proficiency levels across studies), I eventually coded 11 out of 16 moderator variables, using the following criteria. Regarding the excluded moderator variables, I descriptively synthesize previous studies for the purpose of providing some insights into future directions for L2 fluency research.

4.7.1 Speech stimulus preparation

Regarding the phase of speech stimulus preparation, five moderator variables were initially selected, including speakers' background, speech elicitation method, and length of speech stimuli. However, some of the moderator variables failed to reach a satisfactory level of comparability across studies. For those moderator variables, I descriptively synthesize previous studies here instead of conducting moderator analyses (RQ1-2).

Speakers' L1 and L2

As summarized in Table 3, I coded speakers' language background in terms of their L1 and L2. If studies combined speakers of multiple L1 backgrounds, I coded them as Varied. Although there was huge variability of speakers' L1 among studies, previous studies have

extensively researched the relationship between UF and PF in the context of L2 English (k = 16), followed by L2 Dutch (k = 5). Due to the varying L1 backgrounds of studies, variability of L1-L2 pairs was also remarkable. Therefore, I decided to exclude moderator variables of L1 and L1-L2 pairs from the subsequent moderator analyses; only a moderator variable of target L2 was submitted to the moderator analysis. It might be noteworthy that the most researched group of L2 speakers was Japanese-speaking learners of English (k = 6).

Table 3. *Descriptive summary of frequency of researched L1 and L2 of speakers.*

L1 background	*k*	Target L2	*k*	L1-L2 pairs	*k*
Japanese	6	English	16	Japanese - English	6
Mandarin	2	Dutch	5	Mandarin - English	2
English	2	French	2	French - English	1
Korean	2	Japanese	1	Hungarian - English	1
French	1	Spanish	1	Korean - English	1
Hungarian	1			Persian - English	1
Persian	1			Slavic - English	1
Slavic	1			Spanish - English	1
Spanish	1			English - French	1
Varied	8			Korean - Japanese	1
				English - Spanish	1
				Varied - Dutch	5
				Varied - English	2
				Varied - French	1

Note. k = number of studies. The total number of studies = 22. Some studies employed multiple groups of speakers.

Proficiency level

The pooled studies were first tentatively categorised into several proficiency levels of speakers (see Table 4). Although previous studies tended to recruit speakers from a wide range of proficiency levels (coded as Varied in Table 4), there was huge variability of how scholars assessed speakers' proficiency levels, that is, assessment methods such as existing high-stakes tests (e.g., TOEFL) and speakers' enrolled language classroom (see Table 5). Due to the varying methods of assessing speakers' proficiency levels across studies, it can be

argued that there may be the potential lack of validity of the categorization. Therefore, a moderator variable of proficiency level was excluded from the moderator analysis.

Table 4. *Frequency of different proficiency levels of speakers.*

Proficiency level	The number of subgroups
Beginner	6
Beginner-to-Intermediate	1
Intermediate	4
Intermediate to Advanced	3
Varied	9
Not reported	2

Note. The total number of studies = 22.

Table 5. *Trend of assessment methods for proficiency levels of speakers.*

Assessment method	*k*
High-stakes test scores	5
Canadian Language Benchmark	3
TOEFL	2
Placement test scores	1
Existing language class	4
Length of Residence; Overseas experience	4
Raters' assessment	1
Self-reported proficiency	1
Not reported	6

Note. *k* = number of studies. The total number of studies = 22. The studies that did not reported speakers' proficiency levels were excluded from the assessment method.

Education level

Considering the fact that previous studies have reported the age of speakers using different statistics, I decided to group studies into several categories of educational level, following previous meta-analyses in L2 research (e.g., de Vos et al., 2018; Uchihara et al., 2019). However, as summarized in Table 6, due to the huge variability across studies, I only describe the trends of previous studies in terms of targeted groups of speakers. Table 6 shows that as with other lines of L2 research, university students have been most researched,

possibly because they are easy to access for researchers (cf. Plonsky & Kim, 2016). In contrast, elementary and secondary students can be regarded as underresearched groups of speakers, which have not been examined as a single group of speakers. Finally, it should be noted that only one study examined the relationship between UF and PF in the context of L2 teachers as speakers.

Table 6. *Frequency of different education levels of speakers.*

Education level	The number of groups
University students	7
University students and Post-university	5
Post-university (Workers, Immigrants)	4
Mixed (Secondary to Postgraduate)	2
Elementary	1
L2 teachers	1
Pre-university language institute	1
Not reported	4

Note. Only 11 studies reported the mean or range of age; thus, alternatively, education level.

Task type

Considering the assumption that dialogic fluency is theoretically distinctive from monologic fluency (Tavakoli, 2016; van Os et al., 2020; Wright & Tavakoli, 2016), previous studies were categorised into monologic and dialogic speech according to their speech elicitation tasks. Despite the theoretical distinction between monologic and dialogic fluency, to the best of my knowledge, there is no empirical research on how UF-PF link varies according to the interactivity of speech (monologic vs. dialogic speech). I thus decided to take an exploratory approach to decide whether to exclude studies/effect sizes based on dialogic speech data from the meta-analysis (for the result of the heterogeneity test, see Section 4.8). Furthermore, studies using monologic tasks were further categorised into several sub-levels, in accordance with the extent to which listeners can predict the content of speech. As summarized in Table 7, the following labels were created: (a) Controlled (i.e., predefined content and form; e.g.,

read-aloud speech), (b) Closed task (i.e., predefined content; e.g., picture narrative), and (c) Open task (i.e., open-ended content; e.g., argumentative speech, personal narrative).

Table 7. *Frequency of different task types and stimulus type.*

Speech stimuli	The number of subgroups
Task type	
Monologic	26
Controlled production	3
Closed task	13
Open task	10
Dialogic	6
Stimulus type	
Entire speech	15
Excerpt	17

Note. The total number of studies = 22. Some studies employed multiple speaking tasks.

Excerpts vs. entire speech

Regarding the approach to presenting speech stimuli to listeners (see Section 3.5.1), I coded studies as either entire speech or excerpt (see Table 7). However, it is noteworthy that even excerpts had variability in the exact length of excerpts (20 to 300 seconds).

4.7.2 Rater recruitment

Listeners' language background

This variable simply consisted of L1 raters and L2 raters (see Table 8). L1 raters refer to listeners whose first language is the target language of the speakers, while L2 raters are those who speak the target language of the speakers as second languages. It should be noted that L2 raters can either share the same L1 as the speakers (k = 3; Ahmadi & Sadeghi, 2016; Kormos & Dénes, 2004; Negishi, 2012) or have different L1 background other than the speakers' L1 (k = 2; Magne et al., 2019; Rossiter, 2009). Rossiter (2009) also employed one group consisting of both L1 and L2 raters, which was labelled as Mixed.

Relating to listeners' background, I initially attempted to categorize studies by listeners' familiarity with the speakers' L1, including L2 speech accented by the L1s of the speakers and L2 learning experience of the speakers' L1. However, as described by Table 8, there was huge variability in the approach to operationalizing or measuring such kind of familiarity, and thus I decided to exclude this moderator variable from the moderator analysis.

Layperson vs. experienced raters

In addition to language background, Study 1 aims to examine how differently listeners with relevant experiences perceive L2 speech, compared to inexperienced raters. As indicated by Table 8, there were different types of experiences relevant to L2 PF judgements, such as teaching experience and linguistic expertise. However, due to the potential overlap between teaching experience and linguistic expertise, I merged different types of raters' experiences. For instance, raters with teaching experience are arguably likely to have some experience of linguistics and assessment as well. As result, this variable consisted of Inexperienced raters, Experienced raters, and Mixed, which pooled both inexperienced and experienced raters.

Table 8. *Descriptive summary of listeners' background.*

Listener background	The number of subgroups
Language background	
L1 rater	17
L2 rater	5
Mixed	1
Not reported	8
Experience	
Inexperienced raters	11
Experienced raters	17
Language teaching experience	8
Expertise in linguistics	5
Expertise in language assessment	1
Mixed	2

Not reported	1
Familiarity with speakers' L1	
Self-report (e.g., Scaler rating)	6
Teaching experience with the target group of speakers	5
L1-shared L2 speakers	4
Expertise in linguistics of the target language	2
Learning experience of the speakers' L1	1
Not reported	6

Note. The total number of studies = 22. Some studies employed multiple groups of listeners.

4.7.3 Rating procedure

Definition of perceived fluency for raters

Motivated by the distinction between lower- and higher-order fluency (i.e., temporal fluency vs. overall oral proficiency), prior research has operationalized PF by the presence or absence of the definition of fluency from researchers. Thus, the pooled studies with semantic scales were first categorized based on whether researchers provided the definition of fluency to raters: (a) No definition and (b) Researcher's definition. Furthermore, some studies provided rubrics of existing assessment tools (e.g., CEFR) or created for research purposes (e.g., Sato, 2014). Therefore, two categories were added: (c) Rubrics of existing assessment tools and (d) Research-based rubrics (see Table 9).

Number of scale points

Although the number of scale points can be by nature regarded as a continuous variable, I found a limited variation among pooled studies, as summarized in Table 9. I thus decided to deal with this moderator variable as a categorical variable. One study used a sliding bar scale without numerical values on the scale (Saito, Trofimovich, & Isaacs, 2017; see Table 9). However, this type of rating scale can be considered distinctive from traditional rating scales with numerical values, and I thus excluded Saito et al. (2017) from the moderator analysis. No distinction was made in whether studies employed some transformation of scores (e.g.,

Rasch analysis; Negishi, 2012) and whether studies used either semantic scales or rubrics/descriptors.

The amount of rating practice

This variable consisted of two categories: Short practice and Extensive training. Studies were labelled as short practice when researchers only familiarized their raters with the use of rating scale with several speech samples (e.g., three samples; see Table 9), immediately prior to main rating. On the other hand, studies were categorized as extensive training when researchers provided some interventions, such as feedback and discussion among raters, for their raters to reach a consensus on scores either with researchers or among raters.

Table 9. *Descriptive summary of perceived fluency rating procedure.*

Rating procedure	*k*
Definition of fluency	
No definition (raters' intuition)	9
Researchers' definition	11
Existing assessment tools (e.g., CEFR)	6
Researcher-based rubrics	3
No. of scale points	
5	4
6	6
7	5
9	10
10	4
1000 (with no numerical points)	1
Pre-rating training	
Short practice	15
3 samples	5
4 samples	3
5 samples	6
6 samples	1
Extensive training	4
Not reported	8

Note. k = number of studies. The total number of studies = 22

4.7.4 Utterance fluency measure computation

The fourth phase of L2 PF research is the computation of UF measures as predictor variables for PF scores. At this phase, the first decision that researchers need to make is whether temporal features of speech are annotated by either manual or automated coding. Therefore, I decided to include the method of speech annotation as a moderator variable. In order to calculate UF measures either manually or automatically, researchers also need to specify the criteria for target disfluency features such as pauses and repairs. I thus focus on the definition and focus of pauses and the focus of disfluency features.

Speech annotation

This variable has two categories: Manual coding and Automatic annotation (see Table 10). Manual coding refers to studies where researchers manually transcribe and annotate temporal features with some assistance of acoustic analysis software such as Praat (Boersma & Weenink, 2012). On the other hand, studies were coded as automatic annotation when researchers annotate speech or compute UF measures by only using some type of software. In the pooled studies, studies coded as automatic annotation used either De Jong and Wempe's (2009) script in Praat (*Praat Script Syllable Nuclei v2*) or the continuous speech recognizer (Strik et al., 1997).

Table 10. *Descriptive summary of speech annotation methods.*

UF measure analysis	*k*
Speech annotation	
Manual coding	17
Automatic annotation	3
NA	2

Note. *k* = number of studies. The total number of studies = 22. There were two studies without information about annotation methods.

Definition of pauses

Considering the fact that studies can specify pauses differently according to measures, coding for pause measures was conducted at the level of effect sizes rather than the level of studies. As reviewed previously, some studies specified the threshold for silent pauses in terms of not only the minimum length of pauses, but also the maximum length of pauses. However, due to the limited number of studies specifying the upper bound of silent pauses (k = 2), I focused only on the minimum length of silent pauses. Although a threshold for silent pauses is conceptually a continuous variable, considering the limited range of pause lengths (see Table 11), I decided to treat this variable as a categorical variable. Moreover, due to the limited number of effect sizes, I excluded the category of 300ms pauses (k = 2). Eventually, this moderator variable consists of the following pause length categories: 200ms, 250ms, and 400ms. In addition to the threshold for silent pauses, I included pause location as another moderator variable. Pause measures were thus classified by three categories: Both (counting pauses regardless of location), Within clause (pauses in the middle of clauses), and Between clause (pauses at clause boundaries). Finally, Study 1 also aims to examine the moderator effects of pause type—silent and filled pauses. Since some studies counted silent and filled pauses together (e.g., Trofimovich et al., 2017), I first separated studies in terms of whether pause measures were separately computed for silent and filled pauses. Pause measures were then further classified, according to the following categories: Filled pauses, Silent pauses, and Mixed (counting pauses regardless of type).

Table 11. *Descriptive summary of definition and scope of pauses and disfluency features.*

UF measure analysis	*n*
Pause length	
200 ms	39
250 ms	79
400 ms	47

Pause location	
Both	62
Within clause	14
Between clause	7
Pause type	
Filled pauses	17
Silent pauses	79
Mixed	4
Disfluency features	
Mixed	10
Repetition	6
Self-correction	6

Note. *n* = number of effect sizes. The total number of studies = 22.

Focus of disfluency features

As with pause measures, repair fluency measures were also classified according to their scope of disfluency phenomena. Effect sizes were labelled by the targeted disfluency features, while repair fluency measures based on multiple phenomena were labelled as Mixed. The frequency of each category was summarized in Table 11. Due to the limited number of effect sizes, I excluded the measure of false starts ($k = 1$) and thus resulted in three subgroups: Mixed, Repetition, and Self-correction.

4.7.5 Reporting practice of statistics

Following previous meta-analyses, the reporting practice of statistics in primary studies was also examined for descriptive statistics, reliability estimates, and type of regression analyses. Among 22 primary studies, 16 studies reported descriptive statistics for PF scores, and 15 studies included descriptive statistics for UF measures. As shown in Table 12, I found a range of inter-rater reliability indices for PF scores, while only few studies reported inter-coder reliability for UF measures. However, note that the use of automatic annotation for UF measures results in less room for subjectivity, and the reliability can be hypothesized to be perfect. The trend in regression analysis is summarized in Table 13, showing that many of

primary studies relied only on correlation analyses. Meanwhile, the recent use of linear mixed-effects modelling is notable (Bosker et al., 2013; Préfontaine et al., 2016), because it can control for individual raters' variability in ratings as a random-effects predictor variable, as opposed to the simple aggregation of rating scores in traditional multiple regression analyses.

Table 12. *Summary of reliability indices for measures of perceived fluency and utterance fluency.*

Index Type	*k*	*Mdn*	*Range*
Perceived fluency			
Cronbach	13	0.94	.85–.98
Correlation (Pearson, Spearman)	3	0.75	.62–.81
Intraclass correlation	4	0.74	.53–.93
Cohen's kappa	1	0.81	—
Rasch	1	0.76	—
Not reported	3	—	—
Utterance fluency			
Cronbach	3	0.92	.90–.94
Automatic annotation	3	—	—
% agreement	2	0.90	.80–.99
Raw score difference	1	—	—
Not reported	13		

Note. *k* = number of studies; *Mdn* = Median. The total number of studies = 22. Two studies reported multiple reliability indices for perceived fluency.

Table 13. *Summary of types of regression analysis for the utterance-perceived fluency link.*

Type of regression analysis	*k*
Stepwise multiple regression	6
Hierarchical multiple regression	2
Linear mixed-effects modelling	2
Correlation-only	14

Note. k = number of studies. The total number of studies = 22. Some studies reported multiple types of regression analyses.

4.8 Statistical Analysis

All the statistical analyses were implemented in R using *meta* package (Schwarzer, 2007), following Harrer et al. (2019). Prior to the analysis answering RQs, the extent to which the current dataset was influenced by publication bias was examined, using funnel plots and

Egger's tests. The visual inspection of the funnel plots (see Figure 8) as well as the results of Egger's tests (see Table 14 in Section 5.2.1) indicated that there were no substantial influences of publication bias on the findings of the meta-analysis.

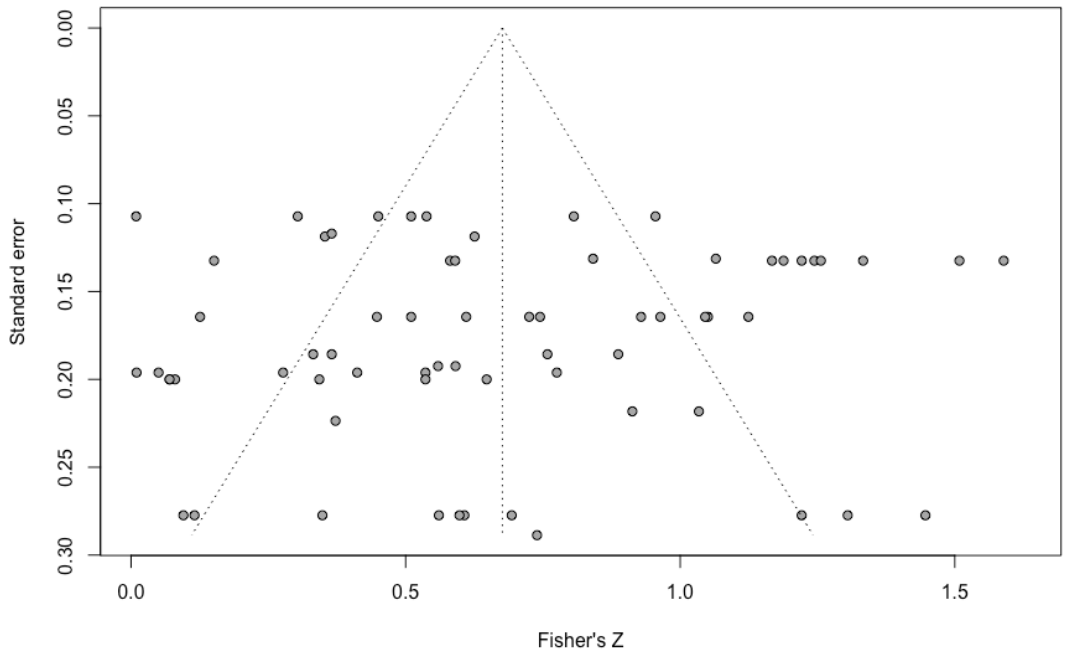


Figure 8. A funnel plot for six selected utterance fluency measures in RQ1, excluding effect sizes based on dialogic speech data.

Motivated by the theoretical distinction between monologic and dialogic fluency, I first checked whether the interactivity (monologue vs. dialogue) moderates the effect size between UF measures and PF scores. Using the inversed effect sizes (i.e., unidirectional correlational coefficients), the heterogeneity test (Cochran's Q test) showed that the effect of interactivity was significant ($Q(1) = 29.16$, $p < .0001$), suggesting that the correlation coefficients in dialogic speech ($k = 19$, $r = .08$, CI[-.10, .25], $p = .3891$) were lower than those in monologic speech ($k = 114$, $r = .54$, CI[.48, .60], $p < .0001$). More importantly, the correlation coefficients in dialogic speech were not significant, indicating the possibility that UF in dialogic speech differently contributes to the establishment of PF judgements. Therefore, I

decided to use effect sizes based on monologic speech data (for the pooled results based on both monologic and dialogic speed data, see Appendix A).

Before addressing research questions, the independence of observations in the pooled effect sizes was examined (Plonsky & Oswald, 2015) and then multiple effect sizes were averaged in the following cases. First, both Saito et al. (2018) and Magne et al. (2019) used the same speech data; however, the former study employed L1 raters, while the latter employed L2 raters. For the sake of independence of observations, I thus averaged their effect sizes across studies (i.e., averaged across L1 and L2 raters' correlation coefficients) for the overall effect size calculation (RQ1-1). Meanwhile, the effect sizes of these two studies were separately included for moderator analyses (RQ1-2), because Study 1 aims to examine the effects of rater background on the strength of the UF-PF link. Similarly, some studies employed different groups of raters using the same speech stimuli. In this case, if studies reported the correlation coefficients by pooling different rater groups, the pooled correlation coefficients were included for RQ1-1. Otherwise, the correlation coefficients were averaged across raters for RQ1-1. In addition, some studies used speech samples elicited through different speaking tasks but from the same speakers and/or at multiple time points. The inclusion of different correlation coefficients for each task and/or for each time point violates the independence of observations for RQ1-1. In such cases, correlation coefficients were thus averaged across tasks and/or time points, whereas those effect sizes were separately included for each category of target moderator variables (RQ1-2).

In order to answer RQ1-1, the inverse-variance weighted overall effect sizes were computed separately for six UF measures (articulation rate, pause frequency, pause duration, disfluency rate, speech rate, mean length of run), using a random-effects model with Restricted

Maximum Likelihood estimation method (Novianti et al., 2014; Veroniki et al., 2016). The rationale behind using random-effects modelling was that it was assumed that the pooled studies potentially come from different populations rather than one single true population. In order to consider the sampling errors at the level of studies, random-effects modelling was thus more appropriate than fixed-effects modelling in Study 1 (Harrer et al., 2019; Plonsky & Oswald, 2015). I also decided to exclude influential cases for the sake of the robust estimates of aggregated effect sizes (Viechtbauer & Cheung, 2010). The exclusion criteria was set based on the prediction intervals of target measures, which suggest the possible range of correlation coefficients in future studies (Higgins et al., 2009; Nagashima et al., 2019). This decision was also motivated by the aim of the current study to suggest a methodological guideline for future L2 fluency studies. A within-group *Q* statistic was employed to detect the potential heterogeneity of the effect sizes across the studies included in the analyses.

In order to address RQ1-2, subgroup analyses were conducted for the moderator variables, all of which were categorical. As with RQ1-1, I used random-effects modelling for pooling the effect within each subgroup. Furthermore, considering the categorization of the moderator variables, it is possible that there are potential different subgroups of most of the moderator variables (e.g., categorization of task types). Statistically speaking, the categorization of subgroups may introduce a new sampling error at the subgroup level, and therefore I decided to use random-effects modelling for between-subgroup comparisons while controlling for such potential sampling errors. The minimum number of studies for each category of moderator variables was set as $k = 3$, following previous meta-analytic studies in L2 research (e.g., Li, 2016; Uchihara et al., 2019). Since all of the moderator variables were categorical variables, a between-group *Q* statistic was calculated to examine the impact of the moderator variables on the effect sizes. Following Plonsky and Oswald (2014), effect sizes of

correlation coefficients were interpreted as follows: Small = |.25–.40|, Medium = |.40–.60|, Strong = |.60–1.00|.

4.9 Summary

This chapter outlined the methodology of Study 1 which synthesized previous research into the relationship between UF measures and PF scores, using a meta-analytic approach. Combining three different literature resources (database searching, journal search, and ancestry search from review papers), Study 1 initially pooled 5,061 studies. In order to select relevant previous studies in a systematic manner, a total of nine inclusion criteria was established, and eventually 22 studies (263 effect sizes) were included for the current meta-analysis. Regarding the UF measures, considering a variety of different UF measures in the pooled studies, Study 1 selected six UF measures with regard to their construct validity and comparability among studies—articulation rate, silent pause frequency, silent pause duration, disfluency rate, mean length of run, and speech rate. Although some of the empirically motivated moderator variables were excluded from the statistical moderator analysis due to the lack of valid comparability across studies (e.g., different criteria for proficiency levels, different reporting practice of raters' background), 11 moderator variables were submitted to the subsequent moderator analysis, considering the different phases of L2 PF research methodologies.

Chapter 5: Results and Discussion of Study 1—Predictive Power of Utterance Fluency for Second Language Perceived Fluency[9]

5.1 Introduction

This chapter reports the results of Study 1 (RQ1-1, RQ1-2) which meta-analysed the relationship between UF and PF. RQ1-1 is concerned with the overall effect sizes of correlation coefficients between six UF measures commonly used in previous studies—articulation rate, pause frequency, pause duration, disfluency rate, mean length of run, and speech rate—and listener-based PF judgements (Section 5.2.1). RQ1-2 enquires into the moderator effects of methodological factors on the UF-PF link. To address RQ1-2, a set of moderator analyses was conducted, using 11 methodological factors identified in the synthesis of previous studies, including speech stimulus preparation, listeners' background, rating procedure, and selection of UF measures (Section 5.2.2). The predictive power of UF measures for PF ratings are discussed with regard to the potential moderator effects of methodological factors (Section 5.3).

5.2 Results

Based on the inspection of the initial forest plots of six utterance fluency measures, Préfontaine et al.'s study (2016) was identified as the influential case in the measure of pause duration (see Figure 9). Notably, Préfontaine et al. (2016) claimed that the positive correlation between mean duration of pauses and PF scores in their study can be explained by the combination of L1 and L2 of their participants (L1 English–L2 French). The current literature search did not find any other studies investigating the UF-PF link in the same population of L2 learners. Therefore, I regarded the effect sizes based on the pause duration

[9] Several sections of this chapter were accepted for publication in *The Modern Language Journal* as Suzuki, Kormos, and Uchihara (in press, 2021).

measure in Préfontaine et al. (2016) as methodologically exceptional cases and thus excluded their averaged effect size (k = 1, from three different tasks) from the effect size aggregation of pause duration (RQ1-1) and their raw effect sizes (k = 3) from the relevant moderator analysis (RQ1-2).

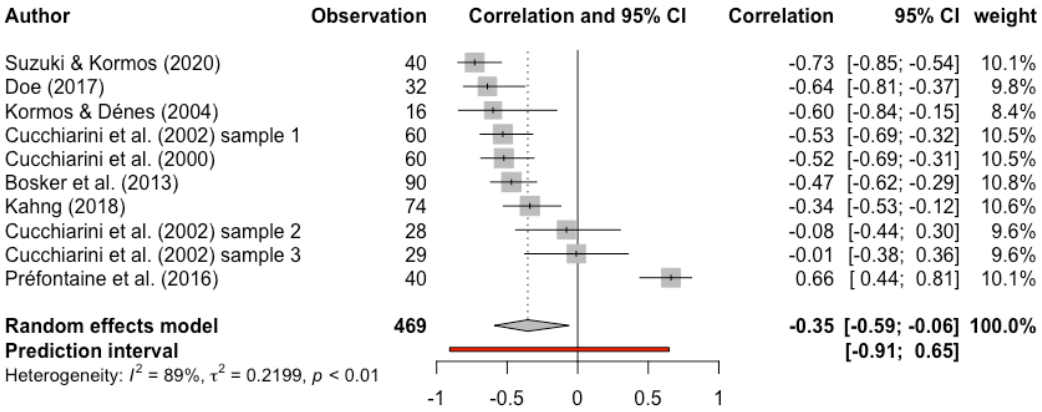


Figure 9. The initial forest plot of pause duration without excluding the influential case. *Note.* The diamond indicates the overall average correlation; and the red line shows a prediction interval.

5.2.1 Effect size aggregation

To answer RQ1-1, which asked about the overall relationship between PF and six selected UF measures, a set of effect size aggregations was conducted to produce the overall effect sizes and their 95% confidence intervals. As summarized in Table 14 below, the results suggested that all the measures were significantly associated with PF ratings (for the forest plots, see Figure 10–15). Both composite measures (mean length of run, speech rate) can be considered as showing strong effects (r = .72, .76, respectively), while the effect size for the speed fluency measure (articulation rate) was slightly smaller than that of the composite measures, but it still indicated a strong effect size (r = .62). Interestingly, within breakdown fluency measures, pause frequency measures (r = -.59) showed a stronger association with PF scores than pause duration measures (r = -.46). Moreover, the 95% confidence interval of

repair fluency measure (r = -.20, CI[-.30, -.09]) suggested that the effect size of repair fluency was significantly lower than all other UF measures. In other words, the association of repair fluency to PF tends to be weaker than that of speed and breakdown fluency. Finally, according to the Q-tests, the aggregated effect sizes for all the UF measures except for repair fluency showed significant heterogeneity among the studies, confirming the possibility that moderator variables may affect the association between PF scores and different UF measures.

Table 14. *Results of effect size aggregations for six utterance fluency measures.*

UF measures	*n*	*Sample size*	*Weighted effect size*	*CI*	*Q(df)*	*p-value*	*Egger's test p-value*
Speed fluency							
Articulation rate	11	525	0.62	[.45, .74]	56.11(10)	< .0001	0.049
Breakdown fluency							
Pause frequency	17	746	-0.59	[-.69, -.48]	70.29(16)	< .0001	0.486
Pause duration	9	429	-0.46	[-.59, -.31]	22.23(9)	0.0045	–
Repair fluency							
Disfluency rate	9	452	-0.20	[-.30, -.09]	7.70(8)	0.464	–
Composite							
Mean length of run	9	335	0.72	[.59, .74]	32.26(8)	< .0001	–
Speech rate	11	365	0.76	[.64, .85]	50.98(10)	< .0001	0.128

Note. n = number of effect sizes; Sample size = total number of observations. Since the minimum number of effect sizes for Egger's test is k = 10, the Egger's tests were not performed for Disfluency rate and Mean length of run. However, a visual inspection of their funnel plots suggested that there was no substantive bias among the effect sizes in both measures.

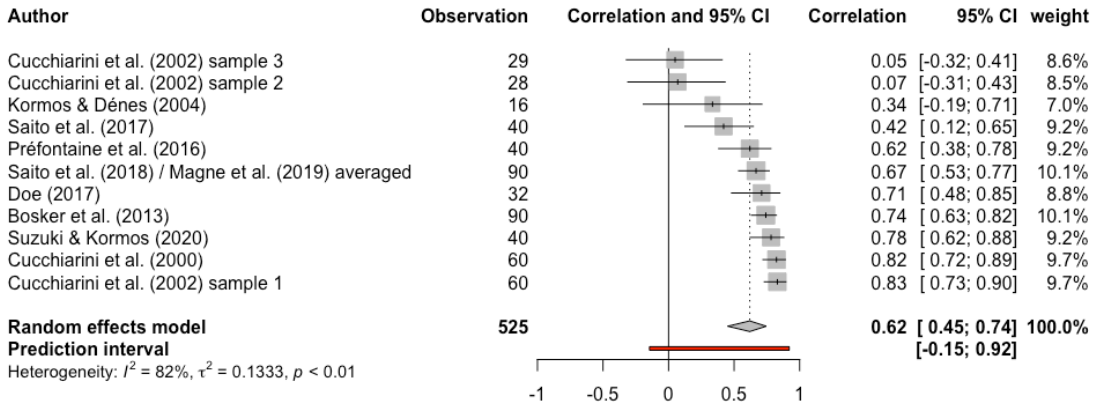


Figure 10. An overall average correlation between perceived fluency scores and articulation rate (indicated by the diamond) and correlations with confidence intervals for each study.

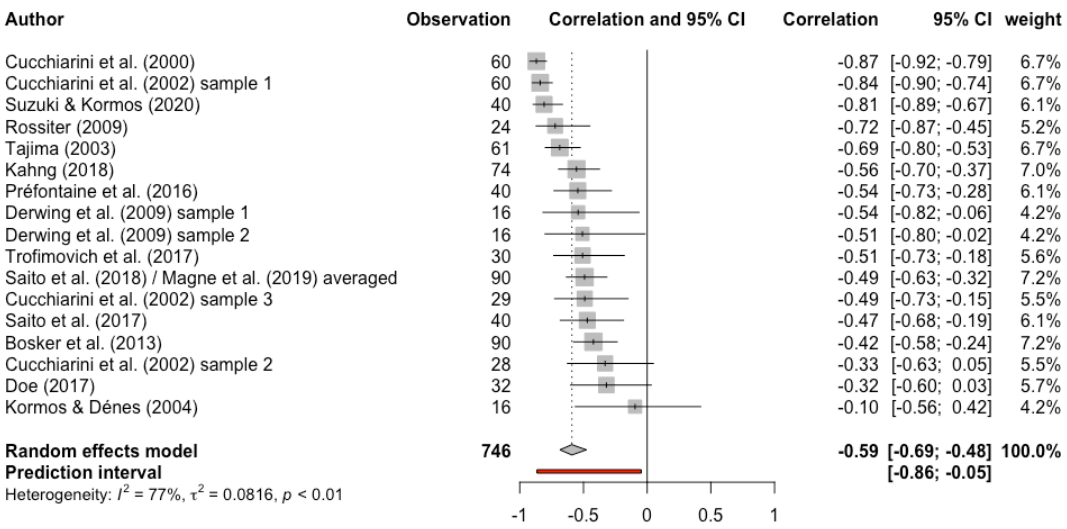


Figure 11. An overall average correlation between perceived fluency scores and silent pause frequency (indicated by the diamond) and correlations with confidence intervals for each study.

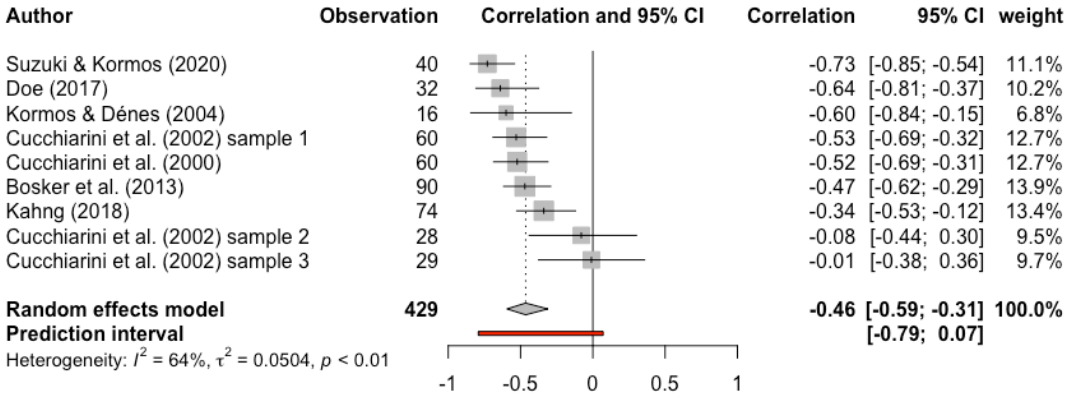


Figure 12. An overall average correlation between perceived fluency scores and mean duration of silent pauses (indicated by the diamond) and correlations excluding one influential case with confidence intervals for each study.

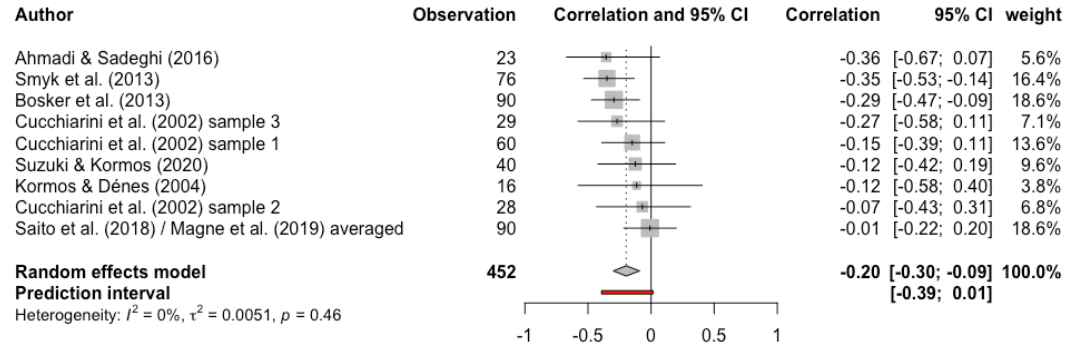


Figure 13. An overall average correlation between perceived fluency scores and disfluency rate (indicated by the diamond) and correlations with confidence intervals for each study.

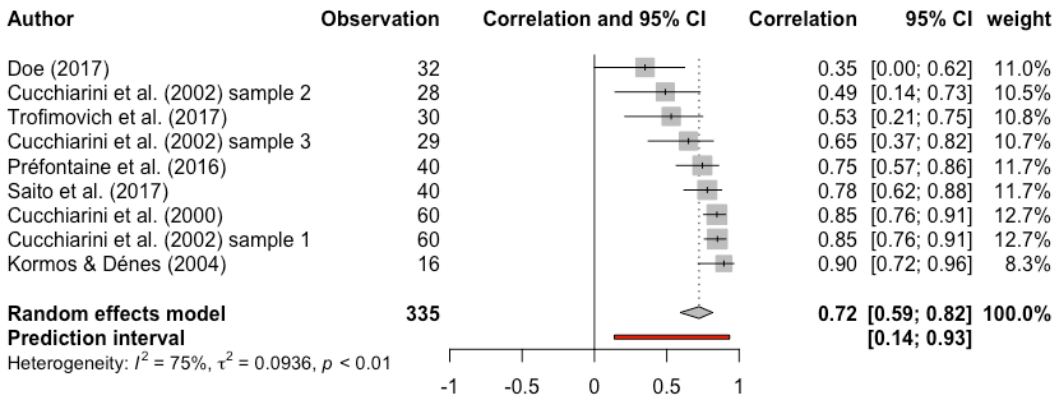


Figure 14. An overall average correlation between perceived fluency scores and mean length of run (indicated by the diamond) and correlations with confidence intervals for each study.

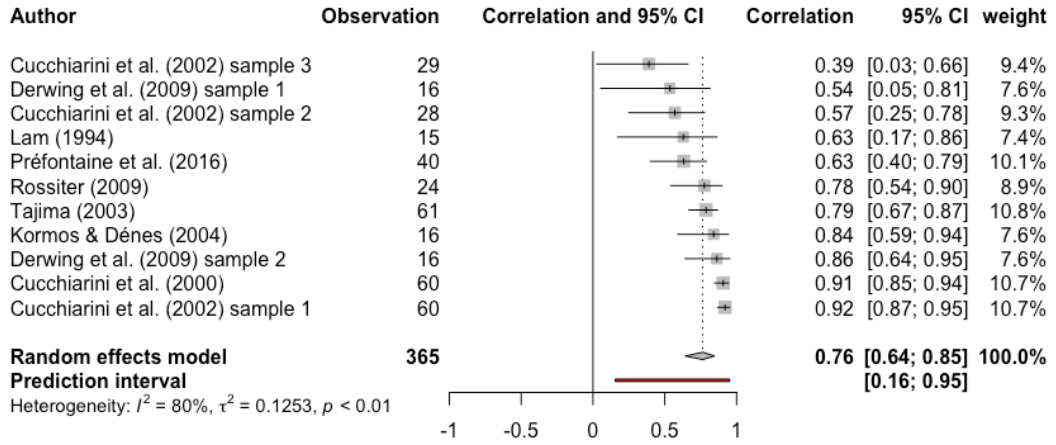


Figure 15. An overall average correlation between perceived fluency scores and speech rate (indicated by the diamond) and correlations with confidence intervals for each study.

5.2.2 Moderator analysis

5.2.2.1 Speech stimulus preparation

Three moderator variables related to speech stimulus preparation were examined (see Table 15). First, despite the non-significant difference between the subgroups ($Q(1) = 3.15$, $p = .076$), studies using entire speech as speech stimuli ($r = .59$) tended to demonstrate slightly higher correlation coefficients than those using excerpts of speech ($r = .50$). Second, I found significant effects of speaking task type on the correlation coefficients between UF and PF measures ($Q(3) = 7.91$, $p < .019$). A set of post-hoc Q tests revealed that effect sizes based on controlled production ($r = .74$) showed higher correlation coefficients than the other two

types of monologic speech (both *ps* < .01). There was no significant difference between closed tasks (*r* = .53) and open tasks (*r* = .51) in the size of the correlation coefficients (*Q*(1) < 0.01, *p* = .983). Third, I found a significant effect of target L2 on the UF-PF connection (*Q*(3) = 28.58, *p* < .0001). A series of post-hoc *Q* tests revealed that there were no significant differences among the subgroups of L2 Dutch, English, and French (*r* = .52–.61), while studies investigating fluency in L2 Japanese (*r* = .77) showed higher correlation coefficients than these three L2 subgroups (all *ps* < .001).

Table 15. *Results of analysis of categorical moderator variables related to speech stimulus.*

Moderator variable	*n*	*r*	*95%*	*Q(df)*	*p*
Target L2				28.58(3)	< .0001
Dutch	44	0.52	[.42, .62]		
English	65	0.54	[.47, .61]		
French	12	0.61	[.55, .67]		
Japanese	4	0.77	[.71, .81]		
Speaking task type				7.91(2)	0.019
Controlled speech	14	0.74	[.60, .83]		
Closed task	61	0.53	[.46, .59]		
Open task	50	0.51	[.43, .58]		
Speech sample				3.15(1)	0.076
Entire speech	65	0.59	[.52, .66]		
Excerpt	60	0.50	[.44, .57]		

Note. n = number of effect sizes.

5.2.2.2 Listeners' background

Regarding the moderator variables related to listeners' background, I examined the effects of listeners' experience (Experienced vs. Inexperienced raters) and language background (L1 vs. L2 speakers) on the UF-PF link (see Table 16). I found no significant effects of listener experience (*Q*(1) = 1.96, *p* = .162). Similarly, a heterogeneity test revealed that listeners' language background did not differentiate the strength of the association between PF and UF measures (*Q*(1) = 0.86, *p* = .355). However, comparing their ranges of 95% confidence

intervals, it should be noted that L1 raters (r = .56, CI[.51, .61]) indicated a narrower range of confidence intervals than L2 raters (r = .48, CI[.29, .64]).

Table 16. *Results of analysis of categorical moderator variables related to listeners' background.*

Moderator variable	*n*	*r*	*95%*	*Q(df)*	*p*
Experience				1.96(1)	0.162
Experienced	70	0.58	[.51, .65]		
Inexperienced	52	0.51	[.44, .58]		
Language background				0.86(1)	0.355
L1 raters	109	0.56	[.51, .61]		
L2 raters	16	0.48	[.29, .64]		

Note. n = number of effect sizes.

5.2.2.3 Rating procedure

According to the heterogeneity tests, none of the moderator variables of rating procedures showed significant effects on the correlation between PF and UF measures (see Table 17). As regards the definition of fluency presented to listeners, however, the category of research-based rubrics suggested a strong effect size (r = .67), while the remaining three categories indicated medium-to-strong effect sizes (r = .51–.59). I thus performed post-hoc Q tests and found that there was a significant difference only between the categories of research-based rubrics and researcher's definition ($Q(1) = 5.38$, p = .020).

Table 17. *Results of analysis of categorical moderator variables related to rating procedure.*

Moderator variable	*n*	*r*	*95%*	*Q(df)*	*p*
Definition of fluency for raters				6.52(3)	0.089
Researcher's definition	59	0.51	[.44, .57]		
No definition	47	0.57	[.47, .66]		
Existing assessment tools	10	0.59	[.49, .68]		
Research-based rubrics	8	0.67	[.55, .77]		
No of scale points				3.41(3)	0.333
5-point	23	0.58	[.44, .69]		
6-point	9	0.63	[.55, .70]		
9-point	58	0.53	[.46, .60]		

10-point	26	0.57	[.42, .69]		
Rater training				1.43(1)	0.232
Short practice	90	0.54	[.48, .60]		
Extensive training	6	0.66	[.47, .79]		

Note. n = number of effect sizes. Due to the limited number of effect sizes, the subgroup of 7-point scales (k = 1) was excluded from the moderator analysis of scale points.

5.2.2.4 Utterance fluency measure computation

With respect to moderator variables related to the selection and calculation of utterance fluency measures (see Table 18), I first examined the impact of speech annotation methods (manual vs. automatic annotation). A heterogeneity test did not reveal a significant difference of effect sizes between annotation methods ($Q(1) = 0.58$, $p = .448$).

Regarding silent pause duration measures, due to the limited number of subgroups of pause location (mid-clause pauses, k = 2; end-clause pauses, k = 1), I only conducted a moderator analysis on minimum pause length. The results revealed that there was no significant effect of pause length on the strength of the association with PF ($Q(1) = 1.93$, $p = .165$). However, it should be noted that effect sizes with a 250 ms threshold for silent pauses ($r = –.60$, CI[–.75, –.39]) can be considered strong, while those with a 200 ms threshold are regarded as medium in size ($r = –.41$, CI[–.59, –.19]).

As for pause frequency measures, moderator analyses were conducted for pause location, pause length, and pause type. As summarized in Table 18, despite the non-significant effect of pause location on the whole ($Q(2) = 4.25$, $p = .119$), the effect size of pauses within clauses was considered strong, while both categories of pauses between clauses and pauses including both locations were regarded as medium effect sizes. Post-hoc Q tests revealed that the frequency of pauses within clauses ($r = –.72$) tended to show higher correlation coefficients than that of pauses between clauses ($r = –.48$; $Q(1) = 4.01$, $p = .045$). Meanwhile,

the difference in effect sizes between pauses within clauses and those with both locations approached statistical significance ($Q(1) = 3.47$, $p = .062$). With regard to pause length, I did not find any significant effects on the correlation with PF scores ($Q(2) < .01$, $p = .999$). However, the range of confidence intervals of the subgroups suggested that the longer threshold of silent pauses tended to have a narrow confidence interval (e.g., 400 ms, $r = –.57$, CI[–.64, –.50] vs. 200 ms, $r = –.56$, CI[–.80, –.19]). In other words, pause length did not affect the predicting power of the measure for listener-based judgements of fluency, while the longer cut-off duration of silent pauses may enhance its stability. Furthermore, I found significant effects of pause type on the correlation coefficients between PF and UF measures ($Q(2) = 32.57$, $p < .0001$). A set of post-hoc Q tests demonstrated that the difference between silent pauses ($r = –.57$) and a combination of both silent and filled pauses ($r = –.47$) did not reach statistical significance ($Q(1) = 3.14$, $p = .076$), while filled pause frequency measures ($r = –.24$) showed significantly lower correlational coefficients than the other two subgroups (both $ps < .01$).

I also conducted a moderator analysis on frequency-based repair fluency measures in terms of the scope of target disfluency features. The results showed that the moderator effects of disfluency features did not reach statistical significance ($Q(2) = 1.29$, $p = .524$), while only the subgroup combining all types of disfluencies (Mixed) indicated a significant weak correlation ($r = –.22$, CI[–.33, –.10]).

Table 18. *Results of analysis of categorical moderator variables related to UF measure analysis.*

Moderator variable	*n*	*r*	*95%*	*Q(df)*	*p*
Speech annotation				0.58(1)	0.448
Manual coding	89	0.54	[.48, .59]		
Automatic annotation	34	0.59	[.48, .68]		

Mean pause duration

Pause location				–	–
Both	8	–0.42	[–.55, –.27]		
Within clauses	2	–0.71	[–.90, –.27]		
Between clauses	1	–0.63	[–.79, –.39]		
Pause length				1.93(1)	0.165
200ms	5	–0.41	[–.59, –.19]		
250ms	6	–0.60	[–.75, –.39]		
400ms	0	–			
Pause frequency					
Pause location				4.25(2)	0.119
Both	23	–0.55	[–.62, –.47]		
Within clauses	6	–0.72	[–.84, –.55]		
Between clauses	4	–0.48	[–.64, –.27]		
Pause length				0.00(2)	0.999
200ms	6	–0.56	[–.80, –.19]		
250ms	13	–0.57	[–.67, –.46]		
400ms	14	–0.57	[–.64, –.50]		
Pause type				32.57(2)	< .0001
Both	5	–0.47	[–.59, –.33]		
Filled	10	–0.24	[–.34, –.14]		
Silent	29	–0.57	[–.67, –.52]		
Disfluency rate					
Type of repair features				1.29(2)	0.524
Mixed	8	–0.22	[–.33, –.10]		
Repetition	3	–0.13	[–.45, .22]		
Self-correction	3	–0.08	[–.30, .13]		

Note. n = number of effect sizes.
he

5.3 Discussion

Despite the extensive examination of the UF-PF link in prior research, the relative strength of the predicting power for PF ratings between speed and breakdown fluency measures has varied across studies, and the contribution of repair fluency measures to PF has been inconsistently observed. A close examination of previous studies also suggested the potential moderator effects of methodological factors on the UF-PF link (see Section 3.5). Motivated by these issues, Study 1 meta-analysed the correlation coefficients between UF measures and

PF judgements and the moderator effects of methodological factors on the effect sizes. The inverse-variance weighted overall effect sizes for six major UF measures—articulation rate, pause frequency, pause duration, disfluency rate, mean length of run, and speech rate—were calculated, using a random-effects model with the Restricted Maximum Likelihood estimation method (RQ1-1). The moderator analysis of methodological factors was also conducted using subgroup analyses with random-effects modelling (RQ1-2). The following sections discuss the extent to which each dimension of UF contributes to the judgements of PF and how methodological features can affect listeners' selective attention to L2 speech in evaluations of fluency.

5.3.1 Overall predictive power of utterance fluency in perceived fluency

With the primary goal of quantifying the overall association strengths of different dimensions of UF with listener-based PF judgements (RQ1-1), I meta-analysed the correlation coefficients between six representative UF measures and PF scores. The results demonstrated strong effect sizes for speed fluency (r = .62) and composite measures (r = .72, .76). The strong predictive power of speed fluency and composite measures for PF judgements align with previous findings that these two measures distinguish performance at different levels of proficiency (e.g., Tavakoli et al., 2020). The results indicate that PF judgements in previous research tend to have been based on what Tavakoli and Hunter (2018) call a narrow definition of fluency. The fact that these composite measures explain a large variance in listeners' judgements suggests that they mostly regard fluency as "ease, flow and continuity of speech" (Tavakoli & Hunter, 2018, p. 343). However, a considerable proportion of variance in fluency judgements still remains unexplained after UF measures are accounted for (i.e., leftover variance ranges from 38.4 to 57.8%). Therefore, the results of the current meta-analyses suggest that listeners do not rely on 'very narrow' conceptualizations of

fluency or take only speed, breakdown, and repair features into account (cf. Tavakoli & Hunter, 2018). To some extent, listeners might also attend to linguistic aspects, such as lexis, grammar, and pronunciation.

As regards breakdown fluency measures, the effect sizes were stronger for pause frequency measures (r = –.59) than pause duration measures (r = –.46), indicating that listeners might pay more attention to the frequency of pauses than their duration. This finding is also supported by De Jong et al.'s (2013) findings showing that pause frequency is associated with a wider range of CF measures than pause duration. Contrary to the inconsistent relationship between repair fluency and PF in previous studies (e.g., Cucchiarini et al., 2002; Kormos & Dénes, 2004; Saito et al., 2018), the aggregated effect sizes in the current study demonstrated a significant but small effect size (r = –.20). These findings are in line with Tavakoli et al. (2020), who investigated the discriminatory role of breakdown fluency measures in the assessment of oral language proficiency. They also found that the frequency of repairs did not differ across levels of proficiency. Repair phenomena tend to be associated with speakers' L1 speaking style (Peltonen & Lintunen, 2016), and consequently they might serve as less reliable cues for listeners than speed, breakdown, and composite measures.

5.3.2 Moderator effects of methodological variables

Motivated by the results of heterogeneity tests as well as the review of literature, I conducted a set of moderator analyses to identify which methodological variables can moderate the UF-PF link. I examined 11 moderator variables, which arise in different phases of L2 UF-PF link research.

5.3.2.1 Speech stimulus preparation

Target L2. I observed medium-to-strong effect sizes in L2 Dutch, English, and French (r = .52–.61), while L2 Japanese showed a stronger effect size than these three languages (r = .77). One possible explanation for this difference may lie in cross-linguistic differences in phonological units. Dutch, English, and French are syllable-based, whereas Japanese is mora-based. The basic form of mora consists of one consonant and one vowel and typically ends with vowel sounds. Accordingly, consonant clusters between units are unlikely to occur in mora-based languages, and the length of basic units tends to be shorter in morae than in syllables (see Collins & Mees, 2003; Vance, 2008). Therefore, I might argue that there are less rhythmic variations in mora-based languages if temporal features (e.g., speed, pauses) are constant, compared to syllable-based languages. Building on these assumptions, fluency judgements of L2 Japanese might be less susceptible to suprasegmental features (stress, rhythm), and thus might be more closely aligned with objective measures of utterance fluency than those of the other three languages. Conversely, particularly in the context of syllable-based languages, such rhythmic/suprasegmental aspects might affect listener-based judgements of fluency (Kormos & Dénes, 2004; S. Suzuki & Kormos, 2020).

Task Type. A stronger effect size was found when speech stimuli were elicited through controlled production tasks (r = .74) than through spontaneous speech tasks, including closed and open tasks (r = .53, .51, respectively). One possible reason for the higher correlation coefficients in controlled speech might be because in controlled speech, there is virtually no variation in content and linguistic expression (e.g., vocabulary, grammar), whereas in spontaneous speech, the content and linguistic forms vary across speakers due to the open-ended nature of tasks. Therefore, when judging the fluency of controlled speech, due to the

lack of such variation in content and form, listeners' attention might exclusively focus on temporal features.

With regard to spontaneous speech tasks, results showed that the UF-PF link might be less influenced by the predefined nature of the content of speech. This finding should be interpreted carefully. Prior research has consistently reported the effects of task design features, suggesting that L2 learners tend to produce more fluent speech in closed tasks than in open ones (for a review, see Tavakoli & Wright, 2020). In other words, UF is supposed to differ between closed and open tasks. However, in terms of the association to PF, such differences in UF tend to disappear. This is possibly because despite different UF performance across task types, listeners may intuitively and flexibly adjust their perceptions about the extent to which utterance features reflect the speaker's CF, according to the speaking context and task (Segalowitz, 2010, 2016). As a result, the association between PF judgements and UF characteristics may tend to be consistent between closed and open tasks. Alternatively, in previous studies, listeners might have been able to predict the content of speech, even when elicited from open tasks. First, it may be possible that open tasks elicit similar speech samples across speakers, as their topic is generally predetermined by task instructions. Second, researchers usually familiarize listeners with the topic and/or discourse of open tasks to avoid familiarity bias (Rossiter, 2009).

Length of Speech Stimuli. Although the *Q*-test showed that effect sizes did not differ between short excerpts and entire speech ($Q(1) = 3.15$, $p = .076$), the difference in the effect sizes between these two types of speech stimuli may be considered meaningful. The effect sizes of entire speech were virtually large ($r = .59$), while those of excerpts were medium ($r = .50$). Entire speech samples might provide listeners with more information for judgements than

excerpts. As raters can listen to the complete discourse and are exposed to longer input, their subjective perceptions might align better with the objective temporal features of speech. In sum, either type of stimulus might be used, but for the sake of more valid assessment (e.g., language assessment contexts), entire speech may be a better choice for fluency judgements (Isaacs & Thomson, 2013).

5.3.2.2 Listeners' background

I examined the moderator effects of two major variables of listeners' background—experience and language background. Although the *Q*-tests revealed that neither of the moderator variables differentiated effect sizes, the aggregated effect sizes were substantively different between the subgroups. Regarding experience, the effect size of experienced raters was virtually large (r = .58), while that of inexperienced raters was medium (r = .51). The slightly closer alignments of fluency judgements with temporal features in experienced raters may be in line with Rossiter's study (2009), in which novice and expert raters tended to pay attention to different temporal features, despite similarities in the severity of judgements. Moreover, in the context of holistic assessment of speaking, professional raters tend to be more sensitive to variability in temporal features when it comes to less fluent speech (Duijm et al., 2018). For a better understanding of the role of experience in PF judgements, the effects of rater experience should be more carefully examined with reference to the overall level of UF. As for language background, a relatively wider 95% confidence interval in the group of L2 raters (r = .48, 95%CI[.29, .64]), compared to L1 raters (r = .56, 95%CI[.51, .61]), indicated that correlation coefficients tend to be more stable in the context of L1 raters. However, a variety of factors may underlie the distinction between L1 and L2 raters. Therefore, it is still unclear what individual difference variables, such as proficiency

and L2 learning experience, contribute to L2 raters' variability in the UF-PF link (for the dynamicity of L2 listeners, see Magne et al., 2019; Saito et al., 2019).

5.3.2.3 Rating procedure

Definition of Perceived Fluency for Raters. Although differences in the definitions of fluency presented to raters did not reach statistical significance (p = .089), I found a significant difference between research-based rubrics (r = .67) and semantic scales with researchers' definitions (r = .51). In the pooled studies, research-based rubrics were either created based on qualitative data obtained in the study (Sato, 2014) or adapted from prior work (Nitta & Nakatsuhara, 2014), and thereby they might demonstrate higher construct validity. The studies classified in this category also adjusted the number of scale points according to the proficiency level of their participants. Therefore, a strong effect size might be derived from this type of adjustment to the rating scale for the target population.

Number of Scale Points. Non-significant results for the number of scale points indicate that the association of listeners' perceptions of fluency with temporal features tends to be consistent, regardless of the number of scale points. The current finding is consistent with prior research (Isaacs & Thomson, 2013). However, considering the preceding potential advantage of adjusting scales of rubrics, it is recommended that an appropriate number of scale points should be decided by taking the range of speakers' proficiency into account.

Rater Training. Although the difference between the two subgroups of rater training did not reach statistical difference, the effect size of the subgroup of extensive training (r = .66) can be considered strong, while that of the subgroup of short practice was medium (r = .54). Considering the possibility that the non-significant difference may have derived from the

small number of effect sizes in the subgroup of extensive training (n = 6), the difference in the effect sizes between extensive training and short practice can be considered meaningful. This finding suggests that the length/amount of rater training may enhance the influence of temporal correlates on fluency judgements. Due to the broad category of extensive training in the current study, it is, however, still unclear what type of rater training would significantly increase the association between UF and PF measures.

5.3.2.4 Utterance fluency measure computation

Speech Annotation Method. The moderator analysis revealed that effect sizes tend to be comparable between manual and automated speech annotation methods when calculating UF measures. This finding is remarkable, because the correlation coefficients between manual and automated annotation methods were reported to fall between .70–.80 (De Jong & Wempe, 2009). In other words, when using automated annotation methods, correlations with PF scores could be expected to be somewhat lower, compared to manual annotations. Accordingly, the variance in PF scores explained by UF measures should not be identical across the two annotation methods. However, the current results indicate that automated speech annotation may sufficiently capture temporal features related to the establishment of perceptions of fluency. Therefore, the results provide additional evidence for the predictive validity of automated speech annotation in PF.

Location of Pauses. Due to the limited number of effect sizes in pause duration measures, I conducted moderator analysis of pause location only for pause frequency measures. There were no significant effects of pause location, possibly due to the small number of subgroups (e.g., n = 4 for pauses between clauses). However, a similar pattern of pause location effects was demonstrated in both pause measures, showing the highly strong effect sizes for the

category of pauses within clauses (r = –.71 for pause duration, r = –.72 for pause frequency). Meanwhile, the remaining categories were regarded as showing medium-to-strong effect sizes (r = –|.42–.63|). From the perspective of L2 speech production, pauses within clauses tend to reflect disruptions in linguistic encoding processes, such as lexical retrieval and syntactic procedures (De Jong, 2016b; Kormos, 2006). Therefore, the findings suggest that listeners' perceptions of fluency are established using pause location as an important cue for speakers' efficiency in L2 speech production (i.e., CF).

Length of Pauses. The moderator analysis revealed that the minimum threshold for silent pauses did not moderate the correlation coefficients between either pause measure with PF scores. Particularly in the case of pause frequency measures, the effect sizes of three categories (200 ms, 250 ms, 400 ms) were virtually identical (r = –|.56–.57|). However, the association of pause duration measures with PF might be enhanced with a threshold of 250 ms (r = –.60), compared to 200 ms (r = –.41). This tendency indicates that the inclusion of pauses shorter than 250 ms may lower the predictive power of pause duration measures for listeners' judgements of fluency. These findings support 250 ms being a threshold for silent pauses, which has been regarded as common practice in L2 fluency research (Bosker et al., 2013; De Jong & Bosker, 2013).

Type of Pauses. The effect size of silent pauses approached strong (r = –.57), while that of filled pauses was regarded as being weak (r = –.24). Possibly due to the weak predictive power of filled pauses, I found a medium effect size when combining both filled and silent pauses (r = –.47). From the perspective of speech production mechanisms, both filled and silent pauses are assumed to reflect breakdowns in speech production processes (Kormos, 2006; Segalowitz, 2010) and the time needed to handle such disruptions (Bui et al., 2019).

However, the current findings suggest that listeners may not always perceive filled pauses as an indication of disruption in speech production. The weak role of filled pauses in PF may be due to the fact that filled pauses can provide listeners with the impression of continuation of speech rather than breakdowns (Clark & Fox Tree, 2002).

Selection of Disfluency Features. The moderator analysis failed to detect significant moderator effects of the focus of disfluency features. Furthermore, the aggregated effect sizes within the subgroups of repetition and self-correction did not reach statistical significance. Meanwhile, the subgroup of disfluency measure which counts all kinds of disfluency features (Mixed) suggested a significant but weak effect size ($r = –.22$). The unstable predictive power of separate disfluency features may be due to the methodological difficulty in categorizing disfluency features reliably (see Kormos, 2006). It is also possible that while the frequency of one specific type of disfluency feature might not be sufficient to negatively impact on listeners' perceptions, the joint overall frequency of these features may lower subjective ratings of fluency.

5.4 Summary

This chapter reported the results and discussion of Study 1 which examined the relationship between PF and UF. Motivated by the fact that the UF-PF link has been relatively extensively examined in the literature of L2 fluency, Study 1 meta-analysed the correlation coefficients between UF measures and PF judgements (263 effect sizes from 22 studies) with respect to the potential moderator effects of methodological factors on the UF-PF link. The findings are visualized in Figure 16.

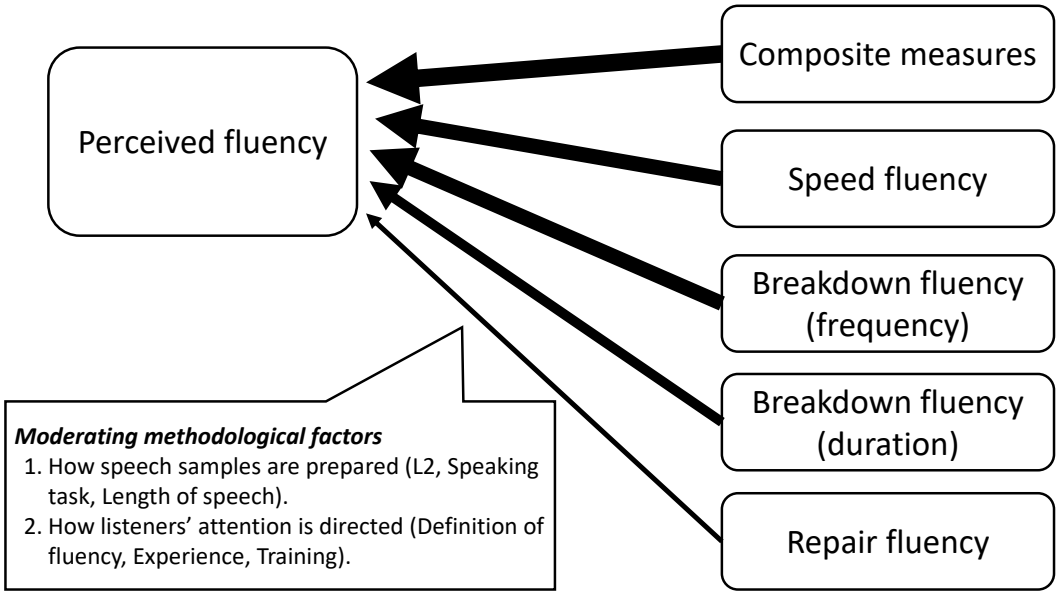


Figure 16. The visualized summary of the findings of Study 1.
Note. The width of the arrows indicates the effect sizes.

RQ1-1 was concerned with the aggregated overall effect sizes of correlation coefficients between UF measures and listener-based PF judgements. The inverse-variance weighted overall effect sizes for six major UF measures—articulation rate, pause frequency, pause duration, disfluency rate, mean length of run, and speech rate—were calculated, using a random-effects model with the Restricted Maximum Likelihood estimation method. The results of effect size aggregations showed that PF was strongly associated with speed and pause frequency (r = .62, .59, respectively), moderately with pause duration (r = |.46|), and weakly but significantly with repair fluency (r = |.20|), while composite measures showed the strongest effect sizes (r = |.72–.76|). These findings suggest that all of the subdimensions of UF were associated with PF ratings, while speed fluency and pause frequency aspect of breakdown fluency tended to have a strong predictive power for PF to an equal extent. However, pause duration aspect of breakdown fluency was likely to have a relatively lower predictive power than those two dimensions of UF. Furthermore, repair fluency in general was found to significantly contribute to PF judgements.

RQ1-2 examined the moderator effects of 11 methodological factors, which were identified through the synthesis of previous studies. A series of moderator analyses was conducted with the help of subgroup analyses with random-effects modelling, using those methodological factors, including speech stimulus preparation, listeners' background, rating procedure, and UF measure analysis. The results revealed that the UF-PF link was moderated by target L2 (syllable- vs. mora-based language), task type (controlled vs. spontaneous speech), and the definition of fluency presented to raters (research-based rubrics vs. semantic scales with researchers' definition of fluency). These findings indicate that listeners' selective attention to temporal features of speech can be influenced by the suprasegmental characteristics of the target language, the variability of language use in speech, and the correspondence between rating tools and listeners' selective attention. In addition, the moderator analyses did not show significant effects of listeners' background on the UF-PF link, indicating that listeners may tend to perceive L2 speech in a relatively consistent manner, regardless of their language status and assessment-related experience. Finally, due to the limited number of effect sizes, the subgroup analyses regarding UF measure analysis factors showed that only pause type (silent vs. filled pauses) made a significant difference in the effect sizes, indicating the larger role of silent pauses in PF judgements. Despite the non-significant difference, breakdown fluency measures specific to mid-clause pauses, however, consistently suggested very strong effect sizes (r = -.71 for pause duration, r = -.72 for pause frequency).

Chapter 6: Methodology for Study 2, Study 3, and Study 4

6.1 Introduction

The current chapter illustrates the methodologies for Studies 2–4. Motivated by the role of PF as the subjective judgements of CF, Study 1 meta-analysed the correlation coefficients between listener-based judgements and objective measures of fluency. The results revealed that all the sub-dimensions of UF—speed, breakdown, and repair fluency—were weakly to strongly associated with listeners' perceptions of fluency. To further clarify the association between CF and UF, the three subsequent studies (Studies 2–4) examine the relationship between L2-specific CF and UF, using three different methodological approaches. Study 2 aims to identify which UF measures are related to different components of speech production processes, comparing UF performance across four speaking tasks which differ in the quality of speech processing demands. Study 3 examines the contribution of L2-specific CF to UF measures and its variability across the four speaking tasks. In Study 3, students' CF was measured through a set of psycholinguistic tests capturing each component of speech production mechanisms (e.g., lexical retrieval speed, sentence construction skills). Finally, assuming that language-general speech processes are shared across learners' L1 and L2 speech production, Study 4 investigates the extent to which L2 UF measures can be predicted by L1 UF measures and also the extent to which the L1-L2 UF association is moderated by L2 proficiency. This chapter begins with the research questions of these studies (Section 6.2), followed by the description of the research context and participants (Sections 6.3–6.4), the ethical procedures (Section 6.5), the speaking tasks (Section 6.6), and the measurements of UF and CF (Sections 6.7–6.9). In addition, the procedures of data collection (Section 6.10) and statistical analyses (Section 6.11) are also illustrated.

6.2 Research Questions

For a better understanding of the construct of L2 oral fluency, prior research has investigated the interrelationship between CF, UF, and PF. Although previous studies have extensively examined the relationship between PF and UF, it is still unclear the extent to which CF contributes to UF (for rare exceptions, De Jong et al., 2013; Kahng, 2020). In addition, a closer look into the literature of L2 fluency research suggested that the contribution of CF to UF can vary, according to the processing demands of speaking tasks (cf. De Jong et al., 2013). Therefore, Study 2 investigates what UF measures are relevant to the demands on conceptualization, the enhanced activation of linguistic representations, and that of phonological representations. These manipulations of speech processing demands were operationalized by adjusting several task design features, including content generation and the availability of relevant linguistic items for speech. Due to the relatively exploratory nature of the research methodology, different speech processing demands across tasks in the current study were cross-validated by students' perception data. Study 2 addresses the following research question:

RQ2. How does L2 utterance fluency performance vary across four types of speaking tasks which differ in the speech processing demands on conceptualization, the activation of linguistic representations, and the activation of phonological representations?

Study 3 examines the relationship between CF and UF in relation to the potential moderator effects of task type on the CF-UF link. Following De Jong et al. (2013) and Kahng (2020), this study operationalized CF as a set of linguistic resources and processing skills involved in speech production. Each subdimension of UF, that is, speed, breakdown, and repair fluency,

was also measured, using a set of existing UF measures identified in Study 1. Furthermore, to examine the variability of the CF-UF link across speaking tasks, Study 3 employed four speaking tasks theoretically differing in speech processing demands. The study is guided by the following research questions:

RQ3-1. To what extent do components of cognitive fluency contribute to subdimensions of utterance fluency?

RQ3-1a. What is the relationship between cognitive fluency measures of lexical, grammatical, and pronunciation knowledge?

RQ3-1b. What is the relationship between utterance fluency measures of speed, breakdown, and repair fluency?

RQ3-2. To what extent is the CF-UF link (RQ3-1) moderated by speaking tasks differing in the quality of speech processing demands?

Study 4 delves into the extent to which L2 UF measures are predicted from the corresponding L1 UF measures. Although prior research has shown that the L1-L2 UF link is cross-linguistically robust (Bradlow et al., 2017), previous studies were largely limited to the picture and personal narrative tasks and also to the combination of stressed-/syllable-timed languages (e.g., Dutch, English, French, and Spanish). To further explore the generalizability of the L1-L2 UF link, Study 4 examines the L1-L2 UF link in the context of argumentative speech produced by L1 Japanese (mora-timed language) learners of English (stress-timed language). In addition, a synthesis of previous studies on the L1-L2 UF link (see Section 3.9) suggested that the strength of association between L1 and L2 fluency can be weakened or enhanced by L2 proficiency. To examine the moderator effects of L2 proficiency on the L1-

L2 UF link, Study 4 was to use factor scores based on the factor structure of CF from the results of Study 3. The following research questions were formulated:

RQ4-1. To what extent are L2 utterance fluency measures predicted from the corresponding L1 utterance fluency measures?

RQ4-2. To what extent are the L1-L2 fluency links of different aspects of utterance fluency (RQ4-1) moderated by L2 proficiency?

6.3 Research Context

The current research project recruited Japanese learners of English as participants at a private university in Tokyo, Japan. I decided to approach this population of L2 learners mainly concerning several demographic characteristics: the pair of L1 and L2, L2 proficiency level, location of study, and school type. Regarding the pair of L1 and L2, for the sake of comparability with previous studies, Study 4 aims to examine the L1-L2 UF link with L2 English learners with mora-timed language as their L1, that is, the Japanese language in the current thesis. Similarly, the research into the CF-UF link has also lacked the empirical studies in the context of L1 Japanese learners of English (cf. De Jong et al., 2013; Kahng, 2020). I thus decided to approach Japanese-speaking learners of English. To this end, due to the fact that the Japanese language is mostly spoken only in Japan, I also decided to recruit participants in Japan.

As for L2 English proficiency, it can be assumed that it is relatively difficult for Japanese learners to develop their English proficiency, because in Japan, they do not have plenty of opportunities to use English outside the classroom (i.e., English-as-a-Foreign-Language [EFL] country). The Japanese government has not surveyed English proficiency of Japanese

college students, probably because higher education in Japan is not included in the compulsory education. Meanwhile, according to the government's survey on secondary students' English proficiency (MEXT, 2017), at the end of Japanese secondary school, only 11.7% of students reached the A2-level on the CEFR scale of speaking, while the proficiency level of the remaining students (87.2%) was at the A1-level. This narrow range and lower levels of oral proficiency may be problematic for the current thesis, because Studies 2–4 use correlational and regression analyses which are based on the variance of variables. In other words, a relatively wide range of oral proficiency may increase the robustness of the research findings of the thesis. Accordingly, the later stage of education, that is, a higher education level, was selected for the target population of the current research project.

Previous research conducted in similar contexts (i.e., Japanese university students) indicated that university students, particularly those who voluntarily participated in the research, have relatively higher proficiency levels, ranging from the B1 to C1 level (Saito, 2017) and the A2 to C1 level (Saito, Suzuki, et al., 2019; S. Suzuki & Kormos, 2020). These previous studies recruited their participants in a private university located in Tokyo (the capital city of Japan). One of the possible reasons for a relatively wide range of L2 proficiency levels (especially, upper levels of proficiency) in these previous studies might be the socio-economic status of families of the students. In Japan, private universities, particularly in Tokyo, tend to charge higher tuition fees than public universities. Therefore, the families of students in private universities are likely to have high socio-economic status and thus can afford to offer extracurricular English learning opportunities for their children. Following these previous studies, I selected Tokyo for the location of study and chose a private university for recruiting the participants for the current research project.

6.4 Participants

To reach an adequate statistical power for the multivariate statistics (SEM for Study 3, Generalized mixed-effects modeling for Studies 2 and 4), the minimum number of sample size was determined with regard to the number of observed variables for the planned SEM models. Although the optimal sample size for the SEM analysis may vary according to different statistical factors (e.g., distributions of variables, estimation methods; see Kyriazos, 2018), I followed the ratio of the sample size to the number of variables as a rule of thumb. Traditionally, the optimal ratio for confirmatory factor analyses (CFA) and SEM can range from five to ten (for a review, see Kyriazos, 2018). As a total of 20 observed variables (11 UF measures and 9 CF measures) was predetermined for Study 3 (see Sections 6.7–6.8), the minimum number of sample size was $N = 100$.

A total of 128 Japanese learners of English voluntarily participated in the current research project. Despite a slightly higher proportion of female participants, the gender balance among the participants was generally equal (Female = 73, Male = 55). Their ages ranged from 18 to 27 years ($M_{age} = 20.43$, $SD_{age} = 1.81$). Regarding their L2 proficiency, their self-reported university placement test scores suggested that most of them could be placed on the B1–B2 levels of the CEFR scale, while some of them seemed to reach the C1-level (for the frequency of each level, see Figure 17). According to the abovementioned government's report (MEXT, 2018), the current group of participants might be more proficient in L2 English than the overall population of Japanese university students. The slightly higher proficiency of the current participants might be due to the fact that they were recruited in a private university and on the basis of voluntary participation, which might have attracted those who were relatively confident of their English skills. Most participants had no overseas experience such as study abroad, while 31 students out of 128 had an experience of residing

in English-speaking countries for longer than one month (M_{month} = 18.6, *Mdn* = 10.0, *Range* = 1–78). However, most of them (*n* = 19) were immersed in English-speaking countries for shorter than one year (e.g., language exchange programs). The current group of students can thus be regarded as a group of L2 English learners whose dominant language is Japanese, with varying proficiency levels ranging from low-intermediate to advanced.

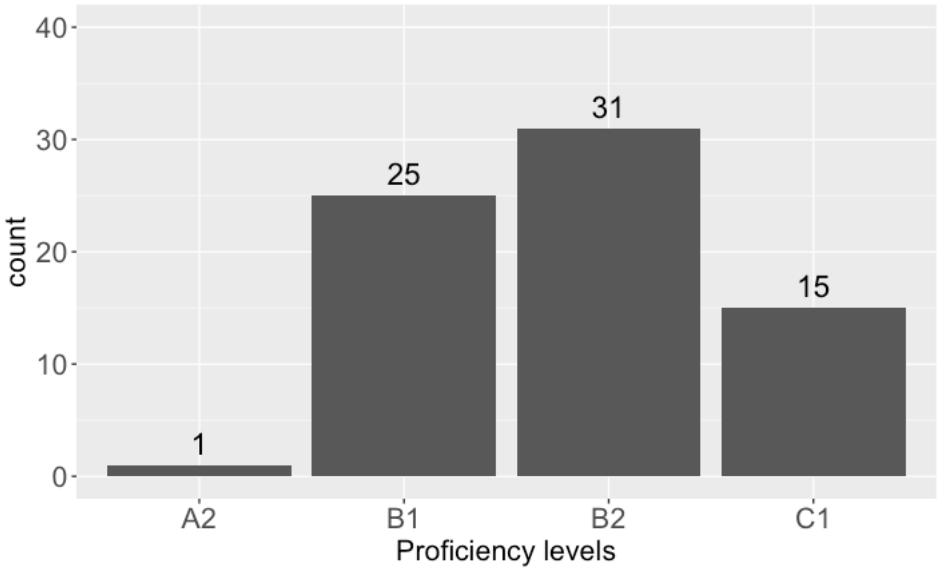


Figure 17. The histogram of proficiency levels judged by the self-reported scores of university placement tests.
Note. Seventy two out of 128 students reported their placement test scores.

Studies 2–4 were conducted using the same dataset, which included a set of CF and UF measures from 128 Japanese learners of English. The dataset was completed in three rounds of data collection, with the first round (January 2019) as the pilot study to check whether the materials are appropriate for the participants from the same population (Japanese learners of English). The pilot data (*n* = 24) confirmed that all the research materials had appropriate psychometric characteristics and were suitable for the target sample. However, I added one additional argumentative speech task in the second and third rounds of data collection (May to June 2019 and November to December 2019, respectively), due to some potential confounding factors (e.g., unexpected familiarity with the topic of the first argumentative

task). Although there were no clear effects of such confounding factors (for the effects of topic, see Appendix B), I decided to include the data of the additional argumentative task for Study 4 to examine the L1-L2 fluency link across different topics. To sum up, Studies 2 and 3 were conducted with the entire dataset (n = 128), while Study 4 was based on the data from the second and third rounds of data collection (n = 104).

6.5 Ethics

The ethical approval for Studies 2–4 was granted by the Faculty of Arts and Social Sciences and Lancaster Management School's Research Ethics Committee of Lancaster University in January 2019. To comply with the ethical guidelines by the university, the electronic version of an information sheet describing the purpose and procedure of the study was provided via email when participants applied for the participation. In addition, at the beginning of the data collection session, participants were given the printed version of the information sheet and an informed consent form (Appendix C; see also Section 6.10).

6.6 Speaking Tasks and Procedure

Study 2 aims to investigate the variability of L2 UF performance across speaking tasks differing in speech processing demands. Study 3 attempts to examine the moderator effects of speaking task design on the relationship between CF and UF. Accordingly, the current research project operationalized task design features based on the framework of speech processing demands (e.g., Préfontaine & Kormos, 2015; Skehan, 2009). To extend this line of research, the current research project focused on the different speech processing demands on conceptualization, and the enhanced activation of two different types of representations in formulation (linguistic and phonological representations), as summarized in Table 19. To manipulate these speech processing demands, the study employed four speaking tasks which

were designed to differ in the quality of processing demands for speech production: (a) an argumentative speech task, (S. Suzuki & Kormos, 2020), (b) a related picture narrative task (Préfontaine & Kormos, 2015), and (c–d) text summary tasks with and without read-aloud assistance (RAA) (cf. Košak-Babuder, Kormos, Ratajczak, & Pižorn, 2019). For the actual prompts, see Appendix D–F.

Table 19. *Summary of the contrast of speaking tasks in relation to different speech processing demands.*

Target speech processing demands	Target features	Contrast
Conceptualization	Whether speakers need to plan the content of speech	Argumentative Picture narrative
Linguistic representations (Formulation)	Whether the relevant linguistic items are activated by task	Picture narrative TS without RAA
Phonological representations (Formulation)	Whether the phonological form of the linguistic items is activated by task	TS without RAA TS with RAA

Note. TS = Text summary task; RAA = Read-aloud assistance.

The first focus was the speech processing demands on conceptualization. The major role of conceptualization in speech production is the process of planning the content of speech (i.e., macroplanning; see Section 2.5). To test the effects of conceptualization demands on speaking performance, I compared the argumentative and picture narrative tasks (see Section 3.8). The argumentative task required the speakers to conceptualize the content of speech and its order of presentation, whereas in the picture narrative task, students retold the predefined content based on a given visual prompt. The prompt of the argumentative task was adopted from S. Suzuki and Kormos (2020). In this task, students were initially provided with a statement—*The Tokyo Olympics in 2020 will bring economic growth to Japan*—and then instructed to argue how far they agree with the statement (for the prompt, see Appendix D).[10] The reason why this topic was selected was that I intended to select a topic which was related and meaningful to speakers themselves but simultaneously abstract to some extent so that

[10] Note that all the speech data were collected before the outbreak of COVID-19; thus, the topic of the argumentative speech was meaningful to participants at the time of the data collection.

participants would be required to connect different pieces of information logically to make a coherent argument. To reduce the demands on conceptualization in the picture narrative task, the information that needs to be communicated and its order of presentation were visually described in a clear time sequence so that learners did not need to plan a substantial part of the content of speech. I selected Préfontaine and Kormos' (2015) picture cartoon as the prompt for the condition with lower demands on conceptualization. The cartoon prompt consists of 11 scenes, describing the story where a businessman had started a house-gardening for pleasure, but gradually his identity as a businessman drove himself to expand it into farming business. The transition between scenes is predictable (Préfontaine & Kormos, 2015), and for this reason, the demands on the coherence of events could be considered low.

The second focus was the speech processing demands on formulation. The condition of reduced formulation demands was operationalized by activating the linguistic representations of relevant items prior to speech. It can be assumed that the activation of the linguistic items can assist L2 learners to retrieve those linguistic items efficiently with fewer attentional resources (i.e., priming effects; see Section 3.8). The current study adopted a text summary task, where speakers first read a short expository text written in L2 English (adapted from Millington, 2019) and then were asked to summarize the text. Since the source text provides students with some relevant linguistic items before they start speaking, it can be assumed that the activation of relevant linguistic items is relatively high, compared to the situation without such a source text. In order to avoid the situation where speakers simply read aloud the source text at the phase of speaking performance, they were not allowed to refer to the text while speaking. The text summary task in the current thesis is characterized by three components: information transfer as the task requirement, the predefined content of speech, and the activation of relevant/necessary linguistic items prior to speech. The first two features

are shared with the picture narrative task, so that the effects of the enhanced activation of linguistic representations could be compared between the picture narrative and text summary tasks. The picture narrative task requires participants to retrieve linguistic items, including vocabulary and grammar, relying on their own knowledge resources without any prior linguistic input. Meanwhile, the text summary task allows participants to utilize some linguistic items, which are embedded in the source text, in their speech with relatively low cognitive demands for retrieval due to the enhanced memory trace. The facilitative role of enhanced memory traces in linguistic retrieval has been suggested by psycholinguistic research on priming effects, which has demonstrated the increased probability and efficiency of selecting a previously activated linguistic item from the input (for a review, see McDonough & Trofimovich, 2008).

Finally, to examine the effects of the enhanced activation of phonological representations of linguistic items, two conditions were provided for the text summary task, that is, *with* versus *without* RAA. With the RAA, L2 speakers read the text while simultaneously listening to the aural recording of the text (i.e., bimodal input). Thus, the two conditions (with vs. without RAA) are assumed to differ in the degree of the activation of phonological representations of linguistic forms in the text during the reading phase (see Košak-Babuder et al., 2019; Liu & Todd, 2014). Comparing text summary speech across these two conditions, the study examined how the enhanced activation of phonological representations of linguistic items by RAA can contribute to UF performance in the text summary task.

As mentioned above, the text summary tasks in Studies 2 and 3 included two elements: source texts and their recordings for RAA. To prepare the source texts for this task type, multiple expository texts were pooled from the Dreamreader.net (Millington, 2019), which is

designed by an experienced EFL instructor based in Japan. Then, to select a pair of two texts (one for each RAA condition), the pooled texts were analysed in terms of lexical complexity and readability as well as text length. Regarding lexical complexity, the frequency levels of words in the texts were examined according to the JACET8000 wordlist which is specifically tailored for Japanese learners of English. Considering that the target population of participants were university students, low-frequency vocabulary items at Level 5 (the upper-intermediate level of university) or above in the list were replaced with synonyms at Level 4 or below (the beginning level of university). Collocational accuracy was checked by a native speaker of English. The readability of the texts was evaluated using the Flesh-Kincaid Reading Ease values which were calculated by the Coh-Metrix software (McNamara et al., 2014). An overall linguistic complexity score was also derived from the TextEvaluator® software (Educational Testing Service, 2013) to select a pair of comparable texts. The textual characteristics of the two selected texts are summarized in Table 20. Although both of these texts were written in an expository genre and had a comparable lexical difficulty, the topic of each text was different; one text (Text A) was about the history of the national flag of the United States, while the other text (Text B) illustrated the history of the flag of the International Committee of the Red Cross. Accordingly, the topic difference across texts might also affect UF performance in the text summary tasks. To avoid the interaction effects between conditions and source texts, two different texts were prepared, and the combination of the order of conditions and the source texts was counterbalanced across participants.

The aural recordings for RAA were recorded by a L1 Canadian English speaker who had 15-year teaching experience of English at universities in Japan. Regarding the speed of the recordings, following Košak-Babuder et al. (2019), I adopted a delivery rate of 120 words per minute, considering the relatively lower level of proficiency of the participants. The

comparability of the delivery speed of recordings was also ensured across texts (see Table 20).

Table 20. *Textual characteristics of two source texts for text summary speech.*

	Text A	Text B
Topic	US flag	Red Cross
Flesh-Kincaid value	71.21	64.79
TextEvaluator® score	380	660
Length in words	324	303
Speed of delivery (words/min)	116.4	119.6

To control for the effects of other task implementation factors on speaking performance, three minutes were consistently provided for pre-task planning to all of the four tasks. Note-taking during the planning time was not allowed for any of the tasks.

To cross-validate the operationalization of different speech processing demands from the students' perspective, a post-speaking questionnaire was employed in the final round of data collection (n = 40). The questionnaire consisted of two parts. The first part asked the participants to retrospectively judge the extent to which they experienced difficulty in five aspects of speech processing–conceptualization, lexical encoding, syntactic encoding, morphological encoding, and articulation–using a six-point scale (1 = *Very easy*, 6 = *Very challenging*). The questionnaire items are presented in Table 21 below. The second part of the questionnaire asked participants to briefly report what they found difficult when they were speaking.

Table 21. *Summary of target constructs and questionnaire items of post-speaking questionnaire.*

Target construct	Questionnaire item (1 = *Very easy*, 6 = *Very challenging*)
Conceptualization	*To think of what to speak (i.e., the content of speech, NOT language use)*
Lexical encoding	*To recall vocabulary and expressions*
Syntactic encoding	*To build sentences with the appropriate order of words*

Morphological encoding	*To choose the appropriate form of words (e.g., tense and plurals)*
Articulation	*To speak with English-like pronunciation*

6.7 Utterance Fluency Measures

Following prior research into L2 UF, Studies 2–4 target three major aspects of UF—speed fluency, breakdown fluency, and repair fluency (Tavakoli & Skehan, 2005). As for speed fluency, there is one measure that solely taps into the construct of speed fluency, that is, *articulation rate*, or its inversed measure, *mean duration of syllables* (Tavakoli et al., 2020). However, since Study 3 aims to examine the construct of UF at the level of latent variables using an SEM approach, UF measures were used to build the measurement model of UF (see Section 6.11). Statistically speaking, more than two observed variables are ideally loaded on the latent variable to avoid an under-identified model (Brown, 2006; for details, see Section 6.10). To this end, in addition to articulation rate, I decided to include two composite measures that mainly capture speed fluency: *speech rate* and *mean length of run*. These two composite measures have been commonly used in previous studies (see Section 4.5). It is thus worthwhile to examine what linguistic knowledge and processing skills underlie these measures (Study 3), because such information may serve as some validity evidence for their use in assessment contexts.

Regarding breakdown fluency, the studies used a fine-grained set of pause-related measures in response to the multidimensional nature of pausing behaviour, that is, *frequency*, *duration*, *type*, and *location* (De Jong, 2016b; Kahng, 2018; S. Suzuki & Kormos, 2020; Tavakoli, 2011). Moreover, care has also been taken to avoid theoretical collinearity among measures (i.e., theoretical overlap; Bosker et al., 2013).

As regards repair fluency, recent studies have shown that distinct cognitive processing might underly different disfluency phenomena, such as self-corrections (Kormos, 1999a, 2000), false starts (Williams & Korko, 2019), and filled pause (Fraundorf & Watson, 2014).[11] Therefore, the current studies also measured these disfluency phenomena separately: self-repair, false starts, and self-repetitions. The selected UF measures are listed below:

Speed fluency

1. *Articulation rate (AR)*. The mean number of syllables produced per second, divided by total phonation time (i.e., total speech duration excluding pauses)

Composite measures

2. *Speech rate (SR).* The mean number of syllables produced per second, divided by total speech duration time, including pauses
3. *Mean length of run (MLR)*. The mean number of syllables produced in utterances between pauses

Breakdown fluency

4. *Mid-clause pause ratio (MCPR)*. The mean number of silent pauses *within* clauses, divided by the total number of syllables produced
5. *End-clause pause ratio (ECPR)*. The mean number of silent pauses *between* clauses, divided by the total number of syllables produced
6. *Filled pause ratio (FPR)*. The mean number of filled pauses, divided by the total number of syllables produced
7. *Mid-clause pause duration (MCPD)*. Mean duration of pauses *within* clauses
8. *End-clause pause duration (ECPD)*. Mean duration of pauses *between* clauses

[11] A filled pause is one type of disfluency phenomena. However, L2 fluency research has traditionally proposed that filled pauses are categorized into breakdown fluency. Accordingly, the current study categorizes the measure of filled pause ratio into breakdown fluency.

Repair fluency

9. *Self-correction ratio (SCR)*. The mean number of self-correction behaviours, divided by the total number of syllables produced
10. *False start ratio (FSR)*. The mean number of false starts/reformulations, divided by the total number of syllables produced
11. *Self-repetition ratio (SRR)*. The mean number of self-repetitions, divided by the total number of syllables produced

All the speech data were transcribed and then annotated for the boundaries of clauses. To minimize the collinearity across different constructs of UF, temporal features for breakdown and repair fluency were standardized by the number of syllables produced in pruned transcripts rather than speech duration, because speech duration can entail the variability of speed fluency. For instance, even within the same length of speech duration, the different numbers of syllables can be produced, depending on the speakers' speed of delivery. To annotate temporal features, such as silent pauses and self-repetitions, the *Praat* software was used (Boersma & Weenink, 2012). Through annotating and excluding the disfluency features, the number of syllables produced in pruned transcripts can be calculated. Following prior research (Bosker et al., 2013; De Jong & Bosker, 2013) as well as the results of Study 1, the threshold of silent pauses was defined as 250 ms. With the assistance of automated detection of silence, the clause boundaries and locations of pauses (i.e., mid- vs. end-clause pauses) were annotated on the TextGrid files of the *Praat*. However, *Praat*'s automatic detection of silences and sounds is based on the amplitude and the fundamental frequency of the sound. Accordingly, some prominent filled pauses were detected as sounds, while high frequency sounds, such as word-initial and word-ending fricatives (e.g., /z/ sound in *because* /bɪˈkɒz/),

were annotated as silences. Therefore, for ensuring the validity of pause identifications, the automatically annotated boundaries of silences and sounds were manually modified.

6.8 Cognitive Fluency Measures

Following the broad conception of CF (e.g., De Jong et al., 2013), the current research operationalized CF as a set of L2-specific linguistic knowledge resources and processing speed. From the perspective of speech production mechanisms, L2-specific resources and processing skills cover the aspects of lexis, syntax, morphology, and pronunciation (i.e., formulation and articulation; see Section 2.6). Accordingly, the study aims to measure the speakers' linguistic resources and processing speed separately at these four linguistic levels. The subsequent sections describe the materials and measurements used in Study 3.

6.8.1 Vocabulary knowledge

Regarding the role of vocabulary knowledge in speech production, a major process is lexical retrieval where the speaker activates and selects the lexical item from the mental lexicon that matches the conceptual meaning of the message (see Section 2.6.1). Therefore, vocabulary knowledge here was assessed in terms of how many different lexical items are stored in the mental lexicon (i.e., vocabulary size) and how quickly those lexical items are retrieved if available (i.e., lexical retrieval speed).

Productive Vocabulary Levels Test

To estimate the speakers' vocabulary size, the study used the Productive Vocabulary Levels Test (PVLT; Laufer & Nation, 1999). In the PVLT, participants were asked to fill the blank in the sentence in the paper format version of the test (for the example items, see Appendix G). Each blank was provided with a few initial letters so that participants would be helped in identifying what the target item was among possible synonyms. To establish the upper limits

of their vocabulary size and to avoid collinearity with the speed dimension of vocabulary knowledge (lexical retrieval speed), they were not provided with a time limit for the response. Considering the expected proficiency levels of the participants, the study administered the tests of 2000, 3000, and 5000 frequency levels (excluding 10,000 level and university word list). The score was computed as the total number of correct responses out of 54 items (18 items from each level). As for the scoring procedure, since the target construct of the PVLT in the study is the range of available lemmas corresponding to the concepts, I focused on whether the participants know the target lemmas. Therefore, following De Jong et al. (2013), inflectional errors and obvious spelling mistakes were ignored.

Picture naming task

The picture naming task was employed to assess the participants' speed of lexical retrieval (De Jong et al., 2013; Leonard & Shea, 2017). Participants were presented with pictures and were instructed to name each picture orally in L2 English as fast and accurately as possible. From a theoretical perspective, the picture naming task is assumed to draw on two cognitive processes: (a) picture recognition (identifying the concept corresponding to the visual stimulus) and (b) name retrieval (retrieving the lemma in the target language corresponding to the concept identified). Therefore, to reduce the demands on picture recognition, care was taken to select an appropriate set of picture stimuli. First, target picture stimuli and names were pooled from Snodgrass and Vanderwart's (1980) list which has been used in various L1 and L2 studies (e.g., Leonard & Shea, 2017). Second, to select picture names (rather than picture stimuli) familiar for the participants, the threshold of frequency level was determined as most frequent 3000 words in the JACET8000 wordlist (JACET, 2003). Third, two indices reported in Snodgrass and Vanderwart's original study were considered to minimize variability in picture recognition difficulty: (a) name agreement (> 95%) and (b) image

agreement (> 3.5 on a 5-point scale). In addition, I took into account the indices of familiarity and visual complexity concerning Japanese culture. The final set of picture stimuli for the study included 50 pictures (for the list of items, see Appendix H).

The current study administered the picture naming task using the *PsychoPy* software package (Peirce, 2007) on a 13-inch Macintosh computer. The participants' response was recorded by the *Praat* software (Boersma & Weenink, 2012) connected to the computer. Following De Jong et al. (2013), participants were first presented with a fixation cross in the middle of the screen for 1500 ms. Afterwards, the picture stimulus appeared with a beep sound with a response deadline set as 10000 ms. A blank screen was presented between trials for 500 ms. The order of picture stimuli was randomized for each participant. Prior to the main 50 trials, three trials were provided as practice trials which were not included in the main dataset. The reaction time (RT) between the onset of the presentation of picture stimuli and the onset of the participants' response was manually annotated using the *Praat* software (Boersma & Weenink, 2012). Incorrect responses and outliers were handled as missing values. Outliers were identified as the RTs below the minimum of 300 ms and the RTs higher than 3 SD above the group mean for each item. As a result, 2.4% of the correct responses (k = 127 out of 5375) were removed. The lexical retrieval speed was computed as the averaged RT for correct responses.

6.8.2 Grammatical knowledge

Grammatical processing in L2 speech production entails a variety of syntactic and morphological processes, such as syntactic procedures and morphological inflections (see Section 2.6.2). Therefore, the study used two experimental tasks: the maze task (Y. Suzuki & Sunada, 2018) and the grammaticality judgement test (GJT; Godfroid et al., 2015). From a

theoretical perspective, the maze task focused solely on the construction of phrasal and clausal structures, while the GJT covered a wide range of syntactic and morphological items. Another difference between these tasks lies in the modality of processing. The maze task can reflect learner's grammatical encoding processes (as a component of formulation), while the GJT may tap into monitoring processes (as a proxy for self-monitoring skills). For each task, both linguistic resources and processing speed were respectively operationalized as accuracy and RT scores. Accuracy scores were computed as the total number of correct responses, and RT scores were measured by the averaged response latency.

Maze task

Focusing on the process of syntactic procedures, the study used the maze task which is designed to measure the automaticity of syntactic processing (Y. Suzuki & Sunada, 2018). In the maze task, participants were asked to construct an entire sentence by choosing from two options (e.g., *The → student* vs. *and → ocean* vs. *took→ the* vs. *dress→ tests.* vs. *organic.*). As depicted in Figure 18, participants were presented with two options of single words on the computer screen and were instructed to select the word which can be grammatically connected to the sentence. One advantage of the maze task over typical sentence construction tasks (e.g., filling the blank of the sentence) is that the maze task can assess participants' real-time incremental processing of the whole sentence rather than the fragments of the sentence (Y. Suzuki & Sunada, 2018).

Study 3 adopted the stimuli from Y. Suzuki and Sunada's (2018) study which was conducted with a similar population of L2 learners (i.e., Japanese college students learning English). The target stimuli consisted of 48 sentence stimuli with the equal numbers of stimuli from four major syntactic structures (n = 12): (a) declaratives (6 simple and 6 complex sentences),

(b) wh-questions, (c) relative clauses (6 subject- and 6 object-relative clauses), and (d) indirect questions (for the examples of items, see Appendix I). The order of these sentence stimuli was randomized for each participant. Prior to the main trials, four sentences were provided as a practice session. The time limit for each response was set as 4300 ms, following Y. Suzuki and Sunada (2018). Participants were instructed to respond as quickly and accurately as possible. The maze task was administered using the DMDX software (Forster & Forster, 2003) on a 13-inch Windows computer. Y. Suzuki and Sunada (2018) were particularly interested in the sentence-level RT performance in the maze task as the indicator of automatization, where the correct responses were counted only when all word items in the sentence were correctly responded. However, as Study 3 intends to capture a more fine-grained variability in the accuracy of syntactic encoding procedures, the word-level analysis was adopted, where the correct responses were counted, regardless of whether the speaker succeeded in constructing the whole sentence. Based on the word-level analysis, the study computed two measures: (a) the number of correct responses in words (accuracy) and (b) the mean duration of the response latency (i.e., RT) of correctly responded trials. Regarding the RT measure, outliers were identified as the RTs below 300 ms or higher than 3 SD above the group mean of the latency of all word-level responses. As a result, 68 RTs (6.6 %) out of 49,406 RTs were removed.

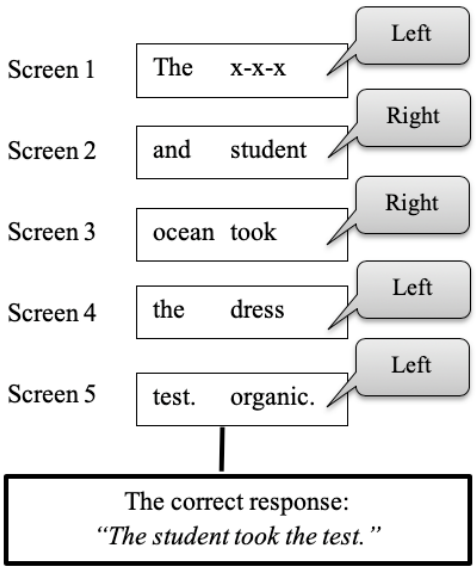


Figure 18. A sample display of the maze task.

Grammaticality judgement test

To capture participants' range of grammatical knowledge, Study 3 employed a GJT. In a GJT, participants are asked to judge the grammaticality of the given sentence stimuli. The conditions of GJTs can be adapted, according to the target type of knowledge and processing (Gass, 2018). As Study 3 aims to examine the role of linguistic knowledge in L2 spontaneous speech production, the timed condition was adopted. Due to the relatively low proficiency of the target population (i.e., EFL learners), I decided to use the written mode of stimuli so that participants would be less likely to fail to understand the stimuli.[12]

The target stimuli were adopted from Godfroid et al.'s (2015) study, which included 17 target grammatical constructions, covering both syntactic and morphosyntactic features. For each grammatical target, four sentence stimuli were devised (in total, 68 sentences). Within the entire set of stimuli, an equal number of grammatical and ungrammatical sentences were created (34 items for each). Using Godfroid et al.'s (2015) target stimuli, several

[12] Ideally, the mode of stimuli should also be compatible with the target situation. In this study, the stimuli might be expected to be provided aurally if participants' listening skills had been sufficient.

administration conditions were adapted for the current study's target population. First, proper nouns including locations and person's names were changed to relatively familiar ones for Japanese learners of English (for the example of items, see Appendix J). Second, Godfroid et al. (2015) set the response time limit for each stimulus sentence as the median response times based on native speakers' pilot data plus 20 % additional time (R. Ellis, 2005; Loewen, 2009). However, considering the relatively lower proficiency level of the present participants, 10 seconds for each stimulus were allowed for the response (cf. Y. Suzuki & DeKeyser 2017).

The timed GJT was administered using the *PsychoPy* software (Peirce, 2007) on a 13-inch Windows computer. Participants were instructed to judge the grammaticality of the sentences as fast and accurately as possible. Prior to the main 68 trials, participants completed eight sentences as practice trials. The practice stimuli were created by the researcher using four grammatical features that were not included in the main trials (irregular plurals, past tense, word order, dummy *do*). For each trial, the term "*Ready?*" was presented in the middle of the screen for 1000 ms, and then the sentence stimulus appeared on the screen for 10000 ms. The main trials were divided into four blocks (17 items for each block), and participants were asked to take a brief break between the blocks.

To compute the accuracy score based on the GJT responses, the current study assigned one point for each correct response, while incorrect responses and no responses within the time limit were assigned no points. Only correct responses were used to compute the RT score excluding outliers whose RT was below 300 ms or higher than 3 SD above the group mean for each sentence stimulus. As a result, a total of 28 RTs (0.4%) was removed from the RT

analysis. The current study calculated the accuracy and RT scores separately for syntactic and morphological features.

6.8.3 Pronunciation knowledge

Although pronunciation knowledge can be evaluated in terms of linguistic resources and processing speed, Study 3 solely focused on the speed aspect of pronunciation knowledge for the following reasons. First, the assessment of pronunciation entails the substantive difficulty of defining what constitutes target-like pronunciation, due to different models of L2 pronunciation learning (e.g., British vs. American English). Second, from the theoretical perspective of speech processing mechanisms, pronunciation errors can be derived from the deviation from the *target-like* pronunciation at any of the three pronunciation-related processes—phonological encoding, phonetic encoding, and articulation (see Sections 2.6.3–2.6.4). However, due to the incremental nature of these encoding processes, it is also difficult to identify the source of such pronunciation errors (e.g., either the phonological representations of the words or the motor articulatory movements), leading to another difficulty with interpreting the results. Third, prior work on phonological encoding (Broos et al., 2018) reported that the language status (L1 vs. L2 speakers) did not differentiate the accuracy and speed of phonological processing.[13] This finding may support the argument that phonological encoding can proceed in a virtually automatic manner, regardless of whether the corresponding L2 phonological categories are established. This is possibly because the representations of both L1 and L2 phonemes are stored in the same place and activated

[13] Broos et al.'s (2018) generalized mixed-effects model suggested that L2 speakers performed better in the accuracy scores of phoneme-monitoring tasks with different conditions of phonological facilitation, which tap into the activation of phonological information of the words. However, they concluded that the relatively lower accuracy of L1 speakers was due to the strong activation of non-target phonemes by the distractors in the conditions, which did not interfere L2 speakers' phonological processes to the same extent. Thus, I here argued that there was virtually no difference between L1 and L2 speakers in the accuracy of phonological encoding in normal conditions.

simultaneously (Poulisse, 1999; Roelofs, 2000). Accordingly, the relatively automatic process of phonological encoding might be achieved by the substitution of the corresponding L1 phonemes. One of the motivations to examine the linguistic resource aspects of linguistic knowledge was that the lack of linguistic resources can result in breakdowns of speech processing, which can be observed at the UF level. However, the above line of argumentation and research findings indicate that the potential impact of the lack of linguistic resources on UF may not be established at the level of pronunciation. Therefore, pronunciation accuracy measures might not be necessary when investigating the CF-UF link.

Despite the potential lack of differences in the speed of phonological encoding processes between L1 and L2 speakers, a significant L2 slow-down in the speed of articulatory movements for L2 pronunciation has been observed (Broos et al., 2018). I thus decided to measure speakers' articulatory speed. However, due to the incremental nature of speech processing, it is difficult to establish a pure measure of articulatory speed, which is separate from the speed of phonological and phonetic encoding. Therefore, I decided to measure the efficiency of these processes holistically, using a controlled speech production task.

Controlled speech production task

The controlled speech was elicited to gauge participants' speed of articulatory gestures. Participants were asked to read a short passage of a direction on shopping silently and then to read it aloud in English (for the script, see Appendix K). The passage was selected from Weinberger's *speech accent archive* (2011) where speech samples of the same passage from a variety of L1 backgrounds are archived. The speech sample data were analysed to compute the articulation rate measure using the same procedure as for the corresponding measure for spontaneous speech (see Section 6.7). The rationale for using controlled speech production,

as opposed to single word production (e.g., delayed picture naming task; De Jong et al., 2013), was that one of the essential processes of phonological encoding, syllabification, is supposed to take place not only within words but also between words, such as linking (Levelt, 1999; see also Section 2.6.3). To capture the speakers' efficiency of pronunciation-related processes including between-word syllabication processes, I decided to select a longer stretch of speech (69 words) as a prompt.

6.9 L1 Utterance Fluency

To measure L2-specific UF, some studies have assessed L2 learners' L1 UF and computed the corrected L2 UF measures as the residuals from the regression models predicting L2 measures from the corresponding L1 measures (e.g., the residuals between L1 and L2 articulation rate; De Jong et al., 2015). This methodological approach is supported by the significant correlations between L1 and L2 corresponding UF measures (e.g., Bradlow et al., 2017; De Jong et al., 2015; De Jong & Mora, 2019; Peltonen, 2018). From a theoretical perspective, the covariance between L1 and L2 fluency may indicate that some idiosyncratic factors can be shared across L1 and L2 speech, such as speaking style (Peltonen, 2018). To the best of my knowledge, the L1-L2 fluency connection, however, has not yet been examined in the context of Japanese learners of English (for the synthesis on this topic, see Section 3.9). It is thus theoretically possible that L2 English UF measures would not be correlated with L1 Japanese corresponding measures in the current studies. Therefore, I computed both L1 and L2 UF measures separately, instead of using the corrected measures.

As reviewed previously (see Section 3.9), it can be hypothesized that the association between L1 and L2 speech production is likely to be observed in open-ended speaking tasks, because a flexible storyline or speech content of the open-ended task can provide more room for

learners' personal speaking style to be reflected. Accordingly, Study 4 elicited L1 speech, using another argumentative speech task with the same task format and procedure as the L2 argumentative task (see Section 6.6). In the L1 argumentative speech task, students were provided with a statement—*Japan should stop its lifetime employment system*—and then instructed to argue how far they agree with the statement. As with the L2 argumentative task, this topic was selected considering its relevance and meaningfulness to the target population as well as its abstractness. Study 4 calculated the same set of L1 UF measures as L2 measures listed in the section of Utterance fluency measures (Section 6.7). Considering the syllable structure and phonological properties of Japanese, I employed a mora rather than a syllable as the standardized unit for the calculation of L1 Japanese UF measures. A mora is fundamentally shorter than English syllables, because the basic structure of morae allows only one consonant at the position of onset of the syllable (Vance, 2008; for details, see Section 2.6.3).

6.10 Data Collection Procedure

Data were collected in two sessions: group testing and individual sessions. In the group sessions, participants were asked to complete linguistic knowledge tests including the paper-based PVLT, the maze task, and the GJT, as well as a language background questionnaire. In the individual sessions, participants performed four English speaking tasks[14], the picture-naming task, and the controlled speech production task with the researcher. The order of these tasks is described in Table 22 below. All participants first joined in the group session, and approximately one week later, they participated in the individual session. I provided all

[14] As mentioned in the first section of this chapter, I used two argumentative tasks in the second and third rounds of data collection (n = 104, in total). Strictly speaking, the first round of data collection asked participants to perform only four speaking tasks. Since the majority of my participants was from the second and third round of data collection, I explain the data collection procedure of the second and third rounds of data collection here. Note that the order of the two topics of the argumentative tasks was counterbalanced.

participants an informed consent form and an information sheet (Appendix C) at the beginning of the group session. I also explained the purpose and procedure of the study and the withdrawal process to all participants. The signed ethics forms were then collected from all the participants. In the group testing session, the order of the PVLT and the grammar tests (the maze task and the GJT as a continued block) was counterbalanced across participants. In the individual session, the order of the argumentative and picture narrative tasks was also counterbalanced across participants. Regarding the text summary tasks, the combination of the order of the conditions (i.e., with vs. without RAA) and the source texts was counterbalanced across participants.

Table 22. *Order and time of the speaking and linguistic knowledge tasks in each of the two sessions.*

Group testing session		Individual session	
Task	Time	Task	Time
Consent form	5 min	Argumentative task	5 min
PVLT	15 min	Picture narrative task	5 min
Maze task, GJT	15 + 5 min	Text summary tasks	8 + 8 min
Background questionnaire	10 min	Controlled production	3 min
		L1 argumentative task	5 min
		Picture naming task	5 min

6.11 Statistical Analysis

As a preliminary analysis, descriptive statistics and correlational analyses were performed to examine the distributions of all the variables and the interrelationship among them. Study 2 aims to compare students' UF measures across tasks differing in the quality of speech processing demands. Considering the potential non-normal distributions of UF measures (e.g., Bosker et al., 2013; Lambert et al., 2020), a Generalized Linear Mixed-effects Model (GLMM) was employed. One statistical advantage of GLMM is that researchers can specify the appropriate probability distribution of the outcome variables (e.g., poisson, gamma distributions). The appropriate distributions of the current UF measures were first specified

by the deviation from a normal distribution (Shapiro-Wilk tests) and the visual inspection of density plots. Depending on the distributions specified, non-positive values (basically 0 values in the current dataset) may prevent the estimation of statistical models (e.g., gamma distribution). Thus, when building GLMMs based on distributions other than a normal distribution, the 0 values were replaced with the -3SD values of the theoretical distributions of the variables, estimated by the Maximum Likelihood (ML) estimation. All the GLMMs reported in the thesis were estimated through the *glmer* function in the *lme4* package (Bates et al., 2015), using R statistical software 4.0.2 (R development Core Team, 2020). To address the RQ of Study 2 (RQ2), the GLMMs were constructed for each UF measure (i.e., outcome variable), using task type as a categorical fixed-effect predictor variable with individual participants as a random-effects predictor. Since task type was a within-subject variable, the random slope of participants may not be distinguished from random error variance (Barr, 2013). Therefore, only the random intercepts of participants were included. Regarding the predictor variable of task type, to minimize the rate of type I errors, Study 2 took a confirmatory approach to comparing the outcome variables (UF measures) between three predetermined contrasts (see Section 6.6). To this end, the GLMMs in Study 2 adopted forward difference contrast coding rather than dummy coding for the categorical variable of task type (for the contrast coding, see Appendix L). The proposed models for Study 2 are described as follows:

RQ2:

$$L2\ UF\ measure \sim Task\ type + (1|Participant)$$

To address the RQs of Study 3, an SEM approach was taken to investigate how CF and UF were associated with each other at the level of latent variables (RQ3-1). Following previous

studies (De Jong et al., 2013; Kahng, 2020), CF was operationalized by a set of measures of linguistic knowledge and processing measures. However, the dimensionality of CF (see Section 3.6) have not been yet addressed in the literature of L2 fluency. In other words, the factor structure of CF has not been specified in prior research. As for the factor structure of UF, since Tavakoli and Skehan (2005) proposed the three-factor structure of UF, the factor structure of UF, to the best of my knowledge, has not been revisited. Therefore, prior to an SEM modelling, Study 3 first tested several factor structures motivated by the theories of speech production and previous studies on fluency development (RQ3-1a, RQ3-1b), using a CFA. After identifying the best-fit factor structures of CF and UF, Study 3 built an SEM model predicting the subconstructs of UF from those of CF. Considering the relatively small sample size of the current dataset (n = 128), the factorability indices (KMO, Bartlett's test of Sphericity) were checked before testing CFA and SEM models, and the goodness-of-fit indices were also inspected for the reliability of extracting latent variables of the present models. As for RQ3-2 addressing the variability of the CF-UF link across tasks, the regression weights were compared across speaking tasks by the standardized coefficients and their 95% confidence intervals, which is analogous to the estimation of t-values in t-tests (i.e., path coefficient t-test; Tabachnick & Fidell, 1996).

In response to the potential non-normal distributions of UF measures, the estimations of all CFA and SEM analyses in Study 3 were conducted with the Robust Maximum Likelihood estimation, as recommended by Hu and Bentler (1998). Considering the relatively small sample size (N < 250) as well as the estimation method (i.e., Maximum Likelihood estimation), Study 3 particularly focused on the model fit indices of SRMR and CFI, because these two indices were reported to be sensitive to the model improvement (Hu & Bentler, 1998). I also reported the indices of chi square/df ratio, TLI, and RMSEA for the sake of

transparency and comparability with future replication studies. The cut-off scores for these model fit indices were predetermined as follows: SRMR (< .08), CFI and TLI (> .95), chi-square ratio/df (< 2.0), and RMSEA (< .06). Although the observed variables of UF were elicited from the same individuals (i.e., repeated measures), I ran separate SEM models for each task rather than longitudinal SEM models which assume different mean scores for the same latent variables across tasks. This decision was motivated by the fact that due to the cross-sectional design of Study 3, the latent variable(s) of UF is supposed to be consistent across tasks.

Regarding the RQs of Study 4, as with the analysis of task effect in Study 2, GLMMs were used due to the potential non-normal distributions of UF measures. To examine the overall associations between L1 and L2 UF measures (RQ4-1), GLMMs were constructed to predict L2 UF measures from the corresponding L1 measures with the random intercepts of individual participants and topics of the L2 argumentative tasks. Meanwhile, to investigate the moderator effects of L2 proficiency on the L1-L2 UF link (RQ4-2), it was tested whether the interaction effects by L1 UF measures and L2 proficiency variable(s) would be significant. The proficiency variable(s) was decided to be calculated as the factor score(s) of CF in the CFA model of Study 3, because these factor scores are independent of measurement errors and also more relevant to UF than the discrete observed variables of CF. The planned models for Study 4 are described as follows:

RQ3-1:

L2 UF measure ~ L1 UF measure + (1|Participant) + (1|Topic)

RQ3-2:

*L2 UF measure ~ L1 UF measure*L2 proficiency + (1|Participant) + (1|Topic)*

6.12 Summary

This chapter outlined the methodology for Studies 2–4, including the description of participants and speaking tasks, the measures of UF and CF, and statistical analysis. Study 2 compared UF performance across four speaking tasks. Study 3 examined the contributions of CF to UF at the level of constructs across four different speaking tasks. Meanwhile, Study 4 investigated the predictive power of L1 UF for L2 UF with the argumentative tasks concerning the moderator effects of L2 proficiency on the L1-L2 UF link. Studies 2 and 3 were conducted with all the participants (n = 128) who were Japanese learners of English, whereas Study 4 proceeded using the data from 104 out of 128 participants who performed two argumentative speech tasks. The rationale for using the two argumentative tasks was that the generalizability of the L1-L2 UF link can be enhanced by controlling its variability across topics. The four speaking tasks were selected to operationalize three different aspects of speech processing demands—conceptualization, the activation of linguistic representations, and the activation of phonological representations. A set of UF measures covered the triad of subconstructs of UF—speed, breakdown, and repair fluency (Tavakoli & Skehan, 2005). CF measures were selected based on previous studies on the CF-UF link (De Jong et al., 2013; Kahng, 2020) as well as theoretical correspondences between CF and speech production mechanisms (Kormos, 2006; Segalowitz, 2010, 2016). In Study 4, L2 proficiency was calculated as the factor score(s) of CF based on the results of Study 3. Studies 2 and 4 employed a mixed-effects modelling approach to control for the random variance of target fixed-effects predictors (task types and L1 UF measures, respectively). In addition, GLMMs were adopted in response to the potential non-normal distributions of UF measures. In Study 3, an SEM approach was employed to examine the relationship between UF and CF at the

level of latent variables. In the following chapters (Chapters 7–9), I report and discuss the results of Studies 2, 3, and 4, respectively.

Chapter 7: Results and Discussion of Study 2—Effects of Speech Processing Demands on Utterance Fluency

7.1 Introduction

This chapter reports the results of Study 2 (RQ2) which is concerned with the variability of UF performance across speaking tasks differing in the demands on particular speech processing stages, including conceptualization, the activation of linguistic representations, and the activation of phonological representations. Prior to answering the RQ, the descriptive statistics of UF measures were inspected (Section 7.2.1). The appropriate distributions of the UF measures were then identified for the subsequent GLMMs to investigate the effects of different speech processing demands on UF performance (Section 7.2.2). Using the post-speaking questionnaire, the speech processing demands of these tasks were also examined from the perspective of students' perceptions (Section 7.2.3). The chapter concludes by discussing the findings of the variability of UF performance according to the different speech processing demands from the perspective of speech production mechanisms (Section 7.3).

7.2 Results

7.2.1 Descriptive statistics of utterance fluency measures

The descriptive statistics of UF measures were first examined as a preliminary analysis for the subsequent multivariate analyses. As summarised in Table 23, the Shapiro-Wilk tests suggested that most of the UF measures were non-normally distributed, while articulation rate seemed to be normally distributed across tasks. Since the distributions of the UF measures would decide the probability distributions of the GLMMs for task effects, density plots for each measure were created to visually inspect the distributions across tasks (see Appendix M). The density plots suggested that the distributions of all the UF measures were positively skewed, characterized by a long tail in the positive direction, except for articulation

rate. Accordingly, in the subsequent GLMMs, the gaussian distribution (i.e., normal distribution) was applied to the GLMM of articulation rate, while the GLMMs of the remaining UF measures adopted the gamma distribution—one of the continuous probability distributions where a possible range of values is from zero to +∞ (Coupé, 2018).

Table 23. *Descriptive summary of utterance fluency measures in Study 2.*

					Shapiro-Wilk test	
UF measures	Task	*M*	*SD*	*SE*	*Statistics*	*p-value*
Articulation rate	Arg	3.140	0.598	0.053	0.980	0.059
	Cartoon	2.824	0.506	0.045	0.986	0.201
	TS.withoutRAA	2.664	0.459	0.041	0.990	0.526
	TS.withRAA	2.697	0.463	0.041	0.984	0.131
Speech rate	Arg	1.742	0.679	0.060	0.969	0.005
	Cartoon	1.499	0.582	0.051	0.970	0.006
	TS.withoutRAA	1.401	0.489	0.043	0.985	0.159
	TS.withRAA	1.376	0.482	0.043	0.959	< .001
Mean length of run	Arg	4.507	2.820	0.249	0.591	< .001
	Cartoon	3.547	1.443	0.128	0.872	< .001
	TS.withoutRAA	3.427	1.211	0.107	0.905	< .001
	TS.withRAA	3.408	1.343	0.119	0.813	< .001
Mid-clause pause ratio	Arg	0.215	0.090	0.008	0.970	0.006
	Cartoon	0.242	0.102	0.009	0.976	0.025
	TS.withoutRAA	0.255	0.104	0.009	0.904	< .001
	TS.withRAA	0.259	0.093	0.008	0.982	0.080
End-clause pause ratio	Arg	0.059	0.021	0.002	0.968	0.004
	Cartoon	0.086	0.023	0.002	0.991	0.603
	TS.withoutRAA	0.076	0.019	0.002	0.984	0.141
	TS.withRAA	0.076	0.019	0.002	0.968	0.004
Filled pause ratio	Arg	0.111	0.099	0.009	0.836	< .001
	Cartoon	0.097	0.094	0.008	0.843	< .001
	TS.withoutRAA	0.116	0.108	0.010	0.803	< .001
	TS.withRAA	0.118	0.095	0.008	0.876	< .001
Mid-clause pause duration	Arg	1.120	0.522	0.046	0.790	< .001
	Cartoon	1.119	0.489	0.043	0.864	< .001
	TS.withoutRAA	1.140	0.503	0.044	0.714	< .001
	TS.withRAA	1.195	0.466	0.041	0.884	< .001
End-clause pause duration	Arg	1.368	1.193	0.105	0.608	< .001
	Cartoon	1.302	0.732	0.065	0.855	< .001
	TS.withoutRAA	1.590	1.334	0.118	0.605	< .001
	TS.withRAA	1.485	1.178	0.104	0.531	< .001

Self-repetition ratio	Arg	0.076	0.070	0.006	0.858	< .001
	Cartoon	0.102	0.072	0.006	0.918	< .001
	TS.withoutRAA	0.090	0.079	0.007	0.838	< .001
	TS.withRAA	0.094	0.072	0.006	0.853	< .001
Self-correction ratio	Arg	0.021	0.016	0.001	0.923	< .001
	Cartoon	0.025	0.017	0.002	0.943	< .001
	TS.withoutRAA	0.025	0.019	0.002	0.827	< .001
	TS.withRAA	0.024	0.016	0.001	0.954	< .001
False start ratio	Arg	0.008	0.011	0.001	0.735	< .001
	Cartoon	0.007	0.008	0.001	0.795	< .001
	TS.withoutRAA	0.012	0.012	0.001	0.853	< .001
	TS.withRAA	0.011	0.010	0.001	0.890	< .001

Note. Arg = Argumentative task (Tokyo Olympics); Cartoon = Picture narrative task; TS.withoutRAA = Text summary task without read-aloud assistance; TS.withRAA = Text summary task with read-aloud assistance.

7.2.2 Task effects on utterance fluency performance

A series of GLMMs were run to examine whether the participants' UF measures differed across four speaking tasks (Argumentative speech task, Picture narrative task, and Text summary task with and without RAA). For each GLMM, UF measures were used as an outcome variable, while four categories of task types were entered as a fixed-effect predictor variable. Moreover, individual participants were included in the GLMMs as a random-effect variable. Since task type was a within-subject variable, the random slope of participants may not be distinguished from random error variance (Barr, 2013) and thus were not included. As mentioned in Section 7.2.1, the gaussian distribution (i.e., normal distribution) was applied to the GLMM of articulation rate, while the gamma distribution was applied to the GLMMs of the remaining UF measures with the log link function. The objective of the GLMM analysis here was to compare UF performance between three predetermined pairs of contrast (for details, see Section 6.6): (a) the argumentative task and the picture narrative task (conceptualizing demands), (b) the picture narrative task and the text summary task without RAA (enhanced activation of linguistic representations of relevant items), and (c) the text summary tasks without and with RAA (enhanced activation of phonological representations

of relevant items). To this end, the GLMMs here adopted forward difference contrast coding rather than dummy coding for the categorical variable of task type. Accordingly, the rate of type I errors can be minimized by avoiding additional multiple comparisons irrelevant to the predetermined objective of statistical analysis. The GLMMs revealed that significant effects of task types on all of the UF measures (see Table 24; for the other estimates of the GLMMs, see Appendix N).

Speed fluency measures (articulation rate, speech rate, and mean length of run) showed a similar pattern of task effects. In all of these measures, the participants' speed of delivery was faster in the argumentative task than in the picture narrative task. Besides, articulation rate and speech rate were higher in the picture narrative task than in the text summary task without RAA. Regarding the contrast between two conditions of text summary tasks, there were no significant differences in either of speed fluency measures.

As for pause ratio measures (breakdown fluency), students produced both mid- and end-clause silent pauses less frequently in the argumentative task than in the picture narrative task. In contrast, the frequency of filled pauses was higher in the argumentative task than in the picture task, indicating that the direction of effects of conceptualization demands was opposite between silent and filled pauses. Meanwhile, participants produced fewer mid-clause pauses and more end-clause pause in the picture narrative task than in the text summary task without RAA. Similar to mid-clause pauses, filled pauses were less frequent in the picture narrative task than in the text summary task without RAA. Finally, as with speed fluency measures, there were no significant differences in pause ratio measures between the RAA conditions.

Regarding pause duration measures, there were no significant differences in the duration of both mid- and end-clause pauses between the argumentative task and the picture narrative task. However, the length of end-clause pauses tended to be shorter in the picture narrative task than in the text summary task without RAA. Meanwhile, mid-clause pauses were longer in the text summary task with RAA than in the text summary task without RAA.

Finally, repair fluency measures showed a nuanced picture of task effects, depending on the type of disfluency phenomena. First, students produced fewer self-repetitions and self-corrections in the argumentative task than in the picture task. Second, the frequency of self-repetitions in the picture narrative task was higher than in the text summary task without RAA. Third, students produced fewer false starts in the picture narrative task than in the text summary task without RAA.

Table 24. *Summary of the effects of three predetermined contrasts of speaking tasks on utterance fluency performance.*

Comparison	*Estimate*	*SE*	*z-value*	*p*	Contrast
Articulation rate					
Arg - PicN	0.316	0.030	10.537	< .001	Arg > PicN
PicN - TS.w/oRAA	0.160	0.030	5.329	< .001	PicN > TS.w/oRAA
TS.w/oRAA - TS.withRAA	-0.032	0.030	-1.077	0.282	*n.s.*
Speech rate					
Arg - PicN	0.152	0.018	8.521	< .001	Arg > PicN
PicN - TS.withoutRAA	0.059	0.018	3.286	0.001	PicN > TS.w/oRAA
TS.w/oRAA - TS.withRAA	0.010	0.018	0.562	0.574	*n.s.*
Mean length of run					
Arg - PicN	0.211	0.019	11.303	< .001	Arg > PicN
PicN - TS.withoutRAA	0.021	0.019	1.148	0.251	*n.s.*
TS.w/oRAA - TS.withRAA	0.009	0.019	0.480	0.631	*n.s.*
Mid-clause pause ratio					
Arg - Cartoon	-0.117	0.023	-4.972	< .001	Arg < PicN
PicN - TS.withoutRAA	-0.072	0.023	-3.068	0.002	PicN < TS.w/oRAA
TS.w/oRAA - TS.withRAA	-0.018	0.023	-0.754	0.451	*n.s.*
End-clause pause ratio					
Arg - Cartoon	-0.381	0.024	-15.821	< .001	Arg < PicN
PicN - TS.withoutRAA	0.117	0.024	4.881	< .001	PicN > TS.w/oRAA

TS.w/oRAA - TS.withRAA	0.005	0.024	0.210	0.834	*n.s.*
Filled pause ratio					
Arg - Cartoon	0.191	0.050	3.819	< .001	Arg > PicN
PicN - TS.withoutRAA	-0.208	0.049	-4.205	< .001	PicN < TS.w/oRAA
TS.w/oRAA - TS.withRAA	-0.065	0.049	-1.312	0.190	*n.s.*
Mid-clause pause duration					
Arg - Cartoon	0.003	0.023	0.112	0.911	*n.s.*
PicN - TS.withoutRAA	-0.027	0.023	-1.161	0.246	*n.s.*
TS.w/oRAA - TS.withRAA	-0.054	0.023	-2.315	0.021	TS.w/oRAA < TS.withRAA
End-clause pause duration					
Arg - Cartoon	0.003	0.032	0.081	0.935	*n.s.*
PicN - TS.withoutRAA	-0.149	0.032	-4.735	< .001	PicN < TS.w/oRAA
TS.w/oRAA - TS.withRAA	0.024	0.032	0.750	0.454	*n.s.*
Self-repetition ratio					
Arg - Cartoon	-0.415	0.064	-6.440	< .001	Arg < PicN
PicN - TS.withoutRAA	0.183	0.064	2.846	0.004	PicN > TS.w/oRAA
TS.w/oRAA - TS.withRAA	-0.110	0.064	-1.722	0.085	*n.s.*
Self-correction ratio					
Arg - Cartoon	-0.164	0.077	-2.133	0.033	Arg < PicN
PicN - TS.withoutRAA	0.039	0.077	0.505	0.614	*n.s.*
TS.w/oRAA - TS.withRAA	-0.007	0.077	-0.086	0.931	*n.s.*
False start ratio					
Arg - Cartoon	0.062	0.115	0.538	0.591	*n.s.*
PicN - TS.withoutRAA	-0.586	0.115	-5.088	< .001	PicN < TS.w/oRAA
TS.w/oRAA - TS.withRAA	0.131	0.115	1.142	0.254	*n.s.*

Note. Arg = Argumentative task (Tokyo Olympics); Cartoon = Picture narrative task; TS.w/oRAA = Text summary task without read-aloud assistance; TS.withRAA = Text summary task with read-aloud assistance.

7.2.3 Students' perceptions of different speech processing demands

To cross-validate the operationalization of speech processing demands in Study 2, the post-speaking performance questionnaire (see Section 6.6) was used to examine how students' perceptions of demands on five processing dimensions—Conceptualization, Lexical encoding, Syntactic encoding, Morphological encoding, and Articulation—across four tasks. The descriptive statistics of participants' responses were summarized in Table 25 and Figure 19. Although the Shapiro-Wilk tests and the histograms of these items suggested that most of the items were not normally distributed (for the histograms of items, see Appendix O), I decided to perform a two-way repeated-measures ANOVA, considering the robustness of

ANOVA to the violation of normal distribution (Schmider et al., 2010). The two-way ANOVA compared their responses across Task (Argumentative task, Picture narrative task, and Text summary tasks without and with RAA) and Dimension (Conceptualization, Lexical encoding, Syntactic encoding, Morphological encoding, and Articulation) as the within-subject independent variables. The results yielded a significant interaction by Task and Dimension, $F(8.524, 332.441) = 4.358$, $p < .001$, $\eta_p^2 = 0.101$ (Greenhouse-Geisser adjustment adopted, due to the violation of sphericity), while either of Task ($F(3, 117) = 0.617$, $p = 0.605$, $\eta_p^2 = 0.016$) or Dimension ($F(2.721, 106.138) = 1.971$, $p = 0.129$, $\eta_p^2 = 0.048$; Greenhouse-Geisser adjustment adopted) did not show a significant effect on the responses. To identify the location of statistically significant differences, post-hoc comparisons were conducted with Holm's rejective Bonferroni adjustment. The results indicated that the participants perceived more difficulty with conceptualization in the text summary task with RAA than in the picture narrative task ($p = .018$, $d = .623$) and that they found that lexical retrieval was more challenging in the argumentative task than in the text summary task with RAA ($p = .027$, $d = .605$).

Table 25. *Descriptive summary of participants' responses on the post-speaking questionnaire.*

Dimension	Task	*Mean*	*Median*	*SD*	*SE*
Conceptualization	Arg	3.40	3.00	1.52	0.24
	Cartoon	2.85	3.00	1.39	0.22
	TS.withoutRAA	3.40	3.00	1.39	0.22
	TS.withRAA	3.75	4.00	1.43	0.23
Lexical encoding	Arg	4.13	4.00	1.45	0.23
	Cartoon	3.98	4.00	1.29	0.20
	TS.withoutRAA	3.68	4.00	1.46	0.23
	TS.withRAA	3.25	3.00	1.37	0.22
Syntactic encoding	Arg	3.88	4.00	1.52	0.24
	Cartoon	3.43	4.00	1.55	0.25
	TS.withoutRAA	3.70	4.00	1.40	0.22
	TS.withRAA	3.55	3.00	1.38	0.22
Morphological encoding	Arg	3.65	4.00	1.35	0.21
	Cartoon	4.00	4.00	1.24	0.20

	TS.withoutRAA	3.50	3.50	1.32	0.21
	TS.withRAA	3.53	3.50	1.36	0.22
Articulation	Arg	3.18	3.00	1.52	0.24
	Cartoon	3.40	3.00	1.68	0.27
	TS.withoutRAA	3.40	3.00	1.60	0.25
	TS.withRAA	3.23	3.00	1.56	0.25

Note. Arg = Argumentative task (Tokyo Olympics); Cartoon = Picture narrative task; TS.withoutRAA = Text summary task without read-aloud assistance; TS.withRAA = Text summary task with read-aloud assistance; a six-point scale was used (1 = very easy; 6 = very challenging).

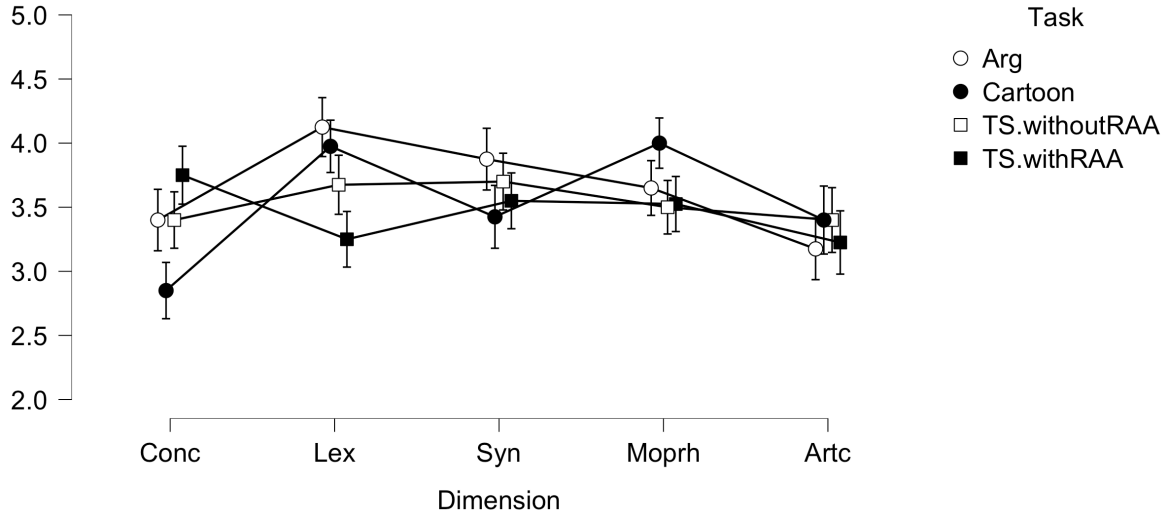


Figure 19. Descriptive plots of responses across Task and Dimension.

In addition to the participants' quantitative responses, their qualitative responses on what they found difficult were also examined. Their responses were coded for different stages of speech processing and grouped into three major processes of speech production, that is, *Conceptualization*, *Formulation*, and *Articulation*. The raw frequency of participants who mentioned each coding label was summarized in Table 26. Overall, participants were rarely conscious of processing demands of articulation (n = 6 out of 104 responses). However, the relative frequency of the number of speakers reporting the demands of conceptualization and formulation seemed to vary across tasks. Specifically, in the argumentative task, the perceived demands on conceptualization and formulation were equally reported, while in the picture narrative task, speakers were more likely to be aware of speech processing demands on formulation than on conceptualization. Furthermore, in both conditions of the text

summary task, conceptualization processes were commonly perceived as more demanding than formulation processes.

Table 26. *Descriptive summary of the speakers' perceived speech processing demand while speaking.*

Category	Arg	Cartoon	Text summary		Total
			w/o RAA	with RAA	
Conceptualization	12	8	17	24	61
Formulation	11	14	5	7	37
Articulation	2	3	0	1	6
Total	25	25	22	32	104

Note. Arg = Argumentative task (Tokyo Olympics); Cartoon = Picture narrative task.

7.3 Discussion

To explore the association between UF measures and different speech processing components, Study 2 operationalized the conceptualizing demands as content generation and discourse organization. Accordingly, the argumentative task was selected as the speaking condition with enhanced conceptualizing demands and was contrasted with the picture narrative task where students were required to produce the predefined content of speech with clearly sequenced events. Meanwhile, the demands on formulation were manipulated by the activation level of relevant linguistic items at the levels of linguistic and phonological representations. Using the picture narrative task as the baseline condition, the activation level of linguistic representations was enhanced in the text summary task where students needed to process relevant linguistic items through reading the source text. The phonological representations were further activated by the bimodal source text (i.e., RAA). The effects of activating phonological representations were thus examined by comparing UF performance across two conditions of the text summary task (i.e., without vs. with RAA). The research goal of Study 2 (RQ2) was addressed by building a set of GLMMs for each UF measure with a fixed-effect predictor variable of task type and the random intercepts of individual participants. The following sections discuss how two major components of speech processing

demands—conceptualization and activation of relevant linguistic items—were reflected in UF performance.

7.3.1 Speech processing demands on conceptualization

Comparing the argumentative and picture narrative tasks, the GLMMs suggested that the enhanced conceptualizing demands resulted in a faster speed of delivery (articulation rate, speech rate, mean length of run), fewer silent pauses (both mid- and end-clause pauses), more filled pauses, and fewer self-repetitions and self-corrections. From a theoretical perspective, the necessity for content generation and organizing different ideas (e.g., opinion, supporting information) in a coherent manner in the argumentative task was hypothesized to elevate the conceptualizing demands and thus to reduce the attentional resources available for the subsequent linguistic processing (see Section 3.8). Surprisingly, students' speech was, however, more fluent in all the subconstructs of UF in the argumentative task than in the picture narrative task.

The overall fluency advantage in the argumentative task might be explained by the open-ended nature of the task. As a result, students had some freedom to plan speech content. Since conceptualization is responsible for content planning, open tasks arguably enhance conceptualizing demands. However, despite the enhanced demands on conceptualization, the open-ended nature of tasks can either enhance or limit students' fluency performance (Préfontaine & Kormos, 2015). More specifically, in open tasks, learners can select only the information that they can express with their own resources. In other words, even if they conceptualized a highly complex or elaborated message, they could modify or simplify the message so as to express it with their own linguistic repertoires. Therefore, it seems plausible to argue that the argumentative task requiring content generation should have forced students

to engage with content elaboration, while students could avoid using difficult or complex linguistic items.

Notably, the frequency of filled pauses was higher in the argumentative task than in the picture narrative task. This finding is in line with previous studies reporting that the conceptualization demands set by providing many alternative choices (Christenfeld, 1994) and by discourse transitions (Fraundorf & Watson, 2014; Greene & Cappella, 1986; Roberts & Kirsner, 2000) led to the increase of filled pauses. In this sense, the current operationalization of conceptualizing demands posed by possible alternative information and/or discourse organization in the argumentative task may have succeeded in elevating the conceptualization demands.

The responses in the post-speaking questionnaire may also support the high conceptualizing demands in the argumentative task. First, the mean scores of conceptualization difficulty were slightly higher in the argumentative task than in the picture narrative task (M = 3.40 in the argumentative task vs. M = 2.85 in the picture narrative task, on a 6-point scale). Second, the qualitative responses also suggested that in the argumentative task, speakers were engaged with the generation of content and the organization of discourse—planning their opinions, examples, and supporting information, while maintaining the coherence of information—as indicated in the following excerpt:

- *I was not able to communicate my reasons very clearly when expressing my opinion. It was very difficult to speak clearly and coherently about my opinions.* (Participant ID 3004)

- *I think I repeated my opinions I wanted to assert over and over again, and then my speech was not coherent as a whole.* (Participant ID 2028)

In contrast, some speakers mentioned that even though it was easy to specify what to express due to the predefined content in the picture narrative task, they had difficulty with retrieving the corresponding vocabulary items. These characteristics of the picture narrative task were observed in the following excerpts:

- *The original story was there, so it was easy to get to the point and decide to speak*. (Participant ID 2039)
- *I had difficulty with translating what I wanted to express into English because I didn't have the relevant vocabulary. I also had difficulty with coming up with the conjunctions.* (Participant ID 3001)

Comparing the argumentative task with the picture narrative task, it can also be hypothesized that in the argumentative task, speakers would have been required to use lexically sophisticated items and syntactically complex structures due to the abstract topic (i.e., economic effects of holding the Olympics). In accordance with this hypothesis, some speakers reported difficulties with retrieving lexical items and expressions due to the difficulty of the topic of the argumentative task. Other speakers found that the picture narrative task was relatively easier particularly in terms of lexical encoding than the argumentative task due to the familiar topic of the picture prompt. These were observed in the following excerpts:

- *It* (the argumentative task) *was a formal subject, so I had difficulty with retrieving the words and such*. (Participant ID 3023)
- *Compared to expressing an opinion* (the argumentative task), *the story* (the picture narrative task) *was easier to do because I could use phrases that I use in everyday life.* (Participant ID 3021)

However, even in the picture narrative task, there were some concepts or events that were challenging to express for the participants who do not have an opportunity to use L2 English in daily conversational settings. This might be supported by the similar scores of perceived lexical encoding demands between the picture narrative task and the argumentative task (M = 4.13 in the argumentative task vs. M = 3.98 in the picture narrative task, on a 6-point scale) and also by the following excerpts:

- *It was difficult to explain some pictures. I didn't know the English translation of Katei Saien* (home gardening). (Participant ID 3005)
- *(What I found challenging was) to describe Noujo ga kakudai siteiku* (the expansion of farm). (Participant ID 3003)

These excerpts above indicate that the respondents who reported high demands on lexical retrieval in the picture narrative task appeared to encode the L2 lemmas that correspond to L1 lemmas in their mental lexicon. This is possibly because due to the lack of the knowledge of appropriate L2 lemmas in their mental lexicon, the obligatory events or concepts for describing the picture prompt activated the corresponding L1 lemmas even in the mode of L2 English. Such activation of L1 lemmas may have enhanced the competition between L1 and L2 lemmas and, if they failed to retrieve L2 lemmas, it led to the modification of the

preverbal message so that it would consist of concepts that can be expressed using their own lexical repertoires. As a result, despite the small room for content generation, participants' speaking performance was less fluent in the picture narrative task than in the argumentative task.

In addition, a close examination of each UF measure can provide a more nuanced picture of how the enhanced conceptualizing demands are reflected in UF performance. Regarding articulation rate, this measure of speed fluency is assumed to tap into the overall efficiency of speech production due to the serial nature of speech production (Kormos, 2006; S. Suzuki & Kormos, 2020). The GLMM showed that articulation rate was enhanced as a function of the conceptualizing demands in the current study. This result is opposed to the findings in Préfontaine and Kormos (2015) where articulation rate was higher in the related picture narrative task (i.e., the condition of reduced conceptualizing demands) than in the unrelated one. This opposite pattern of speed fluency can be explained by the different range of participants' proficiency level between their study and the current study. The participants in the current study were mostly pre-intermediate-level learners (the B1 level on the CEFR scale; see Section 6.4), whereas the proficiency levels of Préfontaine and Kormos' (2015) participants were reported to range from pre-intermediate to proficient levels. It can thus be hypothesized that competent L2 learners may not have difficulty with retrieving lexical items even with the predefined content of speech and thus can experience the same degree of lexical processing demands, regardless of whether the task is open or closed (i.e., related vs. unrelated picture narrative tasks). Accordingly, the overall efficiency of speech production may be subject to the conceptualizing demands and thus be challenged in the open task (i.e, unrelated picture narrative task) in Préfontaine and Kormos' (2015) study. Meanwhile, in closed tasks, lower-level learners can also theoretically benefit from the macrostructure of

speech by its predefined content of speech (Tavakoli & Skehan, 2005). However, due to their limited lexical repertoires, the formulation demands may have been higher in the closed task than in the open task in the current study. The less fluent speech of the current participants in the (related) picture narrative task may indicate that the enhanced formulation demands might have exceeded the beneficial effects of the macrostructure of speech (i.e., reduced conceptualizing demands).

Assuming that mid-clause pauses in utterances are reflective of disruptions in L2-specific speech processing (De Jong, 2016b; Götz, 2013; Lambert et al., 2017; Tavakoli, 2011), the abovementioned interplay of content generation and proficiency (especially, linguistic repertoires) in articulation rate might also be applied to mid-clause pause ratio. Meanwhile, prior research suggests that end-clause pauses are related to conceptualization-related processes (De Jong, 2016b; Tavakoli, 2011). It can thus be hypothesized that end-clause pauses would be more frequent in open tasks where conceptualizing demands were elevated, compared to closed tasks. However, the current study showed the opposite pattern. The frequency of end-clause pauses was higher in the picture narrative task than in the argumentative task. These unexpected behaviours of end-clause pauses may also have derived from the potential limited linguistic repertoire of the current participants. As argued above, the lack of L2 lemmas corresponding to the conceptualized ideas would require speakers to modify the planned message. The predefined content of the picture narrative task in the current study could have provided such an opportunity for the modification of speakers' message and subsequently might have increased the frequency of end-clause pauses due to the shortage of attentional resources by the modification of the preverbal message (cf. Felker et al., 2019).

To sum up, the increased frequency of filled pauses and the students' perceptions suggested that the enhanced conceptualizing demands may have been successfully operationalized by the open-ended nature and task requirement of the argumentative task. However, due to the potential interplay between the predefined content and the relatively lower proficiency of participants in the current study, the formulation demands in the picture narrative task appeared to be higher than in the argumentative task, as indicated by less fluent speaking performance in the picture narrative task. It can thus be suggested that the effects of processing demands may vary, depending on learners' proficiency level.

7.3.2 Enhanced activation of linguistic and phonological representations

To examine which UF measures are linked to the enhancement of activation levels of linguistic and phonological representations, Study 2 also compared students' UF performance between the picture narrative and text summary tasks without RAA and between the text summary tasks without and with RAA, respectively. The results of GLMMs indicated that the enhanced activation of linguistic representations (the picture narrative task vs. the text summary task without RAA) resulted in slower articulation rate, more mid-clause pauses, fewer but longer end-clause pauses, more filled pauses, fewer repetitions, and more false starts. Meanwhile, the enhanced activation of phonological representations (the text summary tasks without vs. with RAA) resulted only in longer mid-clause pauses.

Due to the nature of activation spreading in speech production (Kormos, 2006; Levelt, 1989), the enhancement of activation level of linguistic representations is supposed to facilitate the retrieval of the activated items (see Section 3.8). Accordingly, it was hypothesized that the enhanced activation of linguistic representations would facilitate the retrieval of linguistic items and thus would enhance UF performance. However, participants' speaking

performance in the text summary task (without RAA) was characterised by slower articulation rate and more mid-clause pauses. From a theoretical perspective, articulation rate captures the overall efficiency of speech production (cf. Kormos, 2006; S. Suzuki & Kormos, 2020), and mid-clause pauses are reflective of disruptions in speech processing due to L2-specific problems (Götz, 2013; Lambert et al., 2017). Contrary to the hypothesis, the results here suggested that the retrieval of linguistic items for speech may have been impeded by activating the linguistic representations of relevant items embedded in the source text. One possible scenario for this unexpected finding might be the complex interplay between the source text and the learners' proficiency level. First, the source texts of the text summary task in the study were supposed to provide some useful or even necessary vocabulary items and grammatical structures to accomplish the subsequent rendering of the text. Since participants were engaged with the reading comprehension of the source texts, the linguistic items in the text should have been activated in learners' mental lexicon. Second, I manipulated vocabulary items in the texts so that participants can understand the texts. However, those items may not have been necessarily available for production (i.e., receptive but not productive vocabulary; Nation, 2013). Due to the partially acquired status of these activated items, speakers may have lacked some lexical properties needed to use them in their speech (e.g., syntactic properties, phonological forms). The failure of retrieval of such linguistic information can lead to the slow retrieval of linguistic items or even breakdowns (see Uchihara et al., 2020), which can be observed as mid-clause pauses (cf. De Jong, 2016). As such, the higher activation of partially acquired linguistic items in the text summary task might have increased the frequency of mid-clause pauses.

Despite the predefined content of speech in both the picture narrative task and the text summary task (without RAA), the increased frequency of filled pauses in the text summary

task may indicate that conceptualizing demands might have been slightly higher in the text summary task than in the picture narrative task, because filled pauses might be related to the high demands on discourse/content planning (Fraundorf & Watson, 2014; Greene & Cappella, 1986; Roberts & Kirsner, 2000; see also Section 7.3.1). The slightly higher conceptualizing demands in the text summary task might also be indicated by the scores of the post-speaking questionnaire (M = 2.85 in the picture narrative task vs. M = 3.40 in the text summary task on a 6-point scale). Besides, the following participants' responses also suggested that conceptualizing demands in the text summary task may have derived from the necessity of recalling and selecting the content of source texts:

- *I thought I understood the text largely, but I had difficulty with remembering what it said when I was speaking*. (without-RAA, Participant ID 2028)
- *I wanted to choose the information that would make it easier to understand, but I couldn't.* (without-RAA, Participant ID 3007)

The slightly higher conceptualizing demands in the text summary task could also explain the increase in the number of false starts and the longer duration of end-clause pauses. From the perspective of speech production, false starts may indicate a need for more resources for conceptual processing, including discourse management (Williams & Korko, 2019). Similarly, end-clause pauses are supposed to capture the conceptualization processes such as content planning (De Jong, 2016b; Götz, 2013; Lambert et al., 2017). Accordingly, it seems plausible to argue that in the text summary task, participants might have experienced a shortage of attentional resources and needed more time for content planning due to the active engagement with recalling and selecting information from the source text. However, it should be noted that the *frequency* of end-clause pauses decreased in the text summary task. Albeit

speculative, one possible explanation is that the frequency and duration of end-clause pauses may reflect different components of conceptualization, that is, macroplanning and microplanning (Kormos, 2006; Levelt, 1989). From a theoretical perspective, macroplanning is responsible for the generation and selection of information to communicate, while microplanning transforms such selected information into the propositional form by specifying the informational aspects of the message, such as referents and argument structures (see Section 2.5). Accordingly, the abovementioned demands on selecting information in the text summary task can be closely associated with macroplanning difficulty. It can thus be hypothesized that the longer duration of end-clauses was reflective of the difficulty in macroplanning processes. However, to comprehend the source text of the text summary task, participants were supposed to have parsed the source text and thus to specify the informational aspects of (at least most parts of) the text at the phase of text reading (i.e., the creation of situation model; see van Dijk & Kintsch, 1983). In other words, reading comprehension in the text summary task could have created the memory trace of such informational aspects of the text in students' short-term memory. As a result, due to the memory trace of the information required for microplanning, the breakdowns might have been reduced at clausal boundaries. However, the potential correspondence between macroplanning and duration of end-clause pauses and between microplanning and frequency of end-clause pauses has to be tested in carefully designed experimental studies.

Comparing UF performance between two conditions of the text summary task, it was examined which UF measures are associated with the enhanced activation of phonological representations operationalized by RAA (i.e., reading-while-listening in text comprehension). The bimodal input of the source text was assumed to further activate the relevant linguistic items embedded in the text and thus to facilitate the retrieval of those items. A set of GLMMs

showed that the significant effect of RAA was only found in mid-clause pause duration. However, the direction of the effect was opposite to the hypothesis. Students produced longer mid-clause pauses in the with-RAA condition than in the without-RAA condition. This finding can be explained by the potential interplay of the activation of linguistic items and the relatively lower proficiency level of the current participants. The bimodal source text in the RAA condition should have enhanced the activation level of in-text linguistic items to a larger extent than the written-only source text comprehension (i.e., without-RAA condition). Accordingly, it may also have enhanced the competition in lexical retrieval between learners' productive vocabulary items and the activated in-text items, which were only available for comprehension, subsequently extending the latency of retrieving linguistic items. Assuming that mid-clause pauses are associated with disruptions in the retrieval of L2-specific linguistic knowledge (De Jong, 2016b; Götz, 2013; Lambert et al., 2017), such delayed retrieval could have contributed to the longer duration of mid-clause pauses in the with-RAA condition of the text summary task.

7.4 Summary

This chapter reported the results of Study 2 and discussed the findings from the perspective of speech production mechanisms and task design features. RQ2 was concerned about how L2 UF performance varies across four types of speaking tasks which differ in the quality of speech processing demands. To address RQ2, students' UF performance was compared across four speaking tasks, constructing the GLMMs for each of 11 UF measures with the categorical variable of task type as a fixed-effects factor and individual students as random intercepts. Possibly due to the relatively lower proficiency level of the current participants, the results of the effects of different speech processing demands on UF performance were different from previous studies. The findings are summarized in Table 27.

Table 27. *Summary of findings of Study 2*

	Content generation	Enhanced activation of linguistic representations	Enhanced activation of phonological representations
Speed fluency	+ Articulation rate + Speech rate + Mean length of run	- Articulation rate - Speech rate	–
Breakdown fluency (frequency)	- Mid-clause pause ratio - End-clause pause ratio + Filled pause ratio	+ Mid-clause pause ratio - End-clause pause ratio	–
Breakdown (duration)	–	+ End-clause pause duration	+ Mid-clause pause duration
Repair fluency	- Self-repetition ratio - Self-correction ratio	- Self-repetition ratio + False start ratio	–

Note. The + sign indicates the increase of the utterance fluency measures; the – sign indicates the decrease of the utterance fluency measures.

The increased demands on conceptualization were examined by comparing UF performance between the argumentative and picture narrative tasks. Although the increased frequency of filled pauses and the students' responses confirmed the enhanced conceptualizing demands in the argumentative task, students' speech was more fluent at three subconstructs of UF—speed, breakdown, and repair fluency—in the argumentative task than in the picture narrative task. This enhancement of UF performance with the increased conceptualizing demands showed a potential complex interplay of task design features and proficiency level. The open nature of the argumentative task could have allowed students to avoid difficult lexical items and complex syntactic structures by modifying their preverbal message. In contrast, in the picture narrative task, they could not avoid such linguistic demands due to the predefined content of the task, and this was manifested in disruptions in speech processing (as indicated by the higher frequency of silent pauses). As a result, participants produced less fluent speech in the picture narrative task than in the argumentative task.

The contrasts for the enhanced activation of linguistic and phonological representations also suggested the potential complex interplay of task design features and L2 proficiency. The results showed that the enhanced activation of linguistic representations (the picture narrative task vs. the text summary task without RAA) led to slower articulation rate and more mid-clause pauses, both of which are reflective of the efficiency in L2 speech processing. Meanwhile, the enhanced activation of phonological representations resulted in longer mid-clause pauses, based on the contrast between the two conditions of the text summary tasks. These findings indicate that the linguistic items activated by the source text may have not been fully acquired, and thus students could not have used those items in their output. Furthermore, the activation of partially acquired items, at either linguistic or phonological levels, may have inhibited students from retrieving their own linguistic resources that were readily available for productive use, which was observed as disfluency.

Chapter 8: Results and Discussion of Study 3—Contributions of Cognitive Fluency to Utterance Fluency

8.1 Introduction

This chapter reports the results of Study 3 and discusses them in light of the construct definition of CF and UF as well as the interrelationship between the subconstructs of CF and UF. The first RQ of Study 3 (RQ3-1) is concerned with the contribution of CF to UF at the level of constructs. RQ3-2 examines the variability of the CF-UF link across four speaking tasks which were designed to differ in the quality of speech processing demands. Prior to answering these RQs, the dimensionality of CF and UF (RQ3-1a, RQ3-1b) was also tested. To this end, the descriptive statistics of CF measures (for the descriptive statistics of UF measures, see Section 7.2.1) and intercorrelations of CF and UF measures were explored (Sections 8.2.1, 8.2.3). In order to address RQ3-1a and RQ3-1b, a set of CFA models of CF and UF was tested with regard to the goodness of fit to the current dataset (Sections 8.2.2, 8.2.4). As a preliminary analysis to the SEM analysis of the CF-UF link, the intercorrelations between CF and UF measures were examined across the tasks (Section 8.2.5). Using the best-fit factor structures identified through the CFA models, the SEM model of the CF-UF link was constructed, and its model-fit indices and regression coefficients were reported (Section 8.2.6). These findings were discussed from the perspective of L2 speech production mechanisms, providing some insights into the construct validity of CF and UF components (Section 8.3).

8.2 Results

8.2.1 Descriptive statistics and intercorrelation of cognitive fluency measures

As a preliminary analysis to the subsequent multivariate analyses (CFA and SEM for RQ3-1 and RQ3-2), the descriptive statistics of CF measures and the intercorrelation between CF

measures were examined. The descriptive statistics of the CF measures are summarized in Table 28. According to the results of the Shapiro-Wilk tests, the RT and accuracy scores of the maze task and the accuracy scores of GJT (Morphology and Syntax) were not normally distributed ($p < .05$). Considering the non-normal distributions of these CF measures, the interrelationship among CF measures was examined using Spearman's rank-order correlation coefficients. The correlation matrix is shown in Table 29.

Table 28. *Descriptive summary of cognitive fluency measures in Study 3.*

				Shapiro-Wilk test	
Cognitive fluency measures	*Mean*	*SD*	*SE*	*Statistics*	*p-value*
PVLT	25.55	6.92	0.61	0.993	0.756
Picture Naming RT	1099.05	180.48	15.95	0.990	0.522
Maze Word RT	1164.43	202.93	17.94	0.969	0.005
Maze Word Accuracy	385.45	39.73	3.51	0.861	< .001
GJT Morphology RT	4039.66	969.44	85.69	0.989	0.396
GJT Syntax RT	4291.60	955.92	84.49	0.990	0.461
GJT Morphology Accuracy	20.17	3.12	0.28	0.976	0.024
GJT Syntax Accuracy	31.28	4.13	0.37	0.947	< .001
Articulatory speed	190.26	26.04	2.30	0.990	0.502

N.B. RT measures are expressed in milliseconds. Articulatory speed refers to the mean number of morae per minute.

The results of intercorrelations suggested several association patterns of CF measures within and between linguistic domains (vocabulary, grammar, and pronunciation). The interpretation of the effect sizes r_s was guided by Plonsky and Oswald's (2014) guideline tailored for L2 research contexts (r = |.25–.40| as Small; r = |.40–.60| as Medium; r = |.60–1.00| as Strong).

First, there was a weak association between two vocabulary knowledge measures—PVLT (vocabulary size) and Picture Naming RT (lexical retrieval speed). Although Picture Naming RT was only weakly related to the speed of sentence construction (Maze Word RT) and that of articulation (Articulatory speed), PVLT was correlated with all of the CF measures to a moderate-to-strong degree. Interestingly, PVLT was more strongly correlated with other CF

measures (r_s = |.330–.605|) than with the other lexical measure, that is, Picture Naming RT (r_s = -.259). In other words, grammatical knowledge measures, including sentence construction skills and grammatical error detection skills, were more closely related to how widely learners know vocabulary items (PVLT) than to how efficiently they can retrieve vocabulary items (Picture Naming RT).

Second, within the domain of grammatical knowledge, there was a moderate correlation between the RT and accuracy measures of the maze task (r_s = -.400), indicating that these two dimensions of sentence construction may be related but are relatively distinct. Meanwhile, among the GJT measures, there were no meaningful associations between speed and accuracy dimensions in either morphological (r_s = -.028) or syntactic (r_s = -.063) items. However, as for the between-linguistic levels, the GJT RT scores were strongly correlated between morphological and syntactic items (r_s = .923), while the accuracy scores were moderately correlated between the morphological and syntactic items (r_s = .494). These correlational patterns suggest that in the context of GJT, there is a clear distinction between speed and accuracy dimensions. Syntactic and morphological processing were more closely related to each other in the speed dimension than in the accuracy dimension. Moreover, the GJT RT and accuracy measures were also weakly or moderately correlated with the corresponding measures of the maze task (r_s = .459–.471 for RT; r_s = .299–.544 for accuracy), meaning that despite some variability in the strength of correlation coefficients, the speed and accuracy measures of the maze task and the GJT may tap into the same construct of grammatical knowledge.

Third, articulatory speed was moderately correlated with the different RT measures (r_s = |.299–.477|) and PVLT (vocabulary size; r_s = .356), while weakly or non-significantly with the accuracy measures (r_s = .104–.192). Considering the essential role of vocabulary size in

various speech processing (Kormos, 2006), it seems plausible to argue that the current measure of articulatory speed was directly or indirectly related to the vocabulary size measure. In addition, it is suggested that the articulatory speed measures might be associated with the speed aspects of linguistic knowledge rather than the accuracy aspects.

Finally, strong correlations across linguistic domains were found between PVLT and GJT Syntax accuracy (r_s = .605) and between PVLT and Maze accuracy (r_s = .588). Since both GJT Syntax accuracy and Maze accuracy tap into the target-likeness of syntactic knowledge, these strong correlations across linguistic domains indicate that the more lexical items L2 learners have acquired, the more target-like their syntactic knowledge is.

Table 29. *A correlational matrix of cognitive fluency measures.*

Cognitive fluency measures	2	3	4	5	6	7	8	9
1. PVLT	-0.259**	-0.534***	0.588***	-0.370***	-0.372***	0.330***	0.605***	0.356***
2. Picture Naming RT	—	0.384***	-0.113	0.171	0.120	-0.192*	-0.115	-0.299***
3. Maze Word RT		—	-0.400***	0.471***	0.459***	-0.282**	-0.328***	-0.477***
4. Maze Word Accuracy			—	-0.164	-0.168	0.299***	0.544***	0.104
5. GJT Morphology RT				—	0.923***	-0.028	-0.013	-0.449***
6. GJT Syntax RT					—	-0.036	-0.063	-0.446***
7. GJT Morphology Accuracy						—	0.494***	0.108
8. GJT Syntax Accuracy							—	0.192*
9. Articulatory speed								—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

8.2.2 Confirmatory factor analysis of cognitive fluency

To specify the factor structure of CF best fitting the current dataset (RQ3-1a), three proposed CFA models were tested. The first model (CF Model 1; see Figure 20) was a single-factor model, which assumes that CF is a unitary construct. Statistically speaking, one of the advantages of a single-factor model is that the model is constructed with the minimum number of parameters, meaning that the estimation of the proposed model is relatively robust for a small sample size. From a theoretical perspective, the dimensionality of the construct of CF has not been specified in the literature of L2 fluency (cf. Segalowitz, 2010, 2016). It is thus essential to explore the extent to which a single-factor model fits the current set of CF measures.

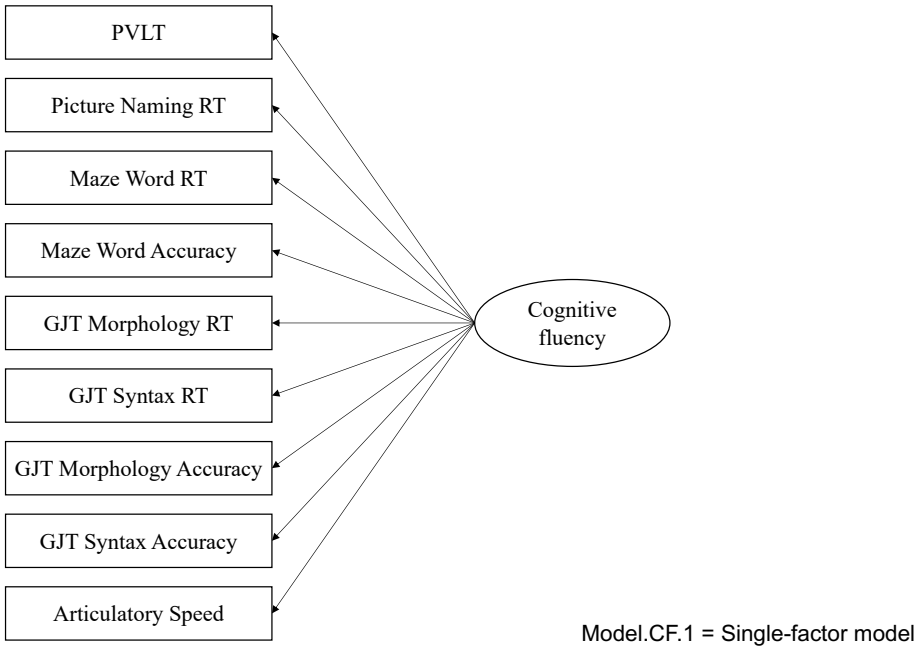


Figure 20. A single-factor model of cognitive fluency (Model.CF.1).
Note. Residuals are omitted for the sake of brevity.

The second model (CF Model 2; see Figure 21) consisted of two subconstructs of CF, namely, *linguistic resource* and *processing speed.* These two subconstructs were, on the one hand, motivated by the distinction made in empirical studies (De Jong et al., 2013; Kahng, 2020). On the other hand, the two-factor model was also in line with the conceptualization of

CF, which regards the association to UF as one of the essential characteristics of the theoretically valid definition of CF (Segalowitz, 2010, 2016). From the perspective of speech production mechanisms, disruptions in speech processing, which are observed as breakdowns at the level of utterances, are assumed to be caused by either lack of linguistic resources or slow processing speed (see Section 3.6). Therefore, the latent variable of linguistic resource consisted of the CF measures capturing the range of linguistic resources (the PVLT scores, the accuracy score of the maze task, and the accuracy scores of the GJT), while the latent variable of processing speed was composed of RT-based measures and the articulatory speed measure.

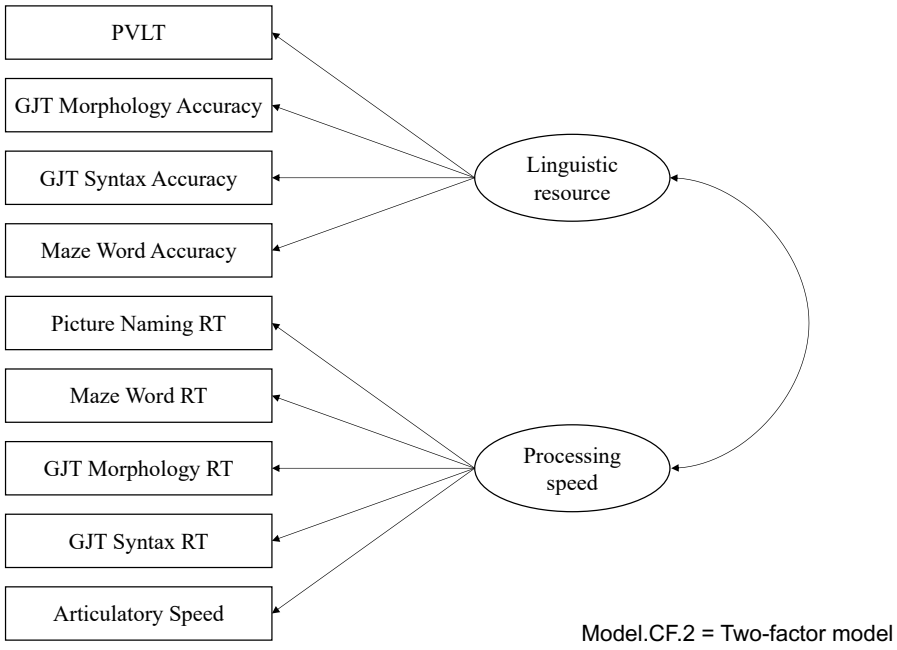


Figure 21. A two-factor model of cognitive fluency (Model.CF.2).
Note. Residuals are omitted for the sake of brevity.

Finally, I proposed a three-factor model which comprises linguistic resource, processing speed, and monitoring speed (CF Model 3; see Figure 22). The rationale for separating monitoring processes from encoding processes is that linguistic encoding processes and monitoring processes can differ in the modality of processing. The former entails the activation and retrieval of linguistic knowledge, while the latter is operated by speakers'

comprehension mechanisms (see Section 2.7). However, from the perspective of speech production mechanisms, the linguistic resources for monitoring processes are identical with those for linguistic encoding processes, particularly at lexical and morphosyntactic levels (Levelt, 1999). In other words, although both encoding and monitoring processes may access the same knowledge resources, their operators are hypothesized to be different—formulator and monitoring loops, respectively (Kormos, 2006; Levelt, 1989). In the current study, the distinction between encoding and monitoring processes was made at the syntactic level; the maze task captures grammatical encoding processes, while the GJT measures tap into monitoring processes (see Section 6.8.2). Considering the potentially identical resources for encoding and monitoring processes, the accuracy measures of the maze task and the GJT task remained to load on the latent variable of linguistic resource. Meanwhile, the RT measures of the GJT task (GJT Morphology RT, GJT Syntax RT) were used to create the third subconstruct of CF, that is, *Monitoring speed*. The remaining speed-related measures loaded on the latent variable of processing speed. Therefore, the observed variables of processing speed are different between the two- and three-factor models. For these proposed CFA models of CF, the residual covariances were set across CF tasks (e.g., Maze task RT and Accuracy measures; for the full R code, see Appendix P).

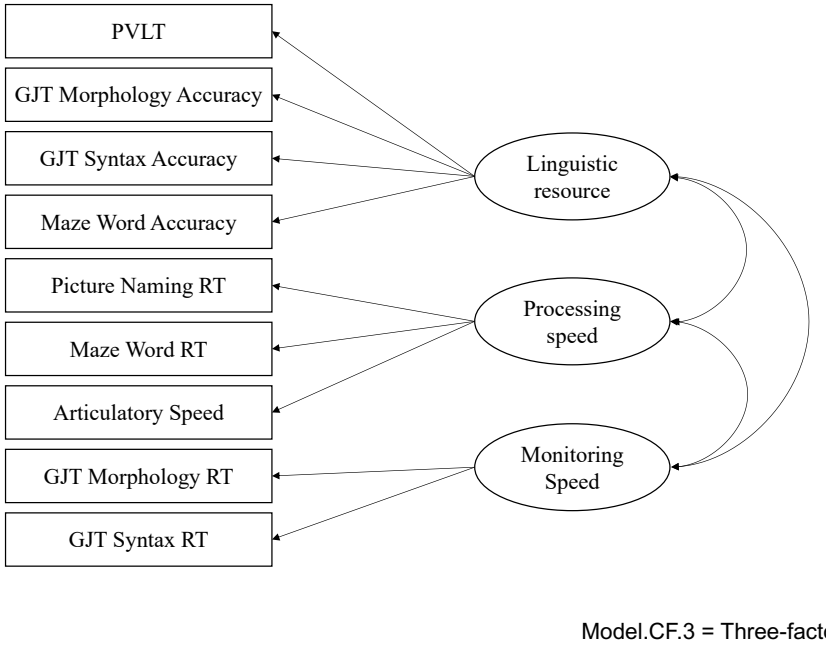


Figure 22. A three-factor model of cognitive fluency (Model.CF.3).
Note. Residuals are omitted for the sake of brevity.

As suggested by the descriptive statistics of CF measures (see Section 8.2.1), the assumption of univariate normality for SEM was violated in some CF measures (GJT Accuracy scores, Maze RT and Accuracy scores). In response to this violation of univariate normality, the CFA analyses used the maximum likelihood parameter estimates with standard errors and a mean-adjusted chi-square test statistic (i.e., Satorra-Bentler corrections; hereafter, robust maximum likelihood estimation), which is considered robust for non-normal distributions of observed variables (Deng et al., 2018). As for the factorability, the results of the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy (KMO = 0.758) suggested that the current set of CF measures can be considered as showing a moderate factorability (KMO > .70; Field, 2009). Note that for the sake of understandability of results, the CF measures based on RT (Picture Naming RT, Maze Word RT, GJT Syntax RT, and GJT Morphological RT) were inversed in the subsequent CFA and SEM analyses.

The model-fit indices of the three proposed models are summarized in Table 30. Due to the relatively small sample size (N < 250), as well as the estimation method (i.e., maximum

likelihood), the evaluation of the CFA models was mainly based on the statistics of SRMR (< .08) and CFI (> .95) (Hu & Bentler, 1998). The SRMR indices indicated a good fit for all three models, while two- and three-factor models (SRMR = 0.051) showed a slightly better fit than the single-factor model (SRMR = 0.078). The better fit of the two- and three-factor models was also indicated by the CFI index. Comparing these two models, it can be argued that the two-factor model has a slightly better model fit than the three-factor model consistently in many model-fit indices. In principle, the more parsimonious the model is (i.e., fewer parameters), the more robust the estimation of the model is in terms of the residuals between the estimated and observed data (Schoonen, 2015). Study 3 thus adopted the two-factor model for the factor structure of CF in the subsequent SEM analysis.

Table 30. *Selected model-fit indices for the three tested CFA models of cognitive fluency.*

Model	*df*	*χ2*	*p-value*	*χ2/df ratio*	CFI	TLI	SRMR	RMSEA [90%CI]
One-factor	20	65.179	< .001	3.259	0.919	0.854	0.078	0.133[0.098, 0.169]
Two-factor	19	32.296	0.029	1.700	0.976	0.955	0.051	0.074[0.024, 0.117]
Three-factor	17	32.286	0.014	1.899	0.973	0.942	0.051	0.084[0.037, 0.127]

Note. CFI = comparative fit index; SRMR = standardized root mean square; RMSEA = root mean square error of association; the cut-off values for good fit: χ2/df ratio < 2.0; SRMR < .08; CFI and TLI > .95, RMSEA < .06.

The standardised regression coefficients (β values) and their 95% confidence intervals in the two-factor model are summarised in Table 31. All the regression paths were found significant. The strengths of regression coefficients can be compared in terms of the confidence intervals, using an analogy to *t*-tests. Regarding the latent variable of linguistic resource, the regression path of the PVLT (vocabulary size) was significantly stronger than that of the GJT accuracy score of morphological items. Meanwhile, there were no statistical differences in the strengths of regression coefficients between lexical (PVLT) and syntactic resources (Maze Word Accuracy, GJT Syntax Accuracy). As for the latent variable of processing speed, the strongest regression path was found in Maze Word RT (syntactic processing speed). Moreover, the regression path of Maze Word RT was stronger than the

other four observed variables. The primary component of processing speed is thus supposed to be the speed of sentence construction. Note that there were no significant differences in the strengths of regression paths among the remaining four observed variables.

Table 31. *Summary of the standardized regression coefficients and their 95% confidence intervals of the finalized CFA model of cognitive fluency*

					95%CI	
Latent variable	Direction	Observed variable	*β*	*p*	Lower	Upper
Covariance between latent variables						
Linguistic resource	–	Processing speed	0.676	< .001	0.515	0.838
Measurement model						
Linguistic resource	→	PVLT	0.867	< .001	0.800	0.933
	→	Maze Word Accuracy	0.689	< .001	0.552	0.826
	→	GJT Morph. Accuracy	0.455	< .001	0.264	0.647
	→	GJT Syn. Accuracy	0.722	< .001	0.610	0.834
Processing speed	→	Pic. Naming RT	0.424	< .001	0.269	0.580
	→	Maze Word RT	0.862	< .001	0.762	0.963
	→	GJT Morph. RT	0.616	< .001	0.472	0.761
	→	GJT Syn. RT	0.590	< .001	0.441	0.739
	→	Articulatory speed	0.589	< .001	0.464	0.714

8.2.3 Intercorrelation of utterance fluency measures

To examine the dimensionality of UF (RQ3-1b), different factor structures of UF were tested. As a preliminary analysis, the interrelationship between the observed variables of UF measures was inspected. Due to the non-normal distributions observed in most of the UF measures in Study 2 (see Section 7.2.1), the intercorrelations among UF measures were examined by Spearman's rank-order correlation coefficients, separately for each task (see Table 32–35). In general, the interrelationship among UF measures did not differ across tasks. However, the strengths of associations differed across the subconstructs of UF. Speed fluency measures strongly correlated with each other (r_s = .684–.924). Overall, there were moderate-to-strong correlations among breakdown fluency measures, with some non-significant correlations between filled pause ratio and end-clause pause measures (duration and ratio). Among repair fluency measures, self-repetition ratio was correlated moderately-

to-strongly with self-correction ratio (r_s = .399–.604) and weakly-to-moderately with false start ratio (r_s = .221–.392).

The strongest correlations between subconstructs of UF may indicate the general interrelationship between speed, breakdown, and repair fluency. Regarding the relationship between speed and breakdown fluency, there were highly strong correlations between mean length of run and mid-clause pause ratio (r_s = -|.975–.980|), suggesting that learners who produced a longer run were less likely to stop in the middle of clauses. Meanwhile, this exceedingly high correlation can also be explained by the fact that the measure of mean length of run mathematically reflects the frequency of pauses, including both mid- and end-clause pauses. Compared to the moderate correlations between mean length of run and end-clause pause ratio (r_s = |.455–.531|), it can be assumed that mid-clause pauses might have been more frequent than end-clause pauses (see also the descriptive statistics in Table 23). However, as with speech rate, mean length of run is a composite measure (see Section 6.7). Their correlations with pause frequency measures are thus expectable or even circular. Despite slightly lower correlation coefficients, articulation rate was also strongly correlated with mid-clause pause ratio (r_s = -|.716–.803|). Considering the construct validity of these speed fluency measures, articulation rate should be regarded as a representative measure of speed fluency, because this measure is methodologically independent of breakdown fluency features (i.e., pausing behaviour). Accordingly, the strong association between articulation rate and mid-clause pauses indicated that the association between speed and breakdown fluency was not entirely due to the methodological procedure of measure calculation and rather was supposed to reflect the close relationship at the level of constructs. As for the association between speed and repair fluency, both articulation rate and mean length or run were moderately-to-strongly correlated with self-repetition ratio (r_s = -|.507–.597|, -

|.413–.630|, respectively). Similarly, regarding the breakdown-repair fluency link, there was a moderate-to-strong correlation between mid-clause pause ratio and self-repetition ratio (r_s = -|.420–.633|).

Table 32. *A correlational matrix of the utterance fluency measures in the argumentative task.*

	2	3	4	5	6	7	8	9	10	11
1. AR	0.825***	0.782***	-0.778***	-0.354***	-0.381***	-0.460***	-0.382***	-0.558***	-0.444***	-0.304***
2. SR	—	0.905***	-0.905***	-0.402***	-0.605***	-0.767***	-0.691***	-0.572***	-0.353***	-0.207*
3. MLR		—	-0.978***	-0.486***	-0.587***	-0.535***	-0.531***	-0.630***	-0.432***	-0.277**
4. MCPR			—	0.329***	0.567***	0.567***	0.517***	0.633***	0.380***	0.275**
5. FCPR				—	0.272**	0.103	0.359***	0.184*	0.351***	0.046
6. FPR					—	0.545***	0.388***	0.486***	0.198*	0.265**
7. MCPD						—	0.586***	0.357***	0.069	0.156
8. FCPD							—	0.227**	0.075	-0.091
9. SRR								—	0.529***	0.392***
10. SCR									—	0.265**
11. FSR										—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 33. *A correlational matrix of the utterance fluency measures in the picture narrative task.*

	2	3	4	5	6	7	8	9	10	11
1. AR	0.834***	0.804***	-0.803***	-0.288***	-0.364***	-0.602***	-0.418***	-0.597***	-0.406***	-0.369***
2. SR	—	0.924***	-0.902***	-0.447***	-0.515***	-0.866***	-0.714***	-0.546***	-0.35***	-0.346***
3. MLR		—	-0.975***	-0.455***	-0.549***	-0.716***	-0.54***	-0.571***	-0.474***	-0.359***
4. MCPR			—	0.273**	0.560***	0.711***	0.482***	0.585***	0.475***	0.381***
5. FCPR				—	0.168	0.308***	0.468***	0.108	0.123	0.066
6. FPR					—	0.494***	0.257**	0.510***	0.397***	0.406***
7. MCPD						—	0.657***	0.368***	0.176*	0.321***
8. FCPD							—	0.295***	0.064	0.103
9. SRR								—	0.604***	0.345***
10. SCR									—	0.251**
11. FSR										—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 34. *A correlational matrix of the utterance fluency measures in the text summary task without RAA.*

	2	3	4	5	6	7	8	9	10	11
1. AR	0.837***	0.730***	-0.729***	-0.298***	-0.303***	-0.576***	-0.518***	-0.518***	-0.251**	-0.227*
2. SR	—	0.906***	-0.903***	-0.396***	-0.455***	-0.843***	-0.713***	-0.487***	-0.249**	-0.052
3. MLR		—	-0.980***	-0.479***	-0.465***	-0.655***	-0.533***	-0.514***	-0.363***	-0.081
4. MCPR			—	0.325***	0.473***	0.678***	0.52***	0.551***	0.348***	0.099
5. FCPR				—	0.121	0.207*	0.332***	0.001	0.255**	-0.111
6. FPR					—	0.366***	0.265**	0.379***	0.239**	0.068
7. MCPD						—	0.600***	0.279**	0.095	-0.035
8. FCPD							—	0.296***	0.037	-0.105
9. SRR								—	0.399***	0.221*
10. SCR									—	0.013
11. FSR										—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 35. *A correlational matrix of the utterance fluency measures in the text summary task with RAA.*

	2	3	4	5	6	7	8	9	10	11
1. AR	0.797***	0.684***	-0.716***	-0.231**	-0.265**	-0.499***	-0.354***	-0.507***	-0.248**	-0.46***
2. SR	—	0.867***	-0.861***	-0.437***	-0.428***	-0.799***	-0.616***	-0.388***	-0.107	-0.322***
3. MLR		—	-0.978***	-0.531***	-0.409***	-0.53***	-0.405***	-0.413***	-0.194*	-0.317***
4. MCPR			—	0.371***	0.401***	0.539***	0.368***	0.420***	0.205*	0.314***
5. FCPR				—	0.229**	0.211*	0.346***	0.126	0.038	0.120
6. FPR					—	0.394***	0.153	0.277**	0.123	0.097
7. MCPD						—	0.489***	0.162	-0.072	0.150
8. FCPD							—	0.111	-0.096	0.038
9. SRR								—	0.471***	0.366***
10. SCR									—	0.141
11. FSR										—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

8.2.4 Confirmatory factor analysis of utterance fluency

As with the CFA models of CF, several CFA models were proposed for UF. Although L2 fluency research has traditionally followed Tavakoli and Skehan's (2005) triad model of UF (speed, breakdown, and repair fluency), the factor structure of UF has not been revisited, even though new UF measures, including pause measures considering pause locations (e.g., mid- vs. end-clause pause ratio), have been more recently introduced. Therefore, Study 3 took an exploratory approach to identifying the factor structure of UF best fitting the current dataset, with regard to the stability across different task types. First, due to its advantage of the minimum number of parameters, a single-factor model was proposed (UF Model 1, see Figure 23). Second, from the theoretical perspective of speech production mechanisms, UF is supposed to reflect the smoothness and disruptions of speech processing. A variety of temporal features of speech can thus be categorised into these two dimensions. Accordingly, I proposed a two-factor model (UF Model 2; see Figure 24) that consists of the two latent variables: *processing smoothness* and *processing disruptions*.

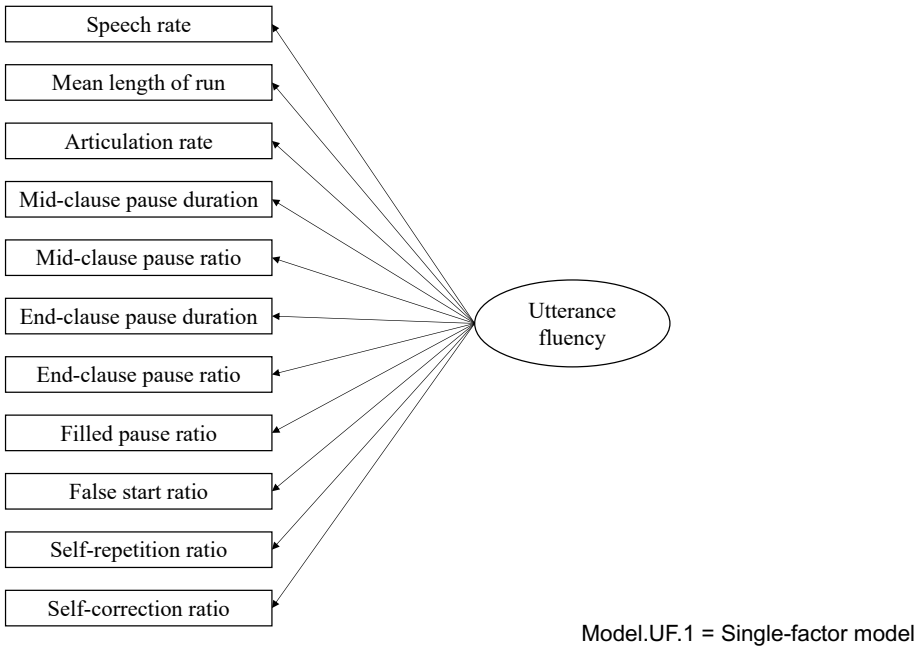


Figure 23. A single-factor model of utterance fluency (Model UF 1).
Note. Residuals are omitted for the sake of brevity.

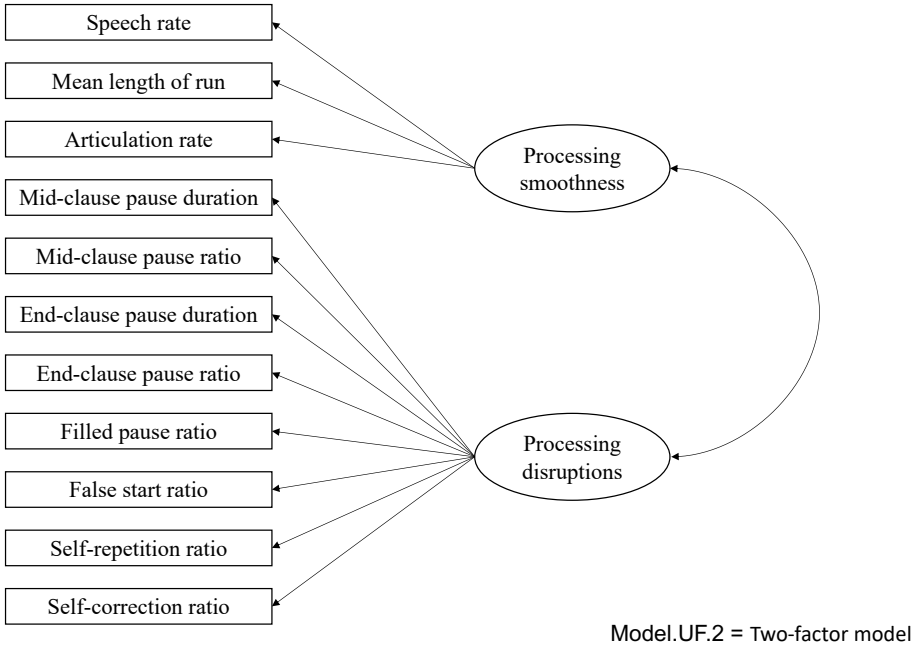


Figure 24. A two-factor model of utterance fluency (Model UF 2).
Note. Residuals omitted for the sake of brevity.

Finally, following Tavakoli and Skehan's (2005) three-dimensional structure of UF, the third model (UF Model 3; see Figure 25) was comprised of three latent variables of speed fluency, breakdown fluency, and repair fluency. Strictly speaking, the measures of speech rate and mean length of run are composite measures, because both measures tap into multiple dimensions of UF (see Section 6.7). However, from the statistical perspective, the number of parameters in the measurement model of CFA (i.e., the relationship between a latent variable and the associated observed variables) is recommended not to exceed the number of data points in the correlation matrix (i.e., under-identified models; Brown, 2006). The number of parameters includes both factor loadings and measurement errors and thus equals the twice of the number of observed variables loaded onto the given latent variable. The number of data points in the correlation matrix (b) can be calculated by the following formula:

$$b = p * (p + 1) / 2$$

Note. p refers to the number of observed variables loaded onto the latent variable.

Although only articulation rate purely taps into the construct of speed fluency (Tavakoli et al., 2020), the composite measures of speech rate and mean length of run were also loaded onto the latent variable of speed fluency to avoid an under-identified model in the CFA model (UF Model 3). Throughout the proposed CFA models of UF, the residual covariances were set between mid- and end-clause pause ratio measures and mean length of run, because among these measures, their measurement errors are commonly attributed to the pause annotation.

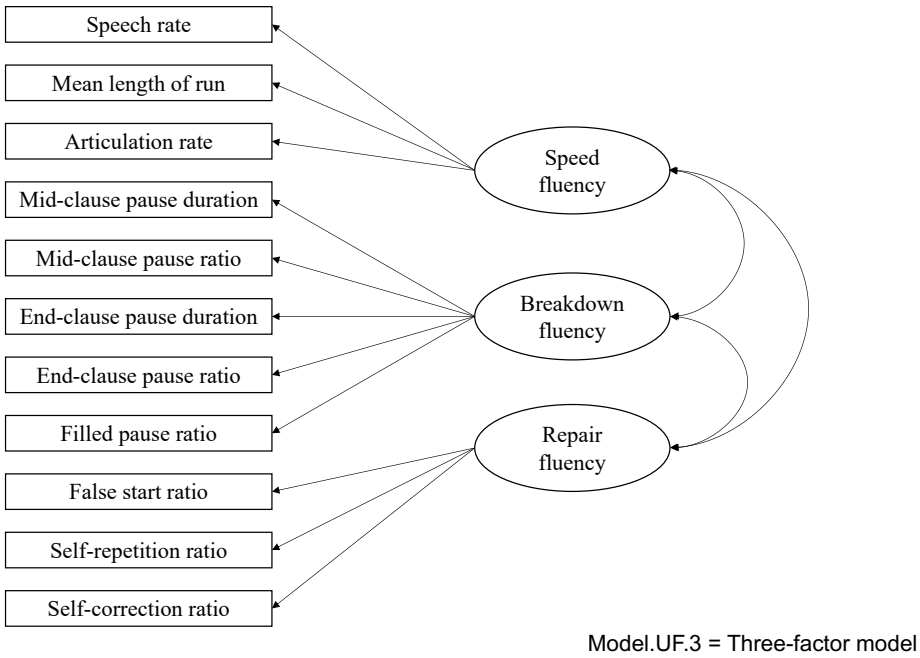


Figure 25. A three-factor model of utterance fluency (Model UF 3).
Note. Residuals are omitted for the sake of brevity.

In response to the non-normal distributions of most of the UF measures (see Section 7.2.1), the CFA models of UF used the robust maximum likelihood parameter estimation (Deng et al., 2018). As a preliminary analysis, the factorability of the current dataset of UF measures was tested separately for each task. The KMO measures of sampling adequacy suggested that the current dataset of UF measures in all of the four tasks (0.772 for the argumentative task; 0.786 for the picture narrative task; 0.790 for the text summary task without RAA; 0.762 for the text summary task with RAA) can be considered as showing a good factorability (KMO

> .70; Field, 2009). For the sake of the interpretability of the findings, the observed variables of breakdown and repair fluency measures were inversed and entered for the subsequent CFA and SEM analyses.

A set of the model-fit indices is summarized in Table 36. Overall, the three-factor model tended to show a relatively better fit across tasks. However, none of the proposed models optimally fit the data (especially, CFI > .95, SRMR < .08), indicating the possibility that there is a better factor structure for the current dataset of UF measures. To explore a better CFA model, a data-driven approach was taken to modify the factor structures. Specifically, the collinearity among observed variables was first inspected.

Table 36. *Selected model-fit indices for the three tested CFA models of utterance fluency.*

Model	*df*	*χ2*	*p-value*	*χ2/df ratio*	CFI	TLI	SRMR	RMSEA [90%CI]
One-factor								
Argumentative	41	264.082	< .001	6.441	0.786	0.713	0.097	0.206[0.183, 0.230]
Pic.Narrative	41	325.263	< .001	7.933	0.775	0.698	0.095	0.233[0.210, 0.257]
TS.withoutRAA	41	331.009	< .001	8.073	0.754	0.670	0.097	0.235[0.212, 0.259]
TS.withRAA	41	285.565	< .001	6.965	0.759	0.676	0.089	0.216[0.193, 0.240]
Two-factor								
Argumentative	40	244.142	< .001	6.104	0.804	0.731	0.079	0.200[0.176, 0.224]
Pic.Narrative	40	324.126	< .001	8.103	0.775	0.690	0.085	0.236[0.212, 0.260]
TS.withoutRAA	40	286.490	< .001	7.162	0.791	0.712	0.079	0.219[0.196, 0.244]
TS.withRAA	40	284.668	< .001	7.117	0.759	0.688	0.096	0.219[0.195, 0.243]
Three-factor								
Argumentative	38	197.022	< .001	5.185	0.848	0.780	0.089	0.181[0.156, 0.206]
Pic.Narrative	38	274.421	< .001	7.222	0.812	0.729	0.092	0.220[0.196, 0.245]
TS.withoutRAA	38	264.358	< .001	6.957	0.808	0.722	0.078	0.216[0.192, 0.241]
TS.withRAA	38	248.775	< .001	6.547	0.792	0.699	0.092	0.208[0.184, 0.233]

Note. CFI = comparative fit index; SRMR = standardized root mean square; RMSEA = root mean square error of association. The cut-off values for good fit: χ2/df ratio < 2.0; SRMR < .08; CFI and TLI > .95, RMSEA < .06.

Although the intercorrelations were checked by non-parametric correlational analyses in the previous section (see Section 8.2.3), the intercorrelation pooled by tasks was tested through

parametric correlational analyses (i.e., Pearson product-moment correlation). This is because an SEM analysis is based on the correlation-matrix based on the parametric correlation coefficients. To inspect the overall intercollinearity among the UF measures, the dataset was pooled over tasks. The correlation coefficients and their heatmap are presented below respectively as Table 37 and Figure 26. To avoid strong collinearity among observed variables, strong correlations, particularly across latent variables (e.g., speed and breakdown fluency measures), were excluded. According to the correlation matrix and the heatmap visualization, speech rate strongly correlated with mid-clause pause ratio (breakdown fluency; r = .845) and articulation rate (speed fluency; r = .859). Although mean length of run also indicated the strong correlations with articulation rate and mid-clause pause ratio, mean length of run showed a relatively weaker correlation with mid-clause pause ratio (r = .731). Considering the fact that both speech rate and mean length of run were the observed variables loaded onto the latent variable of speed fluency, mean length of run would result in relatively weak collinearity, compared to speech rate. In addition, within the observed variables of breakdown fluency, mid-clause pause duration and end-clause pause duration were strongly correlated with each other (r = .735). Although mid- and end-clause pauses are theoretically supposed to represent different underlying processing (De Jong, 2016b; Tavakoli, 2011), the role of pause location in duration may not be statistically distinctive in factor analyses. Therefore, I decided to exclude speech rate from the measurement model of speed fluency and to replace mid-clause pause duration and end-clause pause duration with the mean pause duration measure which was calculated as the mean duration of pauses including both end- and mid-clause pauses. The revised correlation matrix is presented in Table 38.

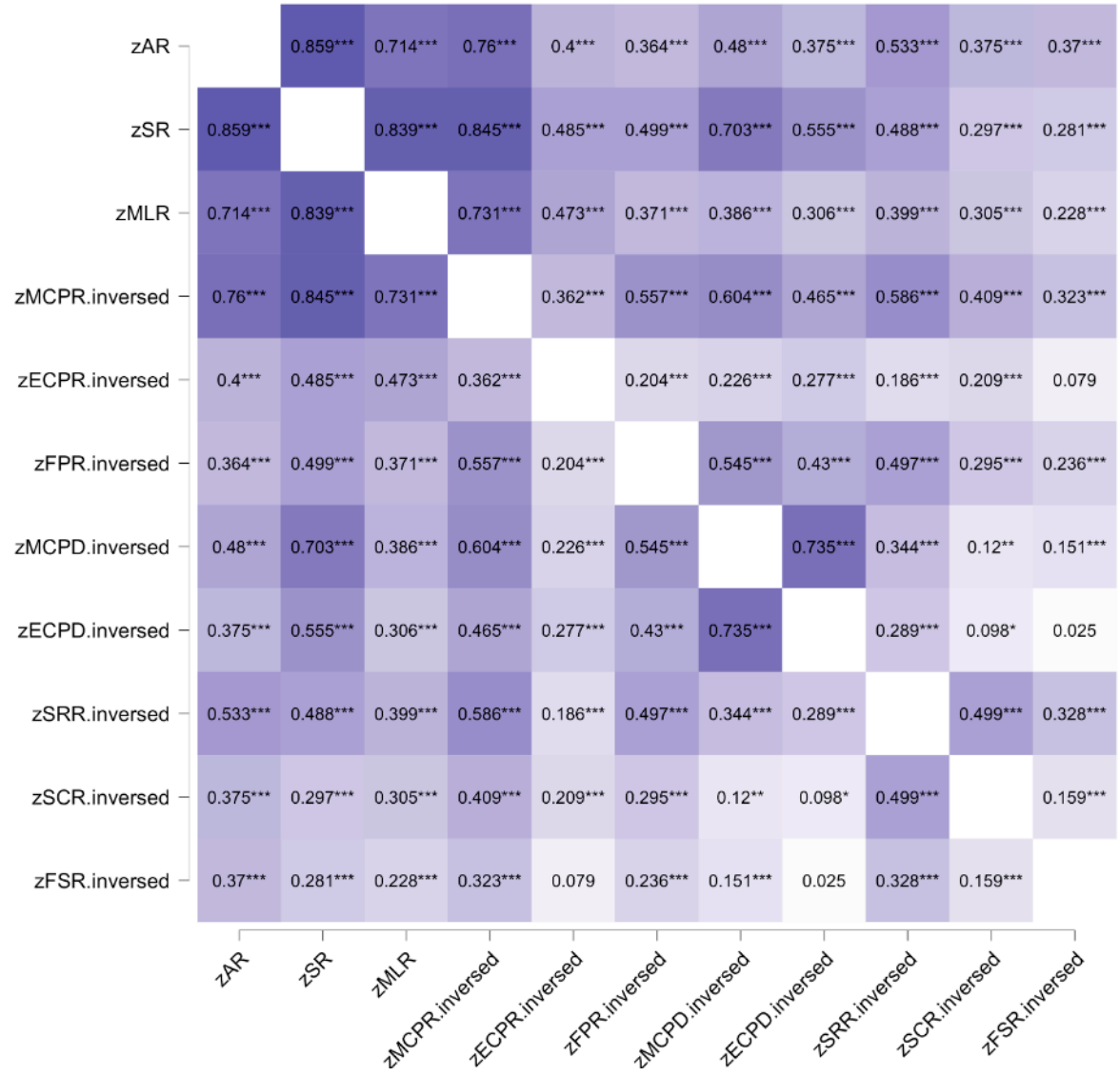


Figure 26. The heatmap visualization of correlaiton coefficients between utterance fluency measures.

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; Each cell refers to each data point of the correlation matrix, and the values in the cells are the correlation coefficients of the data points. The thickness of colour of the cells indicates the strengths of correlation coefficients, meaning that the thicker purple the cell is, the stronger correlation coefficient it shows.

Table 37. *A correlational matrix of the utterance fluency measures pooled across four tasks.*

	2. SR	3. MLR	4. MCPR	5. ECPR	6. FPR	7. MCPD	8. ECPD	9. SRR	10. SCR	11. FSR
1. AR	0.859***	0.714***	0.760***	0.400***	0.364***	0.480***	0.375***	0.533***	0.375***	0.370***
2. SR	—	0.839***	0.845***	0.485***	0.499***	0.703***	0.555***	0.488***	0.297***	0.281***
3. MLR		—	0.731***	0.473***	0.371***	0.386***	0.306***	0.399***	0.305***	0.228***
4. MCPR			—	0.362***	0.557***	0.604***	0.465***	0.586***	0.409***	0.323***
5. ECPR				—	0.204***	0.226***	0.277***	0.186***	0.209***	0.079
6. FPR					—	0.545***	0.430***	0.497***	0.295***	0.236***
7. MCPD						—	0.735***	0.344***	0.120**	0.151***
8. ECPD							—	0.289***	0.098*	0.025
9. SRR								—	0.499***	0.328***
10. SCR									—	0.159***
11. FSR										—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; AR = Articulation rate; SR = Speech rate; MLR = Mean length of run; MCPR = Mid-clause pause ratio; ECPR = End-clause pause ratio; FPR = Filled pause ratio; MCPD = Mid-clause pause duration; ECPD = End-clause pause duration; SRR = Self-repetition ratio; SCR = Self-correction ratio; FSR = False start ratio.

Table 38. *A revised correlational matrix of the utterance fluency measures pooled across four tasks.*

	2. MLR	3. MCPR	4. ECPR	5. FPR	6. MPD	7. SRR	8. SCR	9. FSR
1. AR	0.714***	0.760***	0.400***	0.364***	0.469***	0.533***	0.375***	0.370***
2. MLR	—	0.731***	0.473***	0.371***	0.386***	0.399***	0.305***	0.228***
3. MCPR		—	0.362***	0.557***	0.577***	0.586***	0.409***	0.323***
4. ECPR			—	0.204***	0.285***	0.186***	0.209***	0.079
5. FPR				—	0.532***	0.497***	0.295***	0.236***
6. MPD					—	0.342***	0.113*	0.106*
7. SRR						—	0.499***	0.328***
8. SCR							—	0.159***
9. FSR								—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; AR = Articulation rate; MLR = Mean length of run; MCPR = Mid-clause pause ratio; ECPR = End-clause pause ratio; FPR = Filled pause ratio; MPD = Mean pause duration; SRR = Self-repetition ratio; SCR = Self-correction ratio; FSR = False start ratio.

In addition to reducing the potential collinearity among observed variables, the modification indices were also calculated to explore some residuals that can be replaced with residual covariances to improve the model fit. However, the modification indices only statistically suggest the additional paths that can improve the model fit; the suggested paths were thus accepted only if the residual covariances can be theoretically explained (Raykov & Marcoulides, 2006; Tabachnick & Fidell, 1996). Eventually, the following three residual covariances were adopted.

First, the residual covariance between mean pause duration and filled pause ratio was considered justifiable because when speakers produced relatively longer pauses, they were likely to utilize filled pauses to provide the impression of continuation of speech (Clark & Fox Tree, 2002). This is also supported by a moderate correlation between them in the current dataset (r = .532, see Table 38). Accordingly, some use of filled pauses may be derived from speakers' speaking strategies used for making their speech sound more fluent (Clark & Fox Tree, 2002). In other words, the residual covariance between mean pause duration and filled pause ratio may come from some common idiosyncratic factors other than the construct of breakdown fluency.

Second, the residual covariance between mid-clause pause ratio and self-correction ratio was accepted. From the perspective of speech production mechanisms, mid-clause pauses represent the disruptions in speech processing due to the lack of linguistic resources (De Jong, 2016b; Götz, 2013; Tavakoli, 2011), whereas self-repairs are supposed to indicate overt monitoring processes (Kormos, 2000, 2006). Accordingly, the residual covariance between these two measures can be theoretically explained; when speakers produce breakdowns in the middle of clauses due to the lack of particular linguistic knowledge, they are usually required

to maintain their fluency by modifying their utterances. This possible pattern of self-corrections triggered by mid-clause pauses was also supported by the moderate correlation in the current study (r = .409, see Table 38). Since these two measures belong to different constructs (breakdown and repair fluency, respectively), this shared residual was illustrated as a residual covariance.

Third, the residual covariance between end-clause pause ratio and false start ratio was adopted, because when speakers produce false starts, they are supposed to be engaged with conceptualization processes (Williams & Korko, 2019), which end-clause pauses are also supposed to reflect (De Jong, 2016b; Tavakoli, 2011). More specifically, speakers correct their utterances at the beginning of the utterance (i.e., false start) for the sake of content information appropriacy or correctness. From a theoretical perspective, such content information is specified by conceptualization processes. Therefore, it seems plausible to argue that the causes of both end-clause pauses and false starts are associated with high demands on conceptualization processes (breakdowns vs. overt monitoring for content planning). As with the second residual covariance, this shared residual across constructs was included in the CFA models as a residual covariance.

After these modifications of UF measures and residual covariances, the proposed models (one-, two-, and three-factor models; UF Model 4, UF Model 5, and UF Model 6, respectively) were re-checked for the goodness of fit. As summarised in Table 39, the SRMR indices indicated that all of the models may fit well to the current dataset (< .08), while the other model fit indices (e.g., CFI) consistently showed that the three-factor models better fit the current dataset.

Table 39. *Selected model-fit indices for the three revised CFA models of utterance fluency.*

Model	*df*	*χ2*	*p-value*	*χ2/df ratio*	CFI	TLI	SRMR	RMSEA [90%CI]
One-factor (revised; UF Model 4)								
Argumentative	24	74.682	< .001	3.112	0.903	0.854	0.062	0.128[0.096, 0.162]
Pic.Narrative	24	116.370	< .001	4.849	0.856	0.784	0.070	0.173[0.143, 0.206]
TS.withoutRAA	24	135.293	< .001	5.637	0.822	0.733	0.075	0.190[0.160, 0.222]
TS.withRAA	24	109.895	< .001	4.579	0.837	0.756	0.073	0.167[0.136, 0.200]
Two-factor (revised; UF Model 5)								
Argumentative	23	67.223	< .001	2.923	0.915	0.867	0.062	0.123[0.089, 0.157]
Pic.Narrative	23	112.831	< .001	4.906	0.860	0.781	0.070	0.175[0.143, 0.208]
TS.withoutRAA	23	128.507	< .001	5.587	0.831	0.736	0.074	0.189[0.158, 0.222]
TS.withRAA	23	107.323	< .001	4.666	0.840	0.750	0.073	0.169[0.138, 0.202]
Three-factor (revised; UF Model 6)								
Argumentative	21	57.550	< .001	2.740	0.930	0.880	0.056	0.117[0.081, 0.153]
Pic.Narrative	21	95.357	< .001	4.541	0.884	0.802	0.067	0.166[0.133, 0.201]
TS.withoutRAA	21	110.689	< .001	5.271	0.857	0.754	0.070	0.183[0.150, 0.217]
TS.withRAA	21	92.648	< .001	4.412	0.864	0.767	0.066	0.163[0.130, 0.198]

Note. CFI = comparative fit index; SRMR = standardized root mean square; RMSEA = root mean square error of association. The cut-off values for good fit: χ2/df ratio < 2.0; SRMR < .08; CFI and TLI > .95, RMSEA < .06.

The standardized regression coefficients of the revised three-factor model of UF measures (see Table 40) suggested strong correlations between the latent variables of speed and breakdown fluency (r = .929–.960), indicating the possibility that the distinction between these two latent variables might be redundant. Accordingly, for the sake of the potentially better factor structure of UF measures, I proposed another factor structure with speed and breakdown fluency measures loaded onto one latent variable (UF Model 7; see Figure 27).

Table 40. *Summary of the standardized regression coefficients and their 95% confidence intervals of the three-factor CFA model of cognitive fluency.*

						95%CI	
Latent variable	Direction	Observed variable	Task	*β*	*p*	Lower	Upper
Covariance between latent variables							
Speed fluency	vs.	Breakdown fluency	Arg	0.929	< .001	0.845	1.014
			PicN	0.960	< .001	0.910	1.011
			TS [-RAA]	0.948	< .001	0.868	1.028
			TS [+RAA]	0.945	< .001	0.879	1.011
Speed fluency	vs.	Repair fluency	Arg	0.732	< .001	0.607	0.857

			PicN	0.704	< .001	0.603	0.804
			TS [-RAA]	0.627	< .001	0.489	0.766
			TS [+RAA]	0.600	< .001	0.445	0.755
Breakdown fluency	vs.	Repair fluency	Arg	0.819	< .001	0.688	0.951
			PicN	0.732	< .001	0.616	0.848
			TS [-RAA]	0.762	< .001	0.610	0.914
			TS [+RAA]	0.621	< .001	0.458	0.785
Measurement model							
Speed fluency	→	Articulation rate	Arg	0.863	< .001	0.770	0.957
			PicN	0.865	< .001	0.803	0.927
			TS [-RAA]	0.822	< .001	0.737	0.908
			TS [+RAA]	0.803	< .001	0.707	0.899
	→	Mean length or run	Arg	0.768	< .001	0.693	0.844
			PicN	0.926	< .001	0.903	0.949
			TS [-RAA]	0.926	< .001	0.889	0.963
			TS [+RAA]	0.911	< .001	0.878	0.945
Breakdown fluency	→	Mid-clause pause ratio	Arg	0.979	< .001	0.931	1.027
			PicN	0.960	< .001	0.929	0.991
			TS [-RAA]	0.950	< .001	0.876	1.025
			TS [+RAA]	0.982	< .001	0.934	1.030
	→	End-clause pause ratio	Arg	0.471	< .001	0.315	0.628
			PicN	0.347	< .001	0.203	0.490
			TS [-RAA]	0.399	< .001	0.232	0.566
			TS [+RAA]	0.434	< .001	0.276	0.591
	→	Mean pause duration	Arg	0.499	< .001	0.389	0.610
			PicN	0.658	< .001	0.559	0.756
			TS [-RAA]	0.621	< .001	0.431	0.811
			TS [+RAA]	0.543	< .001	0.335	0.752
	→	Filled pause ratio	Arg	0.601	< .001	0.476	0.725
			PicN	0.564	< .001	0.445	0.684
			TS [-RAA]	0.570	< .001	0.362	0.779
			TS [+RAA]	0.548	< .001	0.399	0.696
Repair fluency	→	False start ratio	Arg	0.445	< .001	0.244	0.645
			PicN	0.473	< .001	0.299	0.648
			TS [-RAA]	0.281	0.011	0.066	0.497
			TS [+RAA]	0.415	< .001	0.208	0.623
	→	Self-repetition ratio	Arg	0.833	< .001	0.729	0.936
			PicN	0.825	< .001	0.731	0.920
			TS [-RAA]	0.877	< .001	0.753	1.001
			TS [+RAA]	0.812	< .001	0.622	1.001
	→	Self-correction ratio	Arg	0.580	< .001	0.423	0.737
			PicN	0.632	< .001	0.506	0.757
			TS [-RAA]	0.591	< .001	0.359	0.822

TS [+RAA]	0.475	< .001	0.314	0.635

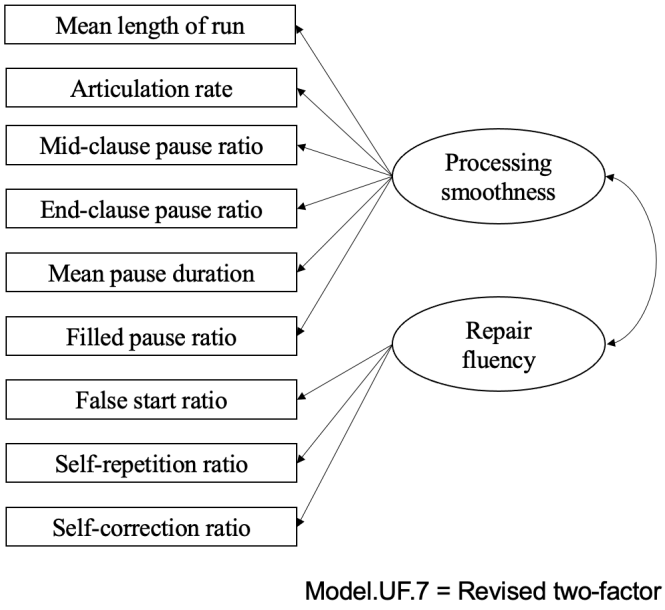


Figure 27. A new two-factor model of utterance fluency (Model UF 7).
Note. Residuals are omitted for the sake of brevity.

The new two-factor model (UF Model 7) was evaluated using a set of the goodness-of-fit indices. As summarized in Table 41, the model-fit of the new model was virtually identical to the revised three-factor model. Although the new two-factor model has fewer parameters than the three-factor model, I decided to adopt the revised three-factor model as the factor structure of UF for the subsequent SEM analysis, considering its theoretical compatibility with Tavakoli and Skehan's (2005) triad of the subconstructs of UF (speed, breakdown, and repair fluency) and L2 speech production mechanisms (Kormos, 2006; Segalowitz, 2010), as well as the differential predictive power of the subconstructs in listener-based judgements of PF, suggested by Study 1 (see Section 5.2.1).

Table 41. *Selected model-fit indices for a new two-factor CFA model of utterance fluency.*

Model	*df*	*χ2*	*p-value*	*χ2/df ratio*	CFI	TLI	SRMR	RMSEA [90%CI]
New two-factor								
Argumentative	23	63.680	< .001	2.769	0.922	0.878	0.058	0.118[0.084, 0.152]
Pic.Narrative	23	98.474	< .001	4.281	0.883	0.816	0.068	0.160[0.128, 0.193]
TS.withoutRAA	23	119.913	< .001	5.214	0.845	0.758	0.070	0.181[0.150, 0.214]
TS.withRAA	23	96.121	< .001	4.179	0.862	0.783	0.066	0.158[0.126, 0.191]

Note. CFI = comparative fit index; SRMR = standardized root mean square; RMSEA = root mean square error of association. The cut-off values for good fit: $\chi 2$/df ratio < 2.0; SRMR < .08; CFI and TLI > .95, RMSEA < .06.

Based on the revised three-factor model (UF Model 6), the standardized regression coefficients of the measurement models of each latent variable indicated which UF measures (i.e., observed variables) primarily contributed to their corresponding constructs (i.e., latent variables). Regarding the latent variable of speed fluency, both measures (articulation rate, mean length of run) seemed to equally contribute to the construct to a large extent, as suggested by the overlap of their confidence intervals across tasks (see Table 40). This equal amount of the contributions to the latent variable may support the statistical decision to classify mean length of run as a measure of speed fluency, despite its composite nature. As for breakdown fluency, mid-clause pause ratio primarily contributed to the latent variable with high standardised regression coefficients (β = .950–.982). Among the other breakdown fluency measures, the regression coefficients did not significantly differ from each other in terms of the overlap of their confidence intervals. Concerning repair fluency, despite the overlap of the confidence intervals between self-repetition ratio and self-correction ratio measures, self-repetition ratio suggested relatively high regression coefficients to the latent variable.

8.2.5 Correlation between cognitive and utterance fluency measures

Study 3 aims to examine the contributions of CF to UF at the level of constructs (statistically, latent variables). As another preliminary analysis for the SEM analysis (RQ3-1, RQ3-2), the general relationship between CF and UF at the level of observed variables was examined (see Table 42). Although observed variables inevitably entail some measurement errors, it is useful to compare the results of the interrelationship between the levels of observed and latent variables. If the results of intercorrelations and an SEM analysis are different, such

deviation of results indicates the extent to which the observed variables entail measurement errors, which may subsequently give some insights into the validity of the observed variables. In this section, I particularly focus on how UF measures of different subconstructs (speed, breakdown, and repair fluency) were associated with CF measures of different linguistic levels (i.e., lexis, syntax, morphology, and pronunciation).

Speed fluency measures tended to correlate with articulatory speed with moderate-to-strong effect sizes (r_s = |.455–.601|) and also with vocabulary and grammatical measures (especially RT-based measures) with moderate effect sizes. Meanwhile, GJT Morphological Accuracy indicated loose relations with speed fluency measures (r_s = |.125–.208|). This tenuous association of morphological knowledge was also observed with the measures of breakdown fluency (r_s = |.012–.203|) and repair fluency (r_s = |.004–.183|). Regarding breakdown fluency measures, the distinctive role of pause location may differ between pause duration and ratio measures. As for pause ratio measures, mid-clause pauses showed a relatively stronger association with CF measures than end-clause pauses. In contrast, both mid- and end-clause pause duration measures appeared to show a similar pattern of associations with the CF measures. For instance, the RT-based CF measures correlated virtually equally with both mid- and end-clause pause duration measures. From the perspective of associations with CF, pause location may play a role only in pause frequency-based measures. Meanwhile, filled pause ratio was found to be only weakly related to the linguistic resource measures of vocabulary and syntax. Finally, among repair fluency measures, self-correction ratio and false start ratio measures did not demonstrate a clear pattern of associations with CF measures. Although self-repetition ratio seemed to correlate with vocabulary measures (PVLT and Picture Naming RT) as well as GJT Syntax measures (RT and Accuracy), these associations were not consistent across tasks.

Table 42. *A correlational matrix of utterance fluency measures and cognitive fluency measures across tasks.*

	Task	AR	SR	MLR	MCPR	ECPR	FPR	MCPD	ECPD	SRR	SCR	FSR
PVLT	Arg	0.395***	0.425***	0.434***	-0.430***	-0.318***	-0.302***	-0.236**	-0.324***	-0.195*	-0.073	-0.135
	PicN	0.400***	0.415***	0.459***	-0.442***	-0.323***	-0.324***	-0.267**	-0.271**	-0.215*	-0.117	-0.107
	TS [-RAA]	0.421***	0.481***	0.506***	-0.508***	-0.326***	-0.370***	-0.276**	-0.363***	-0.244**	-0.251**	-0.017
	TS [+RAA]	0.363***	0.398***	0.406***	-0.393***	-0.238**	-0.398***	-0.250**	-0.219*	-0.167	-0.169	-0.070
Picture Naming RT	Arg	-0.398***	-0.455***	-0.377***	0.404***	0.113	0.264**	0.302***	0.323***	0.273**	0.115	0.055
	PicN	-0.333***	-0.401***	-0.329***	0.333***	0.121	0.166	0.343***	0.323***	0.234**	-0.001	0.182*
	TS [-RAA]	-0.346***	-0.448***	-0.378***	0.392***	0.078	0.161	0.368***	0.401***	0.150	-0.017	0.035
	TS [+RAA]	-0.332***	-0.450***	-0.371***	0.365***	0.243**	0.217*	0.359***	0.308***	0.238**	-0.053	0.087
Maze Word RT	Arg	-0.356***	-0.4***	-0.379***	0.392***	0.141	0.287**	0.253**	0.303***	0.168	-0.001	0.044
	PicN	-0.424***	-0.464***	-0.451***	0.446***	0.186*	0.275**	0.346***	0.325***	0.226*	0.063	0.254**
	TS [-RAA]	-0.418***	-0.501***	-0.469***	0.499***	0.168	0.281**	0.361***	0.393***	0.248**	0.101	0.058
	TS [+RAA]	-0.327***	-0.409***	-0.428***	0.427***	0.244**	0.266**	0.242**	0.236**	0.131	0.022	0.115
Maze Word Acc	Arg	0.168	0.180*	0.197*	-0.198*	-0.123	-0.204*	-0.115	-0.156	-0.087	0.056	-0.183*
	PicN	0.187*	0.204*	0.247**	-0.263**	-0.052	-0.261**	-0.168	-0.105	-0.169	-0.079	-0.194*
	TS [-RAA]	0.208*	0.235**	0.228**	-0.260**	-0.089	-0.242**	-0.158	-0.205*	-0.167	-0.193*	-0.047
	TS [+RAA]	0.177*	0.175*	0.165	-0.174*	0.023	-0.277**	-0.184*	-0.062	-0.197*	-0.173	-0.064
GJT Morph RT	Arg	-0.327***	-0.356***	-0.370***	0.361***	0.240**	0.160	0.256**	0.103	0.190*	0.133	0.175*
	PicN	-0.402***	-0.403***	-0.392***	0.346***	0.323***	0.097	0.229**	0.302***	0.193*	0.101	0.074
	TS [-RAA]	-0.327***	-0.392***	-0.402***	0.357***	0.333***	0.143	0.272**	0.322***	0.146	0.111	-0.041
	TS [+RAA]	-0.275**	-0.402***	-0.378***	0.351***	0.310***	0.131	0.302***	0.302***	0.054	0.158	0.023
GJT Syn RT	Arg	-0.351***	-0.394***	-0.415***	0.407***	0.255**	0.222*	0.288**	0.130	0.208*	0.124	0.133
	PicN	-0.428***	-0.444***	-0.440***	0.386***	0.389***	0.169	0.273**	0.325***	0.249**	0.121	0.049
	TS [-RAA]	-0.355***	-0.432***	-0.448***	0.403***	0.366***	0.195*	0.317***	0.340***	0.193*	0.108	0.001
	TS [+RAA]	-0.271**	-0.427***	-0.423***	0.391***	0.343***	0.171	0.299***	0.322***	0.070	0.116	0.010
GJT MorphAcc	Arg	0.170	0.149	0.142	-0.142	-0.203*	-0.095	-0.051	-0.168	-0.076	-0.068	0.017
	PicN	0.161	0.143	0.127	-0.122	-0.142	-0.070	-0.138	-0.093	0.038	0.021	-0.183*

	TS [-RAA]	0.208*	0.189*	0.181*	-0.199*	-0.080	-0.118	-0.114	-0.177*	-0.074	-0.004	-0.078
	TS [+RAA]	0.187*	0.125	0.159	-0.154	-0.090	-0.106	-0.053	0.012	-0.084	-0.068	-0.017
GJT Syn Acc	Arg	0.321***	0.345***	0.399***	-0.404***	-0.261**	-0.251**	-0.138	-0.284**	-0.168	-0.065	-0.121
	PicN	0.302***	0.355***	0.396***	-0.398***	-0.174	-0.354***	-0.300***	-0.211*	-0.199*	-0.156	-0.228**
	TS [-RAA]	0.313***	0.393***	0.374***	-0.414***	-0.130	-0.319***	-0.293***	-0.300***	-0.297***	-0.208*	-0.070
	TS [+RAA]	0.270**	0.278**	0.332***	-0.314***	-0.190*	-0.315***	-0.175*	-0.140	-0.208*	-0.043	-0.084
Articulatory speed	Arg	0.546***	0.470***	0.455***	-0.459***	-0.190*	-0.150	-0.281**	-0.233**	-0.158	-0.125	-0.075
	PicN	0.601***	0.528***	0.501***	-0.494***	-0.249**	-0.136	-0.377***	-0.353***	-0.218*	-0.061	-0.174*
	TS [-RAA]	0.557***	0.576***	0.522***	-0.533***	-0.153	-0.214*	-0.449***	-0.388***	-0.235**	-0.131	-0.051
	TS [+RAA]	0.533***	0.534***	0.527***	-0.537***	-0.261**	-0.150	-0.328***	-0.322***	-0.071	0.062	-0.229**

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; PVLT = Productive Vocabulary Levels Test; Maze Word Acc = Maze Word accuracy; GJT Morph RT = GJT Morphology RT; GJT Syn RT = GJT Syntax RT; GJT Morph Acc = GJT Morphology Accuracy; GJT Syn Acc = GJT Syntax Accuracy

8.2.6 Structural equation model of cognitive fluency and utterance fluency

Building on the final CFA models of CF (CF Model 2) and UF (UF Model 6), an SEM analysis further examined how the constructs of CF are associated with those of UF. Statistically speaking, the structural models of CF and UF were created separately for four speaking tasks. In the process of integrating the measurement models of CF and UF, one additional residual covariance was included between the articulatory speed measure (CF) and the measure of articulation rate (UF) in the SEM models. The rationale for adding this residual covariance is that both measures were calculated as the mean number of syllables produced per second. Although the articulatory speed measure and the articulation rate measure were elicited via the controlled speech production task and spontaneous speech tasks respectively, measurement errors of these measures can be assumed to be methodologically shared to some extent.

The indices of goodness-of-fit were first inspected to see the extent to which the proposed SEM model of the CF-UF link fitted the current dataset. As with the preceding CFA model testing, due to the relatively small sample size ($N < 250$), the indices of SRMR and CFI were prioritised when evaluating the model fit. As summarised in Table 43 below, the indices of SRMR indicated that the proposed SEM model optimally fitted the current dataset (SRMR < .08), while the CFI indices suggested that there was still room for the improvement in the model fit to the data (CFI < .95). Although the modification indices were calculated, there were no suggested additional paths that were reflective of a theoretical framework of oral fluency and were consistent across tasks. To propose the SEM model of the CF-UF link that is optimally generalizable to different speaking tasks, the proposed model was regarded as the final model based on the current dataset of CF and UF measures. The SEM model with standardized regression coefficients across tasks is visually presented in Figure 28.

Table 43. *Selected model-fit indices for an SEM model of cognitive fluency and utterance fluency.*

Model	*df*	*χ2*	*p-value*	*χ2/df ratio*	CFI	TLI	SRMR	RMSEA [90%CI]
SEM model								
Argumentative	111	207.019	< .001	1.865	0.921	0.891	0.071	0.082[0.065, 0.099]
Pic.Narrative	111	213.012	< .001	1.919	0.924	0.895	0.067	0.085[0.067, 0.102]
TS.withoutRAA	111	196.925	< .001	1.774	0.933	0.908	0.062	0.078[0.060, 0.095]
TS.withRAA	111	214.577	< .001	1.933	0.914	0.882	0.069	0.085[0.068, 0.102]

Note. CFI = comparative fit index; SRMR = standardized root mean square; RMSEA = root mean square error of association. The cut-off values for good fit: χ2/df ratio < 2.0; SRMR < .08; CFI and TLI > .95, RMSEA < .06.

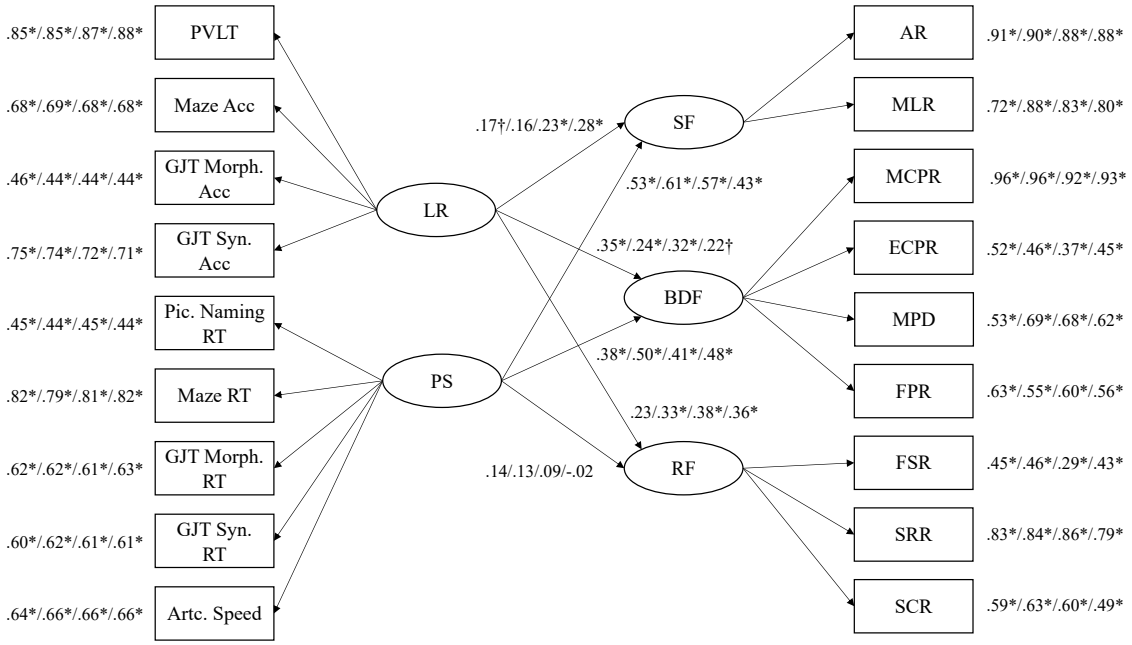


Figure 28. Comparison of the regression coefficients across argumentative speech, picture narratives, and text summary without and with read-aloud assistance.
Note. Residuals are omitted for the sake of brevity. The regression coefficients are presented in the order of the argumentative task, the picture narrative task, the text summary task without RAA, and the text summary task from left to right; LR = Linguistic resource; PS = Processing speed; SF = Speed fluency; BDF = Breakdown fluency; RF = Repair fluency.

8.2.6.1 The structural model of CF and UF

The RQ3-1 of Study 3 investigated how the latent variables of CF were associated with the latent variables of UF. To clarify the contribution of CF to UF, the standardized regression coefficients of the structural model of CF and UF were summarized in Table 44. According to Table 44 as well as Figure 28, speed fluency was associated with linguistic resource only

in the text summary tasks in both RAA conditions and with processing speed in all of the four tasks. Focusing on the confidence intervals of those regression coefficients, there were no differences in the strengths of associations between two conditions of the text summary task. Furthermore, in the argumentative and picture narrative tasks, speed fluency was associated more strongly with processing speed than with linguistic resource.

Meanwhile, breakdown fluency was in general related to both linguistic resource and processing speed, regardless of speaking tasks. The amount of contributions to breakdown fluency did not differ between linguistic resource and processing speed. However, the latent variable of breakdown fluency seemed to show slightly stronger associations with processing speed (β = .376–.502) than with linguistic resource (β = .221–.345).

As for repair fluency, it was found that linguistic resource significantly contributed to the construct of repair fluency in all speaking tasks but the argumentative task. In other words, the significant contribution of linguistic resource to repair fluency was only observed in the speaking tasks where the content of speech is predefined (the picture narrative task and text summary tasks). Meanwhile, processing speed was not related to repair fluency in any of the speaking tasks.

Table 44. *Summary of the standardized regression coefficients and their 95% confidence intervals of the structural model of cognitive fluency and utterance fluency.*

						95%CI	
Latent variable	Direction	Latent variable	Task	*β*	*p*	Lower	Upper
Regression model							
Linguistic resource	→	Speed fluency	Arg	0.168	0.061	-0.008	0.344
			PicN	0.161	0.104	-0.033	0.354
			TS [-RAA]	0.234	0.038	0.013	0.455
			TS [+RAA]	0.276	0.004	0.086	0.465
	→	Breakdown fluency	Arg	0.345	0.001	0.139	0.550
			PicN	0.240	0.025	0.030	0.451

			TS [-RAA]	0.317	0.014	0.064	0.570
			TS [+RAA]	0.221	0.061	-0.010	0.452
	→	Repair fluency	Arg	0.225	0.150	-0.081	0.531
			PicN	0.330	0.049	0.002	0.659
			TS [-RAA]	0.375	0.019	0.062	0.689
			TS [+RAA]	0.360	0.033	0.029	0.692
Processing speed	→	Speed fluency	Arg	0.533	< .001	0.373	0.693
			PicN	0.609	< .001	0.434	0.784
			TS [-RAA]	0.566	< .001	0.371	0.761
			TS [+RAA]	0.431	< .001	0.244	0.617
	→	Breakdown fluency	Arg	0.376	< .001	0.191	0.561
			PicN	0.501	< .001	0.314	0.689
			TS [-RAA]	0.411	0.003	0.144	0.679
			TS [+RAA]	0.480	< .001	0.251	0.710
	→	Repair fluency	Arg	0.136	0.349	-0.149	0.420
			PicN	0.129	0.351	-0.142	0.400
			TS [-RAA]	0.094	0.452	-0.152	0.341
			TS [+RAA]	-0.020	0.906	-0.351	0.311

8.2.6.2 The measurement model of CF

Although the structure of the measurement models of CF and UF in the SEM model were identical to those in the CFA models, the regression coefficients of some paths in the SEM model was adjusted due to the holistic estimation of factor loadings. I thus revisit the strengths of regression coefficients in the SEM model in the current and following sections. The SEM model suggested that there were no significant differences in any of the regression coefficients in the measurement models of CF and UF across tasks, meaning that the factor structures of CF and UF were consistent across speaking tasks. However, the relative importance of linguistic dimensions (Lexis, Syntax, Morphology) was different between the latent variables of CF (i.e., linguistic resource, processing speed) in terms of their range of confidence intervals. Regarding the linguistic resource of CF, the regression coefficients of PVLT (β = .845–.879) were significantly higher than those of Maze Word Accuracy except for the picture narrative task (β = .675–.691), while there were overlaps of confidence intervals between PVLT and GJT Syntax Accuracy (β = .710–.746; see Table 45).

Meanwhile, the regression coefficients of GJT Syntax Accuracy seemed to be higher than those of GJT Morphological Accuracy except for the text summary task without RAA (β = .439–.455).

As regards the measurement model of processing speed, the strongest regression coefficients were found in Maze Word RT (β = .794–.821) which taps into the speed of syntactic procedure processes. According to the 95% confidence intervals, the strengths of coefficients between Maze Word RT and GJT Syntax RT (β = .607–.620) did not reach statistical significance in any of the speaking tasks. Notably, the differences in the coefficients between Maze Word RT and GJT Morphology RT (morphological processing speed; β = .614–.626) and between Maze Word RT and articulatory speed (β = .635–.663) seemed to be approaching statistical significance. Meanwhile, the significant differences in the regression coefficients of processing speed were only found between Maze Word RT and Picture Naming RT (β = .436–.453). Finally, as with the CFA model of CF in Section 8.2.2, the latent variables of linguistic resource and processing speed were strongly associated with each other consistently across tasks (β = .664–.676; see Table 46).

Table 45. *Summary of the standardized regression coefficients and their 95% confidence intervals of the measurement model of cognitive fluency in the final SEM model.*

						95%CI	
Latent variable	Direction	Observed variable	Task	β	p	Lower	Upper
Measurement model of cognitive fluency							
Linguistic resource	→	PVLT	Arg	0.850	< .001	0.788	0.912
			PicN	0.845	< .001	0.783	0.908
			TS [-RAA]	0.870	< .001	0.808	0.933
			TS [+RAA]	0.879	< .001	0.817	0.941
	→	Maze Word Accuracy	Arg	0.680	< .001	0.575	0.785
			PicN	0.691	< .001	0.573	0.810
			TS [-RAA]	0.677	< .001	0.549	0.805
			TS [+RAA]	0.675	< .001	0.552	0.799
	→	GJT Morph. Accuracy	Arg	0.455	< .001	0.305	0.605
			PicN	0.439	< .001	0.279	0.600

			TS [-RAA]	0.441	< .001	0.261	0.621
			TS [+RAA]	0.442	< .001	0.267	0.616
	→	GJT Syn. Accuracy	Arg	0.746	< .001	0.653	0.839
			PicN	0.742	< .001	0.648	0.836
			TS [-RAA]	0.722	< .001	0.617	0.828
			TS [+RAA]	0.710	< .001	0.605	0.814
Processing speed	→	Pic. Naming RT	Arg	0.450	< .001	0.303	0.596
			PicN	0.436	< .001	0.298	0.573
			TS [-RAA]	0.453	< .001	0.295	0.611
			TS [+RAA]	0.439	< .001	0.289	0.589
	→	Maze Word RT	Arg	0.821	< .001	0.738	0.905
			PicN	0.794	< .001	0.702	0.887
			TS [-RAA]	0.813	< .001	0.716	0.909
			TS [+RAA]	0.815	< .001	0.719	0.912
	→	GJT Morph. RT	Arg	0.617	< .001	0.479	0.755
			PicN	0.622	< .001	0.482	0.762
			TS [-RAA]	0.614	< .001	0.480	0.748
			TS [+RAA]	0.626	< .001	0.498	0.754
	→	GJT Syn. RT	Arg	0.604	< .001	0.467	0.741
			PicN	0.620	< .001	0.487	0.754
			TS [-RAA]	0.607	< .001	0.473	0.741
			TS [+RAA]	0.612	< .001	0.489	0.736
	→	Articulatory speed	Arg	0.635	< .001	0.519	0.750
			PicN	0.663	< .001	0.551	0.774
			TS [-RAA]	0.658	< .001	0.543	0.773
			TS [+RAA]	0.659	< .001	0.544	0.775

Table 46. *Summary of the standardized regression coefficients between the latent variables of cognitive fluency and their 95% confidence intervals in the final SEM model.*

						95%CI	
Latent variable	Direction	Latent variable	Task	*β*	*p*	Lower	Upper
Covariance between latent variables							
Linguistic resource	vs.	Processing speed	Arg	0.667	< .001	0.515	0.819
			PicN	0.664	< .001	0.516	0.812
			TS [-RAA]	0.671	< .001	0.531	0.811
			TS [+RAA]	0.676	< .001	0.534	0.817

8.2.6.3 The measurement model of UF

As with the measurement model of CF, the regression coefficients in the measurement model of UF have slightly changed in the SEM model of the CF-UF link due to its holistic nature

(see Table 47). As for speed fluency, both observed variables (articulation rate, mean length of run) considerably contributed to the latent variable of speed fluency, while the regression coefficients of articulation rate (β = .876–.905) seemed to be slightly higher than those of mean length of run (β = .721–.882). Regarding breakdown fluency, mid-clause pause ratio contributed to the construct of breakdown fluency to the largest extent. Moreover, the coefficients of mid-clause pause ratio (β = .919–.963) were significantly higher than those of the other measures—mean pause duration (β = .528–.690), end-clause pause ratio (β = .373–.515), and filled pause ratio (β = .545–.628). There were no significant differences in the regression coefficients among these three measures (mean pause duration, end-clause pause ratio, and filled pause ratio). As regards repair fluency, although the differences in the regression coefficients among repair fluency measures did not reach statistical significance in the CFA model of UF (see Section 8.2.4), the regression coefficients of self-repetition ratio were significantly higher than those of self-correction ratio (except for the text summary task without RAA) and false start ratio. Finally, as with the CFA model of UF in Section 8.2.4, there were strong competitive relationships between the latent variables of speed fluency and breakdown fluency (β = -|.769–.822|; see Table 48) and between those of speed fluency and repair fluency (β = -|.720–.749|), while the latent variables of breakdown fluency were positively associated with those of repair fluency (β = .639–.796).

Table 47. *Summary of the standardized regression coefficients and their 95% confidence intervals of the measurement model of utterance fluency in the final SEM model.*

						95%CI	
Latent variable	Direction	Observed variable	Task	β	p	Lower	Upper
Measurement model of utterance fluency							
Speed fluency	→	Articulation rate	Arg	0.905	< .001	0.836	0.974
			PicN	0.892	< .001	0.838	0.947
			TS [-RAA]	0.876	< .001	0.805	0.948
			TS [+RAA]	0.879	< .001	0.803	0.955
	→	Mean length or run	Arg	0.721	< .001	0.632	0.810
			PicN	0.882	< .001	0.834	0.930

			TS [-RAA]	0.831	< .001	0.773	0.889
			TS [+RAA]	0.800	< .001	0.729	0.872
Breakdown fluency	→	Mid-clause pause ratio	Arg	0.958	< .001	0.911	1.005
			PicN	0.963	< .001	0.922	1.004
			TS [-RAA]	0.919	< .001	0.832	1.005
			TS [+RAA]	0.933	< .001	0.877	0.990
	→	End-clause pause ratio	Arg	0.515	< .001	0.367	0.663
			PicN	0.455	< .001	0.310	0.601
			TS [-RAA]	0.373	< .001	0.205	0.540
			TS [+RAA]	0.449	< .001	0.233	0.664
	→	Mean pause duration	Arg	0.528	< .001	0.379	0.676
			PicN	0.690	< .001	0.595	0.786
			TS [-RAA]	0.681	< .001	0.499	0.862
			TS [+RAA]	0.617	< .001	0.407	0.827
	→	Filled pause ratio	Arg	0.628	< .001	0.510	0.746
			PicN	0.545	< .001	0.427	0.663
			TS [-RAA]	0.598	< .001	0.383	0.813
			TS [+RAA]	0.556	< .001	0.404	0.708
Repair fluency	→	False starts ratio	Arg	0.450	< .001	0.281	0.619
			PicN	0.459	< .001	0.304	0.614
			TS [-RAA]	0.289	0.014	0.059	0.518
			TS [+RAA]	0.427	< .001	0.235	0.620
	→	Self-repetition ratio	Arg	0.827	< .001	0.734	0.919
			PicN	0.837	< .001	0.756	0.917
			TS [-RAA]	0.860	< .001	0.747	0.973
			TS [+RAA]	0.787	< .001	0.648	0.925
	→	Self-correction ratio	Arg	0.587	< .001	0.453	0.721
			PicN	0.632	< .001	0.523	0.741
			TS [-RAA]	0.599	< .001	0.347	0.850
			TS [+RAA]	0.487	< .001	0.337	0.637

Table 48. *Summary of the standardized regression coefficients between the latent variables of utterance fluency and their 95% confidence intervals in the final SEM model*

						95%CI	
Latent variable	Direction	Latent variable	Task	*β*	*p*	Lower	Upper
Covariance between latent variables							
Speed fluency	vs.	Breakdown fluency	Arg	-0.818	< .001	-0.951	-0.686
			PicN	-0.822	< .001	-0.918	-0.726
			TS [-RAA]	-0.800	< .001	-0.919	-0.681
			TS [+RAA]	-0.769	< .001	-0.876	-0.662
Speed fluency	vs.	Repair fluency	Arg	-0.749	< .001	-0.890	-0.608
			PicN	-0.720	< .001	-0.858	-0.583
			TS [-RAA]	-0.720	< .001	-0.899	-0.540

			TS [+RAA]	-0.739	< .001	-0.900	-0.578
Breakdown fluency	vs.	Repair fluency	Arg	0.864	< .001	0.735	1.003
			PicN	0.639	< .001	0.496	0.782
			TS [-RAA]	0.796	< .001	0.642	0.950
			TS [+RAA]	0.716	< .001	0.553	0.879

Note. For the sake of interpretability of the direction of relationship between the latent variables of UF, the regression coefficients in Table 48 were computed without the inversion of the observed variables of breakdown fluency and repair fluency measures.

8.3 Discussion

Although L2 fluency research has extensively examined the relationship between UF and PF, little is known about how CF contributes to UF. Accordingly, it is still unclear which linguistic resources and processing skills enable learners to speak fluently in L2. Motivated by the lack of studies about the CF-UF link at the level of constructs, Study 3 took an SEM approach to examine the CF-UF link (RQ3-1). To this end, Study 3 operationalized CF as a set of linguistic resources and processing skills involved in speech production, and each dimension of UF—speed, breakdown, and repair fluency—was also measured. Furthermore, a close examination of previous studies also suggested that the dimensionality of CF and UF had not been revisited or even specified especially concerning the generalizability across different speaking task types. Accordingly, Study 3 also delved into the factor structure of CF and UF, using a range of CF and UF measures (RQ3-1a, RQ3-1b). Finally, in light of the generalizability and robustness of the CF-UF link, the variability of the association between the subconstructs of CF and UF across different speaking tasks was explored (RQ3-2).

8.3.1 Dimensionality of cognitive fluency

Prior to examining the relationship between CF and UF at the level of constructs (RQ3-1), the factor structure of each construct was examined, using CFA (RQ3-1a, RQ3-1b). Regarding the factor structure of CF, Study 3 tested the single-, two-, and three-factor models of CF. These CFA models were proposed with regard to the components of L2 speech

production (Kormos, 2006; Levelt, 1989; Segalowitz, 2010), in accordance with Segalowitz's (2010) conception of CF. Comparing the model-fit indices (e.g., SRMR and CFI; Hu & Bentler, 1998) of those proposed models, Study 3 adopted the two-factor model which consisted of the latent variables of *linguistic resource* and *processing speed* (CFI = .976, SRMR = .051). The latent variable of linguistic resource involved the PVLT score (vocabulary size), the GJT accuracy scores (syntax and morphology), and the maze task accuracy scores (sentence construction skills), while that of processing speed included the RT measures of the picture naming task (lexical retrieval), the maze task, and the GJT as well as the articulatory speed in the controlled speech production. Compared to the final two-factor model, the single-factor model appeared to show a relatively less adequate fit to the current data (CFI = .919, SRMR = .078), indicating that the construct of CF may not be regarded as a unitary construct. Meanwhile, the two-dimensional construct of CF is in line with the broad definition of CF (see Section 3.6). In light of the construct validity of CF, CF is supposed to explain the rapidity and smoothness of utterances (i.e., UF; Segalowitz, 2010, 2016). Accordingly, Segalowitz (2010, 2016) emphasizes the speed dimension of cognitive processes involved in L2 speech production, such as lexical retrieval speed. However, from the theoretical perspective of L2 speech production, the fluency of utterances is affected not only by the speed of processing skills, but also by the availability of linguistic resources (cf. Kormos, 2006). Similarly, empirical studies on the CF-UF link have also operationalized CF in terms of both processing speed and linguistic resources (De Jong et al., 2013; Kahng, 2020). Therefore, the current finding of two-dimensionality of CF may provide supporting evidence for the broad definition of CF as well as the existing methodological practice of measuring CF components. Finally, it is noteworthy that there was a strong association between these two latent variables (r = .676), indicating that the subdimensions of CF—linguistic resource and processing speed—are interrelated with each other.

Closely looking at the measurement models of the subconstructs of CF, the primary components of linguistic resource and processing speed were different. To interpret the dimensionality of CF in relation to its contributions to UF, the measurement model of CF is mainly discussed based on the one in the final SEM model (see Section 8.2.6.2). As for the latent variable of linguistic resource, PVLT (vocabulary size) had the highest regression coefficients (β = .845–.879). Comparing the boundaries of 95% confidence intervals of them, the regression coefficients of PVLT were significantly higher than those of Maze Word Accuracy except for the picture narrative task (β = .675–.691). However, there were overlaps of confidence intervals between PVLT and GJT Syntax Accuracy (β = .710–.746). As mentioned previously, the maze task was used to capture learners' syntactic encoding skills, while the GJT aimed to tap into their skills of monitoring or detecting linguistic accuracy of stimuli (see Section 6.8.2). Considering the modality difference between these tasks, students' performance in the maze task can be related to the implementation of syntactic encoding rules in L2 (e.g., word order). Meanwhile, the accuracy scores of syntactic items in the GJT may represent their accessibility to the syntactic properties of target lemmas in their mental lexicon. Building on the assumption that syntactic properties of lemmas (e.g., part of speech; see Section 2.6.2) are stored in speakers' mental lexicon (Kormos, 2006; Levelt, 1989), the accessibility of such syntactic properties of lemmas can be regarded as the depth of vocabulary knowledge. Despite the multidimensional nature of vocabulary knowledge (e.g., size, depth, and fluency; Daller et al., 2007; Schmitt, 2008), vocabulary size and depth are arguably closely related to each other (González-fernández & Schmitt, 2020; Zhang & Yang, 2016). This close relationship between vocabulary size and depth may explain the non-significant difference in the regression coefficients between PVLT and GJT Syntactic Accuracy. Moreover, the regression coefficients of GJT Syntax Accuracy seemed to be

higher than those of GJT Morphological Accuracy in the argumentative and picture narrative tasks and also, despite slight overlaps of the confidence intervals, in the text summary tasks in both conditions (β = .439–.455). This may suggest that knowledge of syntactic properties of lemmas might tend to be more important for fluent speech production in L2 than that of morphological accuracy (e.g., articles, plural -*s*). Taken together, these relative strengths of regression coefficients may indicate that lexical resources can be regarded as a primary component of linguistic resource of CF in line with the lexically-driven nature of L2 speech production (Kormos, 2006). Furthermore, lexical resources may tend to play a slightly larger role in light of L2 oral fluency than syntactic resources. Although L2 morphological knowledge also significantly contributes to the construct of linguistic resource, its contributions might be lower than lexical and syntactic resources. From a theoretical perspective, the construct of linguistic resource in CF can thus be defined as the breadth and depth of linguistic repertoires to express speakers' intended message, including vocabulary size and sentence construction skills.

Regarding the latent variable of processing speed of CF, the strongest regression path was Maze Word RT (β = .794–.821) which taps into the speed of sentence construction. Although the regression path of Maze Word RT was stronger than that of the other four observed variables of processing speed in the CFA model of CF, there were slight overlaps in the boundaries of 95% confidence intervals among Maze Word RT, GJT Syntax RT (β = .604–.620), GJT Morphology RT (β = .614–.626), and articulatory speed (β = .635–.663) in the SEM model. However, the regression coefficients of Maze Word RT were still significantly higher than those of Picture Naming RT (β = .436–.453). Despite the slight overlaps in the confidence intervals, it seems plausible to argue that the primary component of processing speed is the speed of sentence construction (measured by Maze Word RT) in

light of L2 oral fluency. Such syntactic processing skills might thus be more important than lexical retrieval speed within the construct of processing speed of CF. Considering the regression coefficients in the measurement model of processing speed, the construct of processing speed can be defined as the efficiency in the manipulation of linguistic knowledge, particularly including the construction of phrases/clauses and the execution of articulatory gestures.

The relative importance of different linguistic domains (here, lexis vs. syntax) was opposite between the constructs of processing speed and linguistic resource. One possible explanation for the primary role of syntactic processing skills in processing speed is that in the current research context, the variability in the speed of linguistic processing might have aligned with the variability in the automaticity of L2 syntactic knowledge (cf. McManus & Marsden, 2019; Morgan-Short et al., 2014). Meanwhile, the current results indicate that the variability in the speed of lexical retrieval reflected the relatively small amount of variance of the latent variable of processing speed. The difference in the amount of contribution to the variance of processing speed might be explained by the difference in the complexity of cognitive processing between the maze task and the picture naming task. In the picture naming task, students are required to identify the concept described in the picture stimulus and to retrieve the L2 lemma corresponding to the concept identified through the visual recognition of the stimulus. It is arguably assumed that there is little individual variability in the speed of visual recognition because of its independence from L2 proficiency. Accordingly, the variability of picture naming latency (i.e., RT scores) should be reflective of the variability in the speed of lexical retrieval in which students select the lemma corresponding to the concept activated. From a neurocognitive perspective, the selection of lemmas in relation to activated semantic information (i.e., concepts) is assumed to largely rely on declarative memory and to be

learned relatively quickly (Ullman, 2015). It can also be argued that due to such a relatively quick acquisition of the concept-lemma mapping, there might not be large variability in the speed of lexical retrieval once the lemmas are acquired. In contrast, the information processing involved in the maze task entails a wider range of cognitive processes. In the maze task, students are asked to construct a sentence by selecting one single word from two options. Accordingly, students are required to access the syntactic properties of the provided words, including the previously selected ones and the candidate words. In addition, they also need to draw on L2 syntactic rules for phrase and clausal construction. Although accessing the syntactic information of lemmas is achieved by declarative memory, the application of syntactic rules to construct phrases according to the syntactic categories of lemmas (i.e., part of speech) is considered relevant to procedural memory (Ullman, 2015). It can thus be assumed that the acquisition of syntactic encoding rules may require a longer time than that of the concept-lemma mappings. Consequently, individual differences in L2 competence might have been captured better by the performance in the maze task.

Taken together, the picture naming task requires students to access their mental lexicon to search for the lexical entry corresponding to the concept, while the maze task includes not only the access to lemma to retrieve its syntactic properties, but also to activate syntactic encoding modules to access L2 syntactic rules. Furthermore, although the picture naming task is largely limited to the retrieval process of information (here, lexical representations), the maze task requires the manipulation of the information retrieved in addition to the retrieval of information (here, syntactic properties of lemma). Therefore, from the perspective of the complexity of cognitive processes, the maze task may include more processing components that entail the variability of automaticity and/or proceduralization than the picture naming task. As a result, the processing speed dimension of performance in the maze

task might have captured such individual variability in processing speed of CF among students more comprehensibly than in the picture naming task.

Another possible reason for the relatively small contribution of lexical retrieval speed to the construct of processing speed may lie in the difficulty of picture names in the picture naming task. To assess students' lexical retrieval speed in a reliable manner, the current study employed a set of relatively familiar lexical items as stimuli, following previous studies. The rationale for this methodological practice is that if learners cannot come up with the name of the given stimulus, the response cannot be used to measure the RT of the stimuli. Accordingly, to secure a sufficient number of responses, scholars commonly employ a set of picture stimuli familiar to their participants. However, due to such relatively familiar lexical items, students' variability in lexical processing speed might have been underestimated.

8.3.2 Dimensionality of utterance fluency

Motivated by the mechanisms of speech production as well as Tavakoli and Skehan's (2005) triad model of UF, the current study tested the single-, two- and three-factor models of UF concerning the fit to the data in four speaking tasks differing in the quality of speech processing demands. A series of CFA revealed that two CFA models—a three-factor model based on Tavakoli and Skehan (2005) (Model UF 6) and a two-factor model with a unitary latent variable consisting of speed and breakdown fluency (Model UF 7) showed a virtually equal model fit to the current dataset. Even in the three-factor model, there were strong associations between the latent variables of speed and breakdown fluency across four speaking tasks (β = |.929–.960| in the CFA model of UF; β = |.769–.822| in the SEM model of the CF-UF link). However, considering the theoretical distinction between speed and breakdown fluency, the three-factor model following Tavakoli and Skehan (2005) was

adopted as the final CFA model of UF, suggesting that the construct of UF consists of speed, breakdown, and repair fluency. The results in the current study confirm the generalizability and robustness of Tavakoli and Skehan's (2005) triad model of UF, which was proposed based on the speech data elicited only from picture narrative tasks. Additionally, the current study showed the optimal model-fit in all of the four different speaking task types (e.g., SRMR = .056–.070). Another methodological advantage of the current study was the inclusion of articulation rate as the measure of speed fluency. Tavakoli and Skehan's (2005) study only included two composite measures—speech rate and mean length of run—as the measures of speed fluency, and these two measures and breakdown fluency measures (e.g., pause frequency, mean pause duration) were loaded on the same latent variable. Accordingly, Tavakoli and Skehan (2005) could only conceptually argue the distinguishability between speed and breakdown fluency. However, thanks to the pure measure of speed fluency (i.e., articulation rate; Tavakoli et al., 2020), the current study statistically proved the distinction between speed and breakdown fluency.

To revisit the construct definition of each dimension of UF, the relative importance of observed variables within each latent variable is discussed. As for speed fluency, both articulation rate and mean length of run considerably contributed to the latent variable of speed fluency, while the regression coefficients of articulation rate (β = .876–.905) seemed to be slightly higher than those of mean length of run (β = .721–.882). This may support the statistical procedure of handling mean length of run as the measure of speed fluency in the SEM analysis and may also suggest that articulation rate is a more representative measure of speed fluency. Comparing the boundary of 95% confidence intervals of regression coefficients, the significant difference between these measures was only found in the argumentative task. However, from the perspective of the construct validity of these

measures, mean length of run is a composite measure and taps into both speed and breakdown fluency (Bosker et al., 2013; Tavakoli et al., 2020). The slightly lower regression coefficients of mean length of run to the latent variable may thus indicate that some amount of variance of mean length of run might have derived from the factor other than the construct of speed fluency (e.g., the construct of breakdown fluency). Therefore, the primary component of speed fluency is arguably represented by the measure of articulation rate. Since articulation rate is regarded as the eventual outcome of the whole range of speech processing (Kormos, 2006; Segalowitz, 2010), the construct of speed fluency can be defined as the overall efficiency of speech production.

Regarding breakdown fluency, the measure of mid-clause pause ratio contributed to the construct of breakdown fluency to the largest extent among various pause frequency and duration measures. Considering the boundaries of 95% confidence intervals, the regression coefficients of mid-clause pause ratio (β = .919–.963) were significantly higher than those of the other breakdown fluency measures—end-clause pause ratio (β = .373–.515), filled pause ratio (β = .545–.628), and mean pause duration (β = .528–.690; except for the text summary task without RAA). There were no significant differences in the regression coefficients among these three measures (mean pause duration, end-clause pause ratio, and filled pause ratio). It is thus plausible that the representative component of breakdown fluency is the frequency of breakdowns in the middle of utterances, while the length of pauses and the frequency of pauses at clausal boundaries and filled pauses might be secondary (Bosker et al., 2013). From the perspective of speech production mechanisms, mid-clause pauses are reflective of disruptions in L2-specific processing, such as lexical retrieval and sentence construction, while end-clause pauses are associated with content-related planning (De Jong, 2016b; Skehan et al., 2016; Tavakoli, 2011). Accordingly, the construct of breakdown

fluency may represent speakers' ability to continue spontaneous speech without disruptions in L2-specific speech processing.

It should be noted that in the course of building different CFA models, the distinction of pause duration measures by pause location was removed due to the strong correlation between mid- and end-clause pause duration measures (r = .735) and then replaced with a single measure of mean pause duration without the distinction of pauses by location. As mentioned previously, the contribution of pause ratio measures to the latent variable of breakdown fluency was differentiated between mid- and end-clause pause ratio measures. In contrast, the distinction of pauses based on location might not be meaningful in pause duration measures. It can thus be argued that the causes of breakdowns—either by linguistic retrieval problems or content-specific problems—could be distinguished based on pause location in a valid manner. Meanwhile, it might not be meaningful to distinguish the length of pauses, which are reflective of the time to repair those breakdowns, based on the location of pauses. This is possibly because the repair of breakdowns can be achieved by either conceptual or linguistic modification of the message (cf. Kormos, 2006; Levelt, 1989). For instance, when a speaker experiences a breakdown in the middle of the utterance due to the lack of lexical expressions for their intended message, the speaker can either conceptually modify the original message or search for some alternative linguistic expressions while keeping the original meaning of the message. In other words, even though pauses take place in the middle of the utterance, the process of repairing the pauses can include both content-relevant and linguistic processing.

As regards the latent variable of repair fluency, the most representative observed variable was the measure of self-repetition ratio. With regard to the boundaries of 95% confidence

intervals, the regression coefficients of self-repetition ratio to the latent variable of repair fluency (β = .787–.860) were overall significantly higher than those of false start ratio (β = .289–.459) and self-correction ratio (β = .487–.632), except for the pair of self-repetition ratio and self-correction ratio in the text summary task without RAA. There were no significant differences in the regression coefficients between self-correction ratio and false start ratio. Accordingly, the primary component of repair fluency can be regarded as the frequency of self-repetitions, while both self-corrections and false starts are of secondary importance. Previous research into repair fluency suggests that the frequency of self-repetitions may be independent of L2 proficiency (Tavakoli et al., 2020) and is reflective of learners' speaking style, as indicated by a strong correlation with the corresponding L1 UF measure (De Jong et al., 2015). In addition, the correlational analysis in the current study (see Section 8.2.5) did not find a consistent pattern of associations between self-repetition ratio and different CF measures across speaking tasks. It might thus be argued that the construct of repair fluency, largely represented by self-repetition ratio, is not directly associated with L2-specific competence. Alternatively, the use of self-repetition can be regarded as part of fluency strategy or problem-solving mechanisms (Dörnyei & Kormos, 1998). Specifically, the use of self-repetitions plays the role of buying time for monitoring or retrieval processes, as lexicalized fillers do. The advantage of self-repetition for fluent speech production is that the repetition of the previously processed utterances would not consume a large amount of attentional resources, because the phonological representations of those utterances are still activated in the speaker's phonological short-term memory. From the perspective of speech production, another important characteristic of repair fluency is that repair fluency features are in a complementary relationship with breakdown fluency (Tavakoli & Wright, 2020). When a speaker experiences the disruptions in speech processing and is required to repair their utterance, the speaker can engage with the repair either by producing no utterances (i.e.,

silent pauses) or repeating the previous utterance (i.e., self-repetition). Therefore, the strategic use of self-repetition is more or less due to the speaker's individual preference, consequently leading to the ambiguous association with L2 competence. However, considering the association of repair fluency with the other constructs of UF (see Table 48), speakers with better speed fluency tend to produce fewer repair fluency phenomena and fewer breakdown fluency features. In other words, higher performance in terms of repair fluency should be characterized by the fewer number of repair fluency phenomena, despite the strategic use of those features. This is presumably because even the strategic use of self-repetitions, as well as the repair of utterances as self-corrections and false starts, only happens when the speaker detects some breakdowns or errors are detected during speech production. Therefore, regardless of the purposes and reasons for the production of repair fluency features, repair fluency as a construct can reflect the ability to produce L2 speech with the fewer number of disfluency features.

8.3.3 Contribution of cognitive fluency to utterance fluency

The SEM model in Study 3 revealed that the multidimensional interrelationship between two components of CF—linguistic resource and processing speed—and the triad of UF—speed fluency, breakdown fluency, and repair fluency—with some variations in the strength of associations across four speaking tasks. The structural model of the SEM model suggested that the latent variable of processing speed of CF contributed to that of speed fluency consistently across speaking tasks (β = .431–.609). Meanwhile, the latent variable of linguistic resource made significant contributions to that of speed fluency only in the text summary tasks regardless of the RAA conditions (β = .234 for without-RAA condition; β = .276 for with-RAA condition). It can thus be argued that the overall efficiency of speech production (speed fluency) is primarily supported by the speed of linguistic processing skills

rather than the breadth and depth of linguistic resources, while linguistic resources may additionally contribute to speed fluency when the relevant linguistic items are to some extent activated prior to speaking. The consistent contributions of the speed dimension of CF to speed fluency in the current study may provide some supporting evidence for Segalowitz's (2016) claim that CF is mainly characterised by the speed of L2-specific linguistic processing. Meanwhile, the task-dependent role of linguistic resources in speed fluency can be interpreted with regard to the effects of the enhanced activation of relevant linguistic items on UF performance. As discussed in Study 2 (see Section 7.3.2), the linguistic items activated in the text summary tasks may impede the retrieval of students' own linguistic resources, because if those linguistic items are not fully acquired for production, despite their higher level of activations, they need to retrieve alternative items mastered for production. Inversely, if students have acquired those activated items for productive use, the enhanced activation of those items can assist students to use the items rapidly (cf. priming effects McDonough & Trofimovich, 2008), subsequently increasing the overall efficiency of speech production (i.e., speed fluency). Therefore, the contributions of linguistic resources to speed fluency may increase, especially in the communicative situation where the acquisition of relevant linguistic items plays an essential role in the completion of the given task.

The latent variable of breakdown fluency was in general associated with both dimensions of CF—linguistic resource and processing speed—consistently across speaking tasks, despite the marginally significant contribution of linguistic resource in the text summary task with RAA (p = .061). Although the latent variable of breakdown fluency seemed to show slightly stronger associations with processing speed (β = .376–.502) than with linguistic resource (β = .221–.345), the difference in the regression coefficients did not reach statistical significance, according to their boundaries of 95% confidence intervals. Therefore, the results

here indicated that the ability to continue speaking without disruptions may be underpinned by both the availability of linguistic resources and the speed of linguistic processing. This finding is in line with the broad definition of CF, which assumes that breakdowns in speech production can be caused by either lack of linguistic resources and slow processing speed (see Kormos, 2006; see also Section 3.6) as well as the methodological practice in the research into the CF-UF link (De Jong et al., 2013; Kahng, 2020). Moreover, the association of breakdown fluency with both dimensions of CF may give some insights into how the constructs of speed fluency and breakdown fluency are theoretically distinguishable. More specifically, despite the large overlap between the latent variables of speed fluency and breakdown fluency suggested by the CFA models of UF, speed fluency was mainly related to the speed dimension of CF, while breakdown fluency was connected to the linguistic resource of CF as well as the processing speed component. These differences in the underlying components of CF between speed fluency and breakdown fluency may also support the statistical decision to adopt the three-factor model of UF, as opposed to the two-factor model which combined speed and breakdown fluency into a unitary construct (Model UF 7).

The SEM model suggested that the latent variable of repair fluency was associated with linguistic resources in all the speaking tasks except for the argumentative task. Meanwhile, the processing speed of CF did not significantly contribute to the latent variable of repair fluency in any of the speaking tasks. In other words, the significant contribution of linguistic resource (β = .330–.375) was only found in the speaking tasks where the content of speech is largely predefined (i.e., closed task; see Pallotti, 2009). Previous studies have argued that the construct of repair fluency is relatively independent of L2 proficiency (Tavakoli et al., 2020) and is reflective of individual speakers' speaking style (De Jong et al., 2015; Peltonen, 2018).

However, the current result of the SEM analysis may suggest that repair fluency is not entirely independent of L2-specific linguistic knowledge in some communicative situations where some constraints on the content of speech are imposed on speakers. Regarding the predefined speech content, one of the crucial task characteristics is that students could not avoid expressing some information to achieve the given task, even if they have not fully acquired the necessary linguistic items for the task. Students are thus required to engage with modifying the intended message or searching for some alternative expressions using their own resources. As discussed previously, students can strategically or subconsciously use self-repetition, which is the representative phenomenon of repair fluency (see Section 8.3.2), to buy time for repairing their utterances either by self-correction or reformulation (Dörnyei & Kormos, 1998). Therefore, the contribution of linguistic resource to repair fluency can be interpreted as the use of and/or engagement with repair due to the lack of linguistic resources needed to express the essential information in the given task. Similarly, the non-significant association between linguistic resource and repair fluency in the argumentative task can be explained by the nature of open tasks. In open tasks including the current argumentative task, students are allowed to avoid using difficult or unfamiliar linguistic items in their speech by modifying their intended message (Préfontaine & Kormos, 2015). They are thus less likely to experience the breakdowns caused by the necessity to retrieve such difficult linguistic items in open tasks than in closed tasks. This pattern was observed in Study 2 as fewer mid-clause pauses in the argumentative task than in the picture narrative task (see Section7.3.1). Consequently, there would be fewer opportunities to use some repair phenomena to buy time for repairing speech in open tasks. Such limited opportunities to repair their utterances may have obscured the association between repair fluency and linguistic resource in the open task (the argumentative task, here). Finally, despite the significant contribution of linguistic resource of CF to repair fluency at the level of constructs, it is noteworthy that there were

inconsistent correlational relationships between the measures of CF and repair fluency at the level of observed variables (see Section 8.2.5). It can thus be argued that to detect the task-specific contribution of CF to repair fluency, more fine-grained measures of linguistic resource and repair fluency might be needed for future studies.

8.4 Summary

This chapter reported the results of Study 3, addressing the contribution of CF (cognitive fluency) to UF (utterance fluency) at the level of constructs (RQ3-1) with regard to its variability across four speaking tasks (RQ3-2). As a preliminary analysis for these RQs, the dimensionality of CF and UF was examined (RQ3-1a and RQ3-1b, respectively). In order to address these RQs, a set of CFA and SEM analyses was conducted. The findings of Study 3 are visually summarized in Figure 29.

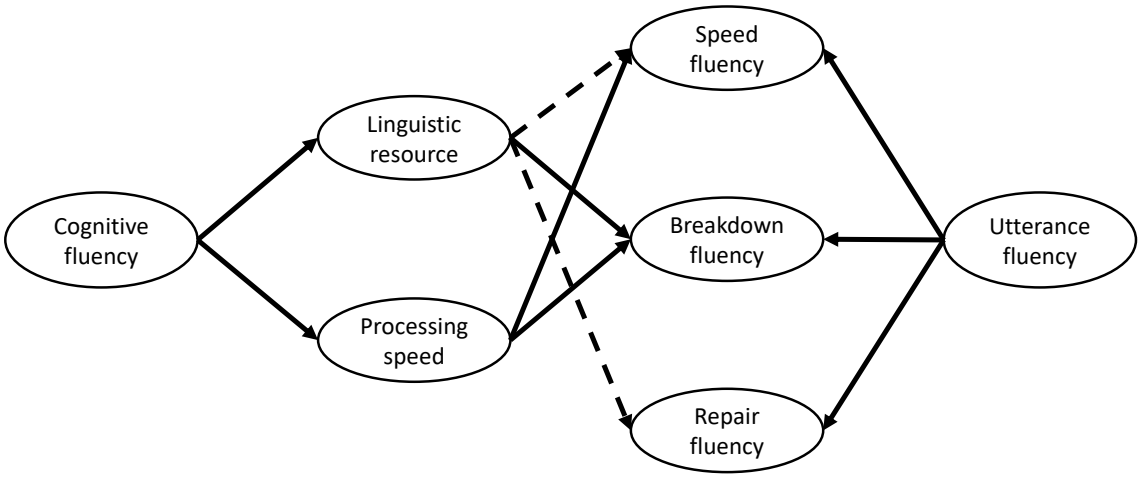


Figure 29. The visualized summary of findings of Study 3.
Note. The arrow lines indicate significant regression paths in all the speaking tasks, and the dotted lines indicate that the significance of the regression paths is task-dependent.

Prior to the SEM analysis regarding the CF-UF link, several CFA models were tested with regard to the goodness of fit to the current dataset to identify the parsimonious factor structure of CF and UF. The final CFA model of CF consisted of two dimensions—*linguistic*

resource and *processing speed*. Building on the factor structure of CF, Study 3 defined the construct of linguistic resource as the breadth and depth of linguistic knowledge to express speakers' intended message and that of processing speed as the efficiency in the manipulation of such linguistic knowledge. The results also suggested that these two dimensions of CF were different in their primary linguistic domains. Vocabulary size was the most representative aspect of linguistic resource, while the speed of constructing phrases and clauses contributed to processing speed to the largest extent.

Meanwhile, the CFA model of UF supported the Tavakoli and Skehan's (2005) triad model of UF which consisted of *speed fluency*, *breakdown fluency*, and *repair fluency*. According to the size of regression coefficients in the CFA model, the representative observed variables were identified for each subconstruct of UF: articulation rate for speed fluency, mid-clause pause ratio for breakdown fluency, and self-repetition ratio for repair fluency. Accordingly, Study 3 also re-defined the construct of speed fluency as the overall efficiency of speech production, that of breakdown fluency as the ability to continue speech production without disruptions in L2-specific linguistic processing, and that of repair fluency as the ability to produce L2 speech with fewer disfluency features.

The SEM analysis revealed the complex interplay between the multidimensionality of CF and UF and speaking task types. Speed fluency was primarily associated with processing speed, while linguistic resources can play a role only when relevant linguistic items are activated in advance (i.e., the text summary tasks). Meanwhile, both linguistic resources and processing speed contributed to breakdown fluency consistently across speaking tasks. Finally, the contribution of linguistic resources to repair fluency was significant only when the content of speech was predefined (i.e., the picture narrative and text summary tasks), while repair

fluency was generally independent of processing speed. These results confirmed that the processing speed of CF showed a consistent pattern of the contributions to UF across speaking task types, whereas the role of linguistic resource of CF in UF may tend to vary, depending on task characteristics, such as the availability of relevant linguistic items and the predefined content of speech.

Chapter 9: Results and Discussion of Study 4—Association of First and Second Language Utterance Fluency

9.1 Introduction

This chapter reports the results of Study 4, which addresses the L1-L2 UF link with respect to the moderator effects of L2 proficiency. As a preliminary analysis, the descriptive statistics and distributions of L1 and L2 UF performance in the argumentative tasks were inspected (Section 9.2.1). To examine how L1 UF behaviours differ from L2 UF behaviours, a set of GLMMs tested the effects of language status (L1 vs. L2) on UF measures, with the random-effects of individual participants and topics of the L2 argumentative tasks (Section 9.2.2). Moreover, correlational analyses were conducted to see the overall tendency of the association between L1 and L2 UF measures (Section 9.2.3). To address RQ4-1, another set of GLMMs were constructed to predict L2 UF measures from the corresponding L1 UF measures with the random-effects of individual participants and topics of the L2 argumentative tasks (Section 9.2.4). For RQ4-2, the factor scores of CF from Study 3 (linguistic resource and processing speed; see Section 8.2.4) and the interaction effects by these two CF scores and L1 UF measures were added to the GLMMs constructed for RQ4-1 (Section 9.2.5). These findings are discussed from the perspective of language-general processes of speech production and individual speaking style, providing insights into the complex interplay between the multidimensionality of L2 proficiency and cross-linguistic divergence in the L1-L2 UF link (Section 9.3)

9.2 Results

9.2.1 Descriptive statistics and distributions of L1 utterance fluency

The descriptive statistics of the L1 and L2 UF measures were summarized in Table 49. Shapiro-Wilk tests suggested that most of the UF measures were not normally distributed,

whereas density plots indicated that articulation rate can be regarded as being normally distributed (see Figure 30; for the other UF measures, see Appendix Q). Accordingly, in the subsequent GLMMs, the gaussian distribution (i.e., normal distribution) was applied to the models of articulation rate, while the gamma distribution was applied to the models of the other UF measures with the log link function.

Table 49. *Descriptive summary of L1 and L2 utterance fluency measures in Study 4.*

					Shapiro-Wilk test	
Utterance fluency measures	Condition	*M*	*SD*	*SE*	*Statistics*	*p-value*
Articulation rate	L1.Arg	8.217	1.561	0.153	0.684	< .001
	L2.Arg.Lib	3.131	0.601	0.059	0.990	0.633
	L2.Arg.Oly	3.142	0.593	0.058	0.976	0.060
Speech rate	L1.Arg	5.584	1.459	0.143	0.965	0.008
	L2.Arg.Lib	1.688	0.688	0.067	0.957	0.002
	L2.Arg.Oly	1.771	0.678	0.067	0.960	0.003
Mean length of run	L1.Arg	16.266	6.068	0.595	0.889	< .001
	L2.Arg.Lib	4.325	2.532	0.248	0.708	< .001
	L2.Arg.Oly	4.596	3.017	0.296	0.567	< .001
Mid-clause pause ratio	L1.Arg	0.044	0.019	0.002	0.975	0.043
	L2.Arg.Lib	0.222	0.111	0.011	0.908	< .001
	L2.Arg.Oly	0.212	0.086	0.008	0.977	0.064
End-clause pause ratio	L1.Arg	0.028	0.008	0.001	0.983	0.202
	L2.Arg.Lib	0.069	0.022	0.002	0.983	0.198
	L2.Arg.Oly	0.058	0.020	0.002	0.961	0.004
Filled pause ratio	L1.Arg	0.034	0.022	0.002	0.928	< .001
	L2.Arg.Lib	0.113	0.093	0.009	0.892	< .001
	L2.Arg.Oly	0.100	0.080	0.008	0.885	< .001
Mid-clause pause duration	L1.Arg	0.804	0.290	0.028	0.920	< .001
	L2.Arg.Lib	1.136	0.499	0.049	0.827	< .001
	L2.Arg.Oly	1.085	0.505	0.050	0.754	< .001
End-clause pause duration	L1.Arg	1.009	0.444	0.044	0.885	< .001
	L2.Arg.Lib	1.338	0.894	0.088	0.741	< .001
	L2.Arg.Oly	1.351	1.241	0.122	0.584	< .001
Self-repetition ratio	L1.Arg	0.007	0.009	0.001	0.747	< .001
	L2.Arg.Lib	0.081	0.084	0.008	0.750	< .001
	L2.Arg.Oly	0.071	0.066	0.006	0.861	< .001

Self-correction ratio	L1.Arg	0.005	0.005	0.000	0.882	< .001
	L2.Arg.Lib	0.020	0.018	0.002	0.861	< .001
	L2.Arg.Oly	0.020	0.015	0.001	0.928	< .001
False start ratio	L1.Arg	0.001	0.002	0.000	0.547	< .001
	L2.Arg.Lib	0.010	0.012	0.001	0.753	< .001
	L2.Arg.Oly	0.008	0.011	0.001	0.746	< .001

Note. L1.Arg = L1 argumentative speech; L2.Arg.Lib = L2 argumentative speech with the topic of Library; L2.Arg.Oly = L2 argumentative speech with the topic of the Tokyo Olympics (for the exact prompts, see Appendix D).

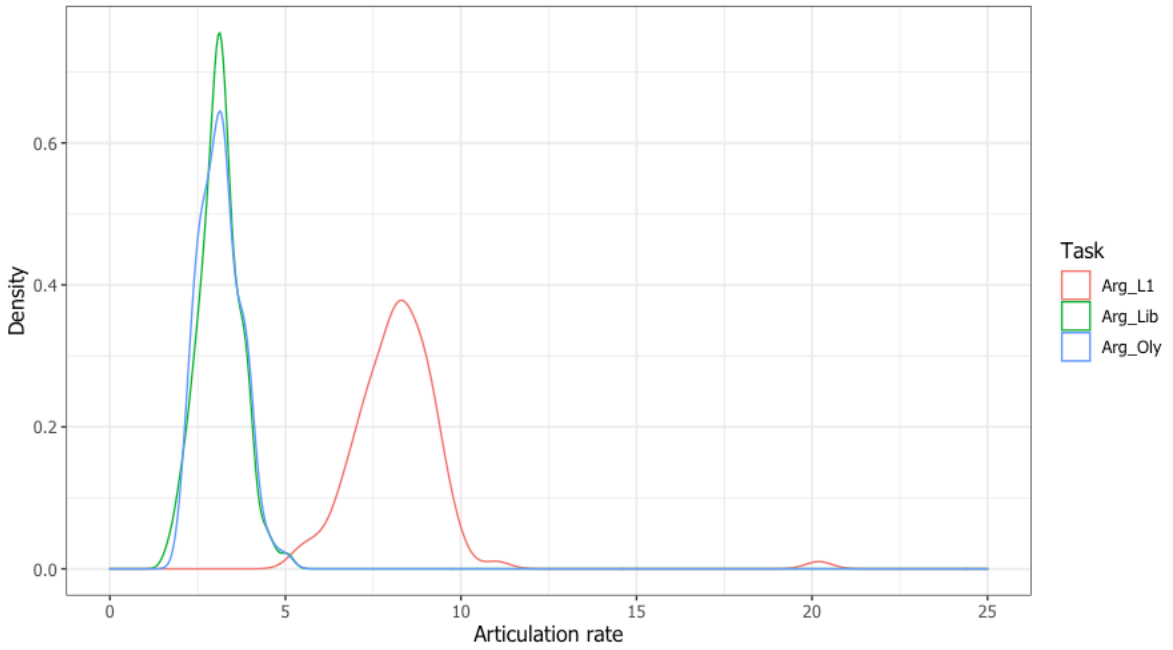


Figure 30. The density plot of articulation rate.
Note. Arg_L1 = the measure in the L1 argumentative task; Arg_Lib = the measure in the L2 argumentative task with the topic of Library; Arg_Oly = the measure in the L2 argumentative task with the topic of the Tokyo Olympics.

9.2.2 Difference between L1 and L2 utterance fluency measures

As another preliminary analysis, a set of GLMMs was built to examine whether UF performance was more fluent in L1 than in L2 in the current dataset. The GLMMs predicted the outcome variable of UF measures from the fixed-effects predictor of Language status (L1 vs. L2) with L1 as the reference level. Considering that this predictor variable was a within-subject independent variable, the inclusion of the random slopes of Participants and Topics for Language status would not be distinguishable from the random error variance (Barr,

2013). Therefore, the GLMMs only included the random intercepts of Participants and Topics of L2 speech. However, in some GLMMs (articulation rate, mid-clause pause ratio, end-clause pause duration, self-correction ratio), the regression models failed to converge with both of the random intercepts. These GLMMs were revised by removing the random intercepts of Topics, because the random intercepts of Topics consistently explained a smaller amount of variance of the outcome variables than the random intercepts of Participants. All the GLMMs indicated the significant simple effects of Language status, showing that participants' UF performance was more fluent in L1 than in L2 (see Table 50). More specifically, their L2 speech was characterized by the slower speed of delivery (i.e., the negative direction of coefficients in AR, SR, and MLR), more and longer breakdowns, and more disfluency features (the positive direction of coefficients in all breakdown and repair fluency measures), compared to their L1 UF performance.

Table 50. *Summary of the effects of Language status on utterance fluency performance.*

		Fixed effects: Language status		Random effects (intercepts)			
UF measures	(Intercept)	*Estimate*	*SE*	Participant	Topic	Marginal R2	Conditional R2
AR	8.217***	-5.080***	0.107	0.254	—	0.846	0.883
SR	1.699***	-1.227***	0.027	0.054	0.065	0.738	0.857
MLR	2.745***	-1.359***	0.028	0.074	0.000	0.725	0.856
MCPR	-3.194***	1.567***	0.040	0.083	—	0.729	0.840
ECPR	-3.601***	0.779***	0.028	0.037	0.002	0.573	0.739
FPR	-3.534***	0.956***	0.063	0.339	0.001	0.244	0.651
MCPD	-0.259***	0.290***	0.001	0.056	0.000	0.127	0.511
ECPD	-0.047	0.189***	0.040	0.126	—	0.026	0.446
SRR	-5.470***	2.581***	0.105	0.535	0.001	0.563	0.766
SCR	-5.445***	1.373***	0.090	0.239	—	0.362	0.567
FSR	-7.267***	2.211***	0.118	0.858	0.012	0.377	0.677

Note. Due to the singular fit, the models of AR, MCPR, ECPD, and SRR excluded the random intercepts of Topics of L2 argumentative speech; the reference level of Language status is L1.

9.2.3 Correlation between L1 and L2 utterance fluency measures

Prior to RQ4-1 (i.e., GLMM), the overall relationship between L1 and L2 UF performance was examined using correlational analyses. Considering the non-normal distributions of most of the UF measures, Spearman's rank-order correlation coefficients were employed to correlate L1 UF measures with the corresponding L2 UF measures in two prompts (the topics of Library and the Tokyo Olympics; see Appendix D). As shown in Table 51, the correlational pattern between L1 and L2 UF performance was, in general, identical across the topics of the L2 argumentative tasks, while two UF measures (end-clause pause ratio, false start ratio) in the topic of Library did not correlate with the corresponding L1 UF measures. From a theoretical perspective, end-clause pauses and false starts are reflective of conceptualization processes (e.g., content planning; De Jong, 2016b; Tavakoli, 2011; Williams & Korko, 2019). It can thus be argued that the differential correlation coefficients between the two topics might have been caused by the difference in the difficulty with content planning or topic development. Although the overall correlational patterns between L1 and L2 UF performance did not substantively differ between the two topics of L2 argumentative tasks, the subsequent GLMMs predicting L2 UF measures from the corresponding L1 UF measures included both of the L2 argumentative tasks and handled the topics as the random-effects variable to control for the variability of the L1-L2 UF link across topics.

Table 51. *Correlation coefficients between L1 and L2 utterance fluency measures.*

	Arg.Lib		Arg.Oly	
Utterance fluency measure	r_s	*p*	r_s	*p*
Articulation rate	0.352	< .001	0.405	< .001
Speech rate	0.359	< .001	0.447	< .001
Mean length of run	0.312	0.001	0.488	< .001
Mid-clause pause ratio	0.253	0.010	0.393	< .001
End-clause pause ratio	0.170	0.085	0.396	< .001
Filled pause ratio	0.478	< .001	0.501	< .001

Mid-clause pause duration	0.410	< .001	0.327	< .001
End-clause pause duration	0.429	< .001	0.486	< .001
Self-repetition ratio	0.312	0.001	0.482	< .001
Self-correction ratio	0.300	0.002	0.243	0.013
False start ratio	0.116	0.241	0.378	< .001

Note. L2.Arg.Lib = L2 argumentative speech with the topic of Library; L2.Arg.Oly = L2 argumentative speech with the topic of the Tokyo Olympics.

9.2.4 Predictive power of L1 utterance fluency for L2 utterance fluency

To examine the predictive power of L1 UF performance in L2 UF performance (RQ4-1), another set of GLMMs was constructed. For all GLMMs, the outcome variable was L2 UF measures, and the fixed-effects predictor variable was the corresponding L1 UF measures, with the random intercepts of Participants and Topics of the L2 argumentative tasks. As summarized in Table 52, the GLMMs suggested that all of the L2 UF measures were significantly predicted from the corresponding L1 UF measures. The positive directions of the coefficients of the predictor variable of L1 UF measures in the GLMMs indicated that participants who spoke fluently in L1 also tended to speak fluently in L2. In addition, the amount of variance of L2 UF measures explained by the corresponding L1 UF measures varied considerably, depending on the constructs of the UF measures. The marginal R^2 value refers to the variance explained by the fixed-effects variable (the corresponding L1 UF measures, here), whereas the conditional R^2 value refers to the variance explained by both random- and fixed-effects predictors altogether. For instance, in the case of the GLMM of mean length of run, the variance explained by the corresponding L1 UF measures was 19.6%. According to Plonsky and Ghanbar's (2018) guideline (R^2 = .10–.18 as Small; R^2 = .18–.51 as Medium; R^2 = .51–.70 as Strong), these variances explained by the fixed-effects predictor (Marginal R^2 = 5.2–24.7%) suggest the small-to-medium effect size of the overall predictive power of L1 UF for L2 UF, while a large amount of variance of L2 UF measures was explained by the random-effects predictors (R^2 = 36.4–66.2%; individual participants and

topics, here). Closely looking at the marginal R^2 values, the predictive power of L1 UF for L2 UF was ignorable in articulation rate (R^2 = 7.1%; speed fluency), self-correction ratio (R^2 = 7.1%; repair fluency), and false start ratio (R^2 = 5.2%; repair fluency). Meanwhile, another repair fluency measure—self-repetition ratio—indicated a nearly medium effect size (R^2 = 17.1%). Regarding the breakdown fluency measures, the effect sizes can be considered medium in filled pause ratio (R^2 = 24.7%) and small in the other breakdown fluency measures (R^2 = 10.8–15.5%).

Table 52. *Summary of the effects of L1 utterance fluency measures on the corresponding L2 utterance fluency measures.*

UF measures	(Intercept)	Fixed effects: L1 UF measure		Random effects (intercepts)		Marginal R^2	Conditional R^2
		Estimate	*SE*	Participant	Topic		
AR	3.137***	0.159**	0.052	0.237	—	0.071	0.732
SR	0.448***	0.139**	0.050	0.087	0.001	0.137	0.760
MLR	1.370***	0.180***	0.048	0.085	0.001	0.196	0.718
MCPR	-1.659***	0.179**	0.064	0.138	0.000	0.136	0.720
ECPR	-2.835***	0.113**	0.038	0.051	0.003	0.108	0.563
FPR	-2.663***	0.459***	0.106	0.452	0.002	0.247	0.778
MCPD	0.018	0.143***	0.043	0.068	0.001	0.155	0.680
ECPD	0.119	0.219***	0.059	0.154	0.000	0.139	0.590
SRR	-2.990***	0.411***	0.001	0.439	—	0.171	0.616
SCR	-4.131***	0.227**	0.080	0.263	0.000	0.071	0.436
FSR	-5.158***	0.292**	0.111	0.767	0.012	0.052	0.531

Note. Due to the singular fit, the models of AR and SRR excluded the random intercepts of Topics of L2 argumentative speech.

9.2.5 Moderator effects of L2 proficiency on L1-L2 utterance fluency link

RQ4-2 aims to examine whether the predictive power of L1 UF for L2 UF is moderated (weaken or enhanced) by the speaker's L2 proficiency. To operationalize L2 proficiency in the context of L2 fluency research, I selected the factor scores of CF in Study 3 due to their relevance to L2 UF performance. To investigate the moderator effects of L2 proficiency on the L1-L2 UF link, two interaction terms by L1 UF and each of the CF factor scores (Linguistic resource [LR] and Processing Speed [PS]) were added to the GLMMs constructed

for RQ4-1. To control for the simple effects of the predictor variables (i.e., L1, LR, and PS) on L2 UF, the GLMMs here included these predictors as well. For all the GLMMs for RQ4-2, the structure of random-effects variables was identical to the one constructed for RQ4-1; most models included the random intercepts of both participants and topics of the L2 argumentative tasks, while some of the models only included the random intercepts of participants for the sake of model convergence. From a statistical perspective, RQ4-2 addresses whether those two interaction terms (i.e., by L1 and LR factor score; and by L1 and PS factor score) are significant in a confirmatory manner. Accordingly, to avoid overly complex models with many predictor variables, the GLMMs for RQ4-2 did not include the two-way interaction between LR and PS and the random-slopes of the individual participants and topics for the target interaction effects. The code used to construct the GLMMs for RQ4-2 is shown below:

*L2UF ~ L1UF * (LR factor score + PS factor score) + (1|Participant) + (1|Topic)*

When the model failed to converge, I decided to reduce the predictor variables through backward methods. Specifically, the first attempts were to reduce one of the two-way interaction terms by L1 and CF factor scores (i.e., either by L1 and LR or by L1 and PS) and to see whether the revised model would converge. If the model continued to fail to converge without those interaction terms, the exclusion of CF factor scores (LR and PS) was considered. As a result, the failure of model convergence was found in the GLMMs of mean length of run (MLR), self-repetition ratio (SRR), and self-correction ratio (SCR). In the case of MLR, when including either of the interaction terms, both of the interaction by L1 UF and LR ($\beta = 0.112$, $p = .039$) and by L1 UF and PS ($\beta = 0.119$, $p = .019$) were significant. However, the inclusion of both interaction terms prevented the model from converging.

Consequently, I compared the model fit to the data between these models with either the L1UF-LR interaction or the L1UF-PS interaction, using the pairwise Likelihood Ratio Test (LRT; Baayen, 2008). The result showed that these two models did not significantly differ in the fit to the current. Despite the non-significant difference in the model fit, the model-fit indices in the model with the L1UF-PS interaction (AIC = 558.6, Log Likelihood = -271.32) indicated a slightly better fit to the data than the model with the L1UF-LR interaction (AIC = 559.74, Log Likelihood = -271.87). Accordingly, the final model of MLR included the fixed-effects predictors of L1, LR, and PS as well as the interaction by L1 and PS. SRR did not converge even without the simple effects of LR and PS. Consequently, the final model of SRR was identical to the model for RQ4-1, which only included the corresponding L1 UF measures as the fixed-effects predictor. Meanwhile, the GLMM of SCR converged by excluding the interaction between L1 UF and LR.

Table 53. *Summary of the interaction effects by L1 utterance fluency measures and linguistic resource and processing speed on the corresponding L2 utterance fluency measures.*

	Fixed effects: L1UF*LR		Fixed effects: L1UF*PS			
UF measures	*Estimate*	*SE*	*Estimate*	*SE*	Marginal R^2	Conditional R^2
AR	-0.198**	0.075	0.306**	0.103	0.321	0.737
SR	-0.120	0.062	0.199**	0.072	0.499	0.822
MLR	—	—	0.119*	0.051	0.442	0.768
MCPR	0.053	0.089	0.147	0.095	0.448	0.777
ECPR	0.019	0.054	0.024	0.062	0.202	0.585
FPR	0.165	0.181	-0.225	0.212	0.347	0.795
MCPD	-0.084	0.069	0.140	0.076	0.352	0.722
ECPD	-0.042	0.101	0.024	0.107	0.301	0.625
SRR	—	—	—	—	0.169	0.622
SCR	—	—	0.080	0.099	0.091	0.436
FSR	-0.107	0.207	0.119	0.199	0.123	0.529

Note. Random-effects predictors include random intercepts of individual participants and two topics of L2 speech; L1UF = the corresponding L1 UF measures; LR = Factor score of linguistic resource; PS = Factor score of processing speed; due to the failure of model convergence, the models of MLR, SRR, and SCR excluded the interaction term by L1UF and LR, while the models of SRR excluded the interaction term by L1UF and PS.

As summarized in Table 53, the significant moderator effects of L2 proficiency on the L1-L2 UF link was found only in speed fluency measures—articulation rate, speech rate, and mean length of run (for the full statistical estimates, see Appendix R). More specifically, the L1-L2 association in articulation rate was weakened by the score of linguistic resource and was also enhanced by the score of processing speed. In other words, for those who acquired a wider range of L2 linguistic resources, L2 articulation rate tended to be relatively independent of L1 articulation rate (see Figure 31). In contrast, the L1-L2 association in articulation rate, speech rate, and mean length of run was enhanced as a function of L2 processing speed (see Figure 32–34). This consistent pattern in speed fluency measures suggests that the more efficiently learners can process L2 knowledge, the closer to L1 their L2 speed fluency is.

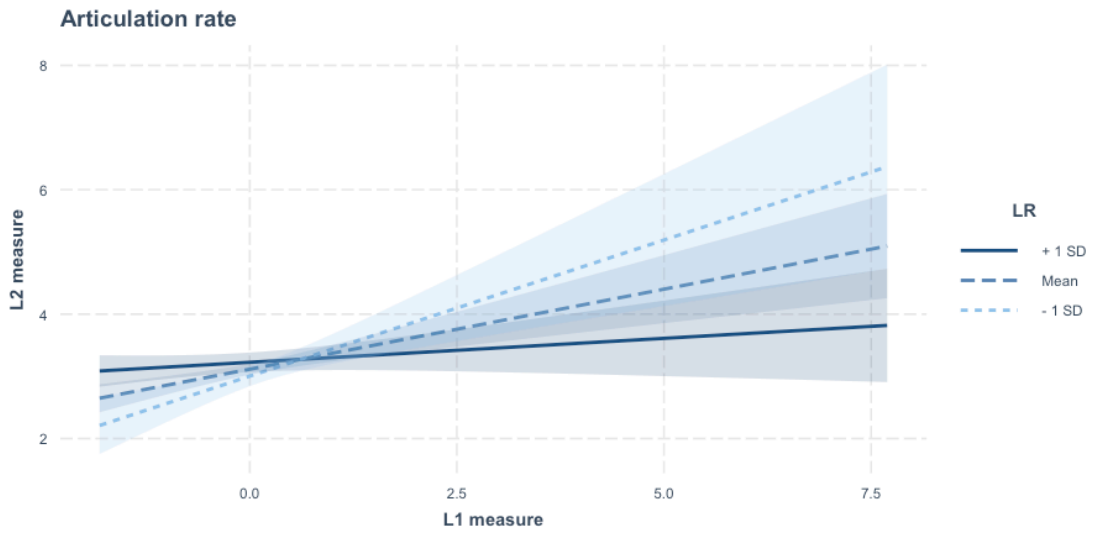


Figure 31. The interaction plot of the relationship between L1 and L2 articulation rate measures, separated by the score of L2 linguistic resource.

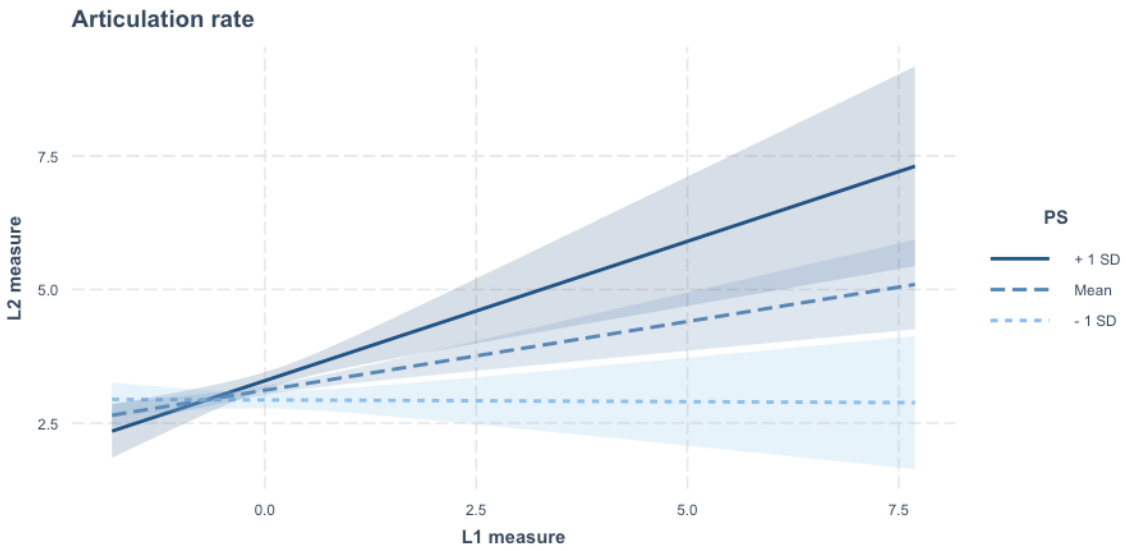


Figure 32. The interaction plot of the relationship between L1 and L2 articulation rate measures, separated by the score of L2 processing speed.

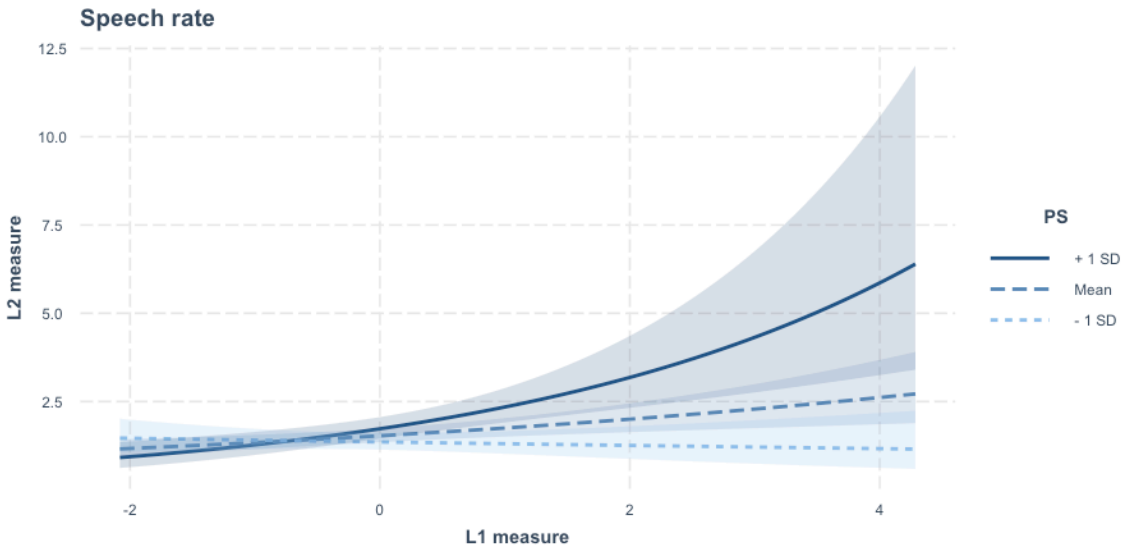


Figure 33. The interaction plot of the relationship between L1 and L2 speech rate measures, separated by the score of L2 processing speed.

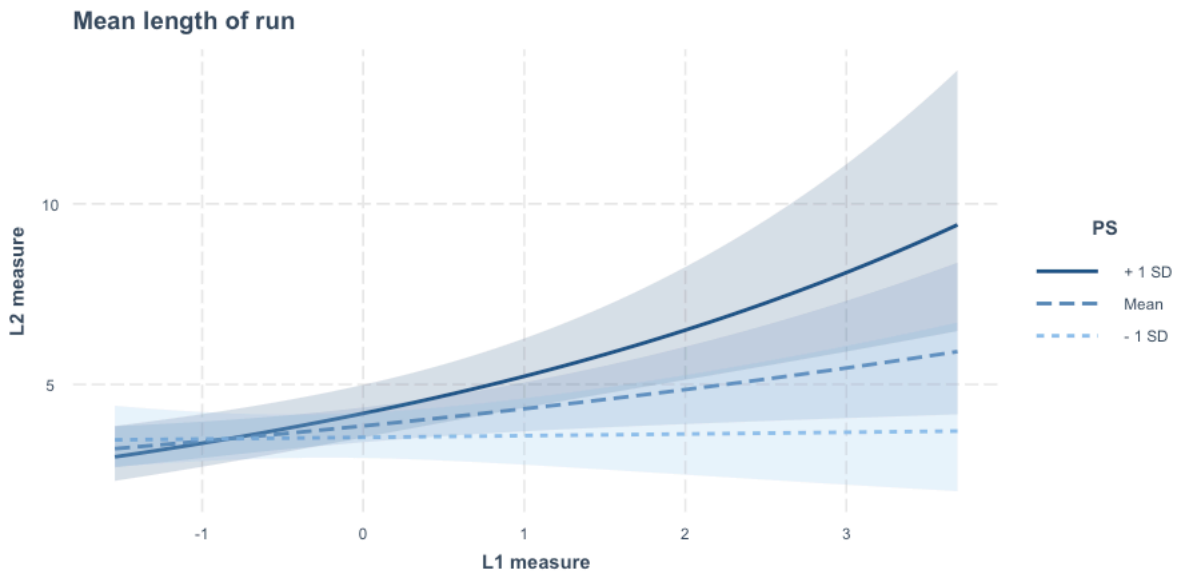


Figure 34. The interaction plot of the relationship between L1 and L2 mean length of run measures, separated by the score of L2 processing speed.

9.3 Discussion

Motivated by the lack of studies about the L1-L2 UF association with the pair of mora- and stress-timed languages, Study 4 investigated the extent to which L2 UF measures can be predicted form the corresponding L1 UF measures (RQ4-1), using L1 and L2 argumentative speech data produced by Japanese-speaking learners of English. A synthesis of prior research also suggested that the association between L1 and L2 UF measures can be moderated by the speaker's L2 proficiency (see Section 3.9). Thus, using the factor scores of CF from Study 3 as the measures of fluency-specific proficiency, Study 4 also examined whether L2 proficiency can significantly moderate the predictive power of L1 UF measures for L2 UF measures (RQ4-2). To this end, I constructed a set of GLMMs, which can take into account the non-normal distributions of L1 and L2 UF measures.

9.3.1 Predicting L2 utterance fluency from L1 utterance fluency

A set of GLMMs predicting L2 UF measures from the corresponding L1 UF measures showed that there were significant associations between L1 and L2 UF measures in all of the

UF measures covering speed, breakdown, and repair fluency as well as composite measures. In addition, the GLMMs suggested the small-to-medium effect sizes of the overall predictive power of L1 UF performance for the L2 counterparts in terms of marginal R^2 values (R^2 = .052–.247). Compared to previous studies on the L1-L2 UF link, the effect sizes in the current study were relatively lower than those previous studies, which reported moderate-to-strong effects sizes of correlation coefficients especially in speed and breakdown fluency (e.g., De Jong et al., 2015; Huensch & Tracy-Ventura, 2017; Peltonen, 2018). The coefficients of determination based on the correlation coefficients in Table 51 also suggested the similar range of the explained variance of L2 UF measures (r^2 = .013–.228 for the topic of Library; r^2 = .059–.251 for the topic of the Tokyo Olympics). Accordingly, the relatively small effect sizes in the current study may not have been caused by the differences in statistical analyses (e.g., gamma distributions, mixed-effects modelling). One plausible explanation for the relatively small effect sizes in the current study may lie in the cross-linguistic effects. As reviewed previously (see Section 3.9), the L1-L2 UF link has been mostly examined with L2 learners of English (stress-timed language) with the background of syllable- or stress-timed language as their L1. Considering the differences in syllable complexity and information density between English and Japanese, the Japanese language may differ from the English language in temporal and rhythmic aspects of speech more divergently than other stress- and syllable-timed languages (cf. Pellegrino et al., 2011). It can thus be assumed that due to such cross-linguistic divergences between English and Japanese, the participants in the current study may have been less likely to transfer temporal aspects of L1 Japanese to their L2 English speech production.

From the perspective of speech production mechanisms, language-general processing is fundamentally limited to conceptualization processes, while the processes of formulation and

articulation are considered language-specific (see Section 2.9). Meanwhile, particularly when the modules of L2-specific processing (e.g., syntactic encoding, phonological encoding) have not been established, L1 linguistic knowledge can also be transferred for L2 speech production (Kormos, 2006). However, the proficiency levels of participants in the current study mostly ranged from the B1 level to the C1 level on the CEFR scale (see Section 6.4). Considering the relatively wide range of proficiency levels, it can thus be assumed that there might be large variability in the degree of establishment of L2 linguistic knowledge among the participants. Therefore, conceptualization-related processing may be more likely to be captured by the covariance between L1 and L2 UF measures than L2-specific linguistic processing. Building on this assumption, the remaining part of this section discusses the current findings with regard to what aspects of L1 speech production may have been transferred in L2 speech production.

In Study 4, the medium effect sizes of the L1-L2 UF association were found in mean length or run, filled pause ratio, and self-repetition ratio. As regards mean length of run, each run is segmented by pauses. Therefore, especially for those who have attained highly automatized linguistic knowledge (e.g., L1 speakers), the length of run might be reflective of the unit of conceptual planning (Kormos, 2006; Levelt, 1989). It has also been suggested that L1 speakers produce pauses at the clause boundaries more frequently than in the middle of utterances (Skehan et al., 2016; Tavakoli, 2011), indicating that breakdowns in L1 speech production are mainly caused by content-related processing which end-clause pauses are supposed to reflect (De Jong, 2016b; Götz, 2013). Considering the language-general nature of conceptualization processes, the L1-L2 link in mean length of run may indicate the speakers' capacity and/or efficiency for content planning shared across L1 and L2 speech.

Meanwhile, previous studies have commonly reported the moderate-to-strong effect sizes of the L1-L2 UF link in filled pause frequency (De Jong et al., 2015; Duran-Karaoz & Tavakoli, 2020; Peltonen, 2018) and self-repetition frequency (De Jong et al., 2015; Huensch & Tracy-Ventura, 2017). It may thus be argued that these disfluency features are reflective of speaking style. The current results also confirmed the medium effect sizes of the L1-L2 UF link in these UF measures, despite the cross-linguistically divergent pair of L1 and L2 (i.e., Japanese-speaking learners of English). In addition, prior research suggested that filled pauses are associated with the demands on content planning (Fraundorf & Watson, 2014; Roberts & Kirsner, 2000; see also Section 7.3.1). Since the L1-L2 UF link can reflect speakers' language-general processing or idiosyncratic factors, the L1-L2 UF link in filled pause frequency may indicate the possibility that speakers may tend to elaborate on speech content similarly in both L1 and L2 speech production.

The current study also found a small effect size of the L1-L2 UF link in the measures of silent pause frequency and duration. The similar effect sizes in the measures of mid- and end-clause pauses suggest that pause location may not differentiate the effect size of the L1-L2 UF link. However, conceptualization-related processes, which end-clause pauses are supposed to reflect (De Jong, 2016b; Tavakoli, 2011), are shared across L1 and L2 speech production. It could thus be hypothesized that end-clause pause measures should have strong effect sizes of the L1-L2 UF link. Similarly, mid-clause pauses may have small or negligible effect sizes due to their association with language-specific linguistic processing. This interrelationship between pause location and the L1-L2 UF link was reported in Peltonen (2018) especially in pause duration measures. The relatively small effect sizes of the L1-L2 UF link in end-clause pause measures in Study 4 might be explained by the range of proficiency levels of the participants. Due to the wide range of proficiency levels (B1 to C1

level; see Section 6.4), there might have been variability among participants in the coverage of speech planning before starting to speak in L2 speech production (i.e. scope of planning; Gilbert et al., 2020). As pauses at clausal boundaries can be regarded as the starting point of planning for the subsequent unit of ideas or information (cf. Foster et al., 2000), end-clause pauses may be reflective of the scope of planning. Interestingly, a variety of individual difference factors associated with L2 proficiency (e.g., self-perceived proficiency, cumulative exposure to L2) can affect how far ahead L2 learners plan their utterances before speaking (see Gilbert et al., 2020). In other words, some students may have prepared not only for content planning but also for some linguistic planning (e.g., vocabulary and pronunciation) at clausal boundaries in L2 speech production. Consequently, such individual variability of L2-specific speech planning at end-clause pauses might have reduced the covariance of end-clause pause measures between L1 and L2 speech in the current study.

In contrast, the L1-L2 UF link in mid-clause pauses could have been expected to be small or even negligible. However, the current results showed that the measures of mid-clause pauses in L1 speech were weakly but significantly associated with the counterparts in L2 speech. This is possibly because of the relatively difficult or abstract topics of the argumentative tasks for both L1 and L2 speech production (see Sections 6.6, 6.9). As a result, even in the L1 argumentative task, infrequent vocabulary items may have been needed to complete the task. It can thus be assumed that such high demands on lexical retrieval in L1 speech might have contributed to the similarity between L1 and L2 speech production, that is, the covariance between the frequency and duration of mid-clause pauses between L1 and L2 speech.

Furthermore, the current results revealed that there were no meaningful L1-L2 UF associations in articulation rate, self-correction ratio, and false start ratio. Regarding

articulation rate, previous studies commonly reported the moderate-to-strong effect sizes of correlation coefficients between the measures in L1 and L2 speech (De Jong et al., 2015; De Jong & Mora, 2019; Huensch & Tracy-Ventura, 2017) as well as the cross-linguistic robustness (Bradlow et al., 2017). However, Study 4 found the negligible predictive power of L1 articulation rate for the L2 counterpart. One possible reason for this may lie in the pair of mora-timed L1 (Japanese) and stress-timed L2 (English) in Study 4. Note that Bradlow et al.'s (2017) cross-linguistic study did not include the speakers of L2 English with a mora-timed L1 background. Since the maintenance of isochronism in Japanese is achieved by every single mora rather than stressed syllables (cf. stress-timed language such as English), the cross-linguistic differences in rhythmic aspects are highly divergent between Japanese and English (e.g., vowel reduction ratio, the variability of syllable length; cf. Pellegrino et al., 2011; Vance, 2008), compared to the pair of syllable-based languages. It can thus be assumed that Japanese-speaking learners of English may have a limited range of rhythmic features that can be transferred from L1 Japanese speech to L2 English speech. Therefore, only the negligible effect size in the L1-L2 UF link might have been found in articulation rate.

Meanwhile, the negligible effect size of the L1-L2 UF link was also found in the frequency of self-corrections and false starts, both of which reflect the process of self-monitoring and repairing utterances (Kormos, 2006; Williams & Korko, 2019). Although the L1-L2 UF link in false starts has been rarely examined, previous studies have reported the significant predictive power of L1 self-correction behaviour for the L2 counterpart (De Jong et al., 2015; Huensch & Tracy-Ventura, 2017; Zuniga & Simard, 2019). Albeit speculative, one possible explanation for the inconsistent results between previous studies and the current study may lie in the cross-linguistic and/or cross-cultural differences in the norms for self-repairs in spontaneous speech (Tavakoli & Wright, 2020). The descriptive statistics of these two self-

repair measures indicated that the participants in Study 4 produced a limited number of self-corrections and false starts (see Section 9.2.1), compared to syllable-based languages in previous studies (De Jong et al., 2015; Huensch & Tracy-Ventura, 2017). It can thus be suggested that self-repair itself should be avoided in Japanese monologues. However, only a few studies have investigated the cross-linguistic and cross-cultural differences in the norm of temporal features in speech (Paulston et al., 2012 for silent pauses; Tian et al., 2017 for filled pauses). Therefore, future studies are needed to examine the cross-linguistic and cross-cultural influences on the L1-L2 UF link, especially in repair fluency features.

9.3.2 Role of L2 proficiency in the association between L1 and L2 utterance fluency

In addition to cross-linguistic effects, L2 proficiency has been regarded as another important factor for the L1-L2 UF link. Previous studies operationalized L2 proficiency in a variety of manners, such as longitudinal changes (Huensch & Tracy-Ventura, 2017) and vocabulary size (De Jong & Mora, 2019; Peltonen, 2018). In Study 4, for the sake of the compatibility with UF performance, two factor scores of CF derived from Study 3—linguistic resource and processing speed—were adopted. A set of GLMMs predicting L2 UF from the corresponding L1 UF measures, these two CF scores, and their interactions with L1 UF measures suggested that the L1-L2 UF link can be moderated by the CF scores only in speed fluency measures. More specifically, the L1-L2 link in articulation rate was weakened as a function of the score of linguistic resource. In other words, for those who acquired a wider range of L2 linguistic resources, L2 articulation rate tended to be relatively independent of L1 articulation rate (see Figure 31). Note that despite the marginally significant level, the score of linguistic resource positively contributed to articulation rate ($\beta = 0.122$, $p = .099$; see Appendix R). The GLMM of articulation rate indicates that the higher linguistic resource score the speaker has, the more fluent their L2 speech tend to be. Therefore, as a function of L2 linguistic resources, students'

L2 articulation rate was enhanced and also was simultaneously dissociated from L1 articulation rate. This finding may be explained by the interplay between the nature of articulation rate measure and L2 proficiency. First of all, articulation rate is assumed to capture the overall efficiency of speech production (cf. Kormos, 2006; Segalowitz, 2010), and thus articulation rate may be reflective of the efficiency of different speech production processes. The dissociative pattern of the L1-L2 UF link in articulation rate as a function of linguistic resources indicates that among learners with limited linguistic resources, relatively large variance of L2 articulation rate is explained by its L1 counterpart. In other words, language-general processes and/or idiosyncratic factors may contribute to the overall efficiency of L2 speech production to a larger extent for learners with limited linguistic resources than for those with extensive linguistic resources. It can thus be hypothesized that the increase in linguistic resources may not only speed-up the overall speech production process but also modify the underlying speech processing mechanisms. Due to the holistic nature of articulation rate, one can only speculate what linguistic knowledge or processing skill components contribute to these potential modifications in speech processing. However, such potential changes in speech processing may include the shift of grammatical knowledge from declarative knowledge to procedural knowledge (DeKeyser, 2015) and the changes in the structure of the mental lexicon (e.g., Revised hierarchical model; French & Jacquet, 2004; Pavlenko, 2009).

In contrast, the L1-L2 link in articulation rate, speech rate, and mean length of run was enhanced as a function of L2 processing speed (see Figure 32–34). This consistent pattern in speed fluency measures suggests that the more efficiently learners can process L2 knowledge, the closer to L1 their L2 speed fluency is. This finding is in line with Huensch and Tracy-Ventura's (2017) and Peltonen's (2018) results. Considering that the measures of speed

fluency may capture the overall efficiency of speech production (cf. Kormos, 2006; Segalowitz, 2010) and that in Study 4, students' speech was more fluent in L1 than in L2 (see Section 9.2.2), the results here indicated that the efficient L2 processing skills may allow L2 speakers to speak in L2 as fluently as they do in their L1. Note that the direction of moderator effects of L2 proficiency scores to the L1-L2 UF link was opposite between linguistic resource and processing speed. These opposite pattens of the role of L2 proficiency in the L1-L2 UF link may expand the understanding of the interplay between cross-linguistic differences and L2 proficiency in the L1-L2 UF link (Derwing et al., 2009; Huensch & Tracy-Ventura, 2017). The current findings showed that in the case of learners with cross-linguistically divergent pair of L1 and L2, the relationship between the L1-L2 UF link and L2 proficiency can vary, depending on the aspects of L2 proficiency. The results here may also confirm the two-dimensionality of CF from the perspective of the L1-L2 UF link (see Section 8.3.1).

In contrast to speed fluency measures, the GLMM did not find significant moderator effects of L2 proficiency (indexed by CF scores) on the L1-L2 UF link in any of the breakdown and repair fluency measures. Hence, the current results indicated that L2 breakdown and repair fluency measures are associated with the L1 counterparts, while those associations tend to independent of L2-specific competence. This view of the L1-L2 UF link may support the use of L1-corrected L2 UF measures of breakdown and repair fluency for a more valid assessment of L2 oral proficiency (De Jong et al., 2015). However, it is worth emphasizing that different weights of correction might be needed for speed fluency measures across proficiency levels, as suggested by the interaction effects by L1 UF measures and L2 proficiency scores.

9.4 Summary

This chapter reported the results of Study 4 which addressed the association between L1 and L2 UF performance, with respect to the moderator effects of L2 proficiency. Motivated by the lack of studies on the L1-L2 UF link in the pair of mora-timed and stress-timed language, Study 4 was conducted using L1 and L2 argumentative speech data elicited from Japanese-speaking learners of English (n = 104).

As a preliminary analysis, participants' UF performance was compared between L1 and L2 speech. The GLMMs confirmed that in all of the UF measures, L1 speech was more fluent than L2 speech. Using another set of GLMMs, L2 UF measures were predicted from the corresponding L1 UF measures, while controlling for the variability of the L1 effects across participants and topics of the L2 argumentative tasks (RQ4-1). The results revealed that all the L1 UF measures were found as a significant predictor of the corresponding L2 UF measures, with the overall small-to-medium effect sizes in terms of the variance explained (Marginal R^2 = 5.2–24.7%). Compared to previous studies, these effect sizes of the L1-L2 UF link were relatively small, subsequently suggesting that the L1-L2 UF link tends to be weak in the pair of cross-linguistically divergent L1 and L2 (i.e., Japanese-speaking learners of English). Furthermore, the medium effect sizes of the predictive power of L1 UF for L2 UF were found in mean length of run, filled pause ratio, and self-repetition ratio. Building on the theoretical assumptions of speech production, these findings suggested that the L1-L2 associations may be reflective of speakers' capacity and/or efficiency in content planning and speaking style. The findings of RQ4-1 are summarized in Table 54.

Table 54. *Summary of findings of Study 4.*

	Medium effect size	Small effect size	Negligible effect size
Speed fluency	Mean length of run	Speech rate	Articulation rate
Breakdown fluency (frequency)	Filled pause ratio	Mid-clause pause ratio End-clause pause ratio	–
Breakdown (duration)	–	Mid-clause pause duration End-clause pause duration	–
Repair fluency	Self-repetition ratio*	–	Self-correction ratio False start ratio

Note. The effect size of self-repetition ratio was small but substantially medium (R^2 = 17.1).

In addition, the moderator effects of L2 proficiency on the L1-L2 UF link (RQ4-2) were also examined by adding the interaction terms between L1 UF and two factor scores of CF based on Study 3—linguistic resource and processing speed—to the GLMMs for RQ4-1. These revised GLMMs suggested that the L1-L2 UF association was moderated by L2 proficiency only in the dimension of speed fluency. More specifically, the strength of the L1-L2 UF association in articulation rate was dissociated as a function of L2 linguistic resources. This finding confirmed Derwing et al.'s (2009) claim that in the case of learners with a cross-linguistically divergent pair of L1 and L2, the negative transfer from L1 speech production to the L2 counterpart tends to be reduced with an increase in L2 proficiency. In contrast, the L1-L2 association in articulation rate, speech rate, and mean length of run was enhanced as a function of L2 processing speed. Considering that their speech was more fluent in L1 than in L2, the results indicated that the enhancement of L2 processing speed (i.e., automatization of L2 knowledge) may allow L2 speakers to speak in L2 as fluently as they do in their L1. These opposite directions of the role of L2 proficiency in the L1-L2 UF link suggested that particularly in the case of learners with a cross-linguistically divergent L1-L2 pair, the

relationship between the L1-L2 UF link and L2 proficiency can vary, depending on the aspects of L2 proficiency.

Chapter 10: Conclusions[15]

10.1 Introduction

In this chapter, I answer the research questions of Studies 1–4 of the thesis, drawing on the main findings of the studies (Section 10.2). Next, I discuss the theoretical and methodological contributions of these findings to L2 fluency research (Section 10.3), followed by the pedagogical implications for L2 assessment, learning, and teaching (Section 10.4). I then conclude this chapter, as well as this thesis, by reporting methodological limitations (Section 10.5) and suggesting future directions for L2 fluency research (Section 10.6).

10.2 Main Findings

10.2.1 L2 utterance-perceived fluency link

RQ1-1. What is the overall relationship between perceived fluency and subdimensions of utterance fluency–speed, breakdown, and repair fluency–as well as composite measures?

Study 1 aggregated the effect sizes of correlation coefficients between PF judgements and six UF measures—articulation rate, pause frequency, pause duration, disfluency rate, mean length of run, and speech rate—from primary studies (k = 22). The results showed that articulation rate (speed fluency; r = .62) and pause frequency (breakdown fluency; r = -.59) were strongly associated with PF judgements. Meanwhile, mean pause duration was related to PF judgements with a moderate effect size (breakdown fluency; r = -.46), while disfluency rate was linked to PF scores with a small but significant effect size (repair fluency; r = -.20). In addition, the composite measures of UF (mean length of run, speech rate) showed strong effect sizes of correlation coefficients with PF ratings (r = .72 and r = .76, respectively). These results confirmed that the association strengths of breakdown fluency with PF

[15] Several sections of this chapter were accepted for publication in *The Modern Language Journal* as Suzuki, Kormos and Uchihara (in press, 2021).

judgements tend to be as strong as those of speed fluency when the frequency of pauses is measured, but to be weaker than those of speed fluency when the duration of pauses is targeted. The aggregated effect sizes also confirmed the significant association between repair fluency and PF judgements, despite the small effect size.

RQ1-2. To what extent does the relationship between perceived fluency and utterance fluency vary, according to methodological factors in different phases of L2 perceived fluency research—speech stimuli preparation, rater recruitment, rating procedure, and selection of utterance fluency measures?

Motivated by the possibility that methodological variables affect the UF-PF link, Study 1 also conducted moderator analyses, using a set of methodological factors related to four major phases of L2 PF research—speech stimulus preparation, listeners' background, rating procedure, and UF measure computation. As for speech stimulus, the results showed that the effect sizes of correlation coefficients tend to be higher in L2 Japanese (r = .77) than in L2 Dutch, English, and French (r = .52–.61). In addition, the effect sizes were also likely to be higher with controlled speech samples (e.g., read-aloud speech; r = .74) than with spontaneous speech samples (r = .53 for closed tasks; r = .51 for open tasks). Finally, the effect sizes with the speech samples of entire speech (r = .59) were, albeit at the marginally significant level (p = .076), slightly higher than those with the speech samples of short excerpts (r = .50). Therefore, the UF-PF link may be moderated by how speech stimuli are prepared.

Regarding listeners' background, the results did not show significant moderator effects of relevant experience (Experienced vs. Inexperienced raters) and language background (L1 vs. L2 raters). However, the groups of experienced raters tended to show slightly higher effect

sizes (r = .58), compared to those of inexperienced raters (r = .51). Meanwhile, comparing the range of 95% confidence intervals of the subgroups, L1 raters (r = .55, CI[.50, .60]) indicated a narrower range of confidence intervals than L2 raters (r = .48, CI[.29, .64]). Although the current moderator analyses may suggest that the UF-PF link overall tends to be consistent across different groups of listeners, listeners' background may potentially affect the UF-PF link.

With regard to rating procedures, the moderator analyses demonstrated that the UF-PF link tends to be higher when judged by research-based rubrics (r = .67) than by a simple semantic scale with a research-based definition of fluency (r = .51). Although the other moderator variables—the number of scale points and rater training—failed to reach statistical significance, PF judgements with extensive training (r = .66) seemed to show a slightly higher correlation coefficient than those with short practice (r = .54). These results indicate that the UF-PF link might be moderated by how listeners selectively pay attention to speech characteristics.

Finally, the moderator analyses of UF measure computation showed several findings. First, there was no significant difference in the effect sizes between manual and automatic coding of temporal features. Second, regarding the minimum length of pause length, despite the lack of overall significant effects, there seemed to be a substantial difference between the minimum thresholds of 250 ms (r = -.60) and that of 200 ms (r = -.41) in pause duration measures. Third, pause location differentiated the effect sizes of correlation coefficients between breakdown fluency measures and PF scores. More specifically, measures based on mid-clause pauses showed strong effect sizes in pause frequency and pause duration (r = -.72 and r = -.71, respectively), which were as strong as the composite measures (i.e., mean length

of run, speech rate). Fourth, pause frequency measures based on silent pauses (r = -.57) showed higher effect sizes than those based on filled pauses (r = -.24). Lastly, the aggregated effect sizes of repair fluency measures were significant only when different types of disfluency features were combined (r = -.41 CI[-.33, -.10]), while the effect sizes of fine-grained measures focusing on one particular type of disfluency features (self-repetition, self-correction) did not reach statistical significance. Taken together, the effect sizes of correlation coefficients between PF and UF measures may not differ according to annotation methods (manual vs. automatic coding) but may be affected by the specification of breakdown and repair fluency measures—including pause location, pause type, the threshold for silent pauses, and target disfluency features.

10.2.2 Effects of speech processing demands on utterance fluency performance

RQ2. How does L2 utterance fluency performance vary across four types of speaking tasks which differ in the speech processing demands on conceptualization, the activation of linguistic representations, and the activation of phonological representations?

Study 2 built a set of GLMMs to test whether UF performance differed between three contrasts of speech processing demands, operationalized by task design features. The first contrast—the argumentative and picture narrative tasks—targeted the speech processing demands on conceptualization (i.e., content generation). The result showed that the enhanced conceptualizing demands increased the frequency of filled pauses, confirming the high demands on content and discourse planning in the argumentative task (Fraundorf & Watson, 2014; Greene & Cappella, 1986; Roberts & Kirsner, 2000). However, despite the enhanced conceptualizing demands, students' speech was more fluent in terms of several UF measures (e.g., articulation rate, mid-clause pause ratio, end-clause pause ratio, self-repetition ratio) in the argumentative task than in the picture narrative task. Contrary to the expected benefits of

predefined speech content in fluency (e.g., Préfontaine & Kormos, 2015; Skehan, 2009; Tavakoli & Skehan, 2005), Study 2 showed that the predefined speech content may not allow students to avoid some difficult vocabulary and syntactic structures, which leads to high demands on formulation, particularly for the current participants who are at the relatively lower proficiency levels compared to previous studies. The results suggested that speech processing demands might be subject to learners' L2 proficiency as well as task design features.

The second contrast focused on the enhanced activation of linguistic representations, comparing the picture narrative task and the text summary task without RAA (i.e., the absence vs. presence of source texts in L2). The results showed that the enhanced activation of linguistic representations led to a slow articulation rate and an increase in mid-clause pauses. Similarly, the effects of the enhanced activation of phonological representations on UF performance was examined by the contrast between two conditions of RAA in the text summary tasks (i.e., the reading-only vs. reading-while-listening modes of source texts). The results showed that mid-clause pauses were longer in the with-RAA condition than in the without-RAA condition. These findings consistently demonstrated that the enhanced activation of relevant linguistic items may lower the efficiency of speech production, against the expected benefits based on the activation spreading theories.

10.2.3 Dimensionality of cognitive fluency and utterance fluency

RQ3-1a. What is the relationship between cognitive fluency measures of lexical, grammatical, and pronunciation knowledge?

Motivated by the lack of evidence regarding the dimensionality of CF, Study 3 conducted CFA and identified the two-factor model of CF, which consisted of the latent variables of linguistic resource and processing speed. These two latent variables were closely related to

each other (r = .676), indicating the interdependence of these two dimensions of CF. The results suggested that CF encompasses not only the speed dimensions of linguistic knowledge but also the resource dimensions of linguistic knowledge. A close examination of the regression coefficients from the observed variables to the latent variables indicated that the primary component of linguistic resource of CF is learners' vocabulary size and the secondary one is their syntactic repertoires. Meanwhile, the primary component of processing speed of CF was the speed of syntactic encoding, while the speed of morphological processing and articulatory gestures may play a secondary role in CF. Furthermore, the speed dimension of lexical knowledge in processing speed (i.e., lexical retrieval speed) was found to be a significant but relatively peripheral component underlying fluent speech production in L2.

RQ3-1b. What is the relationship between utterance fluency measures of speed, breakdown, and repair fluency?

As with RQ3-1a, Study 3 proposed several CFA models for the dimensionality of UF and identified the three-factor model, which consisted of three latent variables—speed fluency, breakdown fluency, and repair fluency, supporting the Tavakoli and Skehan's (2005) triad model of UF consistently across four different speaking tasks. However, it should be noted that the latent variables of speed fluency and breakdown fluency were closely related to each other (r = .929–.960), calling for caution in the theoretical distinctiveness between these two latent variables.

In the final CFA model, the latent variable of speed fluency was composed of the observed variables of articulation rate and mean length of run. In accordance with this theoretical assumption, in the final SEM model, the measure of articulation rate tended to contribute to the latent variable of speed fluency to a slightly larger extent than that of mean

length of run, suggesting that articulation rate is a representative measure of speed fluency. The latent variable of breakdown fluency was composed of mid-clause pause ratio, end-clause pause ratio, mean pause duration (counting both mid- and end-clause pauses), and filled pause ratio. The regression coefficients of mid-clause pause ratio to the latent variable were significantly higher than the other observed variables, indicating that the measure of mid-clause pause ratio is the representative measure of breakdown fluency. The third latent variable, repair fluency, was constructed with the observed variables of self-repetition ratio, self-correction ratio, and false start ratio. The regression coefficients of self-repetition ratio were generally higher than those of false start and self-correction ratio. Accordingly, the primary component of repair fluency can be regarded as the frequency of self-repetitions, while both self-corrections and false starts are of secondary importance.

10.2.4 L2 cognitive-utterance fluency link

RQ3-1 and RQ3-2. To what extent do components of cognitive fluency contribute to subdimensions of utterance fluency, and to what extent is the CF-UF link (RQ3-1) moderated by speech processing demands of speaking tasks?

To address RQ3-1 and RQ3-2, the SEM models were constructed based on the final CFA models of CF and UF, separately for four speaking tasks. In other words, the latent variables of speed fluency, breakdown fluency, and repair fluency were predicted from those of linguistic resource and processing speed. RQ3-1 queried whether the regression paths from the latent variables of CF to those of UF were significant, while RQ3-2 was concerned with the extent to which the strengths of the regression paths varied across speaking tasks. The latent variable of speed fluency was associated with that of processing speed consistently across speaking tasks. Meanwhile, the latent variable of linguistic resource significantly contributed to that of speed fluency only in the two conditions of the text summary tasks. The

latent variable of breakdown fluency was overall associated with both of the latent variables of linguistic resource and processing speed consistently across speaking tasks. Although the latent variable of breakdown fluency seemed to show slightly stronger associations with processing speed (β = .376–.502) than with linguistic resource (β = .221–.345), the difference in the regression coefficients did not reach statistical significance. The current results suggested that breakdown fluency may be underpinned by both dimensions of CF, showing the difference between the constructs of speed fluency and breakdown fluency in their underlying components of CF. The latent variable of repair fluency was associated with linguistic resource in the picture narrative task and two conditions of the text summary tasks, but not in the argumentative task. The processing speed of CF did not significantly contribute to the latent variable of repair fluency in any of the speaking tasks. These findings revealed that the processing speed of CF showed a consistent pattern of contributions to UF across speaking tasks, whereas the role of linguistic resource of CF in UF (especially, speed fluency and repair fluency) may vary, depending on task characteristics.

10.2.5 L1-L2 utterance fluency link

RQ4-1. To what extent are L2 utterance fluency measures predicted from the corresponding L1 utterance fluency measures?

Using L1 and L2 speech elicited from the argumentative tasks, a set of GLMMs was constructed to predict L2 UF measures from the corresponding L1 UF measures. The results showed that there were significant associations between L1 and L2 UF measures in all UF measures covering speed, breakdown, and repair fluency and composite measures. The effect sizes were calculated as the value of marginal R^2, that is, the variance of L2 UF measures explained by the L1 counterparts while controlling for the random variance by individual speakers and prompts. Study 4 revealed the overall small-to-medium effect sizes of L1-L2

UF link based on marginal R^2 values (R^2 = .052–.247). More specifically, the medium effect sizes of L1-L2 UF association were found in mean length or run, filled pause ratio, and self-repetition ratio, while the small effect sizes were observed in mid-clause pause ratio, mid-clause pause duration, end-clause pause ratio, and end-clause pause duration. Furthermore, the current results revealed that there were no meaningful L1-L2 UF associations in articulation rate, self-correction ratio, and false start ratio.

RQ4-2. To what extent are the L1-L2 fluency links of different aspects of utterance fluency (RQ4-1) moderated by L2 proficiency?

To examine the moderator effects of L2 proficiency on the L1-L2 UF link, Study 4 used two factor scores of CF derived from Study 3—linguistic resource and processing speed—as a proxy for fluency-related L2 proficiency. These two CF scores and their interactions with L1 UF measures were added to the GLMMs constructed for RQ4-1. The results suggested that the L1-L2 UF link can be moderated by CF scores only in speed fluency measures. The L1-L2 association in articulation rate was weakened as a function of the score of linguistic resource. Meanwhile, the L1-L2 association in articulation rate, speech rate, and mean length of run was enhanced as a function of L2 processing speed.

10.3 Theoretical and Methodological Contributions

This section describes how Studies 1–4 have addressed theoretical issues and methodological challenges in L2 fluency research, particularly with regard to the construct definition of CF, UF, and PF and their interrelationship.

10.3.1 Contributions to L2 utterance-perceived fluency link research

Using meta-analytic techniques, Study 1 addressed two major inconsistent findings in previous studies about the UF-PF link: (a) the relative strengths of predicting power of speed and breakdown fluency for PF and (b) the significance of the contribution of repair fluency into PF. Study 1 revealed that the predictive power of breakdown fluency for the UF-PF link may vary due to the multidimensionality of pausing behaviour (frequency, duration, and location). The large effect sizes of mid-clause pause measures (r = |.71–.72|) may indicate that listeners are sensitive to speakers' breakdowns caused by linguistic processing problems (Kahng, 2018; Saito et al., 2018; S. Suzuki & Kormos, 2020). This might be due to the fact that mid-clause pauses reflect the disruptions in speech processing arising from the linguistic problems such as lexical retrieval and sentence construction difficulties (De Jong, 2016b; Skehan et al., 2016; Tavakoli, 2011). Moreover, the varying predictive power of repair fluency measures across measurements also suggests that the frequency of one specific type of disfluency feature might not be sufficient to negatively impact listeners' perceptions. Although different monitoring processes are supposed to underlie different types of disfluency features (Kormos, 1999a, 2006; Williams & Korko, 2019), listeners may take into account the frequency of disfluencies regardless of the underlying monitoring processes.

The moderator analyses showed that the UF-PF link tends to be moderated by two major categories of methodological factors: (a) how speech samples are prepared for listeners' judgements (target L2, task type, length of speech stimuli) and (b) how listeners' attention is directed (listeners' experience, rater training, the definition of fluency presented to raters). The former category may indicate that PF, that is, listeners' inference about the speaker's CF through their perceptions of UF (Segalowitz, 2010), is subject to the rhythmic/temporal norm of the target language, the components of L2 competence elicited by speaking prompts, and

the amount of exposure to the speech samples. Meanwhile, the latter category may add another piece of evidence for the assumption that listener-based judgements of fluency (i.e., PF) are established by how listeners selectively pay attention to speech characteristics (Segalowitz, 2010).

Relating to the abovementioned theoretical contributions, Study 1 also provides several methodological contributions to the research into the L2 UF-PF link. First, there were no substantive differences in the effect size between manual and automated annotation methods for temporal features. It may thus be plausible to use automated annotation methods to calculate UF measures or even to check the reliability of manual annotation. Second, comparing L1 and L2 listeners for PF judgements, the non-significant difference in the effect size between these two groups of raters suggested that both L1 and L2 raters tend to behave similarly in judging L2 speakers' fluency. However, the effect size of L2 raters showed a wider confidence interval, indicating the relative instability in the association with temporal characteristics of speech (i.e., UF measures). This is in line with the previous finding that L2 listeners' PF judgements tend to vary, according to their personal experience of L2 learning beyond the speech characteristics (Magne et al., 2019). Therefore, to achieve appropriate level of inter-rater reliability among L2 raters, a larger number of raters might be needed, and researchers may use statistical analyses to control for listeners' variability in PF rating behaviour (e.g., Rasch modelling). Finally, the results of Study 1 clearly showed that the predicting power of breakdown fluency measures for PF judgements varies, according to the dimensions of pauses that the measures capture (frequency, duration, location, and type). The multidimensional nature of pausing behaviour should be taken into account when selecting breakdown fluency measures in fluency research. The results also suggested that the

measures based on mid-clause pauses should be included due to their strong associations with PF ratings.

10.3.2 Contributions to L2 cognitive-utterance fluency link research

To clarify the CF-UF link, three separate studies (Studies 2–4) were conducted with different methodological approaches. In accordance with Segalowitz's (2010, 2016) claim that CF is theoretically underpinned by L2 speech production mechanisms, Study 2 examined which components of speech production are related to different FU measures, manipulating speech processing demands on three components of speech production—conceptualization, linguistic representations, and phonological representations. Following another claim by Segalowitz (2010, 2016) that the validity of UF measures can be evaluated in relation to the association with L2-specific CF as opposed to language-general processes, Study 3 investigated which components of CF contribute to the dimensions of UF at the level of constructs, while Study 4 examined the extent to which UF measures can be explained by language-general processing, operationalized as the covariance between L1 and L2 UF measures.

10.3.2.1 Speech processing demands

In Study 2, the speaking task with high conceptualizing demands was operationalized as the argumentative task and was contrasted with the picture narrative task concerning the different degree of content generation. The results of students' perception of task demands and the increase of filled pauses in the argumentative task confirmed the higher conceptualizing demands in the argumentative task than in the picture narrative task. However, the results suggested that students' speech was more fluent in the argumentative task than in the picture narrative task in all the three dimensions of UF performance (speed, breakdown, and repair fluency). This unexpected pattern of UF performance indicated that in open tasks, the

necessity for content generation can lead to high demands on conceptualization but may simultaneously allow students to avoid using some difficult or unfamiliar linguistic items (Préfontaine & Kormos, 2015). In contrast, closed tasks, such as the current picture narrative task, require students to express some key information, even if difficult or unfamiliar linguistic items are needed. Compared to previous studies reporting the beneficial effects of predefined speech content (i.e., closed tasks) on UF performance (e.g., Préfontaine & Kormos, 2015), the lower UF performance in the picture narrative task in Study 2 suggests that the effects of the predefined content on L2 fluency performance might be subject to students' proficiency levels. Similarly, the contrasts for the activation levels of linguistic and phonological representations showed that the enhanced activation of task-relevant linguistic items resulted in lower fluency performance particularly in articulation rate, mid-clause pause ratio, and mid-clause pause duration, against the expected facilitative effects. These results indicated that students may have experienced the competition for selection between their resources available for productive use and the externally activated items. However, despite the opposite direction, these results showed that formulation processing might be related to articulation rate and mid-clause pause measures.

10.3.2.2 L2 linguistic knowledge underlying UF performance

Prior to the CF-UF link at the level of constructs, Study 3 examined the dimensionality of CF and UF. The two-factor solution for the CFA model of CF (linguistic resource, processing speed) indicated that the construct of CF is not unitary. This finding also supports the broad definition of CF proposed based on L2 speech production mechanisms rather than Segalowitz's (2016) narrow definition (see Section 3.6). To the best of my knowledge, Study 3 provides the first empirical evidence regarding the dimensionality of CF. From a methodological perspective, the current result also supports the existing methodological

practice of selecting CF measures in previous studies (De Jong et al., 2013; Kahng, 2020). According to the measurement models of each latent variable, the most representative component of linguistic resource was vocabulary size (indexed by the PVLT), while that of processing speed was sentence construction speed (indexed by RT scores in the maze task). Considering that the availability of linguistic resources is the prerequisite of processing speed, the current factor structure of CF may partially add evidence for the lexically-driven nature of L2 speech production (Kormos, 2006). Meanwhile, the final CFA model of UF supported the three-factor model consisting of speed fluency, breakdown fluency, and repair fluency, originally proposed by Tavakoli and Skehan (2005). Their original triad model of UF was based on the speech data elicited from different picture narrative tasks. However, Study 3 confirmed the feasibility of the triad model of UF across four different speaking tasks (see Section 8.2.4). As with the CFA model of CF, the representative component of each dimension of UF was identified in terms of their regression coefficients. The most representative measure of speed fluency was articulation rate, while that of breakdown fluency was mid-clause pause ratio. Regarding repair fluency, the measure of self-repetition ratio was identified as the most representative measure. Especially in L2 fluency research, researchers are required to select a set of valid measures from a large number of existing UF measures. Those representative measures of each subconstruct of UF may provide insights into the baseline for measure selection in future L2 fluency studies.

The SEM analysis revealed the complex interplay between the multidimensionality of CF and UF and speaking task type. Speed fluency was primarily associated with processing speed, while linguistic resource can play a role only when relevant linguistic items were activated by the task input (i.e., the text summary tasks). Meanwhile, both linguistic resource and processing speed contributed to breakdown fluency consistently across speaking tasks. This

result may suggest that the construct of breakdown fluency may capture L2-specific competence in a comprehensive manner. Finally, repair fluency was related to linguistic resource, only when the content of speech was predefined (i.e., the picture narrative and text summary tasks), and was consistently independent of processing speed. In other words, in open tasks (here, the argumentative task), repair fluency may not tap into L2-specific competence. However, to further clarify the contribution of L2 competence to repair fluency, new repair fluency measures with high sensitivity to underlying L2-specific processing (e.g., different loops of self-monitoring; Levelt, 1983, 1989) may be needed. These results confirmed that the processing speed of CF showed a consistent pattern of the contributions to UF across speaking task types, whereas the role of linguistic resource of CF in UF may tend to vary, depending on task characteristics, such as the availability of relevant linguistic items and the predefined content of speech.

10.3.2.3 Language-general processes reflected in L2 UF measures

Study 4 investigated the extent to which UF can be explained by language-general processes and factors, operationalized as the covariance between L1 and L2 UF measures. The review of literature in the L1-L2 UF link suggested that no studies have examined the L1-L2 link with L1 Japanese-speaking learners of English. The results showed that there were significant associations between L1 and L2 UF measures in all UF measures. However, compared to previous studies, the effect sizes in Study 4 were relatively small, indicating that the divergent crosslinguistic differences in phonological aspects between L1 and L2 may result in the reduced opportunity for linguistic transfer between individuals' L1 and L2 speech production. Relatively strong effect sizes were found in mean length of run, filled pause ratio, and self-repetition ratio. Building on the assumption that the L1-L2 UF link is reflective of language-general processes and factors, those UF measures may capture speakers' general

cognitive capacity and/or idiosyncratic factors such as personal speaking style. Moreover, the results suggested that the moderator effects of L2 proficiency were only found in the measures of speed fluency. Interestingly, the direction of moderating effects of L2 proficiency scores on the L1-L2 UF link was opposite between linguistic resource and processing speed. The strength of the L1-L2 UF association in articulation rate was dissociated as a function of L2 linguistic resources. This finding may partly confirm Derwing et al.'s (2009) claim that in the case of learners with a cross-linguistically divergent pair of L1 and L2, the negative transfer from L1 speech production to the L2 counterpart tends to be reduced with an increase in L2 proficiency. In contrast, the L1-L2 association in articulation rate, speech rate, and mean length of run was enhanced as a function of L2 processing speed, indicating that with efficient L2 processing skills, L2 speakers may tend to produce L2 speech as fluently as they produce L1 speech. These opposite patterns of the role of L2 proficiency in the L1-L2 UF link may expand the understanding of the complex interplay between cross-linguistic differences and L2 proficiency in the L1-L2 UF link (Derwing et al., 2009; Huensch & Tracy-Ventura, 2017).

10.4 Pedagogical Implications

10.4.1 Implications for L2 speaking assessment

In L2 fluency research, PF judgements are conceptualized as listeners' intuitive judgements of speakers' CF, which is substantially equivalent to L2 oral competence (Segalowitz, 2010). However, it is possible that such impressionistic ratings may fail to capture inconspicuous features that reflect some aspects of L2 competence or that raters might be biased by features that are related to idiosyncratic factors, such as personal speaking style. The current studies jointly indicated such a mismatch between listener-based judgements and L2 linguistic knowledge measures. For instance, Study 1 showed the equal importance of speed fluency

and breakdown fluency (pause frequency) to the association with PF judgements, whereas Study 3 suggested that breakdown fluency may cover components of L2 competence (L2-specific CF) more consistently and comprehensively than speed fluency. Integrating the findings from Studies 1–4, this section provides several pedagogical implications for the research-informed practice of L2 assessment, especially in the domain of assessment rubrics, rater training, and automated scoring systems.

Study 3 demonstrated the variability of underlying components of speed fluency and repair fluency across tasks, indicating that the construct validity of these two dimensions of fluency can be different across tasks. This may encourage language testers to examine how L2 competence is reflected in oral fluency performance for each task type. If some variability of the construct of fluency is found across tasks in the language tests, different rubrics for each task type should be prepared (e.g., independent and integrated speaking tasks in TOEFL iBT). Meanwhile, despite the stable construct of breakdown fluency across task types, the validity of UF measures was found to change due to the multidimensional nature of pausing behaviour. For instance, mid-clause pause ratio may have a strong predictive power for PF judgements (Study 1) and also reflect L2-specific competence consistently across tasks (Study 3). Meanwhile, filled pause ratio may be related to content-related engagement (Study 2) as well as language-general idiosyncratic factors (Study 4). In rater training and rubrics for fluency assessment, examiners might thus be instructed or trained to be aware of the multidimensional nature of pauses for a valid assessment of fluency.

In addition to the construct validity of UF measures, Study 1 has several implications for language assessment in terms of rating procedures. The moderator analyses showed that temporal correlates of listener-based judgements can be enhanced when the rating scales or

rubrics are adjusted to the proficiency level of the target population of speakers. Tavakoli et al. (2020) and Saito et al. (2018) reported that depending on proficiency levels and/or overall fluency scores, the relative importance of temporal characteristics, such as articulation rate and pause frequency, to the rating scores tends to vary. I thus recommend that to enhance the validity of fluency assessment, rubrics and rating scales need to be adjusted to the target population, especially with regard to the range of test-takers' proficiency levels. For instance, Tavakoli et al. (2020) showed that articulation rate distinguished learners between A2, B1, and B2 levels, while mean pause duration distinguished only between A2 and B1 levels. Accordingly, the length of pauses should be prioritized at lower levels on the scale, and the speed of delivery can be highlighted at higher levels of the scale. Another possible application of the current findings (mainly, Study 3) is the combination with a computer-adaptive test of CF to roughly determine the proficiency level of the test-takers. According to the estimated proficiency levels, the corresponding rating scales/rubrics, as well as other proficiency-dependent factors such as speaking prompts, can be selected.

Another language assessment domain that the current findings may have implications for is the development of automated scoring systems. For the purpose of replicating human ratings of L2 fluency, the findings of Study 1 would be useful. Considering the importance of pause location in the predicting power for PF judgements, the addition of the information about pause location to the existing speech annotation systems might be needed to better replicate human ratings of L2 fluency. Accordingly, the integration of speech recognition software and natural language processing tools with automated speech annotation may allow for the calculation of mid-clause pause measures, which have strong predicting power for PF judgements. In addition, Study 1 found a potential cross-linguistic difference in the association strengths between UF and PF measures. Similarly, De Jong et al. (2020b) also

reported that the correlation coefficients between automatically (and manually) calculated UF measures and PF ratings differed between L2 Dutch and English. The crosslinguistic effects in the UF-PF link are particularly important if the scoring of fluency is automated and relies on UF measures alone. For some languages, such as L2 Japanese, UF measures might be more reliable indicators of fluency judgements than for other L2s, such as English, Dutch, and French. It is thus recommended that the weights of temporal measures to the outcome score in algorithms should also be adjusted according to the target language as well as speakers' proficiency levels.

10.4.2 Implications for L2 learning and teaching

Considering the importance of maintaining a certain level of fluency from the listeners' perspective in real-world L2 communication (Lennon, 2000), the findings of Study 1 suggest that the development of speed and breakdown fluency, especially the reduction of pauses in the middle of utterances, might be emphasized when L2 oral fluency is targeted as a curricular objective. Meanwhile, building on the assumption that the development of CF leads to the improvement in UF (Segalowitz, 2010, 2016), the results of Study 3 also give some insights into what aspects of different linguistic objectives should be prioritized in relation to L2 fluency development. More specifically, the CFA model of CF showed that vocabulary size was found as the primary component of linguistic resource of CF, while sentence construction speed as the primary component of processing speed of CF. Accordingly, vocabulary instruction can put emphasis on widening students' lexical repertoires for productive use, while grammatical instruction should focus not only on accuracy aspects but also on fluency. However, as suggested by the literature of L2 speech production, some aspects of grammatical knowledge are stored in the mental lexicon (Kormos, 2006; Levelt, 1989, 1999), meaning that vocabulary and grammatical knowledge

are inseparable. Moreover, articulatory speed was found as another component of processing speed of CF, indicating that training on some suprasegmental features, such as linking and vowel reduction, may also facilitate students' fluent speech production. Similarly, the findings of Study 2 suggested that students' fluency performance tends to be challenged as the control for speaking performance, particularly in content and linguistic items, by the task characteristics increases. Thus, in the task-based language teaching approach, when different speaking tasks can be sequenced from open tasks to closed tasks, students' speaking fluency may develop effectively (Robinson, 2011).

The current findings also have several possible applications for fluency training activities (see Rossiter et al., 2010; Tavakoli & Hunter, 2018). To enhance students' fluency from the perspective of listeners' perceptions, teachers can help students to be aware of how the perceptions of fluency can be differentiated by pause location and how filled pauses including lexical fillers can provide the impression of continuation to listeners. To this end, some consciousness-raising activities, where students listen to their own speech or others' speech analytically, may be effective (Tavakoli et al., 2016). Once students understand how pause location affects listeners' perception of fluency, some fluency strategy training can also be offered. With regard to the strong predicting power of mid-clause pauses for PF judgements, students can practice planning ahead at the clausal boundaries so that they can avoid breakdowns in the middle of utterances. Similarly, teaching multiword sequences (e.g., collocations, fixed expressions) might assist students to continue their speech with fewer mid-clause pauses (Tavakoli & Uchihara, 2020), because multiword sequences can be retrieved as single units with a smaller amount of attentional resources, compared to the rule-based construction of phrases (Kormos, 2006).

Another approach to improving students' oral fluency is the enhancement of proceduralization and automatization of L2 knowledge (Kormos, 2006; Segalowitz, 2010). For instance, sufficient time for pre-task planning may provide an opportunity for students to retrieve vocabulary items which are not yet fully integrated into the mental lexicon and which might be difficult to access during spontaneous communication. Such a scaffolded use of lexical items may serve as the first step toward promoting lexical proceduralization. Different types of task repetition can also help students to improve their UF performance, such as the speed of delivery and the reduction in pauses and hesitations (Lambert et al., 2017), especially with the increasing time pressure (i.e., so-called 4/3/2 techniques; Boers, 2014; Nation, 1989; Thai & Boers, 2016).

10.5 Limitations

In this section, I acknowledge methodological limitations for each study in the current thesis to avoid overinterpretation of the findings. As for Study 1, first of all, the total number of primary studies was relatively small, because of the strict screening procedure, which is crucial for the robustness of findings from meta-analyses (Boers et al., 2020). Meanwhile, the number of effect sizes in some subgroups in moderator analyses was too small to perform some subgroup analyses. Therefore, the limited number of studies included highlights the need for more studies that examine the UF-PF link. Second, the significant moderator effects of target L2 in Study 1 might be subsumed under the effects of the L1-L2 combination, because I could not control for the L1 background of speakers due to the huge variability in L1s across studies. Third, due to the variability in methodological practice, I could not include some empirically motivated methodological variables, such as speakers' L2 proficiency level and listeners' familiarity with the speakers' L1, in the moderator analyses. Similarly, I acknowledge that some categories of moderator variables were broad (e.g.,

listeners' experience, rater training, task type), calling for future studies carefully manipulating specific variables. Finally, due to the limited number of studies reporting reliability estimates for UF measures, I could not correct the aggregated correlation coefficients for reliability estimates (i.e., measurement errors), indicating that the calculated effect sizes in Study 1 might have been slightly attenuated (cf. Saito & Plonsky, 2019).

Although Studies 2–4 were conducted using the same dataset (see Chapter 6), the methodological limitations are introduced here with regard to the RQs of each study. Regarding Study 2, while the enhanced conceptualizing demands were indicated by the increase in filled pauses and students' perceptions, the argumentative task may have simultaneously reduced the demands on formulation possibly due to the open-ended nature of the task (i.e., avoidance of difficult linguistic items; see Section 7.3.1). Comparing the findings with previous studies, the relatively lower level of proficiency of the participants might have caused the unexpected effects of content generation on UF performance. Future studies would thus be needed to replicate Study 2 with a group of more advanced learners or with multiple groups of proficiency levels. In addition, although it is virtually impossible to separate the demands on conceptualization and formulation due to the serial nature of speech production, future studies are also required to deploy a more carefully controlled pair of tasks to examine the effects of conceptualizing demands on UF performance, while keeping an optimal level of ecological validity. According to students' perceptions of task demands, the conceptualizing demands in the text summary tasks might have been slightly higher than those in the picture narrative task. In the text summary tasks, students have to recall the content of the source texts without any cues while speaking, whereas the in picture narrative task, the content to include in speech is presented even while speaking. Therefore, as with the

contrast between the argumentative and picture narrative tasks, future studies would be required to carefully control for non-target task design features.

As regards Study 3, methodological limitations are generally related to the selection of CF and UF measures. First, Study 3 did not include the linguistic knowledge measure of multiword sequences or phrasal expressions (cf. De Jong et al., 2013; Kahng, 2020). Due to the SEM approach, the latent variables of CF in Study 3 were relatively independent of measurement errors and thus might have tapped into a certain amount of covariance between the current CF measures and the potential knowledge measures of multiword sequences. However, the potential contribution of the use of multiword sequences to UF performance may be theoretically distinctive from that of rule-based linguistic encoding of formulation processes. More specifically, the processing advantage of multiword sequences is its single-step retrieval, which enables speakers to produce phrases or even clauses without engaging with syntactic processing (Kormos, 2006; Wray, 2000). Accordingly, the inclusion of knowledge measure of multiword sequences may change the structure of the CFA model of CF. Second, due to the substantive difficulty in identifying target-like pronunciation, Study 3 did not examine the accuracy aspects of pronunciation knowledge. The rationale behind this methodological decision was also supported by the theoretical assumption of the relatively automatic nature of phonological and phonetic encoding processes. In other words, the contribution of L2 pronunciation knowledge might be theoretically negligible in light of L2 UF performance. However, previous studies on the UF-PF link with the focus on higher-order fluency (i.e., overall command of language rather than temporal performance) found some unique contributions of pronunciation measures, such as syllable structure errors (S. Suzuki & Kormos, 2020), to PF ratings. Therefore, for a better understanding of the interrelationship between CF, UF, and PF, future studies may add pronunciation accuracy

measures as another observed variable of CF. Third, two composite measures (mean length of run, speech rate) were used as speed fluency measures for statistical reasons to avoid an under-identified model for the measurement model of speed fluency in the CFA. However, due to the intercollinearity among the observed variables of speed fluency, the measure of speech rate was excluded from the CFA model of UF, and the measurement model of speed fluency was eventually regarded as an under-identified model, because the latent variable of speed fluency consisted of two observed variables. Fourth, for the same statistical reason to avoid an under-identified model, I could not specify the latent variables of breakdown fluency separately for pause frequency and duration in the CFA model of UF. All the observed variables of breakdown fluency except for mid-clause pause ratio were associated with the latent variable of breakdown fluency to an equal extent. It may thus be assumed that the construct of breakdown fluency, at least based on the existing breakdown fluency measures, is a unitary construct. However, Study 1 showed that the predicting power for PF judgements was different between pause frequency and pause duration. Therefore, future studies may be needed to examine the dimensionality of breakdown fluency concerning the association with both CF and PF.

With regard to Study 4, I acknowledge methodological limitations regarding the speech elicitation tasks. Although Study 4 statistically controlled for the effects of the topic of L2 argumentative tasks on the prediction of L2 UF measures from the L1 counterparts, the current results are limited to one single task type, that is, an argumentative task. Similarly, the topic of the argumentative tasks was not counterbalanced between L1 and L2 speech. Accordingly, the differences in UF performance between L1 and L2, at least to some extent, might have been subsumed under the topic effects.

10.6 Future Directions for L2 Fluency Research

10.6.1 L2 utterance-perceived fluency link research

Study 1 also revealed several methodological factors in need of further investigation. First, relating to the abovementioned methodological incomparability across L2 fluency studies, one might develop a comprehensive background questionnaire for listeners (cf. Saito et al., 2019). Scholars should also report speakers' proficiency levels in relation to established benchmarks such as CEFR (for a similar suggestion, see Webb et al., 2020), with some justification for their assessment of proficiency (cf. Plonsky & Kim, 2016). Second, the comprehensive library search did not find studies correcting UF measures using the speakers' L1 UF counterparts. Comparing L1-corrected UF measures with the raw counterparts, future studies can explore listeners' sensitivity to the influence of speakers' personal speaking style on their PF judgements. Third, I encourage researchers to report the reliability estimates for both PF and UF measures, unless automated annotation of temporal features is used. This practice would allow future meta-analyses to calculate the effect sizes more precisely by correcting for reliability estimates. Finally, following the recommended practice in L2 speech perception research (Isaacs & Thomson, 2020), supplementary qualitative data may also provide some insights into how listeners selectively pay attention to discrete speech characteristics (e.g., Magne et al., 2019; S. Suzuki & Kormos, 2020).

10.6.2 L2 speech processing demands research

In addition to the degree of content generation (open vs. closed tasks in Study 2), the complexity of the preverbal message is another indicator of speech processing demands on conceptualization, because even in speaking performance elicited via open-ended tasks, there can be individual variability in the elaboration of content or the complexity of the preverbal message. By controlling for the complexity of the preverbal message, the effects of task

design features on UF performance can be more sophisticatedly investigated. Skehan (2009) proposed that lexical sophistication of speech can be a proxy for the complexity of preverbal message, while lexical sophistication has been regarded as one of the subdimensions of lexical complexity (Bulté & Housen, 2012; Eguchi & Kyle, 2020; Michel, 2017). Meanwhile, in task-based performance research, scholars have developed the construct of propositional complexity, which is commonly operationalized as the amount of elaboration or information to convey (Bulté & Housen, 2012; Vasylets et al., 2017). Therefore, it seems that valid measurements for the complexity of preverbal message have not yet been established in L2 research, meaning that research into speech processing demands could be expanded by developing and validating such measures. Another methodological approach to estimating the effects of conceptualizing demands would be the inclusion of relevant individual difference factors, such as working memory capacity and speaking strategies. In the literature of task-based performance, the effects of task design features can be moderated by such individual difference factors (R. Ellis, 2009). By examining an interaction between task design features and individual difference factors, the effects of task design features on speaking performance can be more carefully examined.

10.6.3 L2 cognitive-utterance fluency link research

Study 3 suggests several possible areas that future research into the CF-UF link may address. First, the view of the usage-based approach to speech production could be tested in relation to L2 oral fluency. Following the assumption that CF reflects the efficiency in L2 speech production processes (Segalowitz, 2010, 2016), CF measures in Study 3 and previous studies (De Jong et al., 2013; Kahng, 2020) largely capture linguistic knowledge and processing skills for rule-based construction (e.g., syntactic encoding). This is partly because L2 speech production models (e.g., Kormos, 2006; Segalowitz, 2010) are based on Levelt's (1989, 1999)

model which postulates the rule-based construction of phrases and clauses (see Kormos, 2006). Alternatively, usage-based or exemplar-based processing for speech production may also be able to explain UF performance. For instance, a usage-based paradigm presupposes that learning of grammar is achieved by associative learning, which abstracts grammatical rules based on the information of co-occurrence and statistical regularity such as collocations rather than the proceduralization of declarative knowledge (N. Ellis, 2017). Utilizing such co-occurrence information of items, speakers can activate and retrieve the associated items to construct the phrases or clauses. Depending on learning and processing theories, different sets of CF measures can be selected. It is also worth examining which learning and processing theories (e.g., rule- vs. usage-based processing) would account for UF performance better, using different sets of theoretically motivated CF measures.

Second, while Study 3 adopted a cross-sectional design as a first step towards understanding the CF-UF link at the level of constructs, the developmental perspective should be explored to extend this line of research. As the speed dimension of CF (i.e., automaticity of L2 knowledge) in particular can vary across learning stages of proceduralization and automatization, future studies can investigate how the interrelationship between CF and UF components changes as a function of CF development. Learning conditions (e.g., formal language instruction vs. immersion context) may also affect the CF-UF link identified in Study 3. Regarding learning experiences, a majority of the current participants were EFL learners who had learned L2 English mainly in classroom settings. Especially when replicating this study with learners in naturalistic contexts, it is worth examining the extent to which the current SEM model is plausible with different groups of L2 learners.

Third, future studies can extend the research into the CF-UF link by developing UF measures that are sensitive to learners' CF. Following the construct definition of UF, the validity of L2 UF measures is evaluated in terms of the extent to which L2 automaticity can be captured (Segalowitz, 2010). One of the possible applications would be the coefficient of variance measures from the strand of psycholinguistics. The coefficient of variance is calculated by dividing the standard deviation of performance by the mean and is assumed to tap into the stability of target processing (Segalowitz & Freed, 2004). Most of the existing UF measures are simply calculated as the average of frequency, density, or length of temporal features such as pauses and length of run. Taking an example of mean length of run, the coefficient of variance measure can be calculated by dividing the standard deviation of length of all runs in the individual speech by their mean length. The development and validation of such coefficient of variance measures of UF with regard to its association with CF measures as well as PF ratings might be another promising area of L2 fluency research.

Finally, the practicality and feasibility of CF assessment should be noted as another area of future directions of the CF-UF link research. As mentioned in the section of implications for L2 assessment, the development of computer-adaptive tests of CF would be beneficial for language assessment contexts. Such a quick test of CF would also be useful in research contexts. Study 3 and previous studies have deployed a battery of tests capturing different linguistic knowledge. Considering the future development of the CF-UF link research by integrating other individual difference factors or multiple speaking tasks, it is not ideal that the tasks for CF measures take much time in data collection procedures. Therefore, the computer-adaptive version of CF test, based on the current results of dimensionality of CF, would also be worth developing for research purposes.

10.6.4 L1-L2 utterance fluency link research

Research into the L1-L2 UF link could be extended by exploring more theoretically oriented interpretations of the L1-L2 UF link. Following previous studies (De Jong et al., 2015; Duran-Karaoz & Tavakoli, 2020; Peltonen, 2018) and L2 speech production models (de Bot, 1992; Kormos, 2006; Segalowitz, 2010), Study 4 assumed that the covariance between L1 and L2 UF measures is related to language-general processes and idiosyncratic factors shared across L1 and L2 speech production. However, this assumption should also be further validated, because several alternative explanations for the L1-L2 UF link are still possible. For instance, speech processing in a broad sense is the retrieval and manipulation of information. Although the information to retrieve and manipulate includes language-specific knowledge and real-world knowledge (Levelt, 1989, 1999), the process of retrieval and information processing is arguably underpinned by cognitive ability and capacity, such as executive function and phonological short-term memory (Baddeley, 2003). In other words, the covariance between L1 and L2 UF performance may largely indicate such general cognitive ability. Alternatively, prior research has regarded the L1-L2 UF link as evidence of the reflection of speakers' personal speaking style in the UF measures. Although personal speaking style is one of the idiosyncratic factors, speaking style has been rarely defined in previous studies, and the idiosyncrasy does not solely represent individual speaking style. Therefore, future studies would be needed to further clarify what underlies the covariance between L1 and L2 UF measures, by correlating it with individual difference factors, such as working memory capacity and personality. In addition, from a methodological perspective, different components of L2 proficiency are ideally measured concerning how those proficiency components differently contribute to UF performance. Although the role of L2 proficiency in the L1-L2 UF link has been advocated, the results of Study 4 suggested that the strength and direction of the moderating role of L2 proficiency in the L1-L2 UF link can

vary, depending on which component of L2 proficiency is measured. However, it may not always be practical to conduct various linguistic tests in addition to L1 and L2 speaking tasks. As with the CF-UF link, the L1-L2 UF link research might also benefit from the development of a quick test of CF.

References

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly*, *13*(4), 341–358.

Anderson, J. R. (2009). *Cognitive psychology and its implications* (7th ed.). Worth.

Baayen, R. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge University Press.

Bachoud-Lévi, A.-C., Dupoux, E., Cohen, L., & Mehler, J. (1998). Where is the length effect? A cross-linguistic study of speech production. *Journal of Memory and Language*, *39*(3), 331–346.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189–208.

Baddeley, A., & Hitch, G. J. (2019). The phonological loop as a buffer store: An update. *Cortex, 112,* 91–106.

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, *47*(4), 707–728.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*(June), 3–4.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1).

Boers, F. (2014). A reappraisal of the 4/3/2 activity. *RELC Journal*, *45*(3), 221–235.

Boers, F., Bryfonski, L., Faez, F., & McKay, T. (2020). A call for cautious interpretation of meta-analytic reviews. *Studies in Second Language Acquisition*, *15*(1), 1–23.

Boersma, P., & Weenink, D. (2012). *Praat: doing phonetics by computer [Computer*

software]. www.praat.org/

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.

Bradlow, A. R., Kim, M., & Blasingame, M. (2017). Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate. *The Journal of the Acoustical Society of America*, *141*(2), 886–899.

Broos, W. P. J., Duyck, W., & Hartsuiker, R. J. (2018). Are higher-level processes delayed in second language word production? Evidence from picture naming and phoneme monitoring. *Language, Cognition and Neuroscience*, *33*(10), 1219–1234.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. The Guilford Press.

Bui, G., Ahmadian, M. J., & Hunter, A.-M. (2019). Spacing effects on repeated L2 task performance. *System*, *81*, 1–13.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 23–46). John Benjamins.

Christenfeld, N. (1994). Options and UMS. *Journal of Language and Social Psychology*, *13*(2), 192–199.

Clark, H., & Fox Tree, J. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Collins, B., & Mees, I. M. (2003). *The phonetics of English and Dutch* (5th ed.). Brill.

Coupé, C. (2018). Modeling linguistic variables with regression models: Addressing non-gaussian distributions, non-independent observations, and non-linear predictors with random effects and generalized additive models for location, scale, and shape. *Frontiers*

in Psychology, *9*(April), 1–21.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862–2873.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge University Press.

de Bot, K. (1992). A bilingual production model: Levelt's speaking model adapted. *Applied Linguistics*, *13*, 1–24.

de Bot, K., & Schreuder, R. (1993). Word production and the bilingual lexicon. In *The bilingual lexicon* (pp. 191–214). Benjamins.

De Clercq, B., & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *The Modern Language Journal, 101*(2), 315–334.

De Jong, N. H. (2016a). Fluency in second language assessment. In D. Tsagari & J. Banerjee (Ed.), *Handbook of Second Language Assessment* (pp. 203-218). Berlin, Boston: De Gruyter Mouton.

De Jong, N. H. (2016b). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113–132.

De Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, *15*(3), 237–254.

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. *DiSS 2013. Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech*, *January 2013*, 17–20.

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 behavior. *Applied Psycholinguistics*, *36*(2), 223–243.

De Jong, N. H., & Mora, J. C. (2019). Does having good articulatory skills lead to more fluent speech in first and second languages? *Studies in Second Language Acquisition*, *41*(1), 227–239.

De Jong, N. H., Pacilly, J., & Heeren, W. (2020). *Praat scripts to measure speed fluency and breakdown fluency in speech automatically*. Retrieved from osf.io/w3r7t

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*(5), 893–916.

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.

de Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input. *Language Learning*, *68*(4), 906–941.

DeKeyser, R. (2000). The Robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.

DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 94–112). Routledge.

DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds), *The Routledge Handbook of Instructed Second Language Acquisition*. (pp. 15–32), Routledge.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many

variables: A systematic review of issues and developments. *Frontiers in Psychology*, *9*(April).

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*(4), 533–557.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgements on different tasks. *Language Learning*, *54*(4), 655–679.

Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, *34*(2), 183–193.

Dóczi, B., & Kormos, J. (2016). *Longitudinal developments in vocabulary knowledge and lexical organization*. Oxford University Press.

Doe, T. (2017). *Oral fluency development activities: A one-semester study of EFL students*. Temple University.

Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in l2 communication: A psycholinguistic perspective. *Studies in Second Language Acquisition*, *20*(3), 349–385.

Dressler, A. M., & O'Brien, M. G. (2019). Rethinking perceptions of fluency. *Applied Linguistics Review*, *10*(2), 259–280.

Duijm, K., Schoonen, R., & Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing*, *35*(4), 501–527.

Duran-Karaoz, Z., & Tavakoli, P. (2020). Predicting L2 fluency from L1 fluency behavior. *Studies in Second Language Acquisition, 42*(4), 671–695.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*(2), 163–188.

Eguchi, M., & Kyle, K. (2020). Continuing to Explore the Multidimensional Nature of

Lexical Sophistication : The Case of Oral Proficiency Interviews. *The Modern Language Journal, 104*(2), 381–400.

Ellis, N. C. (2017). Cognition, corpora, and computing: triangulating research in usage-based language learning. *Language Learning*, *67*(S1), 40–65.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*(2), 141–172.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production. *Applied Linguistics*, *30*(4), 474–509.

Felker, E., Klockmann, H., & de Jong, N. H. (2019). How conceptualizing influences fluency in first and second language speech production. *Applied Psycholinguistics*, *40*(1), 111–136.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). SAGE Publications.

Fillmore, C. J. (1979). On fluency. In D. Kempler & W. S. Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85–102). Academic Press.

Forster, K., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 116–124.

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning*, *59*(4), 866–896.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375.

Fraundorf, S. H., & Watson, D. G. (2014). Alice's adventures in um-derland: Psycholinguistic sources of variation in disfluency production. *Language, Cognition and Neuroscience*, *29*(9), 1083–1096.

French, R. M., & Jacquet, M. (2004). Understanding bilingual memory: Models and data. *Trends in Cognitive Sciences*, *8*(2), 87–93.

Gass, S. (2018). SLA elicitation tasks. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology* (pp. 313–337). Palgrave Macmillan.

Gilbert, A. C., Cousineau-Perusse, M., & Titone, D. (2020). L2 exposure modulates the scope of planning during first and second language production. *Bilingualism: Language and Cognition*, *23*(5), 1093–1105.

Godfroid, A., Loewen, S., Jung, S., Park, J. H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition*, *37*(2), 269–297.

González-fernández, B., & Schmitt, N. (2020). Word Knowledge: Exploring the Relationships and Order of Acquisition of Vocabulary Knowledge Components. *Applied Linguistics*, *41*(4), 481–505.

Götz, S. (2013). *Fluency in native and nonnative English speech*. John Benjamins.

Graham, C. R., & Belnap, R. K. (1986). The acquisition of lexical boundaries in English by native speakers of Spanish. *IRAL : International Review of Applied Linguistics in Language Teaching*, *24*(4), 275–286.

Greene, J., & Cappella, J. N. (1986). Cognition and talk: The relationship of semantic units of temporal patterns of fluency in spontaneous speech. *Language and Speech*, *29*(2), 141–157.

Harada, T. (2007). The production of voice onset time (VOT) by English-speaking children in a Japanese immersion program. *IRAL - International Review of Applied Linguistics in Language Teaching*, *45*(4), 353–378.

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing meta-Analysis in R: A hands-on Guide*. PROTECT Lab.

Hawkins, P. R. (1971). The syntactic location of hesitation pauses. *Language and Speech*,

14(3), 277–288.

Hermans, D., Bongaerts, T., de Bot, K., & Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language and Cognition*, *1*(3), 213–229.

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 137–159.

Hoshino, N., & Thierry, G. (2011). Language selection in bilingual word production: Electrophysiological evidence for cross-language competition. *Brain Research*, *1371*, 100–109.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473.

Hu, L.-T., & Bentler, P. M. (1998). Fit Indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453.

Huensch, A., & Tracy-Ventura, N. (2017). Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics, 38*(4), 755–785.

In'nami, Y., & Koizumi, R. (2010). Database selection guidelines for meta-analysis in applied linguistics. *TESOL Quarterly*, *44*(1), 169–184.

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1–2), 101–144.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159.

Isaacs, T., & Thomson, R. I. (2020). Reactions to second language speech. *Journal of Second*

Language Pronunciation, *6*(3), 402–429.

Iwashita, N., Brown, A., McNamara, T., O'Hagan, S., & De Jong, N. H. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49.

JACET. (2003). *JACET List of 8000 Basic Words*. JACET English Vocabulary SIG.

Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences*, *10*(11), 480–486.

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.

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*(3), 569–591.

Kahng, J. (2020). Explaining second language utterance fluency: Contribution of cognitive fluency and first language utterance fluency. *Applied Psycholinguistics*, *41*(2), 457–480.

Korko, M., & Williams, S. A. (2017). Inhibitory control and the speech patterns of second language users. *British Journal of Psychology*, *108*(1), 43–72.

Kormos, J. (1999a). Monitoring and self-repair in L2. *Language Learning*, *49*(2), 303–342.

Kormos, J. (1999b). The effect of speaker variables on the self-correction behaviour of L2 learners. *System*, *27*(2), 207–221.

Kormos, J. (2000). The timing of self-repairs in second language speech production. *Studies in Second Language Acquisition*, *22*(2), 145–167.

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum Associates.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164.

Košak-Babuder, M., Kormos, J., Ratajczak, M., & Pižorn, K. (2019). The effect of read-aloud

assistance on the text comprehension of dyslexic and non-dyslexic English language learners. *Language Testing*, *36*(1), 51–75.

Kroll, J. F., Bobb, S. C., Misra, M., & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica*, *128*(3), 416–430.

Kroll, J. F., Bobb, S. C., & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism*, *9*(2), 119–135.

Kroll, J. F., Sumutka, B. M., & Schwartz, A. I. (2005). A cognitive view of the bilingual lexicon: Reading and speaking words in two languages. *International Journal of Bilingualism*, *9*(1), 27–48.

Kyriazos, T. A. (2018). Applied psychometrics: Sample size and sample power considerations in factor analysis (EFA, CFA) and SEM in general. *Psychology*, *09*(08), 2207–2230.

La Heji, W. (2005). Selection processes in monolingual and bilingual lexical access. In J. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 289–307). Oxford University Press.

Ladefoged, P. (2015). *A course in phonetics* (K. Johnson (ed.); 7th ed.). Cengage Learning.

Lambert, C., Aubrey, S., & Leeming, P. (2020). Task preparation and second language speech production. *TESOL Quarterly*.

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, *35*(5), 607–614.

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167–196.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability.

Language Testing, *16*(1), 33–51.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387–417.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). University of Michigan Press.

Leonard, K. R., & Shea, C. E. (2017). L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors. *The Modern Language Journal*, *101*(1), 179–193.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, Mass: MIT Press.

Levelt, W. J. M. (1993). The architecture of normal spoken language use. In G. Blanken, J. Dittman, H. Grimm, J. C.Marshall, & C.-W. Wallesch (Eds.), *Linguistic Disorders and Pathologies: An International Handbook* (pp. 1–15). Walter de Gruyter.

Levelt, W. J. M. (1999). Language production: A blueprint of the speaker. In C. Brown & P. Hagoort (Eds.), *Neurocognition of language* (pp. 83–122). Oxford University Press.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(01), 1–75.

Li, S. (2016). The construct validity of language aptitude. *Studies in Second Language Acquisition*, *38*(4), 801–842.

Lindqvist, C., Bardel, C., & Gudmundson, A. (2011). Lexical richness in the advanced learner's oral production of French and Italian L2. *IRAL - International Review of Applied Linguistics in Language Teaching*, *49*(3), 221–240.

Liu, Y. T., & Todd, A. G. (2014). Dual-modality input in repeated reading for foreign language learners with different learning styles. *Foreign Language Annals*, *47*(4), 684–

706.

Loewen, S. (2009). Grammaticality judgment tests and the measurement of implicit and explicit L2 knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 94–112). Multilingual Matters.

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

Magne, V., Suzuki, S., Suzukida, Y., Ilkan, M., Tran, M., & Saito, K. (2019). Exploring the dynamic nature of second language listeners' perceived fluency: A mixed-methods approach. *TESOL Quarterly*, *53*(4), 1139–1150.

Maie, R., & Dekeyser, R. M. (2020). Conflicting evidence of explicit and implicit knowledge from objective and subjective measures. *Studies in Second Language Acquisition*, *42*(2), 359–382.

McDonough, K., & Trofimovich, P. (2008). *Using priming methods in second language research*. Routledge.

McManus, K., & Marsden, E. (2019). Signatures of automaticity during practice: Explicit instruction about L1 processing routines can improve L2 grammatical processing. *Applied Psycholinguistics*, *40*(1), 205–234.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

MEXT. (2017). 「平成29年度 英語力調査結果（高校３年生）の概要」. Retrieved from https://www.mext.go.jp/a_menu/kokusai/gaikokugo/__icsFiles/afieldfile/2018/04/06/1403470_03_1.pdf

Michel, M. C. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 50–68). Taylor & Francis.

Millington, N. (2019). *Dreamreader.net*. http://dreamreader.net/

Morgan-Short, K., Faretta-Stutenberg, M., Brill-Schuetz, K. a., Carpenter, H., & Wong, P. C. M. (2014). Declarative and procedural memory as individual differences in second language acquisition. *Bilingualism: Language and Cognition*, *17*(February 2016), 56–72.

Myers-Scotton, C., & Jake, J. L. (2000). Four types of morpheme: Evidence from aphasia, code switching, and second-language acquisition. *Linguistics*, *38*(370), 1053–1100.

Nagashima, K., Noma, H., & Furukawa, T. A. (2019). Prediction intervals for random-effects meta-analysis: A confidence distribution approach. *Statistical Methods in Medical Research*, *28*(6), 1689–1702.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, P. (1989). Improving speaking fluency. *System*, *17*(3), 377–384.

Negishi, J. (2012). Relationships between L2 speakers' development and raters' perception on fluency in group oral interaction. *Jounral of Pan-Pacific Association of Applied Linguistics*, *15*(2), 1–26.

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing*, *31*(2), 147–175.

Novianti, P. W., Roes, K. C. B., & van der Tweel, I. (2014). Estimation of between-trial variance in sequential meta-analyses: A simulation study. *Contemporary Clinical Trials*, *37*(1), 129–138.

Oppenheim, G. M., & Dell, G. S. (2010). Motor movement matters: The flexible abstractness of inner speech. *Memory and Cognition*, *38*(8), 1147–1160.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601.

Paradis, M. (2009). *Declarative and procedural determinants of second languages (Vol. 40)*. John Benjamins.

Paulston, C. B., Kiesling, S. F., & Rangel, E. S. (2012). *The handbook of intercultural discourse and communication*. Wiley-Blackwell.

Pavlenko, A. (2009). Conceptual representation in the bilingual lexicon and second languagevocabulary learning. In A. Pavlenko (Ed), *The bilingual mental lexicon: Interdisciplinary approaches* (pp. 125–160). Multilingual Matters.

Pawley, A., & Syder, F. (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). Longman.

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13.

Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language*, *87*(3), 539–558.

Peltonen, P. (2018). Exploring connections between first and second language fluency: A mixed methods approach. *The Modern Language Journal*, *102*(4), 676–692.

Peltonen, P., & Lintunen, P. (2016). Integrating quantitative and qualitative approaches in L2 fluency analysis: A study of Finnish-speaking and Swedish-speaking learners of English at two school levels. *European Journal of Applied Linguistics*, *4*(2), 209–238.

Pigott, T. D., & Polanin, J. R. (2019). Methodological guidance papers: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46.

Plonsky, L. (2015). *Advancing quantitative methods in second language research*.

Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*(2), 267–278.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 Research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, *102*(4), 713–731.

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, *36*, 73–97.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.

Plonsky, L., & Oswald, F. L. (2015). Meta-anlyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). Routledge.

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition, 39*(3), 579–592.

Postma, A., & Kolk, H. (1993). The covert repair hypothesis: prearticulatory repair processes in normal and stuttered disfluencies. *Journal Of Speech And Hearing Research*, *36*(3), 472–487.

Poulisse, N. (1999). *Slips of the tongue: Speech errors in first and second language production*. J. Benjamins.

Poulisse, N., & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics*, *15*(1), 36–57.

Préfontaine, Y., & Kormos, J. (2015). The relationship between task difficulty and second language fluency in French: A mixed methods approach. *The Modern Language Journal*, *99*(1), 96–112.

Préfontaine, Y., & Kormos, J. (2016). A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 151–169.

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing*, *33*(1), 53–73.

Qiu, Y., & Zhou, X. (2012). Processing temporal agreement in a tenseless language: An ERP study of Mandarin Chinese. *Brain Research*, *1446*, 91–108.

R development Core Team (2020). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. http://www.r-project.org/

Raykov, T., & Marcoulides, G. (2006). *A first course in structural equation modeling* (2nd ed.). Lawrence Erlbaum Associates.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423.

Roberts, B., & Kirsner, K. (2000). Temporal cycles in speech production. *Language and Cognitive Processes*, *15*(2), 129–157.

Robinson, P. (2011). *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance*. John Benjamins.

Roelofs, A. (2000). WEAVER++ and other computational models of lemma retrieval and word-form encoding. *Aspects of Language Production*, 71–114.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, *65*(3), 395–412.

Rossiter, M. J., Derwing, T. M., Manimtim, L. G., & Thomson, R. I. (2010). Oral fluency: The neglected component in the communicative language classroom. *Canadian Modern Language Review*, *66*(4), 583–606.

Saito, K. (2017). Effects of sound, vocabulary, and grammar learning aptitude on adult second language speech attainment in foreign language classrooms. *Language Learning*, *67*(3), 665–693.

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.

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708.

Saito, K., Suzuki, S., Oyama, T., & Akiyama, Y. (2019). How does longitudinal interaction promote second language speech learning? Roles of learner experience and proficiency levels. *Second Language Research*, 026765831988498.

Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? *Studies in Second Language Acquisition*, *41*(5), 1133–1149.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439–462.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second Language Acquisition and Language Testing approaches. *System*, *45*, 79–91.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust?: Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology*, *6*(4), 147–151.

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363.

Schoonen, R. (2015). Structural equation modeling in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 213–242). Routledge.

Schwarzer, G. (2007). meta: An R package for meta-Analysis. *R News*, *7*, 40–45.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

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.

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, *26*(2), 173–199.

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1–14.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532.

Skehan, P. (2014a). Limited attentional capacity, second language performance, and task-based pedagogy. In P. Skehan (Ed.), *Processing perspectives on task performance*. John Benjamins.

Skehan, P. (2014b). *Processing perspectives on task performance*. John Benjamins.

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 97–111.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215.

Sprenger, S. A., Levelt, W. J. M., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, *54*(2), 161–184.

Stemberger, J. P. (1985). An interactive activation model of language production. In A. W. Ellis (Ed.), *Progress in the psychology of language* (pp. 143–186). Lawrence Erlbaum Associates.

Stoehr, A., Benders, T., van Hell, J. G., & Fikkert, P. (2017). Second language attainment

and first language attrition: The case of VOT in immersed Dutch? German late bilinguals. *Second Language Research*, 026765831770426.

Strik, H., Russel, A., Van Den Heuvel, H., Cucchiarini, C., & Boves, L. (1997). A spoken dialog system for the Dutch public transport information service. *International Journal of Speech Technology*, *2*(2), 121–131.

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167.

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency : A meta-analysis of correlational studies. *The Modern Language Journal*, *105*(2).

Suzuki, Y., & DeKeyser, R. (2017). The interface of explicit and implicit knowledge in a second language: Insights from individual differences in cognitive aptitudes. *Language Learning*, *67*(4), 747–790.

Suzuki, Y., & Sunada, M. (2018). Automatization in second language sentence processing: Relationship between elicited imitation and maze tasks. *Bilingualism: Language and Cognition*, *21*(1), 32–46.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*. Harper Collins.

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal*, *65*(1), 71–79.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 133–150.

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*, *50*(2), 447–

471.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, *58*(2), 439–473.

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.

Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of Fluency Across Assessed Levels of Speaking Proficiency. *The Modern Language Journal*, *104*(1), 169–191.

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. 239–273). John Benjamins.

Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*(2), 506–547.

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

Team, R. D. C. (2020). *R: A language and environment for statistical computing* (4.0.2). R Foundation for Statistical Computing.

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *TESOL Quarterly*, *50*(2), 369–393.

Tian, Y., Maruyama, T., & Ginzburg, J. (2017). Self addressed questions and filled pauses: A cross-linguistic investigation. *Journal of Psycholinguistic Research*, *46*(4), 905–922.

Trofimovich, P., Kennedy, S., & Blanchet, J. (2017). Development of second language French oral skills in an instructed setting: A focus on speech ratings. *Canadian Journal of Applied Linguistics*, *20*(2), 32–50.

Uchihara, T., Saito, K., & Clenton, J. (2020). Re-examining the relationship between

productive vocabulary and second language oral ability. In J. Clenton & P. Booth (Eds.), *Vocabulary and the four skills: Pedagogy, practice, and implications for teaching vocabulary* (pp. 146–165). Routledge.

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559–599.

Ullman, M. T. (2015). The declarative/ procedural model. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: an introduction* (pp. 135–160). Routledge.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Academic Press.

van Os, M., de Jong, N. H., & Bosker, H. R. (2020). Fluency in dialogue: The effect of turn-taking behaviour on perceived fluency in native and non-native speech. *Language Learning*.

Vance, T. J. (2008). *The sounds of Japanese*. Cambridge University Press.

Vasylets, O., Gilabert, R., & Manchón, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning*, *67*(2), 394–430.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P. T., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*(1), 55–79.

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125.

Weinberger, S. (2011). *Speech accent archive*. George Mason University. http://accent.gmu.edu

Wheeldon, L., & Levelt, W. J. M. (1995). Monitoring the time course of phonological encoding. *Journal of Memory and Language*, *34*, 311–334.

Williams, S. A., & Korko, M. (2019). Pause behavior within reformulations and the proficiency level of second language learners of English. *Applied Psycholinguistics*, *40*(3), 723–742.

Wolter, B., & Yamashita, J. (2017). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing. *Studies in Second Language Acquisition*, 1–22.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, *21*, 463–489.

Wright, C., & Tavakoli, P. (2016). New directions and developments in defining, analyzing and measuring L2 speech fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 73–77.

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, *44*(4), 647–668.

Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? *Second Language Research*, *23*(2), 123–153.

Zhang, D., & Yang, X. (2016). Chinese L2 learners' Depth of vocabulary knowledge and its role in reading comprehension. *Foreign Language Annals*, *49*(4), 699–715.

Zuniga, M., & Simard, D. (2019). Factors influencing L2 self-repair behavior: The role of L2 proficiency, attentional control and L1 self-repair behavior. *Journal of Psycholinguistic Research*, *48*(1), 43–59.

Appendices

Appendix A: The pooled results of meta-analysis including both monologic and dialogic speech data

UF measures	*k*	*n*	*Weighted effect size*	*CI*	*Q(df)*	*p-value*	*Egger's test p-value*
Speed fluency							
Articulation rate	12	660	0.58	[.39, .72]	114.73(11)	<.0001	0.860
Breakdown fluency							
Pause frequency	18	881	-0.54	[-.67, -.38]	220.11(17)	<.0001	0.301
Pause duration	11	604	-0.36	[-.57, -.10]	79.83(10)	<.0001	0.674
Repair fluency							
Dysfluency rate	10	587	-0.14	[-.27, -.01]	22.76(9)	0.007	0.257
Composite							
Mean length of run	14	573	0.63	[.49, .74]	76.31(13)	<.0001	0.656
Speech rate	15	677	0.69	[.55, .79]	161.53(14)	<.0001	0.175

Note. k = number of studies; n = number of effect sizes.

Appendix B: The effects of topic on utterance fluency in the argumentative task

Results of Wilcoxon Signed-rank tests for utterance fluency measures between topics

Utterance fluency measure	*W*	*p*	*rs*
Articulation rate	2839.0	0.725	0.040
Speech rate	2041.0	0.026	0.252
Mean length of run	2228.0	0.104	0.184
Mid-clause pause ratio	2866.0	0.660	0.050
End-clause pause ratio	4342.0	< .001	0.590*
Filled pause ratio	3506.0	0.012	0.284
Mid-clause pause duration	3358.0	0.042	0.230
End-clause pause duration	3039.0	0.317	0.113
Self-repetition ratio	3082.0	0.184	0.151
Self-correction ratio	2240.0	0.328	0.113
False start ratio	1958.5	0.236	0.151

Note. The alpha level was adjusted using the Bonferroni correction ($\alpha < 0.004$).

Appendix C: Ethics documents

Participant information sheet

プロジェクト名 (Project Title)

第二言語発話運用における潜在的言語知識の役割

The role of underlying linguistic knowledge in second language speaking performance

研究者について (Researcher)

研究者である鈴木駿吾は、英国ランカスター大学博士課程に在籍しており、日本人英語学習者のスピーキング能力についての研究を行なっています。そこで日本人の皆さんに研究協力をお願いしています。

I am a PhD student at Lancaster University, and I would like to invite you to take part in a research study about Japanese English learners' speaking performance.

研究へ参加するかどうか決めて頂く前に、以下の研究内容について注意してお読みください。

Please take time to read the following information carefully before you decide whether or not you wish to take part.

研究の目的（What is the study about?）

本研究は、外国語学習における実際の発話運用能力が、学習者個人の持つ言語知識との間に、どのような関係があるのかを見るものです。

This study finds out how different types of knowledge are related to foreign language speaking proficiency.

研究に参加して頂きたい理由（Why have I been invited?）

外国語学習における実際の発話運用能力と言語知識との関係性を解明するために、日本人英語学習者の皆様にご協力をお願いしています。その関係性の調査のために、皆様のスピーキングを録音させて頂きます。また幾つかの言語知識に関するテストへの回答をお願いいたします。

I have approached you because I am trying to understand how different types of language knowledge play a role in foreign language speaking. In order to examine the relationship between speaking and language knowledge, I would like to record your speaking performance and ask you to complete several tests of English.

もし研究に参加して頂けたら幸いです。

I would be very grateful if you would agree to take part in this study.

ご協力頂く内容について（What will I be asked to do if I take part?）

参加に同意して頂いた際には、研究協力希望者の都合の良い時間に合わせて、個別セッションと集団セッションのそれぞれに参加して頂きます。個別セッションでは、参加者は５つの発話課題と、口頭による回答を行う２つの言語知識課題を、参加者1人ずつに対して研究者（私）が実施します。参加者の方々の発話課題への音声及び言語知識テストへの回答は全て録音されます。録音されたデータは全て、匿名化された後に分析されます。個別セッションは途中の休憩を含め、およそ５０分間で終了いたします。集団セッションでは、スケジュールの合う参加者ごとにグループで実施し、３つの言語知識テストと言語学習経歴に関するアンケートを実施致します。集団セッションは休憩を含めて、およそ７０分で終了いたします。

If you decided to take part, you would attend both one individual session and one group session. In the individual session, you would complete five speaking tasks and two English knowledge tasks which require you to respond orally. Your speech and oral responses would be audio-recorded, and then transcribed and anonymised for analyses. The individual session takes approximately 50 minutes including a break. In the group session, you would take three other English knowledge tasks and also fill in a language a language background questionnaire. The group session takes approximately 70 minutes including a break.

研究参加から参加者が得られる利益（What are the possible benefits from taking part?）

上記の個別セッション・集団セッションの両方を終了した方には、**謝礼金2,000円**をお支払い致します。また参加者の方々に実施する発話課題や言語知識テストは外国語教育や第二言語習得理論の最新の知見を応用したものであり、研究協力を通して参加者の方々のその後の英語学習に役立つものと考えています。

If you take part in this study, you will be awarded JPY 2,000 in the form of local voucher as your participating rewards. Furthermore, your speech data will offer an insight into how foreign language speaking is assisted by language knowledge. Consequently, your participation would contribute to improve classroom teaching for English as a foreign language in Japan.

研究参加の任意について（Do I have to take part?）

本研究への参加は任意であり、参加の可否は早稲田大学との関係や成績等に一切関係はございません。参加を辞退しても、それによって不都合が生じることは一切ありません。

No. It’s completely up to you to decide whether or not you take part. Your participation is voluntary. Your participation or decision not to take part will not affect your studies and grades. If you choose not to take part this will not disadvantage you in any way.

研究参加の可否及び取りやめについて（What if I change my mind?）

一度参加に同意されても、実験中であっても参加希望者の方がそれを取り消すことはいつでも可能です。もし参加を取りやめたい場合には、研究者まで気軽にお伝え

ください。ただし研究セッション終了から二週間が経ってしまうと、参加者の方々のデータは匿名化されており、個人のデータを取り出し、消去することが困難になってしまうため、セッション終了から二週間以内にお申し付けください。

If you change your mind, you are free to withdraw at any time during your participation in this study. If you want to withdraw, please let me know, and I will extract any data you contributed to the study and destroy it. However, it is difficult and often impossible to take out data from one specific participant when the data have already been anonymised. Therefore, you can only withdraw up to 2 weeks after the session.

研究参加によって生じ得る危害について（What are the possible disadvantages and risks of taking part?）

この実験で危険や不快感等を感じることはございません。上述の通り、研究協力者の方々には研究課題を遂行して頂くために、およそ合計２時間を本研究に費やして頂くことになります。

There will be no major disadvantages to taking part. As mentioned above, participation in the study, however, expects you to invest around 2 hours in total for the speaking tasks, knowledge tests and a questionnaire.

個人情報の扱いについて（Will my data be identifiable?）

データ収集の際は、参加者の名前を全て参加者番号に置き換えて、データの収集ならびに保管を行います。この研究で得られる全てのデータは、研究者である私、鈴木駿吾と、私の指導教官であるJudit
Kormos教授以外アクセスできないよう、厳重に保存いたします。上記２名以外に、データにアクセスする可能性としては、発話評価者が参加者の方々の発話を聞き、評価を行う、あるいは、文字起こしを委託された者が、参加者の方々の発話を聞き書きおろすことが想定されますが、いずれの場合にも、秘密保持に関する契約書に同意ならびに署名を実施します。

After the recording session, only I, the researcher conducting this study and my supervisor will have access to the data you share with me. The only other persons who will have access to the data are (a) speech raters who will listen to the recordings and evaluate the quality of speech, and (b) professional transcribers who will listen to the recordings and produce a written record of what you and others have said. Both speech raters and transcribers will sign a confidentiality agreement.

I will keep all personal information about you (e.g., your name and other information about you that can identify you) confidential, that is I will not share it with others.

収集したデータの扱いについて（How will we use the information you have shared with us and what will happen to the results of the research study?）

採集させて頂いたデータは、私が在籍をしているランカスター大学での博士論文の研究の一部に使用されます。また、博士論文を基に学術誌に研究結果の一部を掲載

する予定です。ただし、いずれの論文上においても、大学名や学生の個人名などが特定されることがないよう匿名化して記述いたします。

I will use the data you share with me only in the following ways:

I will use it for academic purposes only. This will include my PhD dissertation or journal publications. I may also present the results of my study at academic conferences or in my future teaching.

収集したデータの一部（録音データを除く）を後続研究や他の研究者にも使用できるようにランカスター大学のデータレポジトリに保管したいと考えていますが、その際には匿名化されたデータのみを保管します。

I would like to make part of my data available for future research and use by other researchers. We will only share anonymised data in this way and will exclude all personal data and sound files from archiving. I intend to archive/share the data via Lancaster University's Data Repository System.

収集したデータの保管について（How my data will be stored?）

あなたのデータは研究者である私以外がアクセスできないように暗号化され、パスワード保護されたコンピュータに保存されます。紙媒体のデータに関しては、私のオフィス内の鍵付きのキャビネットに厳重に保管されます。いずれのデータ形式においても、個人が特定されることのないように保管されます。ランカスター大学のガイドラインに従い、データは10年間保管されます。

Your data will be stored in encrypted files (that is no-one other than me, the researcher will be able to access them) and on password-protected computers. I will store hard copies of any data securely in locked cabinets in my office. I will keep data that can identify you separately from non-personal information (e.g. your views on a specific topic). In accordance with University guidelines, I will keep the data securely for a minimum of ten years.

ランカスター大学における研究目的でのデータの扱いや研究協力者の権利に関して、更に詳細が必要な場合は、以下のウェブページをご参照ください。
For further information about how Lancaster University processes personal data for research purposes and your data rights please visit our webpage: www.lancaster.ac.uk/research/data-protection

質問や問い合わせの場合（What if I have a question or concern?）

質問や懸念等ございましたら、研究者である私、あるいは私の指導教官であるJudit Kormos教授、学科長であるUta Papen教授にご連絡ください。

If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself s.suzuki@lancaster.ac.uk **(Phone: +44 (0)7706 013370) or my supervisor, Professor Judit Kormos** j.kormos@lancaster.ac.uk **(Phone: +44 (0)1524 593039). If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact: the Head of**

the Department of Linguistics and English Language at Lancaster University, Professor Uta Papen. Her email address is u.papen@lancaster.ac.uk and her phone number is +44 (0)1524 593245.

You can also send a letter to Prof Kormos or Prof Papen (Department of Linguistics and English Language, Lancaster University, Lancaster, LA1 4YL, United Kingdom).

本研究は、英国ランカスター大学、Faculty of Arts and Social Sciences and Management School Research Ethics Committeeより審議及び認可されています。

This study has been reviewed and approved by the Faculty of Arts and Social Sciences and Lancaster Management School's Research Ethics Committee.

Shungo Suzuki (本研究調査者)
Department of Linguistics and English Language
Lancaster University
+44(0) 7706 013370
s.suzuki@lancaster.ac.uk

Prof. Judit Kormos (指導教官)
Department of Linguistics and English Language
Lancaster University
+44 (0) 1524 593039
j.kormos@lancaster.ac.uk

Prof. Uta Papen (研究者所属学科長)
Department of Linguistics and English Language
Lancaster University
+44 (0) 1524 593245
u.papen@lancaster.ac.uk

私の研究に興味を持って頂き、心より感謝申し上げます。

Thank you for considering your participation in this project.

参加同意書（CONSENT FORM）

プロジェクト名: 第二言語発話運用における潜在的言語知識の役割

研究者名: 鈴木駿吾（Shungo Suzuki）
Email: s.suzuki@lancaster.ac.uk

以下の項目を読んで、チェックをいれてください。

1. 私は、上記の研究概要書を読み、理解した。私は、それらの情報について考え、質問をする機会を与えられ、それらについて十分な回答を得られた。	☐
2. 私は、研究への参加が任意であること、いかなる理由であってもセッション終了後から二週間以内であれば、実験中であっても参加を取りやめできることを理解した。終了後二週間以内に取りやめた場合は、私のデータは消去されることになることを理解した。	☐
3. 私は、収集された私に関するデータが将来の報告書や論文、学会発表などで使用される可能性があるが、その場合に個人情報が含まれないことと、私個人が特定されることがないことを理解した。音声データを除く、匿名化されたデータがランカスター大学のデータレポジトリに保管され、研究目的にのみ再利用されることを理解した。	☐
4. 私は、いかなる報告書や論文、口頭発表において、私の同意無しに個人名が公表されることがないことを理解した。	☐
5. 私は、発話や口頭による回答が録音、文字起こしされること、それらデータが暗号化され厳重に保管されることを理解した。また、それらのデータがランカスター大学のデータレポジトリにおいても、公開されることがないことを理解した。	☐
6. 私は、ランカスター大学のガイドラインに基づき、データが研究終了後最低 10 年間保管されることを理解した。	☐
7. 私は、本研究に参加することを同意する。	☐

_____ _____ _____

参加者氏名 (Name of Participant)　　日付（Date）　　署名（Signature）

I confirm that the participant was given an opportunity to ask questions about the study, and all the questions asked by the participant have been answered correctly and to the best of my ability. I confirm that the individual has not been coerced into giving consent, and the consent has been given freely and voluntarily.

Signature of Researcher /person taking the consent_____ **Date** _____ Day/month/year

（研究者名）　　（日付）

One copy of this form will be given to the participant and the original kept in the files of the researcher at Lancaster University

Appendix D: Argumentative speech tasks

The prompt used for Study 2 (n = 128)

Argumentative task A

Statement:

The Tokyo Olympics in 2020 will bring economic growth to Japan.

How far do you agree?
Give some specific examples and explain why or why not.

The prompt used for Study 3 (n = 104)

Argumentative task B

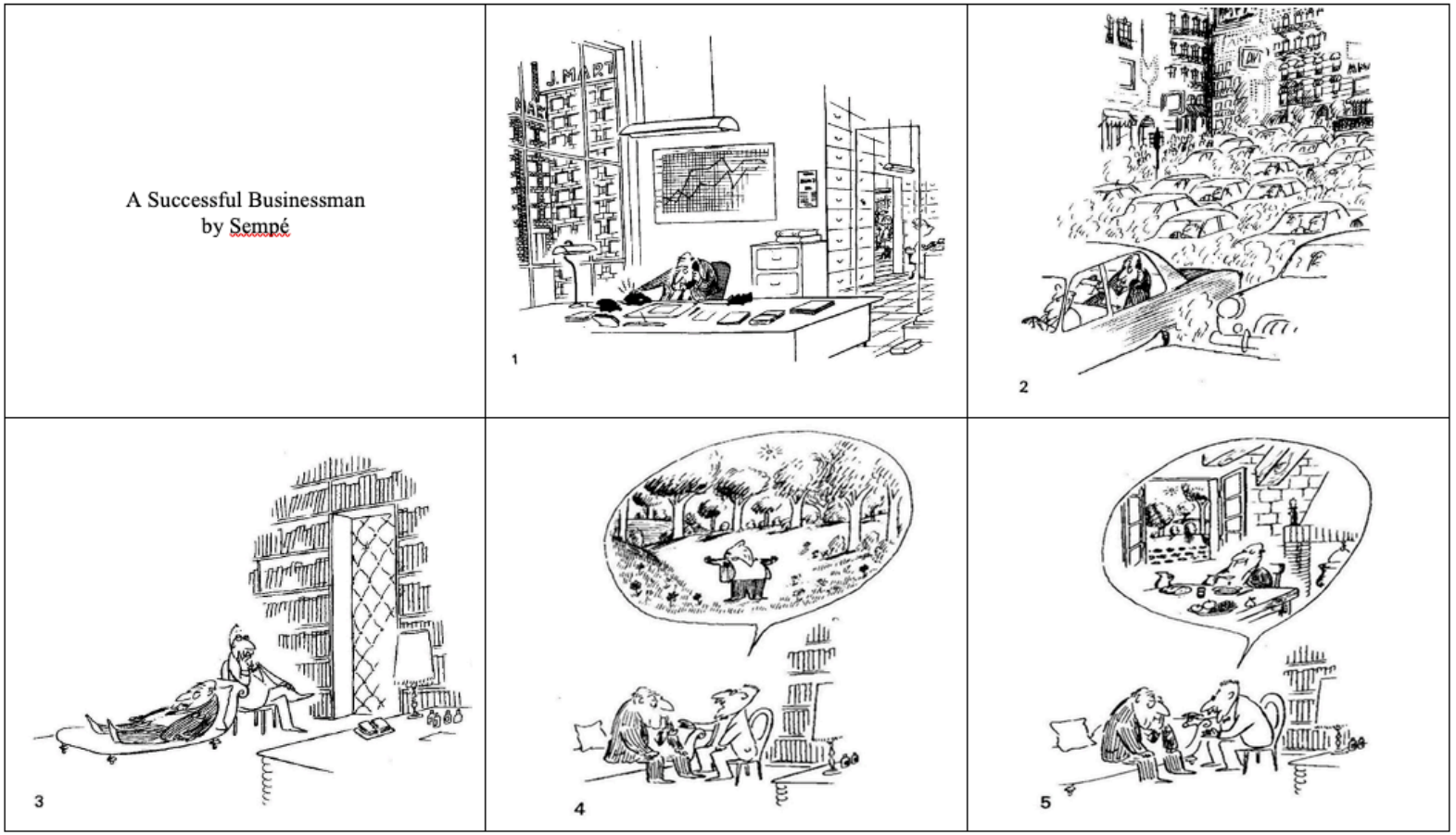
Statement:

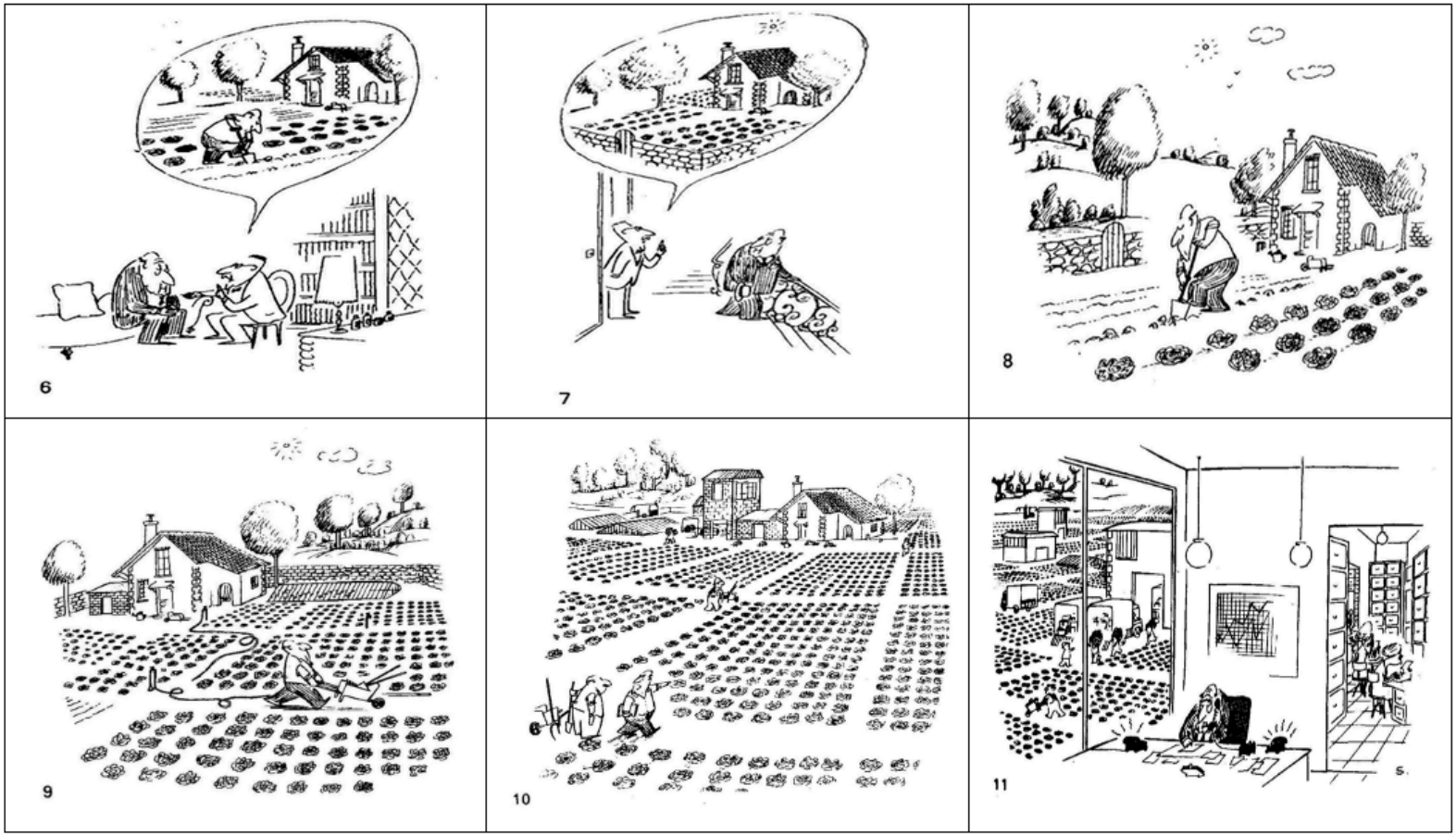
With the development of technology and the Internet, libraries will be unnecessary and disappear.

How far do you agree?
Give some specific examples and explain why or why not.

Appendix E: Picture narrative task (Préfontaine & Kormos, 2015)

Your task is to tell the story of this cartoon strip. You may also add stages not shown by the pictures.
You have **three minutes** to think before you start.





Source: Sempé, Jean-Jacques. (1962). Tout Se Complique. Éditions Denoël.

Appendix F: Text summary task's source texts

Text A: Red-cross

The International Committee of the Red Cross, which is also called the "ICRC" for short, is an organization that helps people around the world. The roots of the ICRC go back to 1859, when a Swiss businessman named Henry Dunant watched a battle while traveling in Italy. After the battle ended, Mr. Dunant was shocked to see the wounded and dying soldiers left on the field of battle. Almost no one could help the dying soldiers. The suffering was terrible and tragic. Mr. Dunant tried to organize some assistance. He asked the local people to care for the wounded and dying soldiers.

After he returned home to Switzerland, Mr. Dunant wrote to the leaders of Europe. He told them what he had seen in Italy. He urged them to create an organization that could help the wounded in times of war. Mr. Dunant also formed a committee of friends, doctors, and lawyers. They held a conference in October, 1863. Governments from around Europe sent their representatives to the conference. By the end of it, they all agreed to help provide for better care to those wounded in wars. People who would help the wounded would also be protected. They would wear a white armband with a red cross to clearly show that they were not soldiers.

In 1864, Mr. Dunant and his committee held another conference. This time, representatives of governments outside of Europe came too. They signed an agreement that listed 10 rules which are called "articles". This guaranteed that all wounded soldiers would be treated with respect, even in times of war. This list of articles later became known as The Geneva Convention. Eventually, Mr. Dunant's organization became known as The International Committee of the Red Cross. It still exists today and it has helped millions of people around the world.

Text B: The First American Flag

Elizabeth "Betsy" Ross is known throughout America as the woman who designed the first flag of the United States. Ross was born in Philadelphia in 1742. She got famous in New York and Philadelphia as an excellent dress-maker. She used high quality materials such as silk. Moreover, she had a good sense of how colors mixed together. She especially enjoyed using bright primary colors in her designs. She decorated many public places and private businesses such as hotels and theaters.

Her work became so famous that the leaders of the American Revolution asked Betsy Ross to design the first flag of the United States. She chose her favorite colors – red, white, and blue. She made thirteen red and white stripes and a blue square field in the upper left corner. Upon the field was a picture of an eagle with thirteen stars. Each star on the flag represented a state. Betsy Ross chose red for the flag because it meant passion. The white meant being pure and innocent, and the blue represented unity and power.

Upon the new flag, Ross marked "The United States of America". Congress approved Ross's flag design before the new country was even called "The United States of America". Before that, it was called "Columbia" and the states were called "colonies". The new flag was finished and signed on July 4th, 1776. It was put up in the Hall of Independence for everyone

to see. The flag was a major part of the celebrations of American independence. The flag was officially adopted as the United States flag on June 14th, 1777.

The flag has changed over the years. A star has been added to the flag for each of the new states that has joined the country since then. Now there are 50 stars on the flag. Betsy Ross continued making flags for the United States until her death in 1832. She is widely regarded as an American hero.

Appendix G: Productive Vocabulary Levels Test (Laufer & Nation, 1999)

The 2000-word level

1. I'm glad we had this opp_____ to talk.

2. There are a doz_____ eggs in the basket.

3. Every working person must pay income t_____.

4. The pirates buried the trea_____ on a desert island.

5. Her beauty and cha_____ had a powerful effect on men.

6. La_____ of rain led to a shortage of water in the city.

7. He takes cr_____ and sugar in his coffee.

8. The rich man died and left all his we_____ to his son.

9. Pup_____ must hand in their papers by the end of the week.

10. This sweater is too tight. It needs to be stret_____.

11. Ann intro_____ her boyfriend to her mother.

12. Teenagers often adm_____ and worship pop singers.

13. If you blow up that balloon any more it will bur_____.

14. In order to be accepted into the university, he had to impr_____ his grades.

15. The telegram was deli_____ two hours after it had been sent.

16. The differences were so sl_____ that they went unnoticed.

17. The dress you're wearing is lov_____ .

18. He wasn't very popu_____ when he was a teenager, but he has many friends now.

Appendix H: Picture naming task item list

Apple	Eye	Onion
Banana	Fish	Pencil
Bed	Foot	Pipe
Belt	Football	Rabbit
Book	Fork	Ruler
Bowl	Frog	Sandwich
Bus	Glass	Shirt
Butterfly	Glove	Snake
Button	Guitar	Sock
Carrot	Hammer	Star
Cat	Hat	Sun
Chain	Heart	Swing
Door	Horse	Thumb
Drum	Iron	Tree
Duck	Key	Turtle
Ear	Lemon	Umbrella
Elephant	Nose	

Appendix I: Four sentence types in the Maze task

Declaratives.

- *The population of the world increases every year.*
- *My sister usually gets to school at eight in the morning.*

Wh-questions

- *Where are you going to put all the old pictures?*
- *How did the lady manage to learn four different languages?*

Relative clauses

- *The boy who is kissing the girl goes to a famous school.*
- *The lady knows the shop which is popular in Tokyo.*

Indirect question*s*

- *The boy wondered if he should take three classes at school.*
- *A lot of foreigners asked whether the train was going south.*

Appendix J: Item list of the GJT (Godfroid et al., 2015)

Syntax–Grammatical.

- *They enjoyed the party very much.*
- *I think that he is nicer and more intelligent than James.*

Syntax–Ungrammatical

- *Tom wanted to know whether was I going.*
- *The bird that my brother caught it has died.*

Morphology–Grammatical

- *I can cook Chinese food very well.*
- *Nate left some pens and pencils at school.*

Morphology–Ungrammatical

- *I must to brush my teeth now.*
- *Anthony live with his friend Kevin.*

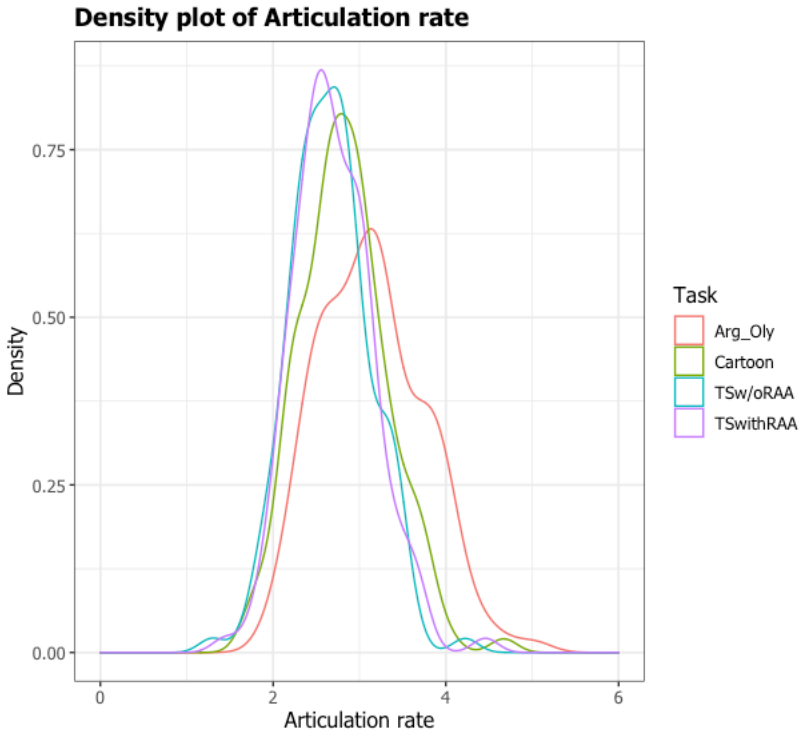
Appendix K: Script of the controlled speaking task (Weinberger, 2011)

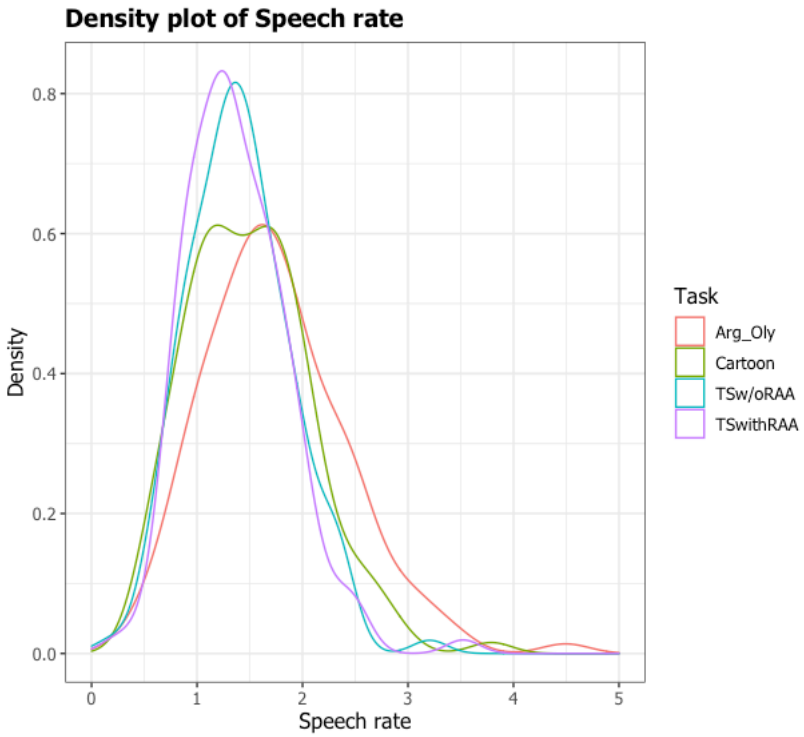
Please call Stella.
Ask her to bring these things with her from the store:
Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob.
We also need a small plastic snake and a big toy frog for the kids.
She can scoop these things into three red bags.
And we will go meet her Wednesday at the train station.

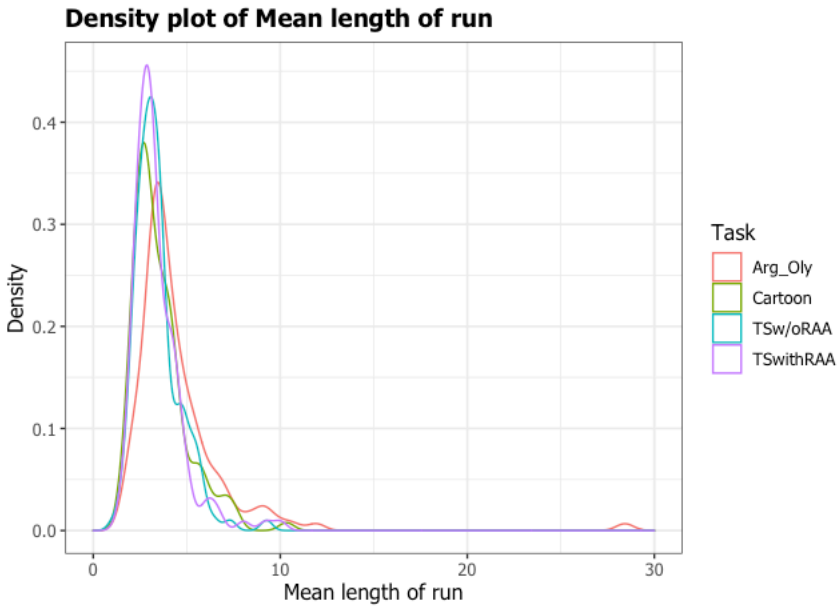
Appendix L: The contrast coding for the categorical variable of Task

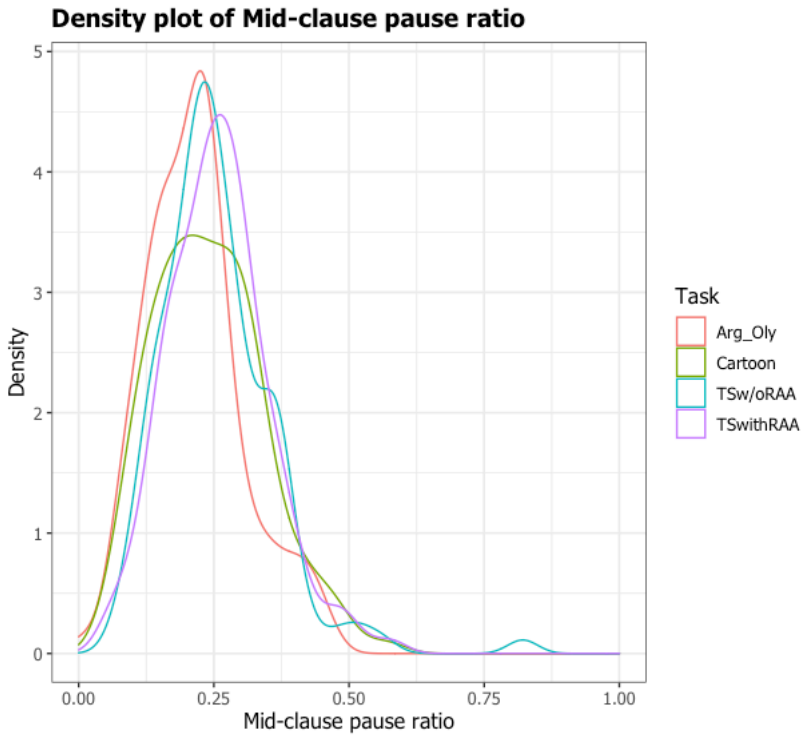
Task	Level 1 vs. 2	Level 2 vs. 3	Level 3 vs. 4
Level 1: Argumentative task	3/4	1/2	1/4
Level 2: Picture narrative task	-1/4	1/2	1/4
Level 3: Text summary task without RAA	-1/4	-1/2	1/4
Level 4: Text summary task with RAA	-1/4	-1/2	-3/4

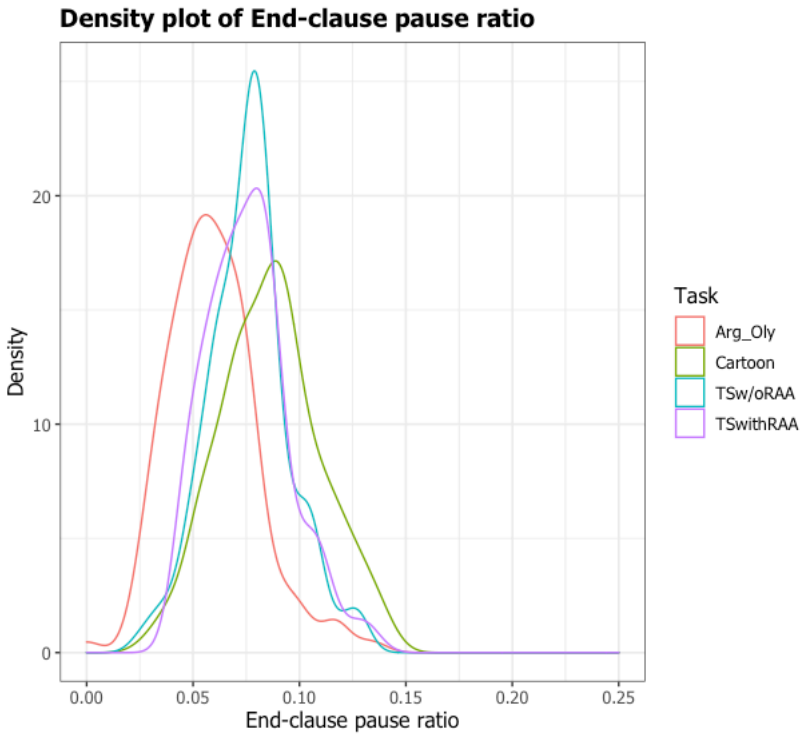
Appendix M: Density plots for L2 utterance fluency measures

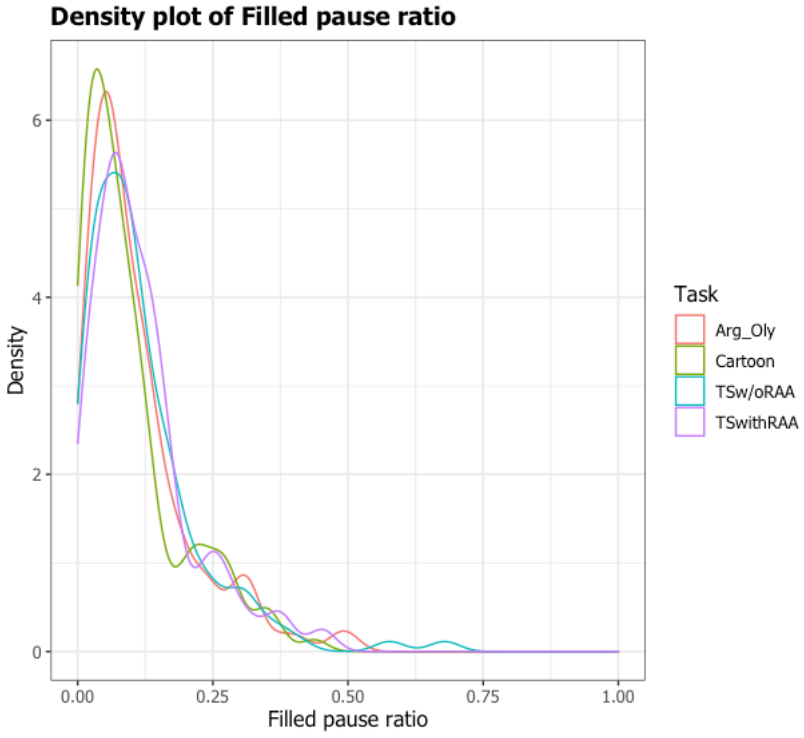


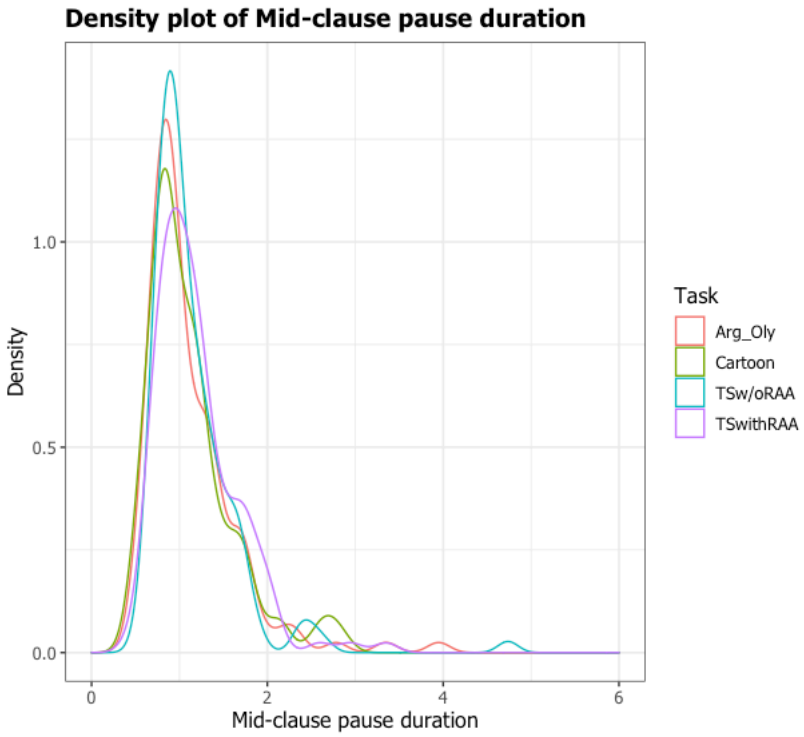


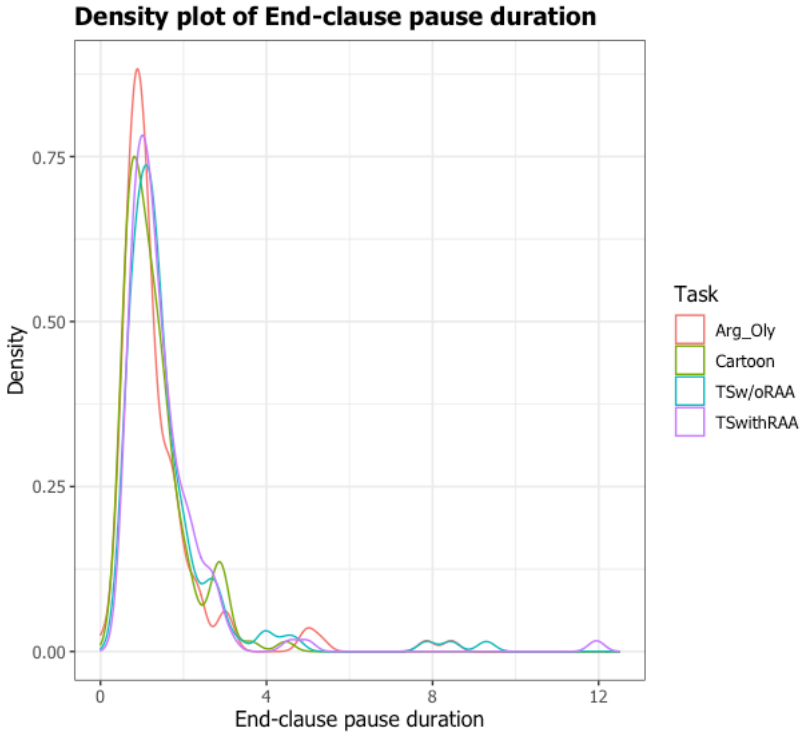

Density plot of Mean length of run
Density
0.0
0.1
0.2
0.3
0.4
0
10
20
30
Mean length of run
Task
Arg_Oly
Cartoon
TSw/oRAA
TSwithRAA

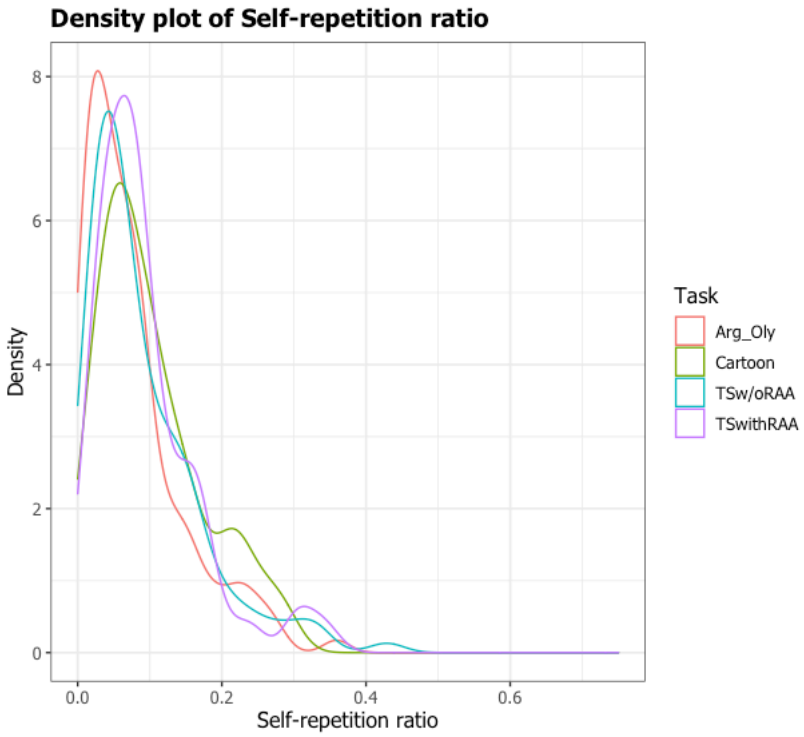

Density plot of Mid-clause pause ratio
Density
0
1
2
3
4
5
0.00
0.25
0.50
0.75
1.00
Mid-clause pause ratio
Task
Arg_Oly
Cartoon
TSw/oRAA
TSwithRAA

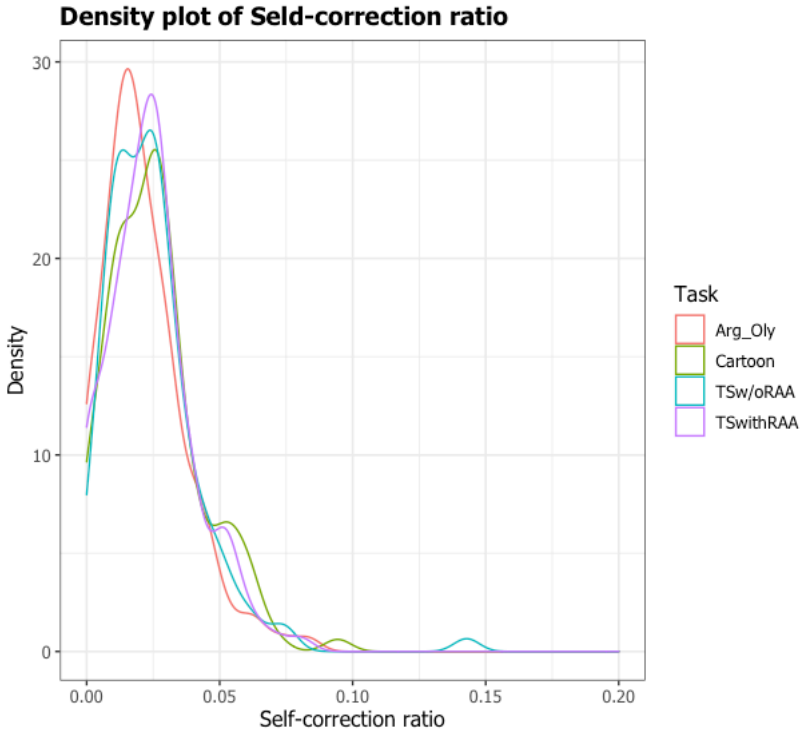

Density plot of End-clause pause ratio
Density
End-clause pause ratio
0
10
20
0.00
0.05
0.10
0.15
0.20
0.25
Task
Arg_Oly
Cartoon
TSw/oRAA
TSwithRAA

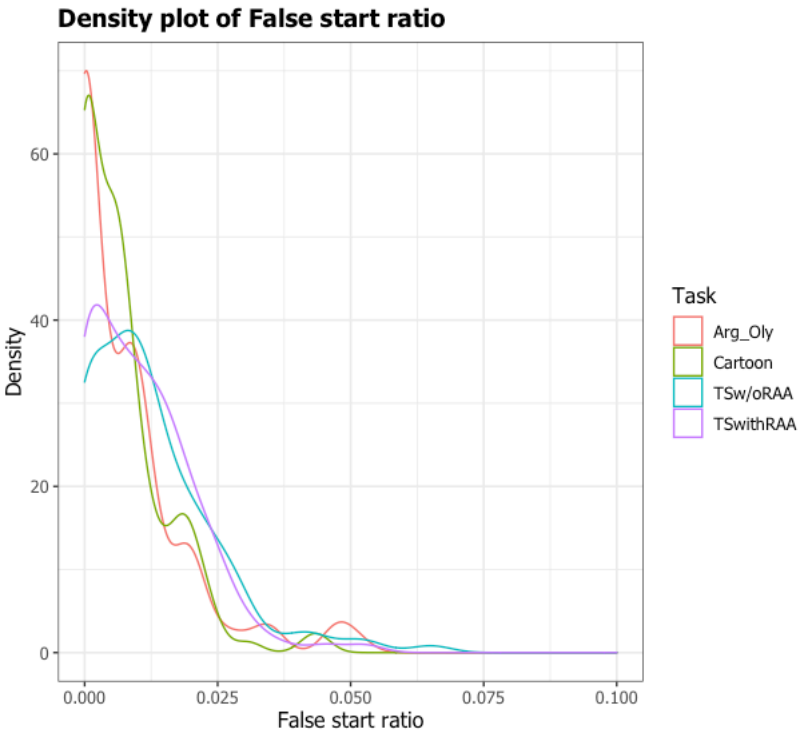

Density plot of Filled pause ratio
Density
Filled pause ratio
0
2
4
6
0.00
0.25
0.50
0.75
1.00
Task
Arg_Oly
Cartoon
TSw/oRAA
TSwithRAA


Density plot of Mid-clause pause duration
Density
1.0
0.5
0.0
0
2
4
6
Mid-clause pause duration
Task
Arg_Oly
Cartoon
TSw/oRAA
TSwithRAA


Density plot of End-clause pause duration
Density
0.75
0.50
0.25
0.00
0
4
8
12
End-clause pause duration
Task
Arg_Oly
Cartoon
TSw/oRAA
TSwithRAA


Density plot of Self-repetition ratio
Density
Self-repetition ratio
0
2
4
6
8
0.0
0.2
0.4
0.6
Task
Arg_Oly
Cartoon
TSw/oRAA
TSwithRAA


Density plot of Seld-correction ratio
Density
Self-correction ratio
0
10
20
30
0.00
0.05
0.10
0.15
0.20
Task
Arg_Oly
Cartoon
TSw/oRAA
TSwithRAA


Density plot of False start ratio
Density
60
40
20
0
0.000
0.025
0.050
0.075
0.100
False start ratio
Task
Arg_Oly
Cartoon
TSw/oRAA
TSwithRAA

Appendix N: Summary of statistical estimates of the GLMMs predicting utterance fluency measures from task type (Study 2)

Model summary for articulation rate

Gaussian without link formula			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	2.831	0.041	2.751	2.912	68.806	< .001
Arg vs. PicN	0.316	0.030	0.257	0.375	10.537	< .001
PicN vs. TS.withoutRAA	0.160	0.030	0.101	0.219	5.329	< .001
TS.withoutRAA vs. TS. With.RAA	-0.032	0.030	-0.091	0.026	-1.077	0.282
Random effects (intercepts)	Variance	SD				
Participants	0.202	0.450				
Information criterion						
LogLikelihood	-167.2					
DIC	334.4					
AIC	346.4					
BIC	371.8					
R2	Estimate					
Marginal	0.120					
Conditional	0.805					

Model summary for speech rate

Gamma (link = log)			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	0.328	0.045	0.240	0.416	7.279	< .001
Arg vs. PicN	0.152	0.018	0.117	0.187	8.521	< .001
PicN vs. TS.withoutRAA	0.059	0.018	0.024	0.094	3.286	0.001
TS.withoutRAA vs. TS. With.RAA	0.010	0.018	-0.025	0.045	0.562	0.574
Random effects (intercepts)	Variance	SD				
Participants	0.058	0.241				
Information criterion						
LogLikelihood	-35.8					
DIC	71.5					
AIC	83.5					
BIC	109.0					
R2	Estimate					
Marginal	0.078					
Conditional	0.653					

Model summary for MLR

Gamma (link = log)			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	1.234	0.040	1.155	1.313	30.655	< .001
Arg vs. PicN	0.211	0.019	0.175	0.248	11.303	< .001
PicN vs. TS.withoutRAA	0.021	0.019	-0.015	0.058	1.148	0.251
TS.withoutRAA vs. TS. With.RAA	0.009	0.019	-0.028	0.046	0.480	0.631
Random effects (intercepts)	Variance	SD				
Participants	0.053	0.231				
Information criterion						
LogLikelihood	-510.5					
DIC	1021.0					
AIC	1033.0					
BIC	1058.4					
R2	Estimate					
Marginal	0.098					
Conditional	0.619					

Model summary for Mid-clause pause ratio

Gamma (link = log)			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-1.507	0.048	-1.601	-1.413	-31.407	< .001
Arg vs. PicN	-0.117	0.023	-0.163	-0.071	-4.972	< .001
PicN vs. TS.withoutRAA	-0.072	0.023	-0.118	-0.026	-3.068	0.002
TS.withoutRAA vs. TS. With.RAA	-0.018	0.023	-0.064	0.028	-0.754	0.451
Random effects (intercepts)	Variance	SD				
Participants	0.068	0.261				
Information criterion						
LogLikelihood	776.7					
DIC	-1553.3					
AIC	-1541.3					
BIC	-1515.9					
R2	Estimate					
Marginal	0.052					
Conditional	0.594					

Model summary for End-clause pause ratio

Gamma (link = log)			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-2.640	0.028	-2.695	-2.585	-94.164	< .001
Arg vs. PicN	-0.381	0.024	-0.428	-0.334	-15.821	< .001
PicN vs. TS.withoutRAA	0.117	0.024	0.070	0.164	4.881	< .001
TS.withoutRAA vs. TS. With.RAA	0.005	0.024	-0.042	0.052	0.210	0.834
Random effects (intercepts)	Variance	SD				
Participants	0.029	0.170				
Information criterion						
LogLikelihood	1389.7					
DIC	-2779.4					
AIC	-2767.4					
BIC	-2742.0					
R2	Estimate					
Marginal	0.204					
Conditional	0.507					

Model summary for Filled pause ratio

Gamma (link = log)			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-2.541	0.087	-2.711	-2.371	-29.274	<
Arg vs. PicN	0.191	0.050	0.093	0.289	3.819	< .001
PicN vs. TS.withoutRAA	-0.208	0.049	-0.305	-0.111	-4.205	< .001
TS.withoutRAA vs. TS. With.RAA	-0.065	0.049	-0.161	0.032	-1.312	0.190
Random effects (intercepts)	Variance	SD				
Participants	0.315	0.561				
Information criterion						
LogLikelihood	919.8					
DIC	-1839.6					
AIC	-1827.6					
BIC	-1802.2					
R2	Estimate					
Marginal	0.019					
Conditional	0.607					

Model summary for Mid-clause pause duration

Gamma (link = log)			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	0.065	0.038	-0.010	0.140	1.711	0.087
Arg vs. PicN	0.003	0.023	-0.043	0.049	0.112	0.911
PicN vs. TS.withoutRAA	-0.027	0.023	-0.073	0.019	-1.161	0.246
TS.withoutRAA vs. TS. With.RAA	-0.054	0.023	-0.100	-0.008	-2.315	0.021
Random effects (intercepts)	Variance	SD				
Participants	0.053	0.230				
Information criterion						
LogLikelihood	-9.4					
DIC	18.8					
AIC	30.8					
BIC	56.2					
R2	Estimate					
Marginal	0.010					
Conditional	0.502					

Model summary for End-clause pause duration

Gamma (link = log)			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	0.201	0.056	0.092	0.311	3.606	< .001
Arg vs. PicN	0.003	0.032	-0.059	0.065	0.081	0.935
PicN vs. TS.withoutRAA	-0.149	0.032	-0.211	-0.087	-4.735	< .001
TS.withoutRAA vs. TS. With.RAA	0.024	0.032	-0.038	0.086	0.750	0.454
Random effects (intercepts)	Variance	SD				
Participants	0.124	0.353				
Information criterion						
LogLikelihood	-244.3					
DIC	488.7					
AIC	500.7					
BIC	526.1					
R2	Estimate					
Marginal	0.020					
Conditional	0.552					

Model summary for Self-repetition ratio

Gamma (link = log)			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-2.679	0.075	-2.826	-2.533	-35.810	< .001
Arg vs. PicN	-0.415	0.064	-0.542	-0.289	-6.440	< .001
PicN vs. TS.withoutRAA	0.183	0.064	0.057	0.309	2.846	0.004
TS.withoutRAA vs. TS. With.RAA	-0.110	0.064	-0.236	0.015	-1.722	0.085
Random effects (intercepts)	Variance	SD				
Participants	0.253	0.503				
Information criterion						
LogLikelihood	905.7					
DIC	-1811.4					
AIC	-1799.4					
BIC	-1774.0					
R2	Estimate					
Marginal	0.045					
Conditional	0.505					

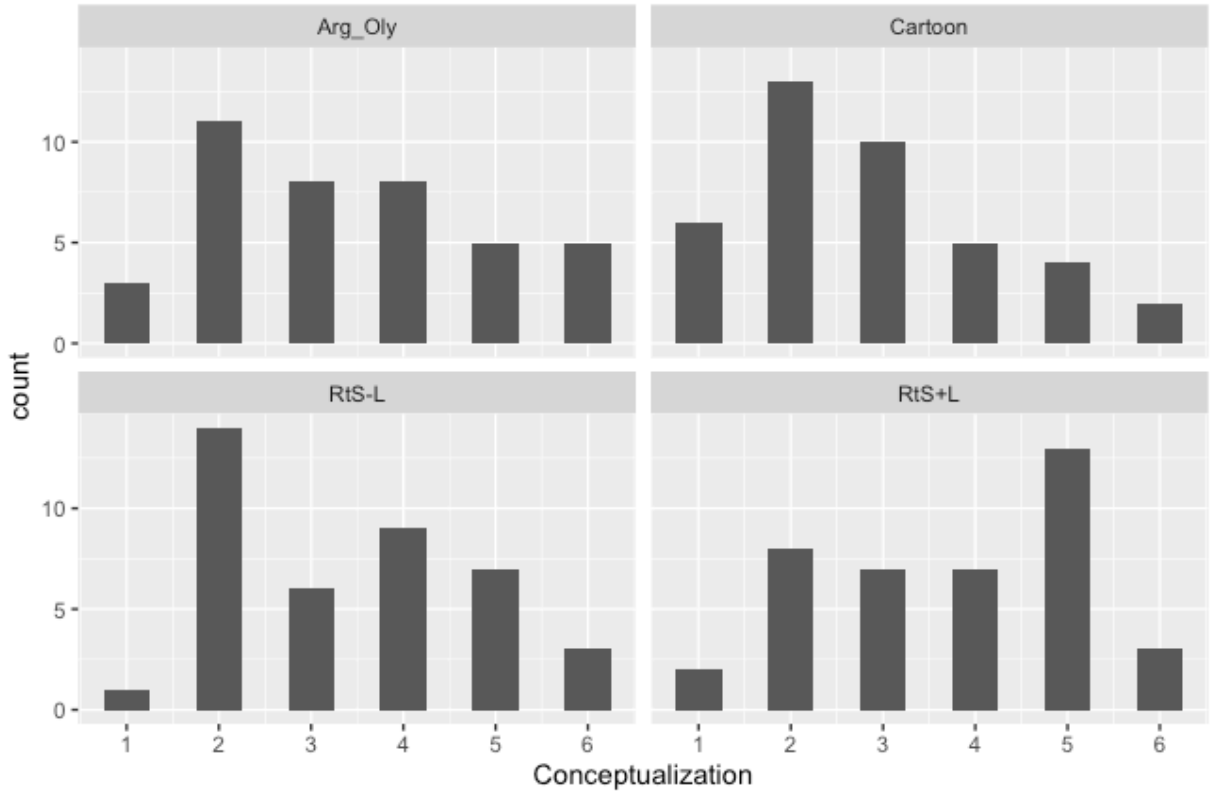
Model summary for Self-correction ratio

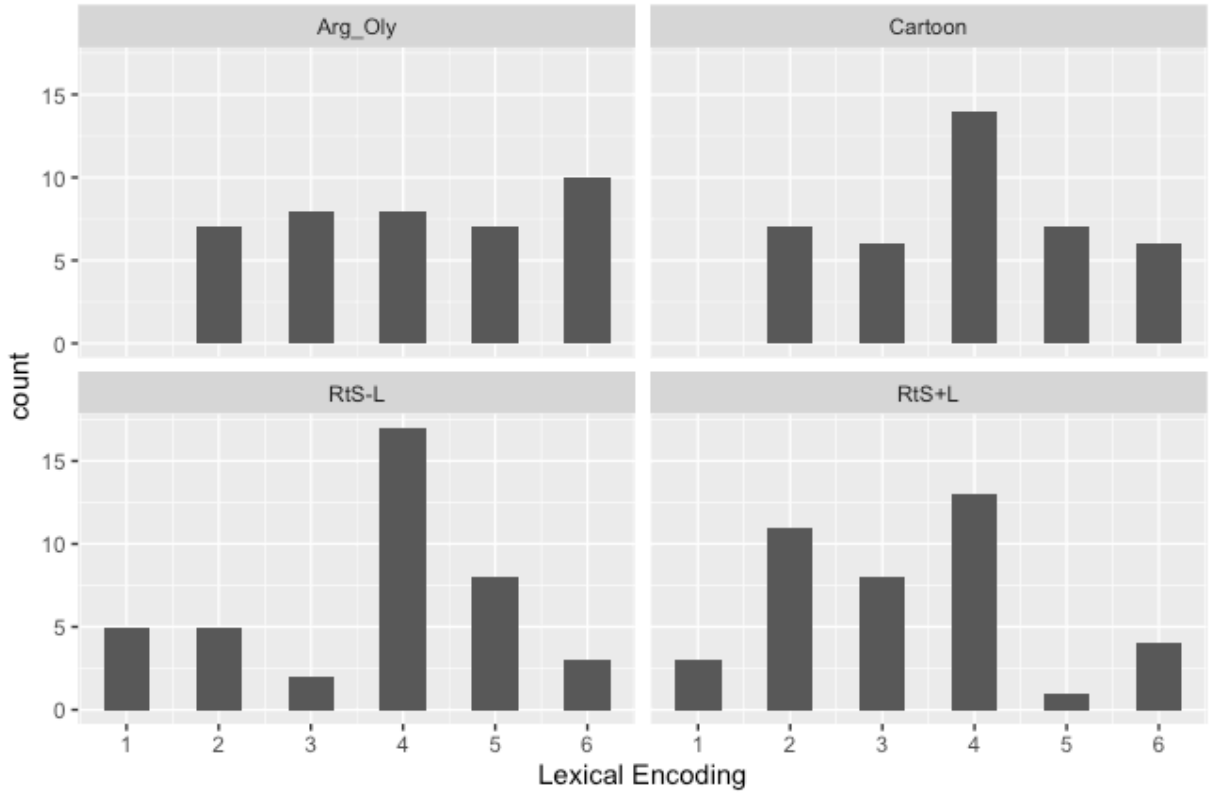
Gamma (link = log)			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-3.834	0.049	-3.931	-3.737	-77.518	< .001
Arg vs. PicN	-0.164	0.077	-0.315	-0.013	-2.133	0.033
PicN vs. TS.withoutRAA	0.039	0.077	-0.112	0.189	0.505	0.614
TS.withoutRAA vs. TS. With.RAA	-0.007	0.077	-0.157	0.144	-0.086	0.931
Random effects (intercepts)	Variance	SD				
Participants	0.106	0.325				
Information criterion						
LogLikelihood	1487.3					
DIC	-2974.6					
AIC	-2962.6					
BIC	-2937.2					
R2	Estimate					
Marginal	0.009					
Conditional	0.243					

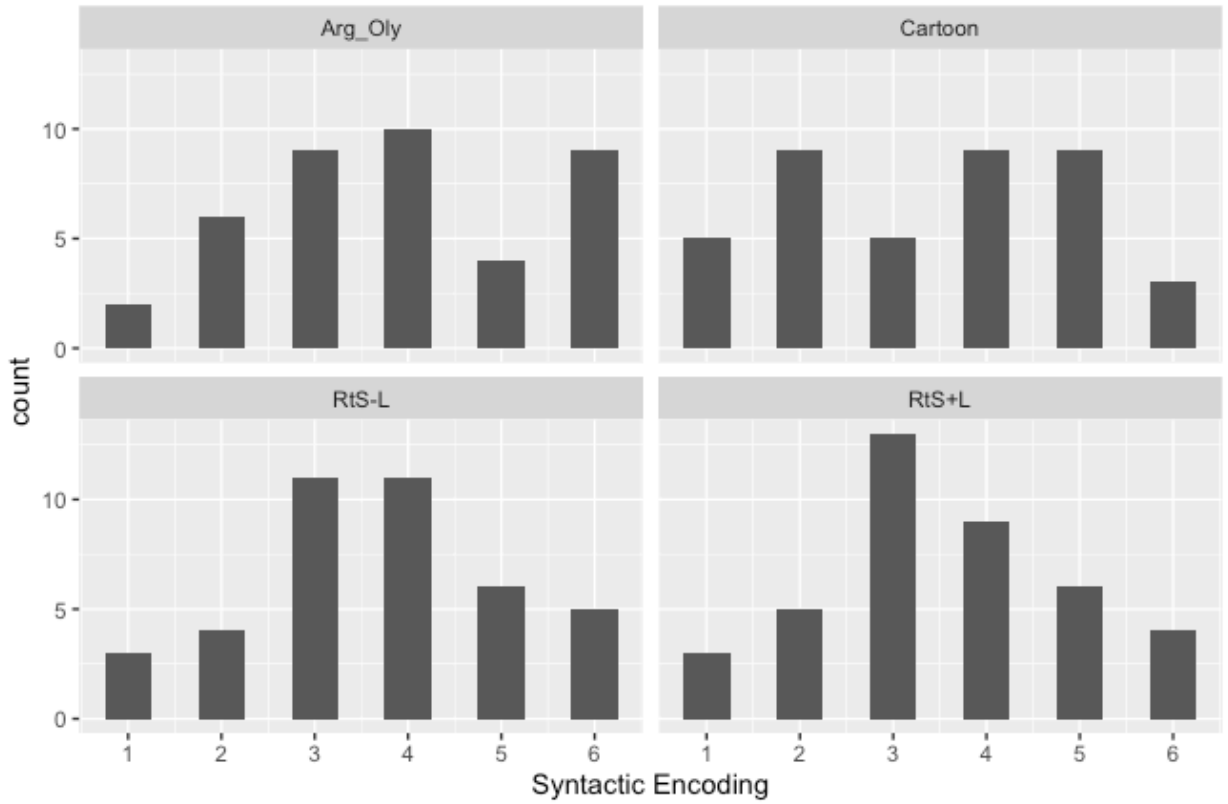
Model summary for False start ratio

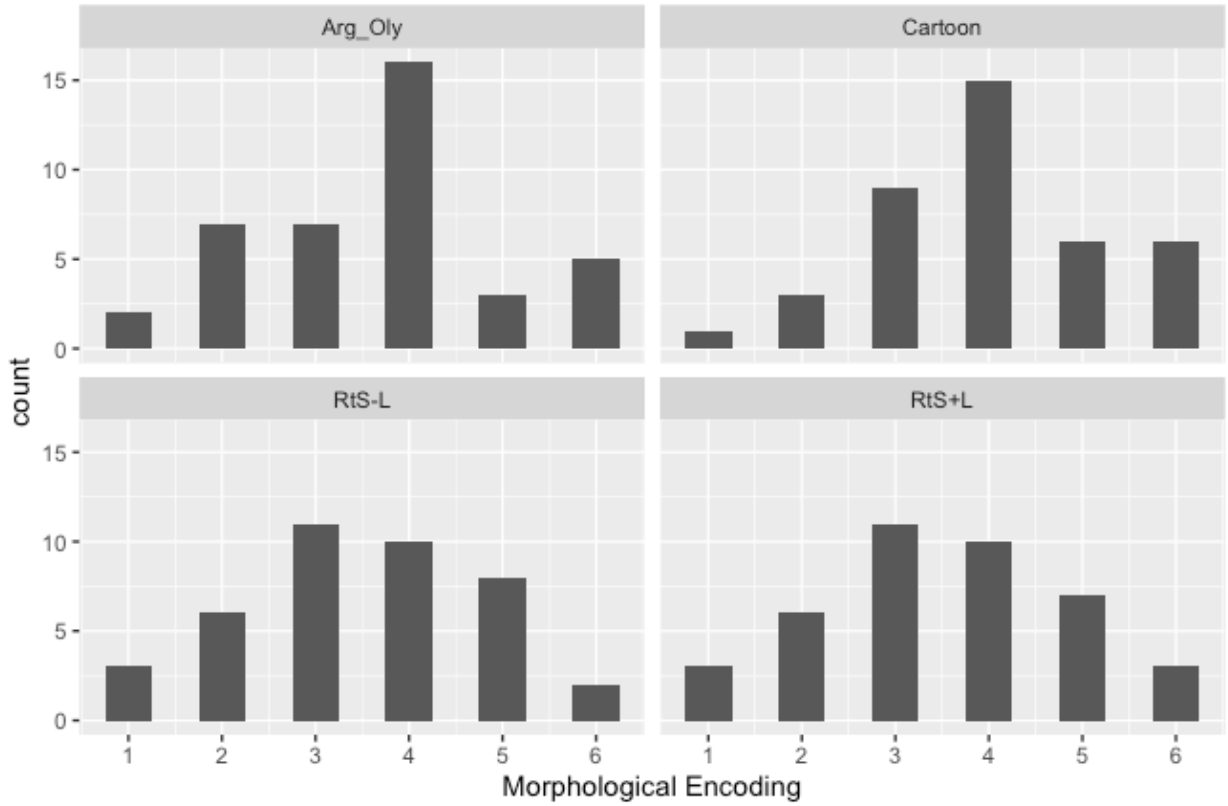
Gamma (link = log)			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-4.832	0.064	-4.958	-4.706	-75.150	< .001
Arg vs. PicN	0.062	0.115	-0.163	0.287	0.538	0.591
PicN vs. TS.withoutRAA	-0.586	0.115	-0.812	-0.360	-5.088	< .001
TS.withoutRAA vs. TS. With.RAA	0.131	0.115	-0.094	0.355	1.142	0.254
Random effects (intercepts)	Variance	SD				
Participants	0.257	0.507				
Information criterion						
LogLikelihood	1882.1					
DIC	-3764.2					
AIC	-3752.2					
BIC	-3726.7					
R2	Estimate					
Marginal	0.057					
Conditional	0.289					

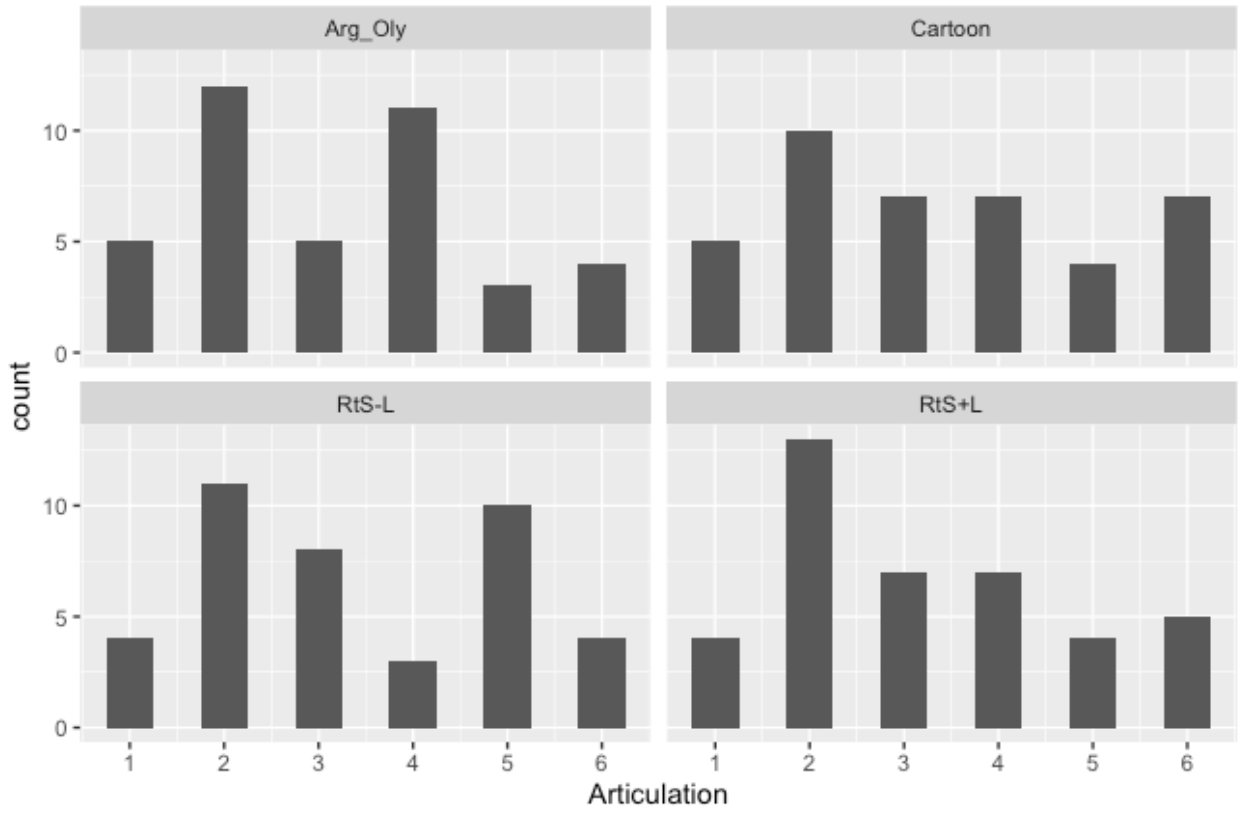
Appendix O: The histograms of post-speaking performance questionnaire items






Arg_Oly
Cartoon
RtS-L
RtS+L
count
0
5
10
1
2
3
4
5
6
Syntactic Encoding


Arg_Oly
Cartoon
RtS-L
RtS+L
count
0
5
10
15
1
2
3
4
5
6
Morphological Encoding


Arg_Oly
Cartoon
RtS-L
RtS+L
count
0
5
10
1
2
3
4
5
6
Articulation

Appendix P: The lavaan syntax for the CFA models of CF and UF, and the SEM model of CF-UF link in Study 3

Model.CF.1

```
CF.cfamodel.1 <- '
CF =~ zPVLT +
      zPicNamingRT.cleaned.inversed +
      zArtcSpeed +
      zMazeWordRT.inversed +
      zMazeWordAccuracy +
      zMorphRT.inversed +
      zSynRT.inversed +
      zMorphAccuracy +
      zSynAccuracy
# Residual covariances across tasks: Maze task
zMazeWordRT.inversed ~~ zMazeWordAccuracy
# Residual covariances across tasks: GJT
zMorphAccuracy ~~ zSynAccuracy
zMorphAccuracy ~~ zMorphRT.inversed
zMorphAccuracy ~~ zSynRT.inversed
zSynAccuracy ~~ zMorphRT.inversed
zSynAccuracy ~~ zSynRT.inversed
zSynRT.inversed ~~ zMorphRT.inversed
'
```

Model.CF.2 (The final CFA model of CF)

```
CF.cfamodel.2 <- '
CF1 =~ zPVLT +
      zMazeWordAccuracy +
      zMorphAccuracy +
      zSynAccuracy
CF2 =~ zPicNamingRT.cleaned.inversed +
      zArtcSpeed +
      zMazeWordRT.inversed +
      zMorphRT.inversed +
      zSynRT.inversed
CF1 ~~ CF2

# Residual covariances across tasks: Maze task
zMazeWordRT.inversed ~~ zMazeWordAccuracy
# Residual covariances across tasks: GJT
zMorphAccuracy ~~ zSynAccuracy
zMorphAccuracy ~~ zMorphRT.inversed
zMorphAccuracy ~~ zSynRT.inversed
zSynAccuracy ~~ zMorphRT.inversed
zSynAccuracy ~~ zSynRT.inversed
zSynRT.inversed ~~ zMorphRT.inversed
'
```

Model.CF.3

```
CF.cfamodel.3 <- '
CF1 =~ zPVLT +
      zMazeWordAccuracy +
      zMorphAccuracy +
      zSynAccuracy
CF2 =~ zPicNamingRT.cleaned.inversed +
      zArtcSpeed +
      zMazeWordRT.inversed
CF3 =~ zMorphRT.inversed +
      zSynRT.inversed
CF1 ~~ CF2
CF1 ~~ CF3
CF2 ~~ CF3
# Residual covariances across tasks: Maze task
zMazeWordRT.inversed ~~ zMazeWordAccuracy
# Residual covariances across tasks: GJT
zMorphAccuracy ~~ zSynAccuracy
zMorphAccuracy ~~ zMorphRT.inversed
zMorphAccuracy ~~ zSynRT.inversed
zSynAccuracy ~~ zMorphRT.inversed
zSynAccuracy ~~ zSynRT.inversed
zSynRT.inversed ~~ zMorphRT.inversed
'
```

Model.UF.1

```
UF.cfamodel.1 <- '
UF =~ zSR +
      zAR +
      zMLR +
      zMCPR.inversed +
      zFCPR.inversed +
      zMCPD.inversed +
      zFCPD.inversed +
      zFilledPauseRatio.inversed +
      zRepetitionRatio.inversed +
      zSelfRepairRatio.inversed +
      zFalseStartRatio.inversed
# residual covariances within the same subconstruct
#Pause freq.
zMCPR.inversed ~~ zFilledPauseRatio.inversed
zFCPR.inversed ~~ zFilledPauseRatio.inversed
zMCPD.inversed ~~ zFCPD.inversed
'
```

Model.UF.2

```
UF.cfamodel.2 <- '
SF =~ zSR +
      zAR +
```

```
      zMLR
DysF =~ zMCPR.inversed +
      zFCPR.inversed +
      zMCPD.inversed +
      zFCPD.inversed +
      zFilledPauseRatio.inversed +
      zRepetitionRatio.inversed +
      zSelfRepairRatio.inversed +
      zFalseStartRatio.inversed
SF ~~ DysF
# residual covariances within the same subconstruct
#Pause freq.
zMCPR.inversed ~~ zFilledPauseRatio.inversed
zFCPR.inversed ~~ zFilledPauseRatio.inversed
zMCPD.inversed ~~ zFCPD.inversed
'
```

Model.UF.3

```
UF.cfamodel.3 <- '
SF =~ zSR +
      zAR +
      zMLR
BDF =~ zMCPR.inversed +
      zFCPR.inversed +
      zMCPD.inversed +
      zFCPD.inversed +
      zFilledPauseRatio.inversed

RF =~ zRepetitionRatio.inversed +
      zSelfRepairRatio.inversed +
      zFalseStartRatio.inversed
SF ~~ BDF
SF ~~ RF
BDF ~~ RF
# residual covariances within the same subconstruct
#Pause freq.
zMCPR.inversed ~~ zFilledPauseRatio.inversed
zFCPR.inversed ~~ zFilledPauseRatio.inversed
zMCPD.inversed ~~ zFCPD.inversed
'
```

Model.UF.4

```
UF.cfamodel.1.1 <- '
UF =~ zAR +
      zMLR +
      zMCPR.inversed +
      zFCPR.inversed +
      zMPD.inversed +
      zFilledPauseRatio.inversed +
```

```
      zRepetitionRatio.inversed +
      zSelfRepairRatio.inversed +
      zFalseStartRatio.inversed
# residual covariances within the same subconstruct
#Pause freq.
zMPD.inversed ~~ zFilledPauseRatio.inversed
zMCPR.inversed ~~ zSelfRepairRatio.inversed
zFCPR.inversed ~~ zFalseStartRatio.inversed
'
```

Model.UF.5

```
UF.cfamodel.2.1 <- '
SF =~ zAR +
      zMLR
DysF =~ zMCPR.inversed +
      zFCPR.inversed +
      zMPD.inversed +
      zFilledPauseRatio.inversed +
      zRepetitionRatio.inversed +
      zSelfRepairRatio.inversed +
      zFalseStartRatio.inversed
SF ~~ DysF
# residual covariances within the same subconstruct
#Pause freq.
zMPD.inversed ~~ zFilledPauseRatio.inversed
zMCPR.inversed ~~ zSelfRepairRatio.inversed
zFCPR.inversed ~~ zFalseStartRatio.inversed
'
```

Model.UF.6 (The final CFA model of UF)

```
UF.cfamodel.3.1 <- '
SF =~ zAR +
      zSR
BDF =~ zMCPR.inversed +
      zFCPR.inversed +
      zMPD.inversed +
      zFilledPauseRatio.inversed
RF =~ zRepetitionRatio.inversed +
      zSelfRepairRatio.inversed +
      zFalseStartRatio.inversed
SF ~~ BDF
SF ~~ RF
BDF ~~ RF
# residual covariances within the same subconstruct
#Pause freq.
zMPD.inversed ~~ zFilledPauseRatio.inversed
zMCPR.inversed ~~ zSelfRepairRatio.inversed
zFCPR.inversed ~~ zFalseStartRatio.inversed
'
```

Model.UF.7

```
UF.cfamodel.4 <- '
OF =~ zMLR +
      zAR +
      zMCPR.inversed +
      zFCPR.inversed +
      zMPD.inversed +
      zFilledPauseRatio.inversed
RF =~ zRepetitionRatio.inversed +
      zSelfRepairRatio.inversed +
      zFalseStartRatio.inversed
OF ~~ RF
# residual covariances within the same subconstruct
#Pause freq.
zMPD.inversed ~~ zFilledPauseRatio.inversed
zMCPR.inversed ~~ zSelfRepairRatio.inversed
zFCPR.inversed ~~ zFalseStartRatio.inversed
'
```

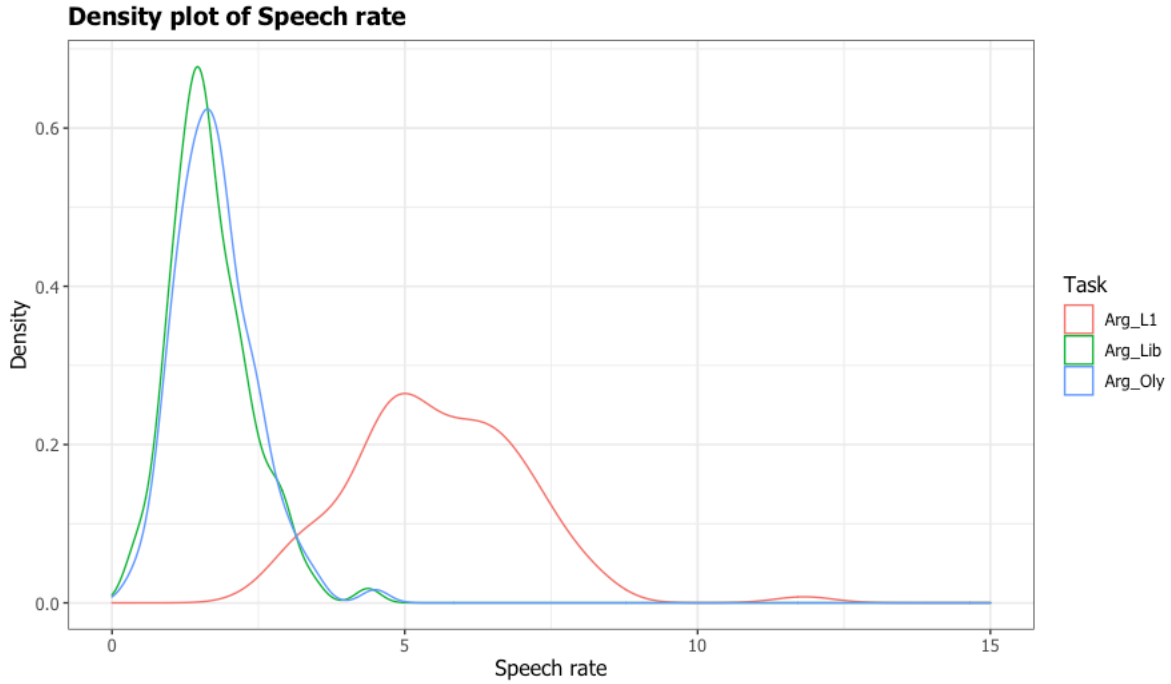
The final SEM model

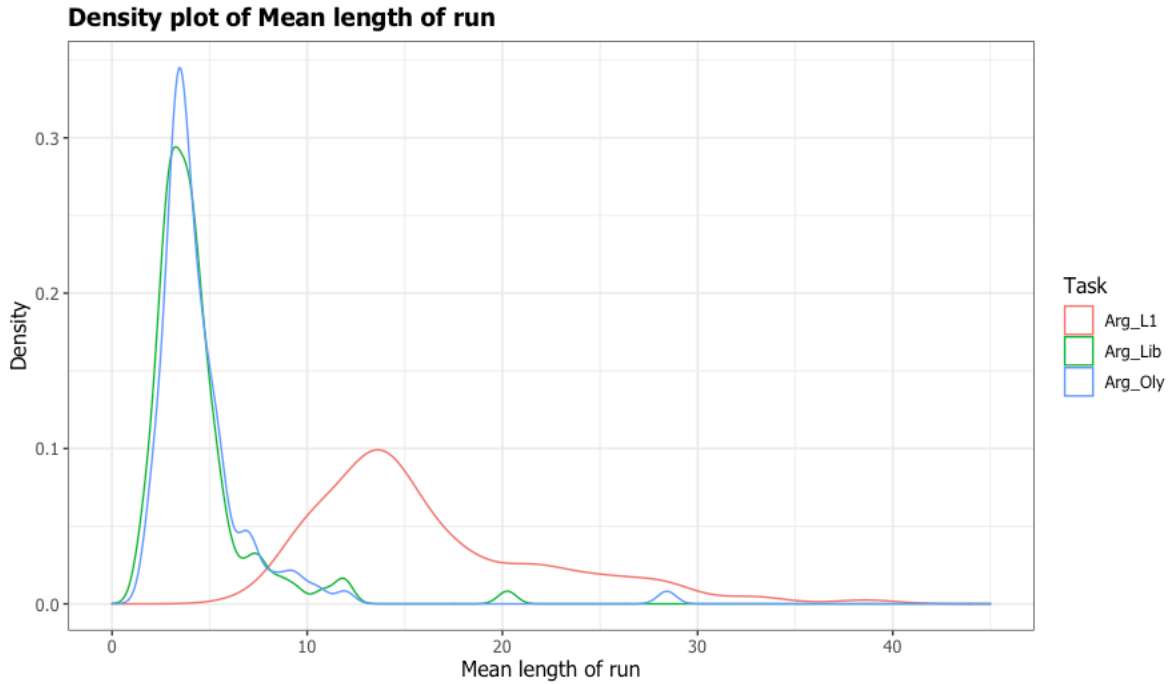
```
SEM.cfamodel.1 <- '
CF1 =~ zPVLT +
  zMazeWordAccuracy +
  zMorphAccuracy +
  zSynAccuracy
CF2 =~ zPicNamingRT.cleaned.inversed +
  zArtcSpeed +
  zMazeWordRT.inversed +
  zMorphRT.inversed +
  zSynRT.inversed
SF =~ zAR +
      zMLR
BDF =~ zMCPR.inversed +
      zFCPR.inversed +
      zMPD.inversed +
      zFilledPauseRatio.inversed
RF =~ zRepetitionRatio.inversed +
      zSelfRepairRatio.inversed +
      zFalseStartRatio.inversed
SF ~ CF1 + CF2
BDF ~ CF1 + CF2
RF ~ CF1 + CF2
# residual covariances within the same subconstruct
#Pause freq.
zMPD.inversed ~~ zFilledPauseRatio.inversed
zMCPR.inversed ~~ zSelfRepairRatio.inversed
zFCPR.inversed ~~ zFalseStartRatio.inversed
# Residual covariances across tasks: Maze task
```

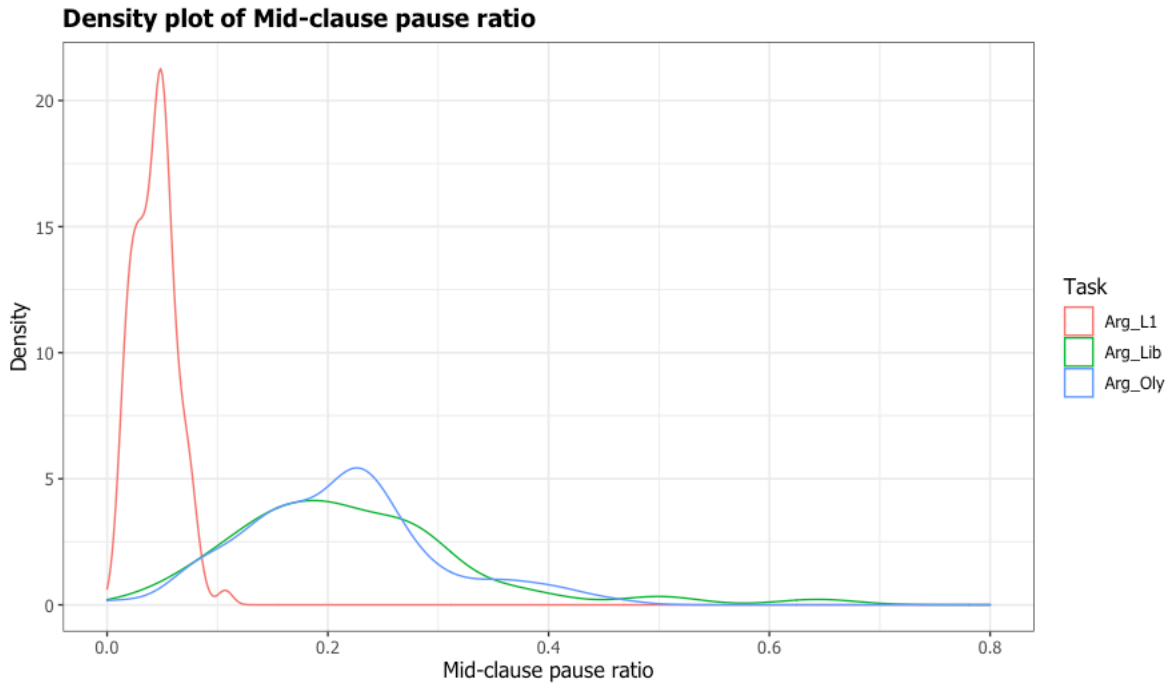
```
zMazeWordRT.inversed ~~ zMazeWordAccuracy
# Residual covariances across tasks: GJT
zMorphAccuracy ~~ zSynAccuracy
zMorphAccuracy ~~ zMorphRT.inversed
zMorphAccuracy ~~ zSynRT.inversed
zSynAccuracy ~~ zMorphRT.inversed
zSynAccuracy ~~ zSynRT.inversed
zSynRT.inversed ~~ zMorphRT.inversed
# Added from MI index (>10)
zMLR ~~ zFCPR.inversed
zMCPR.inversed ~~ zFCPR.inversed
zMLR ~~ zMCPR.inversed
zArtcSpeed ~~ zAR
'
```

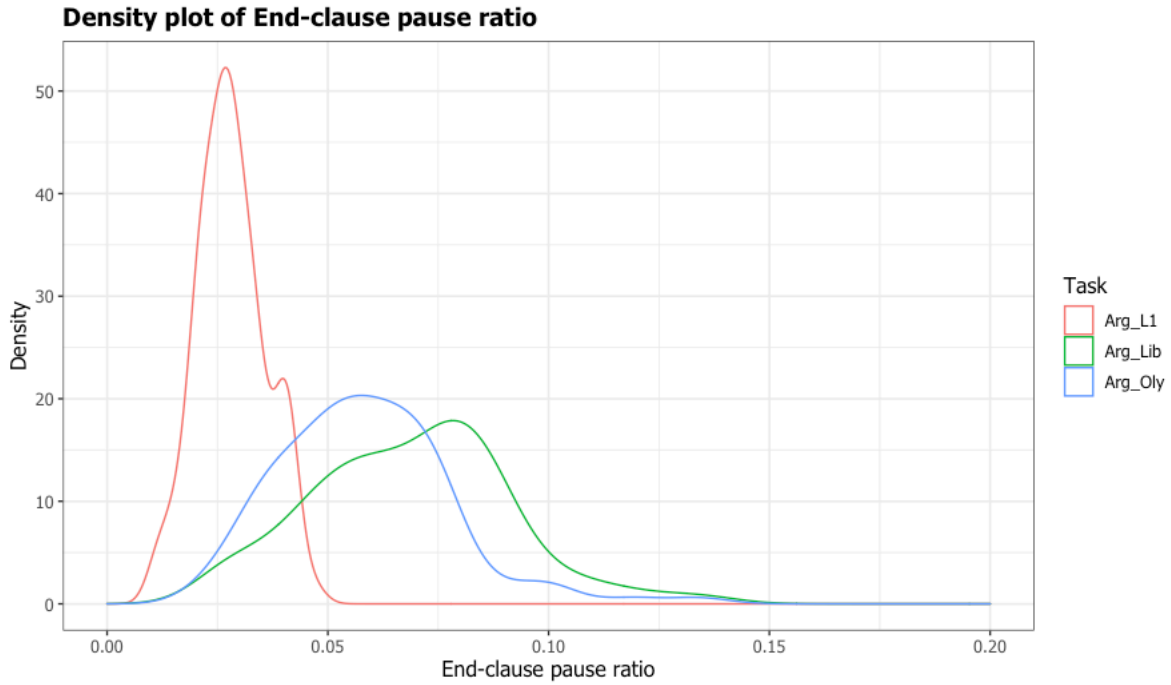
Appendix Q: Density plots for L1 and L2 utterance fluency measures

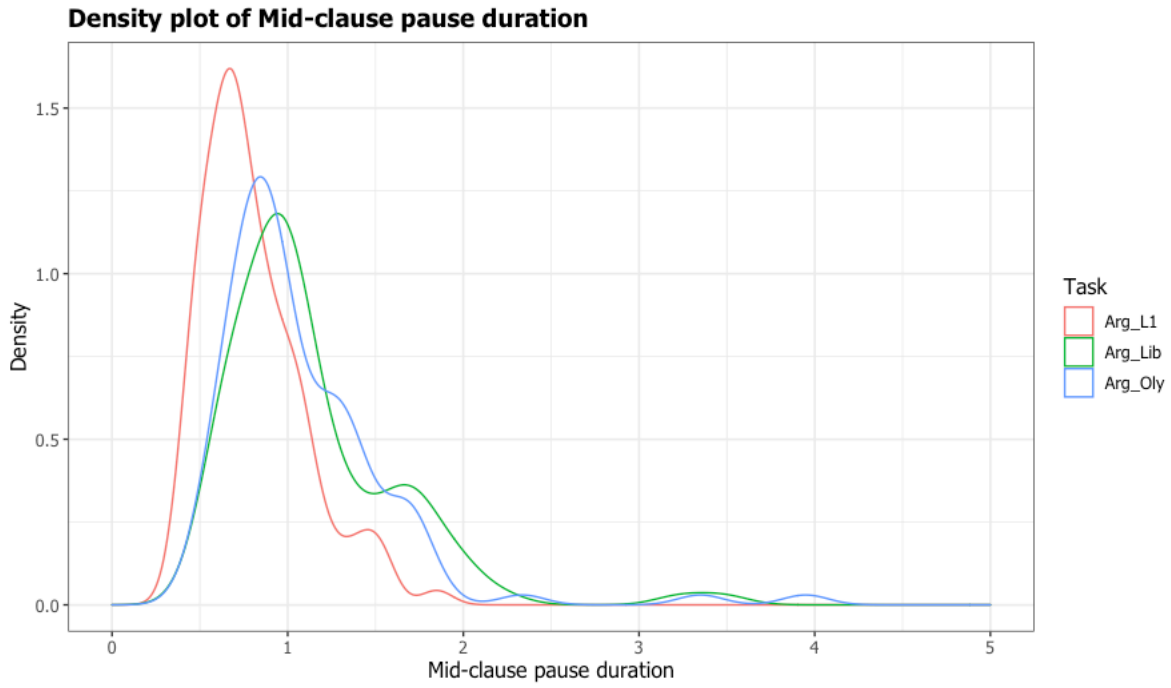
For the density plot of articulation rate, see Figure 30.

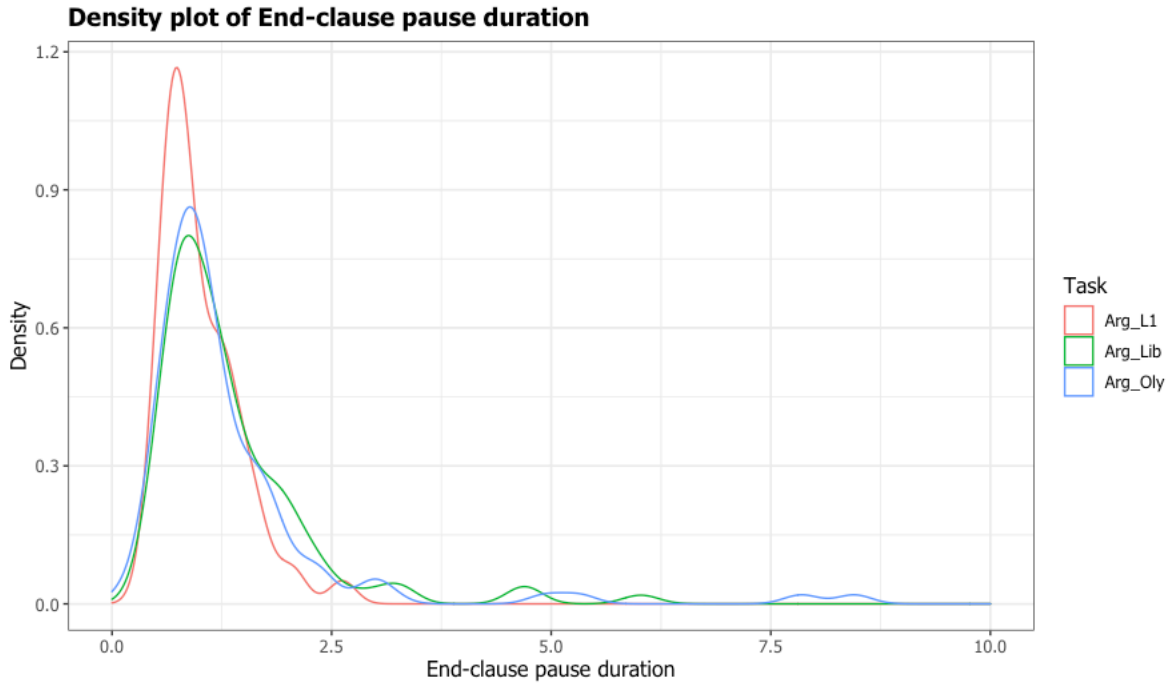


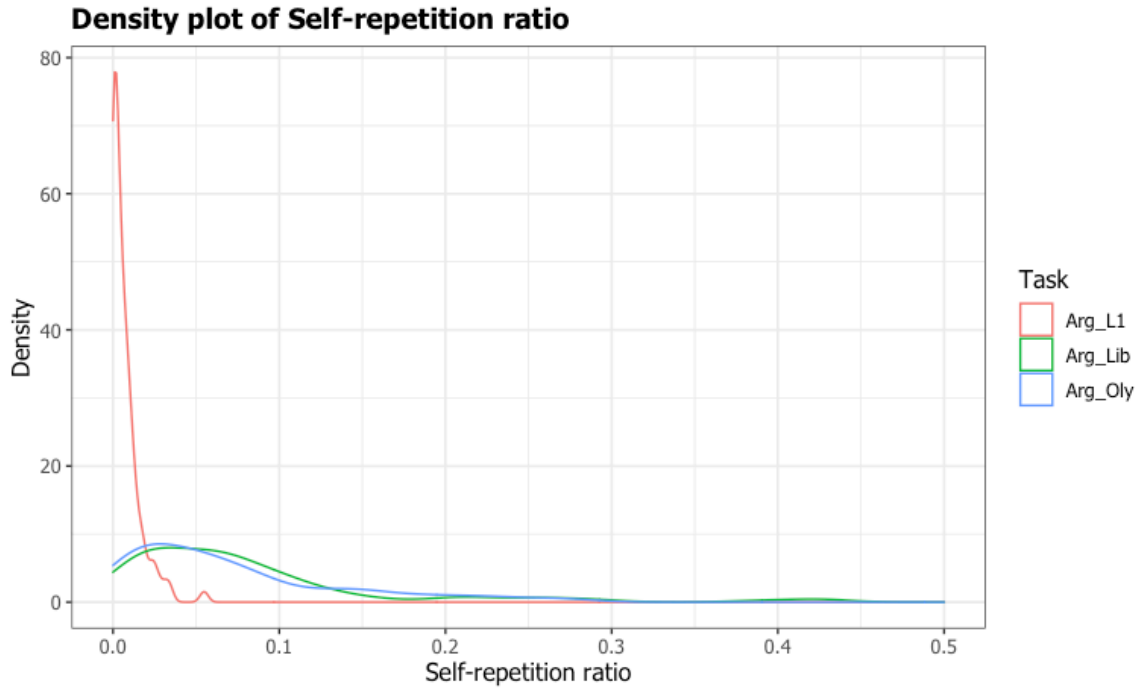


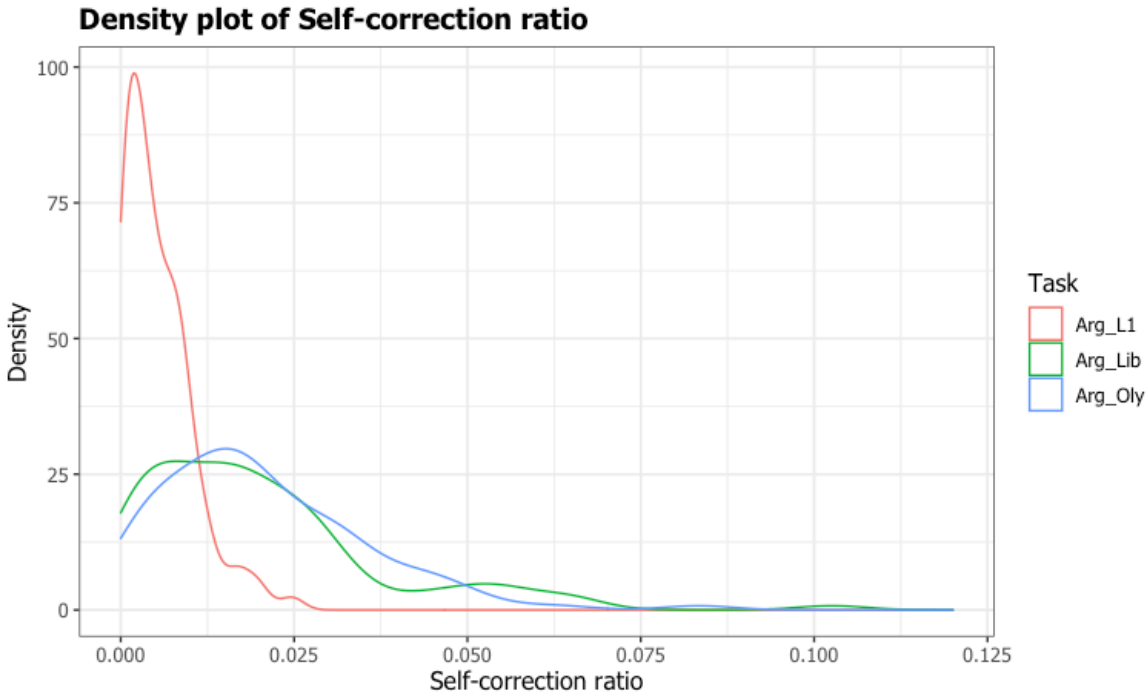

Density plot of Mid-clause pause ratio
Density
Mid-clause pause ratio
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
Task
Arg_L1
Arg_Lib
Arg_Oly


Density plot of End-clause pause ratio
Density
End-clause pause ratio
0
10
20
30
40
50
0.00
0.05
0.10
0.15
0.20
Task
Arg_L1
Arg_Lib
Arg_Oly


Density plot of Mid-clause pause duration
Density
1.5
1.0
0.5
0.0
0
1
2
3
4
5
Mid-clause pause duration
Task
Arg_L1
Arg_Lib
Arg_Oly


Density plot of End-clause pause duration
Density
1.2
0.9
0.6
0.3
0.0
0.0
2.5
5.0
7.5
10.0
End-clause pause duration
Task
Arg_L1
Arg_Lib
Arg_Oly


Density plot of Self-repetition ratio
Density
Self-repetition ratio
0
20
40
60
80
0.0
0.1
0.2
0.3
0.4
0.5
Task
Arg_L1
Arg_Lib
Arg_Oly


Density plot of Self-correction ratio
Density
Self-correction ratio
0
25
50
75
100
0.000
0.025
0.050
0.075
0.100
0.125
Task
Arg_L1
Arg_Lib
Arg_Oly


Density plot of False start ratio
Density
600
400
200
0
0.00
0.02
0.04
0.06
0.08
False start ratio
Task
Arg_L1
Arg_Lib
Arg_Oly

Appendix R: Summary of statistical estimates of the GLMMs predicting L2 utterance fluency measures from the corresponding L1 measures and two scores of cognitive fluency (RQ4-2)

Model summary for Articulation rate

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	3.102	0.044	3.013	3.191	69.721	< .001
L1 fluency measure	0.244	0.054	0.137	0.352	4.543	< .001
LR	0.122	0.073	-0.024	0.269	1.667	0.099
PS	0.209	0.077	0.055	0.363	2.709	0.008
L1 fluency measure by LR	-0.198	0.075	-0.348	-0.048	-2.635	0.010
L1 fluency measure by PS	0.306	0.103	0.100	0.512	2.972	0.004
Random effects (intercepts)	Variance	SD				
Participants	0.152	0.390				
Topic	—	—				
ICC	0.61					
Information criterion						
LogLikelihood	-122.5					
DIC	244.9					
AIC	260.9					
BIC	287.6					
R2	Estimate					
Marginal	0.321					
Conditional	0.737					

Model summary for Speech rate

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	0.416	0.064	0.288	0.545	6.479	< .001
L1 fluency measure	0.125	0.041	0.044	0.206	3.088	0.002
LR	0.135	0.065	0.004	0.266	2.065	0.039
PS	0.141	0.070	0.000	0.282	1.999	0.046
L1 fluency measure by LR	-0.120	0.062	-0.244	0.003	-1.944	0.052
L1 fluency measure by PS	0.199	0.072	0.055	0.343	2.756	0.006
Random effects (intercepts)	Variance	SD				
Participants	0.052	0.227				
Topic	0.001	0.029				
ICC	0.64					
Information criterion						
LogLikelihood	-45.4					
DIC	90.8					
AIC	108.8					
BIC	138.8					
R2	Estimate					
Marginal	0.499					
Conditional	0.822					

Model summary for Mean length of run

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	1.341	0.063	1.214	1.467	21.198	< .001
L1 fluency measure	0.110	0.045	0.021	0.199	2.460	0.014
LR	0.143	0.065	0.012	0.274	2.191	0.029
PS	0.100	0.072	-0.044	0.243	1.388	0.165
L1 fluency measure by LR	—	—	—	—	—	—
L1 fluency measure by PS	0.119	0.051	0.018	0.221	2.347	0.019
Random effects (intercepts)	Variance	SD				
Participants	0.057	0.239				
Topic	0.001	0.030				
ICC	0.59					
Information criterion						
LogLikelihood	-271.3					
DIC	542.6					
AIC	558.6					
BIC	585.3					
R2	Estimate					
Marginal	0.442					
Conditional	0.768					

Model summary for Mid-clause pause ratio

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-1.624	0.056	-1.735	-1.512	-29.174	< .001
L1 fluency measure	0.113	0.055	0.002	0.224	2.037	0.042
LR	-0.158	0.085	-0.328	0.012	-1.861	0.063
PS	-0.174	0.090	-0.354	0.006	-1.931	0.054
L1 fluency measure by LR	0.053	0.089	-0.125	0.230	0.592	0.554
L1 fluency measure by PS	0.147	0.095	-0.044	0.337	1.541	0.123
Random effects (intercepts)	Variance	SD				
Participants	0.087	0.295				
Topic	0.000	0.012				
ICC	0.60					
Information criterion						
LogLikelihood	311.0					
DIC	-621.9					
AIC	-603.9					
BIC	-573.9					
R2	Estimate					
Marginal	0.448					
Conditional	0.777					

Model summary for End-clause pause ratio

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-2.820	0.099	-3.018	-2.622	-28.446	< .001
L1 fluency measure	0.083	0.037	0.009	0.157	2.235	0.025
LR	-0.088	0.062	-0.211	0.035	-1.426	0.154
PS	-0.039	0.065	-0.168	0.091	-0.599	0.549
L1 fluency measure by LR	0.019	0.054	-0.090	0.128	0.352	0.725
L1 fluency measure by PS	0.024	0.062	-0.099	0.148	0.393	0.695
Random effects (intercepts)	Variance	SD				
Participants	0.043	0.208				
Topic	0.003	0.052				
ICC	0.48					
Information criterion						
LogLikelihood	578.8					
DIC	-1157.5					
AIC	-1139.5					
BIC	-1109.5					
R2	Estimate					
Marginal	0.202					
Conditional	0.585					

Model summary for Filled pause ratio

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-2.645	0.116	-2.878	-2.413	-22.756	< .001
L1 fluency measure	0.499	0.109	0.282	0.716	4.593	< .001
LR	-0.249	0.171	-0.592	0.094	-1.453	0.146
PS	-0.107	0.187	-0.480	0.267	-0.570	0.569
L1 fluency measure by LR	0.165	0.181	-0.196	0.526	0.913	0.361
L1 fluency measure by PS	-0.225	0.212	-0.649	0.199	-1.060	0.289
Random effects (intercepts)	Variance	SD				
Participants	0.390	0.625				
Topic	0.002	0.039				
ICC	0.69					
Information criterion						
LogLikelihood	379.6					
DIC	-759.2					
AIC	-741.2					
BIC	-711.2					
R2	Estimate					
Marginal	0.347					
Conditional	0.795					

Model summary for Mid-clause pause duration

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	0.034	0.052	-0.071	0.139	0.647	0.517
L1 fluency measure	0.144	0.037	0.069	0.219	3.854	< .001
LR	-0.082	0.061	-0.203	0.039	-1.358	0.174
PS	-0.091	0.064	-0.219	0.036	-1.431	0.152
L1 fluency measure by LR	-0.084	0.069	-0.221	0.053	-1.223	0.221
L1 fluency measure by PS	0.140	0.076	-0.012	0.293	1.839	0.066
Random effects (intercepts)	Variance	SD				
Participants	0.049	0.222				
Topic	0.001	0.024				
ICC	0.57					
Information criterion						
LogLikelihood	21.8					
DIC	-43.5					
AIC	-25.5					
BIC	4.5					
R2	Estimate					
Marginal	0.352					
Conditional	0.722					

Model summary for End-clause pause duration

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	0.139	0.061	0.017	0.260	2.285	0.022
L1 fluency measure	0.231	0.055	0.120	0.341	4.185	< .001
LR	-0.166	0.089	-0.344	0.013	-1.852	0.064
PS	-0.115	0.093	-0.302	0.072	-1.227	0.220
L1 fluency measure by LR	-0.042	0.101	-0.243	0.159	-0.415	0.678
L1 fluency measure by PS	0.024	0.107	-0.190	0.237	0.221	0.825

Random effects (intercepts)	Variance	SD
Participants	0.111	0.334
Topic	0.000	0.022
ICC	0.46	

Information criterion	
LogLikelihood	-121.8
DIC	243.5
AIC	261.5
BIC	291.6

R2	Estimate
Marginal	0.301
Conditional	0.625

Model summary for Self-repetition ratio

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-3.005	0.121	-3.247	-2.764	-24.911	< .001
L1 fluency measure	0.410	0.100	0.211	0.610	4.111	< .001
LR	—	—	—	—	—	—
PS	—	—	—	—	—	—
L1 fluency measure by LR	—	—	—	—	—	—
L1 fluency measure by PS	—	—	—	—	—	—
Random effects (intercepts)	Variance	SD				
Participants	0.446	0.668				
Topic	0.004	0.061				
ICC	0.54					
Information criterion						
LogLikelihood	374.5					
DIC	-749.0					
AIC	-739.0					
BIC	-722.3					
R2	Estimate					
Marginal	0.169					
Conditional	0.622					

Model summary for Self-correction ratio

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-4.128	0.080	-4.289	-3.967	-51.367	< .001
L1 fluency measure	0.236	0.079	0.077	0.394	2.978	0.003
LR	-0.162	0.129	-0.421	0.096	-1.257	0.209
PS	0.123	0.137	-0.151	0.397	0.898	0.369
L1 fluency measure by LR	—	—	—	—	—	—
L1 fluency measure by PS	0.080	0.099	-0.118	0.278	0.807	0.420
Random effects (intercepts)	Variance	SD				
Participants	0.247	0.497				
ICC	0.38					
Information criterion						
LogLikelihood	630.5					
DIC	-1261.0					
AIC	-1247.0					
BIC	-1223.7					
R2	Estimate					
Marginal	0.091					
Conditional	0.436					

Model summary for False start ratio

			95%CIs for Estimate			
Fixed effects	Estimate	SE	Lower	Upper	*z*-value	*p*
(Intercept)	-5.151	0.141	-5.433	-4.868	-36.497	< .001
L1 fluency measure	0.342	0.118	0.107	0.578	2.905	0.004
LR	-0.530	0.174	-0.877	-0.183	-3.053	0.002
PS	0.347	0.184	-0.022	0.716	1.883	0.060
L1 fluency measure by LR	-0.107	0.207	-0.522	0.307	-0.518	0.604
L1 fluency measure by PS	0.119	0.199	-0.279	0.517	0.599	0.549
Random effects (intercepts)	Variance	SD				
Participants	0.638	0.799				
Topic	0.012	0.111				
ICC	0.46					
Information criterion						
LogLikelihood	788.0					
DIC	-1576.1					
AIC	-1558.1					
BIC	-1528.0					
R2	Estimate					
Marginal	0.123					
Conditional	0.529					