The Multidimensionality of Second Language Oral Fluency: Interfacing Cognitive Fluency and Utterance Fluency

Shungo Suzuki and Judit Kormos

To be published in Studies in Second Language Acquisition

Abstract

The current study examined the extent to which cognitive fluency (CF) contributes to utterance fluency (UF) at the level of constructs. A total of 128 Japanese-speaking learners of English completed four speaking tasks—argumentative task, picture narrative task, reading-to-speaking task, and reading-while-listening-to-speaking task—and a battery of linguistic knowledge tests, capturing vocabulary size, lexical retrieval speed, sentence construction skills, grammaticality judgements, and articulatory speed. Their speaking performance was analyzed in terms of speed, breakdown, and repair fluency (i.e., UF), and scores on linguistic knowledge tests were used to assess students' L2 linguistic resources and processing skills (i.e., CF). Structural equation modelling revealed a complex interplay between the multidimensionality of CF and UF and speaking task types. L2 processing speed consistently contributed to all aspects of UF across speaking tasks, whereas the role of linguistic resources in speed and repair fluency varied, depending on task characteristics.

Introduction

Oral fluency is one of the most robust indicators of L2 proficiency (Tavakoli et al., 2020). Therefore, in the context of learning, teaching, and assessment of second language (L2) speaking skills, oral fluency is commonly regarded as one of the major learning goals. For a better understanding of L2 fluency as a construct and as an important language learning target, it is essential to examine how underlying linguistic knowledge contributes to students' fluent speech production. Insights into how L2 users' linguistic resources and processing mechanisms contribute to the efficiency of speech production may also assist language teachers and materials designers and inform language teaching policymakers what linguistic knowledge areas and skills to develop so that L2 learners may become fluent speakers. In L2 fluency research, underlying linguistic knowledge and the temporal characteristics of speech are termed *cognitive fluency* (CF) and *utterance fluency* (UF), respectively (Segalowitz, 2010, 2016). Specifically, CF refers to "the efficiency of the speaker's underlying processes responsible for fluency-relevant features of utterance" (Segalowitz, 2010, p. 50). UF is concerned with "the oral features of utterances that reflect the operation of underlying cognitive processes" (Segalowitz, 2010. p. 50), including speed of delivery and hesitations. Although few in number, previous studies have examined the relationship between CF and UF (henceforth, CF-UF link), providing important insights into the CF-UF link (De Jong et al., 2013; Kahng, 2020). However, previous studies analyzed the measures of CF and UF only at the level of observed variables, meaning that the findings may entail measurement errors. As any observable phenomena are produced not only by the underlying target constructs, but also by some unpredictable random factors (i.e., measurement error), scholars in the human sciences have adopted the concept of latent variables and calculate them based on the covariance of multiple observed variables (for an overview, see Bollen, 2002; Kline,

2016). To further clarify the CF-UF link, the current study, therefore, examines the CF-UF link at the level of constructs by means of latent variable analyses.

Literature Review

Cognitive Fluency

CF is concerned with how efficiently L2 speakers operate their systems of speech production (Segalowitz, 2010). The validity of operationalization of CF can thus be discussed in terms of how different cognitive and linguistic processes in L2 speech production are reflected in utterances. L2 speech production models (Kormos, 2006; Segalowitz, 2010) are commonly based on Levelt's (1989, 1999) work and assume that L2 speech production entails three major phases—*conceptualization*, *formulation*, and *articulation*—which are executed serially in this order. Conceptualization is responsible for the generation of the preverbal message which includes selected information to convey and its manner of communication. Formulation transforms the preverbal message into corresponding linguistic forms through different linguistic encoding processes (e.g., lexical retrieval, syntactic procedures). Articulation proceeds by moving the speech organs to produce speech sounds. In addition to these major processes of speech production, the *self-monitoring* function examines the interim content and eventual outcome of the preceding processes in terms of appropriacy and linguistic correctness. Among these speech production processes, conceptualization is assumed to be relatively independent of L2 proficiency, because conceptualization is responsible for the manipulation of conceptual information prior to linguistic encoding processes (Kormos, 2006). In contrast, formulation and articulation draw on L2 knowledge and skills, and thus are categorized as L2-specific components of CF (Kahng, 2020; Segalowitz, 2016). One clear distinction between formulation and articulation is the level of representations of processing. Formulation involves several linguistic encoding

modules, all of which manipulate different types of linguistic representations (e.g., lexical and phonological representations). Articulation is considered to be purely motoric, meaning that the execution of articulation involves the use of gestural movements rather than information processing. Meanwhile, the self-monitoring function is related to both conceptual and L2-specific aspects of L2 speech production, because it is driven by either content accuracy or linguistic errors identified in the course of speech production. Building on the notion of L2-specific CF (Kahng, 2020; Segalowitz, 2016), valid measurements of CF should therefore tap into formulation and articulation processes and self-monitoring processes triggered by language-related problems.

Looking closely at the literature on CF research, one may argue that previous studies have used both broad and narrow definitions of CF. In a narrow sense, in accordance with Segalowitz's (2016) original conceptualization, CF refers to the speed and efficiency 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 skills. For instance, lexical processing in L2 speech production is related to the range of available lexical resources (i.e., breadth and depth of vocabulary knowledge) as well as the speed of lexical retrieval (i.e., lexical fluency) (see Kormos, 2006). Following the narrow sense, only lexical fluency is regarded as a lexical component of CF, while the broad definition of CF concerns both the breadth and depth of vocabulary knowledge and lexical fluency as the lexical component of CF. According to Segalowitz's (2010, 2016) framework, CF is conceptualized as a construct that can explain observable temporal features of utterances (i.e., UF). From the perspective of L2 speech production mechanisms, breakdowns in utterances can be caused by both a lack of linguistic resources and a slow processing speed. The valid operationalization of CF may thus involve both linguistic resources and processing speed. Therefore, the current study follows the broad

definition of CF and subsequently operationalizes CF as linguistic resources and processing efficiency at the level of vocabulary, grammar and pronunciation (for a similar methodological decision, see De Jong et al., 2013; Kahng, 2020). However, to the best of our knowledge, the dimensionality of CF (i.e., the number of subconstructs) has not yet been empirically examined. Accordingly, the current study also aims to test different factor structures of L2-specific CF.

Utterance Fluency

Within Segalowitz's (2010, 2016) framework, UF refers to observable temporal features, such as speed of delivery, pauses, and hesitations, which reflect the speaker's CF. There has been a consensus that UF is composed of three subcomponents—speed, breakdown, and repair fluency (Skehan, 2003; Tavakoli & Skehan, 2005). *Speed fluency* is concerned with the density of information or speed of delivery and thus is typically measured by articulation rate (i.e., the number of syllables produced over speech duration excluding pauses). *Breakdown fluency* refers to pausing behaviour and is commonly operationalized in terms of the frequency, duration, type, and location of pauses (S. Suzuki et al., 2021; Tavakoli & Wright, 2020). Among the different dimensions of pausing behaviour, recent studies have recognized the importance of pause location as an indicator of underlying speech processing. Pauses in the middle of utterances are hypothesized to reflect disruptions in L2-specific linguistic processing, while pauses at clausal boundaries are supposed to capture breakdowns in conceptualization-related processes, such as content planning (De Jong, 2016; Tavakoli, 2011). Finally, *repair fluency* covers, by definition, a range of disfluency phenomena, including self-corrections, false starts, and verbatim repetitions. Some scholars argue that repair fluency is in a supplementary relationship with breakdown fluency, because both breakdown and repair fluency are assumed to reflect the operation of self-monitoring

processes (i.e., covert and overt repairs; see Kormos, 2000, 2006) and are regarded as opportunities for speakers to buy time to deal with disruptions in speech processing (De Jong et al., 2015). As such, 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). The triad model was empirically validated to examine the extent to which fluency is distinguishable from other constructs of L2 oral proficiency such as accuracy and complexity, and to establish the robustness of the model across four different prompts of picture narrative tasks. The results of factor analysis in Tavakoli and Skehan's (2005) study indicated two separate factors of UF: one including both speed and breakdown fluency and the other repair fluency. The finding that speed and breakdown fluency were indistinguishable might have been due to the lack of any measure that taps solely into speed fluency, that is, articulation rate. Following Tavakoli and Skehan's (2005) study, different UF measures with high construct validity, such as articulation rate and mid-clause pause frequency, have been employed in L2 fluency research. In addition, even though the triad model of UF was validated only with speech data from picture narrative tasks, the model has been applied to a variety of speaking tasks, going beyond picture narratives. Therefore, to test the validity of Tavakoli and Skehan's (2005) model of UF in diverse research contexts, it is essential to revisit the dimensionality of UF, using a comprehensive set of UF measures based on different speaking tasks.

The Cognitive-Utterance Fluency Connection

According to Segalowitz's (2010, 2016) framework, a speaker's CF is assumed to underlie the UF of their speech. Although few in number, previous studies have examined what

cognitive and linguistic processes can explain variability in UF. Even before Segalowitz's (2010) work, Segalowitz and Freed's (2004) pioneering research investigated the role of L2-specific cognitive ability in L2 oral fluency with 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 by partialing out corresponding L1 measures. They found that the length of run without fillers in L2 speech was positively associated with the speed of L2 lexical access. Meanwhile, L2 speech rate correlated negatively with the processing stability of L2 attention control, measured by the coefficient of variance (CV) index. 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 resource and processing measures to predict different UF measures. Their data were collected from 179 learners of L2 Dutch from various L1 backgrounds. Their CF measures covered vocabulary knowledge (vocabulary size, lexical retrieval speed), grammatical knowledge (grammatical knowledge, sentence construction speed), and pronunciation knowledge (phonetic accuracy, articulatory speed). Their UF measures captured speed, breakdown, and repair fluency. Correlational analyses showed that relevant components of CF varied across UF measures. For instance, mean syllable duration, that is, the inverse measure of articulation rate (speed fluency), correlated with a whole range of CF measures. Meanwhile, breakdown fluency measures were related to more specific dimensions of CF. Mean duration of pauses correlated weakly with lexical retrieval speed only. Moreover, both silent and filled pause ratio measures correlated with lexical and grammatical measures. In addition, their linear mixed-effects modelling included a random slope of speaking task types for all UF measures, except self-repetition ratio, indicating the moderating role of task type in the CF-UF link.

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 corresponding L1 UF measures as another predictor variable to partial out the covariance between L1 and L2 UF measures. In her study, CF measures covered vocabulary size for single words and multi-word phrases, lexical retrieval speed, grammatical resources and processing speed, and articulatory speed, largely following De Jong et al. (2013). Their stepwise multiple regression analyses resulted in 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 indicates that the primary component of CF can vary across the dimensions of UF. Second, the regression models of mid-clause pause ratio and self-correction ratio did not include corresponding L1 UF measures as predictor variables. This finding suggests that pauses in the middle of clauses and self-repair may specifically reflect 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 language-general idiosyncratic factors than to L2-specific CF.

Taken together, previous studies suggest two common patterns with regard to the CF-UF link. First, different components of CF may be associated with different dimensions of UF to varying degrees. 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 the 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) treated speaking tasks as random-effects predictors in their regression models. Meanwhile, previous studies employed different measurements of CF and analyzed measured scores only at the level of observed variables. It can thus be argued that the findings of previous studies may entail measurement errors to some extent. Therefore, L2 fluency research may be extended by examining the CF-UF link at the level of constructs by means of latent variable analyses.

The Current Study

Motivated by the scarcity of studies examining the CF-UF link, the current study examines the relationship between CF and UF across four speaking tasks at the level of constructs, as well as the dimensionality of CF and UF. Accordingly, the study employed a cross-sectional design to investigate the factor structure of CF and UF. Building on Segalowitz's (2010) original framework, we predicted the latent variables of UF (i.e., outcome variables) from those of CF (i.e., predictor variables), using structural equation modelling (SEM). We also included one moderator variable, that is, speaking task at four levels (Argumentative, Picture narrative, Reading-to-speaking, and Reading-while-listening-to-speaking). Following De Jong et al. (2013) and Kahng (2020), this study operationalized CF as a set of linguistic resources and processing skills. Each dimension of UF, that is, speed, breakdown, and repair fluency, was also measured. Furthermore, to examine the moderating role of speaking tasks in the CF-UF link, we employed four speaking tasks: argumentative task, picture narrative task, reading-to-speaking task, and reading-while-listening-to-speaking task. The current study is guided by the following research questions:

RQ1. What is the relationship between cognitive fluency measures of lexical, grammatical, and pronunciation knowledge?

RQ2. What is the relationship between utterance fluency measures of speed, breakdown, and repair fluency?

RQ3. To what extent do components of cognitive fluency contribute to different dimensions of utterance fluency?

RQ4. To what extent is the cognitive-utterance fluency link (RQ3) moderated by speaking tasks?

Method

Participants

To reach adequate statistical power, the minimum number of for the sample size was determined by the ratio of the sample size to the number of variables. Traditionally, the optimal ratio for confirmatory factor analysis (CFA) can range from five to ten (Kyriazos, 2018). As a total of 20 observed variables (11 UF measures and 9 CF measures) was predetermined for the current study (see the Analysis section), the minimum number of sample size was set at $N = 100$. A total of 128 Japanese learners of English, ranging from 18 to 27 years of age ($M_{age} = 20.43$, $SD_{age} = 1.81$), participated voluntarily in the current study (Female = 73, Male = 55). Their self-reported university placement test scores suggested that most of them could be placed at the B1–B2 levels of the Common European Framework of Reference (CEFR; Council of Europe, 2001) scale, while some of them seemed to have reached C1 level.

Speaking tasks

The current study aims to examine the moderator effects of speaking task design on the CF-UF link. Given the mechanisms of speech production as theoretical underpinnings of CF (Segalowitz, 2010, 2016), we selected task design features based on the framework of speech processing demands (e.g., Préfontaine & Kormos, 2015; Skehan, 2009), targeting three speech processing characteristics: content planning (i.e., conceptualization), the pre-emptive activation of relevant linguistic items, and the availability of phonological information. To manipulate these speech processing components, the study employed four speaking tasks: (a) an argumentative speech task, (b) a related picture narrative task, (c) a reading-to-speaking (RtoS) task, (d) a reading-while-listening-to-speaking (RwLtoS) task. All the task prompts are available via OSF (https://osf.io/zrwmn/?view_only=0eeb1c966cb64afc9834acf80a42ad7e). In the argumentative task, students were provided with a statement and argued to what extent they agree/ disagree with it (S. Suzuki & Kormos, 2020), while in the picture narrative task, they were asked to describe an 11-frame cartoon adopted from Préfontaine and Kormos (2015; available in the IRIS database, https://www. iris-database.org). In both RtoS and RwLtoS tasks, students were instructed to read a 300-word long expository text written in English and to retell the content of the text (for details of the tasks, see Kormos et al., forthcoming). However, these tasks differed in the modality of the source text presentation. The RtoS task offered a written text (i.e., reading-only), while a bimodal text (reading-while-listening) was provided in the RwLtoS task. In order to minimize the effects of source texts, we prepared two comparable texts adapted from Millington (2019), and the audio-input for the bimodal source text was recorded by a L1 Canadian English speaker with 15 years of English teaching experience at universities in Japan. There are three intended contrasts between these tasks. First, comparing the argumentative task with the other three tasks, the moderating role of the

necessity for content planning in the CF-UF link can be examined. Second, the contrast between the picture narrative task and both RtoS and RwLtoS tasks may offer insights into how the pre-emptively enhanced activation of linguistic items by means of the source text presentation affects the CF-UF link. Third, the impact of the availability of phonological information on the CF-UF link can be examined by contrasting the RtoS and RwLtoS tasks.

Utterance fluency measures

Following previous studies, we targeted three major aspects of UF—speed, breakdown, and repair fluency (Tavakoli & Skehan, 2005). There is one measure that only taps into the construct of speed fluency, that is, *articulation rate*, or its inverse measure, *mean duration of syllables* (Tavakoli et al., 2020). However, to construct a latent variable, more than two observed variables are ideally loaded onto the latent variable to avoid an under-identified model (Brown, 2006). We thus included two composite measures—*speech rate* and *mean length of run*—as the measures of speed fluency. 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.

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 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 variability in speed fluency. To annotate temporal features, *Praat* software was used (Boersma & Weenink, 2012). After annotating and excluding disfluency features, the number of syllables produced in pruned transcripts was calculated. Following prior research (Bosker et al., 2013; De Jong & Bosker, 2013; S. Suzuki et al., 2021), the threshold of silent pauses was defined as 250 ms. With the assistance of automated detection of silence, clause boundaries and pause locations were

annotated in TextGrid files of *Praat*. To ensure the validity of pause identifications, automatically annotated boundaries of silences and sounds were manually checked and, if necessary, modified.

Vocabulary knowledge

In L2 speech production, vocabulary knowledge mainly plays a role in lexical retrieval where the speaker activates and selects lexical items from the mental lexicon that match the conceptual meaning of the message (Kormos, 2006). We thus assessed speakers' vocabulary size and lexical retrieval speed.

Vocabulary size

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 in a blank in a sentence in the paper format version of the test. Considering the expected proficiency levels of the participants, the study administered tests of 2,000, 3,000, and 5,000 frequency levels (excluding the 10,000 level and university word list). To avoid collinearity with lexical retrieval speed, participants were not given a time limit for their responses.

The score for vocabulary size was computed as the total number of correct responses out of 54 items (18 items from each level). Following De Jong et al. (2013), inflectional errors and obvious spelling mistakes were ignored.

Lexical retrieval speed

To assess the speakers' speed of lexical retrieval, a picture naming task was employed (De Jong et al., 2013; Leonard & Shea, 2017). Participants were presented with pictures and instructed to name each picture orally in English as fast and accurately as possible. Target

stimuli were selected from Snodgrass and Vanderwart (1980). The final set of picture stimuli for the study included 50 pictures (for the selection procedure, see Supplementary Information).

The current study administered the picture naming task using the *PsychoPy* software package (Peirce, 2007). Following De Jong et al. (2013), participants were first presented with a fixation cross in the middle of the screen for 1,500 ms, followed by a picture stimulus with a 10,000-ms response deadline. The order of the picture stimuli was randomized for each participant. Prior to the main trials, three practice trials were conducted.

Lexical retrieval speed was computed as the average reaction time (RT) for correct responses. RT was calculated as the response latency between the onset of the presentation of picture stimuli and that of the participants' response. Incorrect responses and outliers were treated as missing values. Outliers were identified as RTs below the minimum of 300 ms and RTs higher than 3 SD above the group mean for each item. As a result, 2.4% of correct responses (k = 127 out of 5375) were removed.

Grammatical knowledge

Grammatical processing in L2 speech production entails a variety of syntactic and morphological processes, such as syntactic procedures and morphological inflections (Kormos, 2006). Accordingly, we evaluated students' grammatical knowledge in terms of their accuracy and efficiency in syntactic encoding skills and grammatical monitoring processes.

Syntactic encoding skills

The study used the maze task which is designed to measure the automaticity of syntactic processing (Y. Suzuki & Sunada, 2018). In this task, participants were presented with two

options for single words on a computer screen and instructed to select the word which can be grammatically connected to the sentence being constructed from two options (e.g., *The* → *student* vs *and* → *ocean* vs *took* → *the* vs *dress*→ *tests.* vs *organic.*).

Stimuli were adapted from Y. Suzuki and Sunada's (2018) study, which consisted of 48 sentences with 12 sentences including four major syntactic structures each: (a) declaratives, (b) wh- questions, (c) relative clauses, and (d) indirect questions. The order of sentence stimuli was randomized for each participant. Prior to the main trials, four practice sentences were provided. The time limit for each response was set at 4,300 ms, following Y. Suzuki and Sunada (2018). Participants were instructed to respond as quickly and accurately as possible. The maze task was administered using DMDX software (Forster & Forster, 2003).

The study computed two measures: (a) the number of correct responses in words and (b) the mean duration of the response latency (i.e., RT) of trials correctly responded to. Regarding the RT measure, outliers were identified as 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.

Grammatical monitoring processes

To capture participants' grammatical knowledge in the monitoring mode, we employed a timed grammaticality judgement test (GJT; Godfroid et al., 2015). Target stimuli were adapted from Godfroid et al.'s (2015) study, which included 17 target grammatical features. For each grammatical target, four sentence stimuli were devised (68 sentences in total) with two for each of the grammatical and ungrammatical conditions. Considering the relatively low proficiency of the target population, we used written stimuli.

Timed GJT was administered using *PsychoPy* software (Peirce, 2007). 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. For each trial, the term "*Ready?*" was presented in the middle of the screen for 1,000 ms, and then the sentence stimulus appeared on the screen for 10,000 ms.

To compute accuracy scores based on GJT responses, we 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 RT scores, excluding outliers whose RT was below 300 ms or higher than 3 SD above the group mean for each sentence stimulus. Eventually, 28 RTs (0.4%) were removed from the RT analysis. We calculated accuracy and RT scores separately for syntactic and morphological features.

Articulatory skills

The current study solely focused on the speed aspect of pronunciation knowledge, given the substantive difficulty of defining what constitutes target-like pronunciation (Harding, 2018).[1] Moreover, prior work reported that a significant slow-down in L2 oral production may result from the speed of articulatory movements rather than the accuracy and speed of phonological processing (Broos et al., 2018). However, due to the incremental nature of speech processing (Kormos, 2006), we measured the efficiency of pronunciation-related processes holistically, using a controlled speech production task. 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).

Participants were asked to read a 69-word passage of an instruction on shopping silently and then read it aloud in English. The passage was adapted from Weinberger's (2011) speech

accent archive (see http://accent.gmu.edu/index.php). Based on the speech data we computed the articulation rate measure applying the same procedure as the one for speed fluency.

Data collection procedure

Data were collected in two sessions: group and individual sessions. Both sessions were conducted in a research laboratory and lasted for approximately one hour. In the group session, participants worked individually and completed CF tests including the paper-based PVLT, the maze task, and the GJT. In the individual sessions, participants performed four English speaking tasks, the controlled speech production task, and the picture-naming task in this order. All participants first took part in the group session, and approximately one week later they participated in individual sessions. In the group testing session, the order of the PVLT and the grammar tests (the maze task and the GJT) was counterbalanced across participants. In the individual sessions, the order of the argumentative and picture narrative tasks was also counterbalanced across participants. Regarding the RtoS and RwLtoS tasks, the combination of the text presentation mode and source texts as well as its order was counterbalanced across participants.

Analysis

The current study investigates the CF-UF link at the level of latent variables (RQ3) and its variability across tasks (RQ4), using SEM. Prior to SEM analysis, we constructed several theoretically motivated CFA models of CF and UF and tested their model fit to identify the optimal factor structure of CF and UF (RQ1, RQ2). A SEM model was built to predict the latent variables of UF from those of CF. In response to the non-normal distributions of many UF measures (for descriptive statistics, see Supplementary Information), estimations of all CFA and SEM models were made using Robust Maximum Likelihood estimation (Hu &

Bentler, 1998). Considering the relatively small sample size (N < 250) as well as the estimation method (i.e., Maximum Likelihood estimation), we focused on the model fit indices of SRMR and CFI (Hu & Bentler, 1998), while reporting the indices of chi square/df ratio, TLI, and RMSEA for the sake of 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). To address RQ3, the statistical significance of the regression paths from the latent variables of CF to those of UF was tested. As for RQ4, the regression coefficients of paths were compared across four speaking tasks via 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). For the sake of interpretability of results, CF measures based on RT and breakdown and repair fluency measures were inversed in the CFA and SEM analyses. All the CFA and SEM models were estimated through the *cfa* function in the *lavaan* package (Rosseel, 2012), using R statistical software 4.0.2 (R development Core Team, 2020).

Results

Confirmatory factor analysis of cognitive fluency

To specify the factor structure of CF (RQ1), three proposed CFA models were tested. For these proposed CFA models of CF, residual covariances were set across CF tasks (e.g., the RT and accuracy measures of the maze task). The R code and anonymised data set will be made available on the IRIS database (https://www.iris-database.org/iris/app/home/index).

The first model (CF Model 1; see Fig. 1) was a single-factor model, which assumes that CF is a unitary construct. One statistical advantage 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.

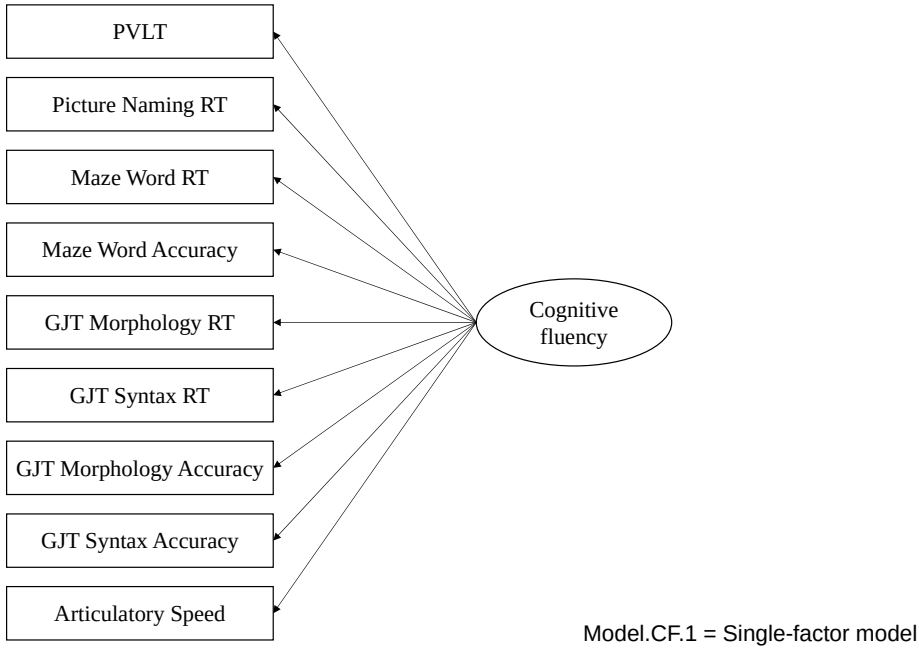

Figure 1. A single-factor model of cognitive fluency (CF Model 1).

Note. Residuals are omitted for the sake of brevity.

The second model (CF Model 2; see Fig. 2) consisted of two subconstructs of CF, namely, *linguistic resource* and *processing speed.* These two subconstructs were conceptualized in accordance with the distinction made in empirical studies (De Jong et al., 2013; Kahng, 2020) and the theoretical assumption of causes for breakdowns in utterances (see the section on Cognitive fluency). The latent variable of linguistic resource consisted of CF measures capturing the range of linguistic resources (the PVLT score, 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.

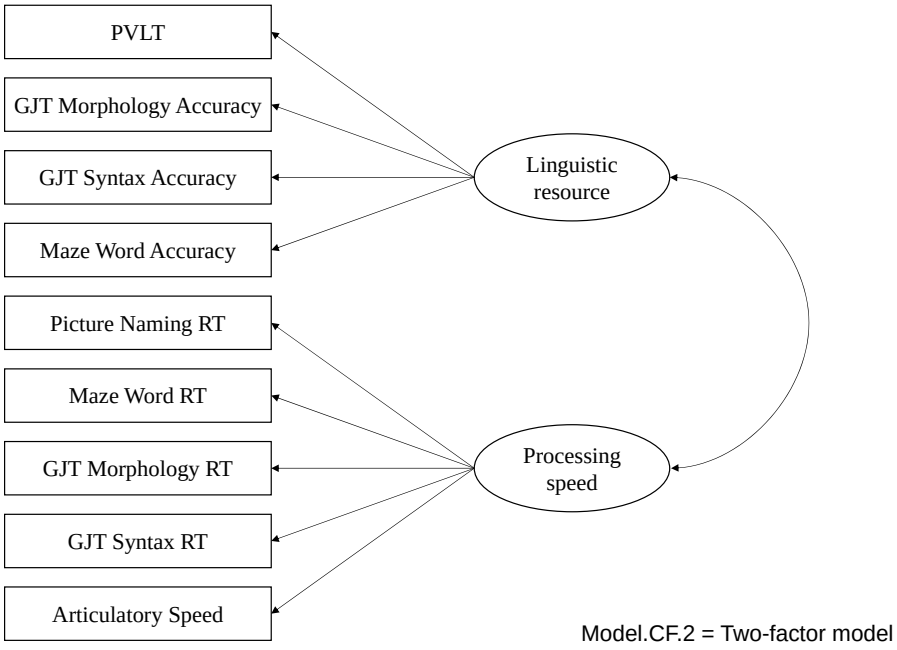

Figure 2. A two-factor model of cognitive fluency (CF Model 2).

Note. Residuals are omitted for the sake of brevity.

Finally, we proposed a three-factor model which comprises linguistic resources, processing speed, and monitoring speed (CF Model 3; see Fig. 3), separating monitoring processes from encoding processes. In L2 speech production, linguistic resources for monitoring processes are identical to those for linguistic encoding processes (Kormos, 2006; Levelt, 1999). Therefore, only the RT measures of the GJT (GJT Morphology RT, GJT Syntax RT) were used to create the third latent variable of CF, that is, *Monitoring speed*.

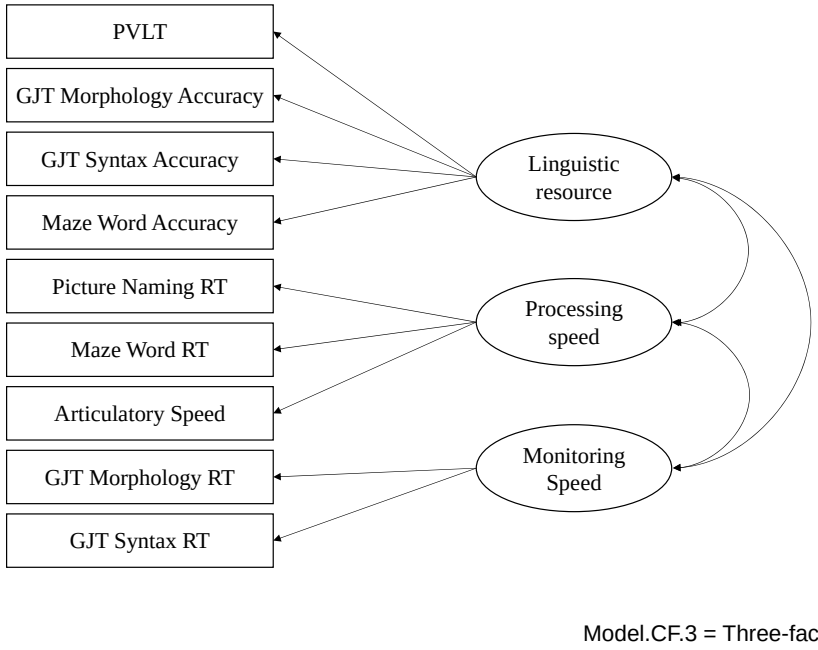

Figure 3. A three-factor model of cognitive fluency (CF Model 3).

Note. Residuals are omitted for the sake of brevity.

The indices of SRMR and CFI indicated an optimal fit for all three models, while two- and three-factor models showed a slightly better fit than the single-factor model (see Table 1). In principle, the more parsimonious the model is (i.e., fewer parameters), the more robust is the estimation of the model (Schoonen, 2015). We thus adopted the two-factor model for the factor structure of CF (for the model parameters of the CF Model 2, see Supplementary Information).

Table 1. *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]

Confirmatory factor analysis of utterance fluency

We proposed several CFA models for UF. First, due to its advantage of statistical robustness, a single-factor model was proposed (UF Model 1, see Fig. 4). Second, motivated by speech production mechanisms, we proposed a two-factor model (UF Model 2; see Fig. 5) by categorizing the temporal features of speech into *processing smoothness* and *processing disruptions*.

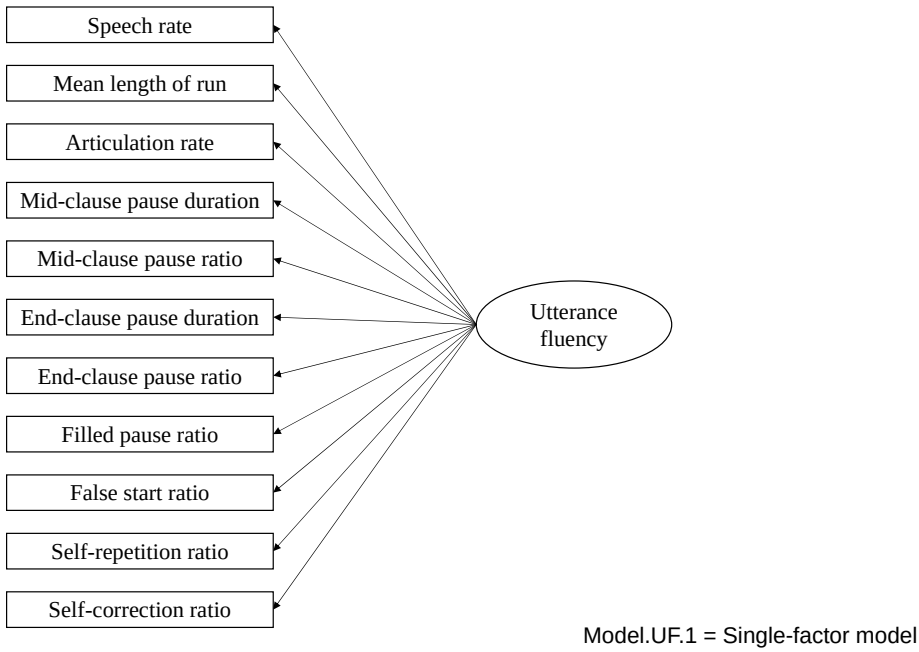

Figure 4. A single-factor model of utterance fluency (UF Model 1).

Note. Residuals are omitted for the sake of brevity.

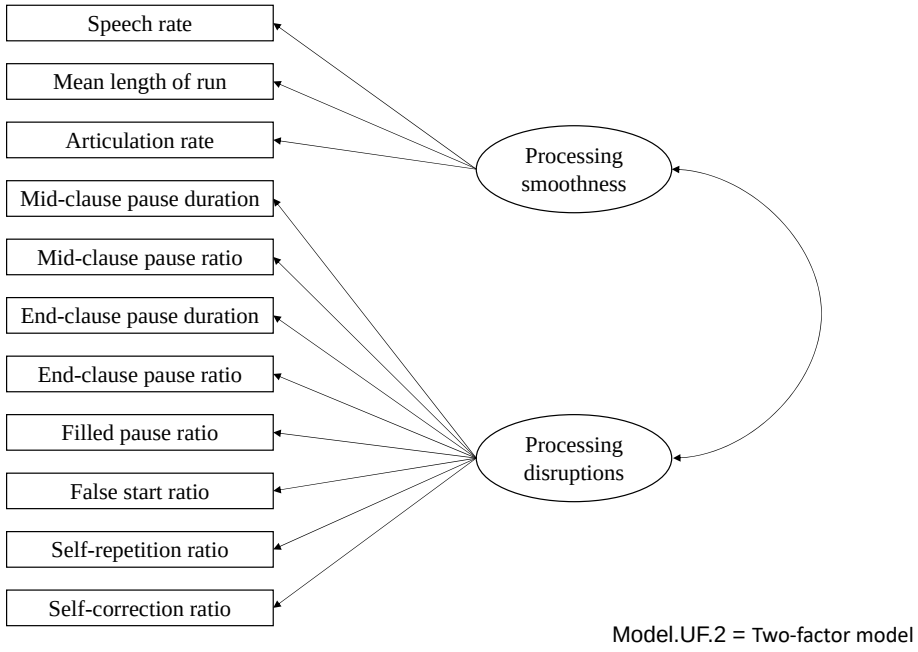

Figure 5. A two-factor model of utterance fluency (UF Model 2).

Note. Residuals are omitted for the sake of brevity.

Finally, following Tavakoli and Skehan (2005), a three-factor model (UF Model 3; see Fig. 6) was proposed, consisting of speed, fluency, and repair fluency. In the proposed CFA models of UF, residual covariances were set between mid- and end-clause pause ratio measures and mean length of run, because their measurement errors are commonly attributed to pause annotation.

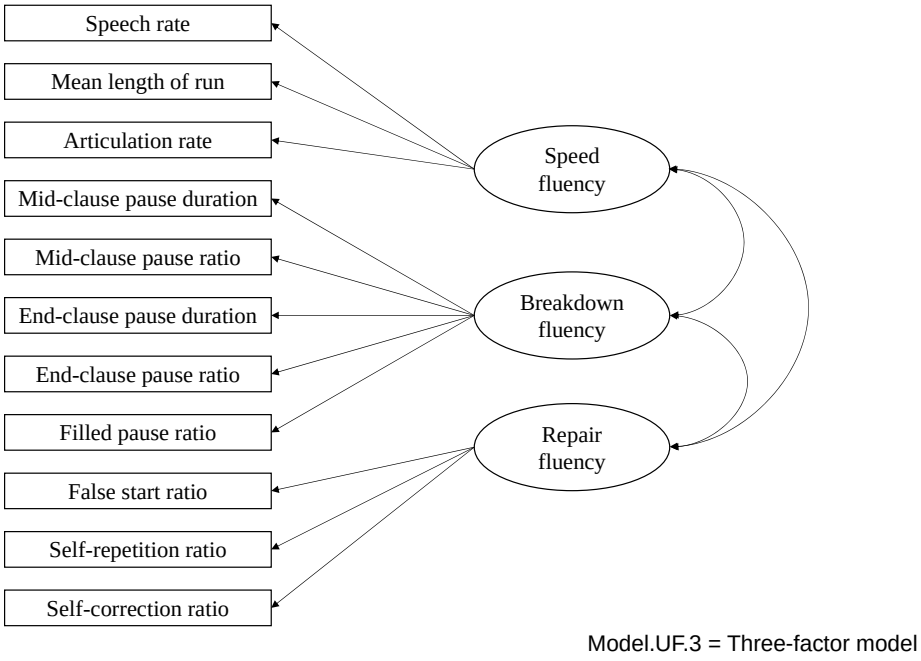

Figure 6. A three-factor model of utterance fluency (Model UF 3).

Note. Residuals are omitted for the sake of brevity.

Although the three-factor model showed a relatively better fit across tasks, none of the proposed models optimally fit the data. To explore a better CFA model, a data-driven approach was taken to modify the factor structures. First, overall intercollinearity among the UF measures was inspected by means of correlation coefficients pooled over tasks. We then excluded speech rate due to its strong correlation with mid-clause pause ratio (r = .845). In addition, motivated by the strong correlation between mid- and end-clause pause duration (r = .735), we also replaced mid- and end-clause pause duration with a single measure of mean pause duration without the distinction of pause locations. Second, modification indices were calculated to explore potential residual covariances and improve the model fit. The following three residual covariances were adopted: (a) between mean pause duration and filled pause ratio, (b) between mid-clause pause ratio and self-correction ratio, and (c) between end-clause pause ratio and false start ratio (for details of model modification, see Supplementary Information).

The revised models of UF (one-, two-, and three-factor models; UF Model 4, UF Model 5, and UF Model 6, respectively) were inspected for goodness of fit. SRMR indices indicated that all the models may fit well to the current data set, while the other model fit indices (e.g., CFI) consistently showed that the three-factor models better fit the current data set (see Table 2).

Table 2. *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]
RtoS task	24	135.293	< .001	5.637	0.822	0.733	0.075	0.190[0.160, 0.222]
RwLtoS task	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]
RtoS task	23	128.507	< .001	5.587	0.831	0.736	0.074	0.189[0.158, 0.222]
RwLtoS task	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]
RtoS task	21	110.689	< .001	5.271	0.857	0.754	0.070	0.183[0.150, 0.217]
RwLtoS task	21	92.648	< .001	4.412	0.864	0.767	0.066	0.163[0.130, 0.198]

The revised three-factor model of UF measures also suggested strong correlations between latent variables of speed and breakdown fluency (r = .929–.960), indicating redundancy in the distinction between these two latent variables. We thus proposed another factor structure

with speed and breakdown fluency measures loaded onto one latent variable (UF Model 7; see Fig. 7).

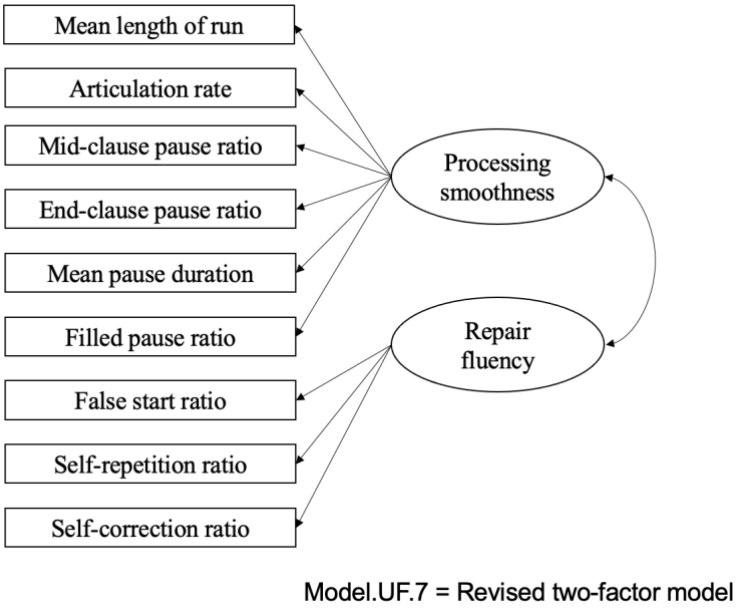

Figure 7. A new two-factor model of utterance fluency (Model UF 7).

Note. Residuals are omitted for the sake of brevity.

Although the model-fit of the new model was virtually identical to the revised three-factor model (SRMR = .058–.070; CFI = .845–.922; see also Supplementary Information), we decided to adopt the revised three-factor model (UF Model 6), considering its theoretical compatibility with Tavakoli and Skehan's (2005) triad model of UF and L2 speech production mechanisms (Kormos, 2006; Segalowitz, 2010).

Structural equation model of the cognitive-utterance fluency link

Building on the CFA models of CF (CF Model 2) and UF (UF Model 6), an SEM model was constructed to predict the latent variables of UF (speed, breakdown, and repair fluency) from those of CF (linguistic resource, processing speed) separately for four speaking tasks.

One additional residual covariance was included between the articulatory speed measure of CF and the articulation rate measure of UF in the SEM model, because measurement errors of these measures can be methodologically shared.

The indices of goodness-of-fit were first inspected. The proposed SEM model optimally fitted the current data set (SRMR < .08), with some potential room for improvement in the model's fit to the data (CFI < .95; see Table 3). The modification indices did not suggest paths that can be verified by a theoretical framework of oral fluency and were consistent across tasks. We thus regarded the model as the final model of the CF-UF link. The SEM model with standardized regression coefficients across tasks is visually presented in Figure 8.

Table 3. *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]
RtoS task	111	196.925	< .001	1.774	0.933	0.908	0.062	0.078[0.060, 0.095]
RwLtoS task	111	214.577	< .001	1.933	0.914	0.882	0.069	0.085[0.068, 0.102]

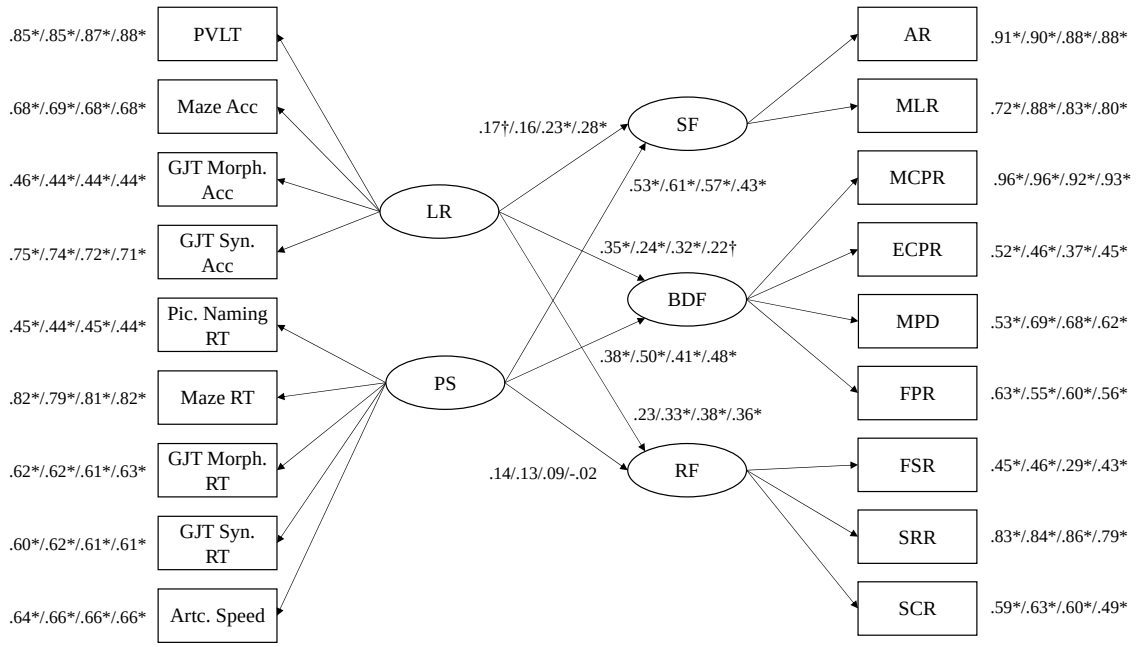

Figure 8. Comparison of the regression coefficients across speaking tasks.

Note. Residuals are omitted for the sake of brevity. Regression coefficients are presented in the order of the argumentative task, the picture narrative task, the RtoS task, and the RwLtoS task from left to right; LR = Linguistic resource; PS = Processing speed; SF = Speed fluency; BDF = Breakdown fluency; RF = Repair fluency.

RQ3 is concerned with how the latent variables of CF are associated with the latent variables of UF. As summarized in Figure 8, speed fluency was associated with linguistic resources only in the RtoS and RwLtoS tasks, and with processing speed in all four tasks. Meanwhile, breakdown fluency was overall consistently related to both linguistic resources and processing speed across tasks. Despite the lack of significant differences, the latent variable of breakdown fluency seemed to show slightly stronger associations with processing speed (β = .376–.502) than with linguistic resources (β = .221–.345). As for repair fluency, linguistic resources significantly contributed to the construct of repair fluency only in speaking tasks where the content of speech was predefined (the picture narrative task, the

RtoS task, the RwLtoS task). Meanwhile, processing speed was not related to repair fluency in any of the speaking tasks.

The SEM model suggested that the relative importance of linguistic dimensions differed between the latent variables of CF in terms of their range of confidence intervals (see Supplementary Information). Regarding linguistic resources, 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). As regards processing speed, the highest regression coefficients were found in Maze Word RT (β = .794–.821). 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. Significant differences in the regression coefficients were only found between Maze Word RT and Picture Naming RT (β = .436–.453). The latent variables of linguistic resources and processing speed were strongly associated with each other consistently across tasks (β = .664–.676).

Looking closely at the measurement models of UF constructs, the regression coefficients of articulation rate (β = .876–.905) to the latent variable of speed fluency seemed to be slightly higher than those of mean length of run (β = .721–.882). Regarding breakdown fluency, 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). As regards repair fluency, the regression coefficients of self-repetition ratio were significantly higher than those of self-correction ratio (except for the RtoS task) and false start ratio. Finally, there were strong competitive relationships between the latent variables of speed fluency and breakdown fluency (β = -|.769–.822|) 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).[2]

Discussion

Motivated by the lack of studies on the CF-UF link at the level of constructs, the current study examined the CF-UF link (RQ3), using SEM. We 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 using four different speaking tasks. Furthermore, in L2 fluency research, the dimensionality of CF and UF had not been revisited, or even specified, especially concerning generalizability across different speaking tasks. Therefore, we also examined the factor structure of CF and UF by means of CFA (RQ1, RQ2). Finally, in light of the generalizability and robustness of the CF-UF link, we explored the variability in the association between the subconstructs of CF and UF across different speaking tasks (RQ4).

Dimensionality of cognitive fluency

We tested the single-, two-, and three-factor models of CF, all of which were proposed based on L2 speech production mechanisms (Kormos, 2006; Levelt, 1989; Segalowitz, 2010) and Segalowitz's (2010) conception of CF. We 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 resources involved the PVLT score (vocabulary size), GJT accuracy scores (syntax and morphology), and the maze task accuracy score (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 articulatory speed in controlled speech production. The strong association between these two latent

variables (r = .676) indicates that the subdimensions of CF—linguistic resources and processing speed—are interrelated. Compared to the final two-factor model, the single-factor model showed 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. The current finding of two-dimensionality of CF may thus provide supporting evidence for the broad definition of CF as well as the existing methodological practice of measuring CF components (De Jong et al., 2013; Kahng, 2020).

The measurement models of the subconstructs of CF suggested that 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 in the final SEM model is discussed. As for the latent variable of linguistic resources, PVLT (vocabulary size) had the highest regression coefficients (β = .845–.879). 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). Students' performance in the maze task can be explained with reference to their efficiency in the application of syntactic encoding procedures in L2 (e.g., word order) as well as accessibility to the syntactic properties of lemmas in their mental lexicon. Meanwhile, the accuracy scores of syntactic items in the GJT may only represent the mastery of syntactic properties of target lemmas. Building on the assumption that the syntactic properties of lemmas (e.g., part of speech) are stored in speakers' mental lexicon (Kormos, 2006; Levelt, 1989), the accessibility of such syntactic properties of lemmas can be regarded as part of the construct of depth of vocabulary knowledge. As vocabulary size and depth are arguably closely related to each other (González-Fernández & Schmitt, 2020), the non-significant difference in the regression coefficients between PVLT and GJT Syntactic Accuracy may be explained by the potential

overlap between vocabulary size and depth. The relative strengths of those regression coefficients suggests that lexical resources can be regarded as a primary component of linguistic resources of CF in line with the lexically-driven nature of L2 speech production (Kormos, 2006). The construct of linguistic resources in CF can thus be defined as the breadth and depth of linguistic knowledge to express speakers' intended message.

Regarding the latent variable of processing speed, the strongest regression path was Maze Word RT (β = .794–.821) which taps into the speed of sentence construction. Despite the slight overlaps of the boundaries of 95% confidence intervals, the regression path of Maze Word RT seemed stronger than that of Syntax RT (β = .604–.620), GJT Morphology RT (β = .614–.626), and articulatory speed (β = .635–.663) in the SEM model. Note that the regression coefficients of Maze Word RT were clearly higher than those of Picture Naming RT (β = .436–.453). Therefore, the current results indicate that the primary component of processing speed may be the speed of sentence construction (measured by Maze Word RT). Such syntactic processing skills might also be more important than lexical retrieval speed within the construct of processing speed of CF (for a different pattern, see Kahng, 2020). One possible explanation for the primary role of syntactic processing skills in processing speed is that variability in the speed of linguistic processing might be aligned with variability in the automaticity of L2 syntactic knowledge (cf. McManus & Marsden, 2019; Morgan-Short et al., 2014). Taken together, the construct of processing speed can be defined as the automaticity of accessing and manipulating linguistic knowledge.

Dimensionality of utterance fluency

Motivated by theoretical conceptualizations of speech production mechanisms as well as Tavakoli and Skehan's (2005) triad model of UF, the current study tested single-, two- and three-factor models of UF. Considering the theoretical distinction between speed and

breakdown fluency, the three-factor model, following Tavakoli and Skehan (2005), was adopted as the final model of UF, suggesting that the construct of UF consists of speed, breakdown, and repair fluency. The optimal model fit in all four tasks (e.g., SRMR = .056–.070) indicated the generalizability and robustness of Tavakoli and Skehan's (2005) triad model of UF across different speaking tasks. Moreover, Tavakoli and Skehan's (2005) study only included two composite measures (speech rate, mean length of run) as measures of speed fluency, and these two measures and breakdown fluency measures loaded on the same latent variable in their study. Tavakoli and Skehan (2005) could thus only conceptually argue for distinguishability between speed and breakdown fluency. Meanwhile, the current study statistically has proved the distinction between speed and breakdown fluency by including the pure measure of speed fluency, that is, articulation rate (Tavakoli et al., 2020).

The construct definition of each dimension of UF can be revisited with regard to the relative importance of observed variables within latent variables. As for speed fluency, 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 a measure of speed fluency in the SEM analysis, despite its composite nature (Bosker et al., 2013; Tavakoli et al., 2020). The slightly lower regression coefficients of mean length of run to the latent variable may indicate that some variance in mean length of run can be derived from factors other than the construct of speed fluency, such as the construct of breakdown fluency. The primary component of speed fluency is thus arguably represented by the measure of articulation rate, which captures the whole range of speech processing mechanisms (Kormos, 2006; Segalowitz, 2010). Therefore, the construct of speed fluency can be defined as the overall efficiency of speech production.

Regarding breakdown fluency, the regression coefficients of mid-clause pause ratio (β = .919–.963) were significantly higher than those of 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 RtoS task). There were no significant differences in the regression coefficients among these three measures (mean pause duration, end-clause pause ratio, and filled pause ratio). Therefore, 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). Mid-clause pauses are reflective of disruptions to L2-specific processing, such as lexical retrieval and sentence construction (De Jong, 2016; Tavakoli, 2011). Accordingly, the construct of breakdown fluency may represent L2 users' ability to continue speaking without disruptions to L2-specific speech processing.

As regards repair fluency, the regression coefficients of self-repetition ratio to the latent variable of repair fluency (β = .787–.860) tended, overall, to be significantly higher than those of false start ratio (β = .289–.459) and self-correction ratio (β = .487–.632). Accordingly, the frequency of self-repetitions can be regarded as the primary component of repair fluency, while both self-corrections and false starts are of secondary importance. The frequency of self-repetitions may be independent of L2 proficiency (Tavakoli et al., 2020) and reflective of learners' speaking style (De Jong et al., 2015). Alternatively, self-repetition can be used as a fluency strategy or problem-solving mechanism (Dörnyei & Kormos, 1998). Specifically, the use of self-repetitions can buy time for monitoring or retrieval processes, as lexicalized fillers do. From the perspective of speech production, another important assumption is that repair fluency is in a complementary relationship with breakdown fluency (Tavakoli & Wright, 2020). When a speaker experiences disruption to speech processing and is required to repair their utterance, the speaker can engage with the repairing process either by producing no speech (i.e., silent pauses) or repeating the previous utterance (i.e., self-repetition). The strategic use of self-repetition may be determined by the speaker's individual

preference and might consequently obscure the association with L2 competence. Taken together, the construct of repair fluency reflects the ability to produce L2 speech with fewer disfluency features.

Contribution of cognitive fluency to utterance fluency

The SEM model revealed the multidimensional interrelationship between CF and UF with some variations across four speaking tasks. 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 RtoS task (β = .234) and the RwLtoS task (β = .276). Therefore, the overall efficiency of speech production (speed fluency) can be primarily supported by the speed of linguistic processing skills. 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 characteristics of RtoS and RwLtoS tasks, that is, the enhanced activation of relevant linguistic items by the source texts. 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 their overall efficiency of speech production (i.e., speed fluency). Therefore, the contributions of linguistic resources to speed fluency may increase when the mastery of relevant linguistic items plays a particularly important role in the completion of a given task.

The latent variable of breakdown fluency was associated with both dimensions of CF consistently across speaking tasks, despite the marginally significant contribution of

linguistic resources in the RwLtoS task (p = .061). The results indicated that the ability to continue speaking without disruption 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 a lack of linguistic resources or a slow processing speed (see Kormos, 2006; see also the section of Cognitive fluency). 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, despite the strong correlation between them. Speed fluency was mainly related to the speed dimension of CF, while breakdown fluency was connected to the linguistic resources of CF as well as the processing speed component.

The significant contribution of linguistic resources to repair fluency was only found in the picture narrative task, the RtoS task, and the RwLtoS task (β = .330–.375). Meanwhile, the processing speed of CF was not associated with the latent variable of repair fluency in any of the speaking tasks. Previous studies have shown that the construct of repair fluency is relatively independent of L2 proficiency (Tavakoli et al., 2020) and reflective of individual speakers' speaking styles (De Jong et al., 2015; Peltonen, 2018). However, the current result may suggest that repair fluency is not entirely independent of L2-specific linguistic knowledge in some communicative situations where the content of speech is mostly predefined (i.e., closed task; see Pallotti, 2009). One essential characteristic of closed tasks is that students cannot avoid expressing some information to achieve the given task, even if they have not fully acquired the necessary linguistic items to convey the intended information. Students are thus required to engage with modifying the intended message or search for some alternative expressions using their own resources. As discussed previously, students can strategically or subconsciously use self-repetition to buy time to repair their

utterances (Dörnyei & Kormos, 1998). Therefore, the contribution of linguistic resources to repair fluency may reflect engagement with repair due to the lack of linguistic resources needed to express task-essential information.

Conclusion

Our study is the first one to examine the CF-UF link at the level of constructs and offers novel insights into how the subconstructs of CF contribute to those of UF across different speaking tasks. Our research has demonstrated that the construct of CF consists of two dimensions—linguistic resources and processing speed—and confirmed the robustness of Tavakoli and Skehan's (2005) three-dimensional model of UF (speed, breakdown, and repair fluency) across tasks. Based on our analyses, we have argued that key components of linguistic resources in CF are the breadth and depth of linguistic knowledge needed for encoding speakers' intended message. This suggests that similar to L1 speech production (Levelt, 1989, 1999), semantic knowledge is essential to ensure the efficiency of encoding L2 speech. We also found that the speed of sentence construction was a key component of the construct of processing speed, which highlights the important role of automaticity of syntactic encoding processes in L2 spoken performance (cf. Kormos, 2006). The SEM analysis also revealed a complex interplay between the multidimensionality of CF and UF and speaking task types. Speed fluency was primarily associated with processing speed, while linguistic resources might only play a role when relevant linguistic items are activated in advance by the input task (i.e., RtoS and RwLtoS tasks). Meanwhile, both linguistic resources and processing speed contributed to breakdown fluency consistently across speaking tasks, suggesting that encoding problems can occur due to both a lack of resources or challenges in accessing and processing linguistic knowledge in real time. Finally, the contribution of linguistic resources to repair fluency was significant only when the content of

speech was predefined (i.e., picture narrative task, RtoS task, RwLtoS task), while repair fluency was generally independent of processing speed. These results confirmed that the processing speed of CF showed a consistent pattern of contributions to UF across speaking task types, whereas the role of linguistic resources of CF in UF tends to vary, depending on task characteristics.

The current findings offer some insights into what linguistic objectives should be prioritized in relation to L2 fluency development. The CFA model of CF showed that vocabulary size was found to be the primary component of linguistic resources, while sentence construction speed was the primary component of processing speed. Accordingly, vocabulary instruction should emphasize widening students' lexical repertoires for productive use (Webb et al., 2020), and grammar instruction should focus not only on accuracy but also on the speed and efficiency of grammatical encoding which can be enhanced through meaningful and engaging practice activities (Y. Suzuki & DeKeyser, 2017). Articulatory speed was also found to be another component of the processing speed of CF, indicating that training on some suprasegmental features, such as linking and vowel reduction, may also facilitate students' fluent speech production (Saito et al., 2019). In addition, our SEM model showed that the construct of breakdown fluency may be consistent across tasks, while that of speed fluency and repair fluency could vary, depending on task characteristics. Therefore, breakdown fluency measures, such as mid-clause pause ratio (for predictive validity in perceived fluency, see S. Suzuki et al., 2021), could be adopted as a representative feature in automated scoring systems for oral proficiency.

Two significant methodological limitations need to be acknowledged in interpreting the current findings. First, we did not include measures of multiword sequences and pronunciation accuracy (cf. De Jong et al., 2013; Kahng, 2020). The processing advantage of multiword sequences in L2 speech production has been advocated in L2 fluency research

(Tavakoli & Uchihara, 2020). Similarly, despite the substantive difficulty in identifying target-like pronunciation, previous studies have found some unique contributions of pronunciation, such as syllable structure errors, to listener-based judgements of fluency (S. Suzuki & Kormos, 2020). Due to the SEM approach, the latent variables of CF in the current study may encompass a certain amount of potential covariance with phraseological competence and pronunciation skills. However, future studies can replicate the current study with additional CF measures of multiword sequences and pronunciation accuracy. Second, two composite measures (mean length of run, speech rate) were used as speed fluency measures for statistical reasons to avoid an under-identified model in CFA analyses. However, due to the intercollinearity among observed variables of speed fluency, the measure of speech rate was excluded from the CFA model of UF. Eventually, the measurement model of speed fluency was regarded as an under-identified model.

References

Boersma, P., & Weenink, D. (2012). *Praat: doing phonetics by computer [Computer software]*. www.praat.org/

Bollen, K. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*, 605–634.

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177/0265532212455394

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. https://doi.org/10.1080/23273798.2018.1457168

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. The Guilford Press.

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*.

De Jong, N. H. (2016). 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. https://doi.org/10.1515/iral-2016-9993

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.

https://doi.org/10.1017/S0142716413000210

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. https://doi.org/10.1017/S0142716412000069

Dörnyei, Z., & Kormos, J. (1998). Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in Second Language Acquisition*, *20*(3), 349–385. https://doi.org/10.1017/S0272263198003039

Forster, K., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 116–124. https://doi.org/10.3758/BF03195503

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. https://doi.org/10.1017/S0272263114000850

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. https://doi.org/10.1093/applin/amy057

Harding, L. (2018). Validity in pronunciation assessment [Bookitem]. In *Assessment in Second Language Pronunciation* (1st ed., pp. 30–48). Routledge. https://doi.org/10.4324/9781315170756-3

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. https://doi.org/10.1037/1082-989X.3.4.424

Kahng, J. (2020). Explaining second language utterance fluency: Contribution of cognitive fluency and first language utterance fluency. *Applied Psycholinguistics, 41*(2), 457–480.

https://doi.org/10.1017/S0142716420000065

Kang, O., & Ginther, A. (2018). *Assessment in second language pronunciation*. Routledge.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (Fourth edi).

Kormos, J. (2000). The timing of self-repairs in second language speech production. *Studies in Second Language Acquisition*, *22*(2), 145–167. https://doi.org/10.1017/S0272263100002011

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum Associates.

Kormos, J., Suzuki, S., & Eguchi, M. (forthcoming). The role of input modality and vocabulary knowledge in alignment in reading-tospeaking tasks. *System*.

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. https://doi.org/10.4236/psych.2018.98126

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

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. https://doi.org/10.1111/modl.12382

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, Mass: MIT Press.

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.

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. https://doi.org/10.1017/S0142716418000553

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. https://doi.org/10.1017/S1366728912000715

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. https://doi.org/10.1093/applin/amp045

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13.

Peltonen, P. (2018). Exploring connections between first and second language fluency: A mixed methods approach. *The Modern Language Journal*, *102*(4), 676–692. https://doi.org/10.1111/modl.12516

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. https://doi.org/10.1111/modl.12186

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

Saito, K., Suzukida, Y., & Sun, H. (2019). Aptitude experience, and second language pronunciation proficiency development in classroom settings. *Studies in Second Language Acquisition*, *41*(1), 201–225. https://doi.org/10.1017/S0272263117000432

Schoonen, R. (2015). Structural equation modeling in L2 research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 213–242). Routledge.

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. https://doi.org/10.1515/iral-2016-9991

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1–14. https://doi.org/10.1017/S026144480200188X

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532. https://doi.org/10.1093/applin/amp047

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. https://doi.org/10.1037/0278-7393.6.2.174

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. https://doi.org/10.1017/S0272263119000421

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), modl.12706. https://doi.org/10.1111/modl.12706

Suzuki, Y., & DeKeyser, R. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, *21*(2), 166–188. https://doi.org/10.1177/1362168815617334

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. https://doi.org/10.1017/S1366728916000857

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. https://doi.org/10.1093/elt/ccq020

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. https://doi.org/10.1111/modl.12620

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. https://doi.org/10.1111/lang.12384

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

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. *The Modern Language Journal*, *104*(4), 715–738. https://doi.org/10.1111/modl.12671

Weinberger, S. (2011). *Speech accent archive*. George Mason University. http://accent.gmu.edu

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. https://doi.org/10.1017/S0142716418000802

[1] As the accuracy or accent of pronunciation is evaluated as a deviation from a target-like benchmark, it is necessary to define target-like pronunciation for the assessment of pronunciation accuracy. However, due to the fact that there are different models of L2 pronunciation learning, especially in English, the assessment of pronunciation entails a substantive difficulty in defining what constitutes target-like pronunciation (Kang & Ginther, 2018). Given the potential challenge to the validity of pronunciation accuracy, we thus decided not to include a cognitive fluency measure for pronunciation accuracy.

[2] For the sake of interpretability in the direction of the relationship between the latent variables of UF, these regression coefficients were computed without inversion of the observed variables of breakdown fluency and repair fluency measures.