Ch.10 Measuring speaking and writing fluency

A methodological synthesis focusing on automaticity

Shungo Suzuki, *Waseda University*

Andrea Révész, *University College London*

Abstract

Second language (L2) researchers regard automaticity as an important facet of proficiency, examining how underlying knowledge contributes to efficiency in language production. To discuss the validity of speech and writing measures as indicators of automatised L2 knowledge, this chapter reports two separate systematic reviews, focusing on the relationship between measures of linguistic knowledge and proficiency and those of speaking and writing fluency. Building on theoretical models of speaking and writing production, the chapter describes the methodological trends in L2 speaking and writing research and also offers practical guidelines for measuring L2 automaticity through spontaneous production data.

Introduction

Second language (L2) researchers regard automaticity as an important facet of proficiency. Thus, it is an important consideration in L2 research how automaticity can be validly and reliably assessed in oral and written production. Previous research has shown that various methodological factors need to be taken into account when establishing the construct validity of measures used to gauge L2 automaticity in speaking and writing. This chapter therefore aims to discuss a list of methodological issues around the use of speech and writing measures as indicators of automatized L2 knowledge, by conducting two separate systematic reviews for automaticity in speaking and writing. Based on the findings from the methodological syntheses, the chapter also offers practical guidelines for measuring L2 automaticity through spontaneous production data.

In L2 speaking research, scholars have assumed that efficiency in speech production processes, a crucial indicator of L2 oral proficiency (Segalowitz, 2010), can be observed as "fluency" of speech performance. More specifically, speech production entails a range of cognitive and linguistic processes, and the efficiency of these processes is termed as *cognitive fluency* (CF; Segalowitz, 2010). CF is originally defined as "the efficiency of the speaker's underlying processes responsible for fluency-relevant features of utterances" (Segalowitz, 2010, p. 50). Here "fluency-relevant features" include a range of observable temporal speech features, such as speed of delivery and pausing behaviours, which has also been termed as *utterance fluency* (Segalowitz, 2010). From the perspective of L2 learning and acquisition, the construct of CF is assumed to capture some aspects of automatized L2 knowledge that is responsible for fluent speech production. However, the efficiency of L2 speech production is subject to not only the degree of L2 automaticity but also some general cognitive skills such as creative thinking (Suzuki et al., 2022). Accordingly, Segalowitz (2010, 2016) narrowed down the notion of CF specific to L2 proficiency as *L2-specific cognitive fluency*, which can be equivalent to the automatized productive knowledge and processing skills (see also Kormos, 2006; Van Moere, 2012). For the sake of the construct validity of measurements, one approach to measuring automaticity in L2 speaking is thus to measure utterance fluency in speaking performance by means of temporal features that are reflective of L2-specific cognitive fluency (see also Tavakoli, 2019).

Although writing, as compared to speaking, is relatively slow, fluency or automaticity is also a key determinant of L2 writing proficiency. Like speakers, writers are constrained by their working memory capacity, and even though access to the evolving text on paper or computer screen may be seen as an extension of writers' working memory, linguistic encoding processes are expected to put pressure on less advanced L2 writer's attentional resources and thus result in disfluency in writing (Schoonen et al., 2019). To capture such breakdowns in fluency, L2 writing researchers have traditionally used product-based fluency measures, that is, defined fluency as a function of the length of the resulting text (e.g., Wolfe-Quintero et al., 1998). This practice has been adopted from research on speaking. The use of product-based fluency measures, however, may not always be applicable to writing. In writing, in contrast to speaking, the number of words in the final written product might be dependent on factors such as the word limit specified by the task instructions or the decision of the writer about whether to write a shorter or longer text (Latif, 2009). Thus, the efficiency of writing processes, or automaticity in writing, is better captured by process-based indices that are derived from observations of the writers' online text production processes (Latif, 2009). Taking a process-based perspective to assess writing fluency, Van Waes and Leijten (2015) identified four dimensions of writing fluency: *production* (i.e., amount of text produced within a time interval), *process variance* (i.e., variation in production rate across time intervals), *revision*, and *pausing behaviour*. Not all of these dimensions, however, seem equally relevant to measuring automaticity in L2 writing. Similar to speaking, writing involves a range of cognitive processes, including general problem-solving skills. Importantly, such general skills are not dependent on L2-specific automaticity in written production or what Segalowitz (2010, 2016) termed as L2-specific cognitive fluency in the context of speaking. In sum, parallel to speaking, the valid measurement of automaticity in L2 writing should entail the use of process-based indices of writing fluency that reflect efficient access to and retrieval of L2 linguistic knowledge during written production.

With a view to informing future research on automaticity in L2 speaking and writing, this chapter first reviews theoretical models of speech and written production, followed by a state-of-the-art overview of existing methods to assess L2 fluency in the two modalities. Through conducting a systematic review of the relationships between cognitive fluency and speaking/writing fluency, the chapter aims to identify potentially valid temporal measures of automaticity and highlight gaps in the literature to help improve current measurement practises. We will end the chapter by offering some methodological guidelines for assessing automaticity in productive skills, as well as practical exercises to engage readers in the process of obtaining fluency indices of L2 speaking and writing.

Models of L2 speaking and writing

To define L2-specific CF or the construct of automaticity in L2 speaking and writing, it is first essential to understand the cognitive and linguistic processes underlying speech and written production. L2 speech production (Kormos, 2006; Levelt, 1989, 1999) entails three major phases—conceptualization, formulation, and articulation—and proceeds in this order. *Conceptualization* is responsible for planning the content of speech, including the communicative intention of the speech and the manner of presentation. Notably, the processes in conceptualization are "pre-verbal" or "language-general", meaning that conceptualization processes do not draw on L2 knowledge. Meanwhile, the second phase, *formulation*, consists of a range of linguistic encoding processes, such as lexical retrieval and syntactic procedures, which altogether convert the message generated from conceptualization into the corresponding linguistic forms, by retrieving speakers' own linguistic knowledge. At the phase of *articulation*, the linguistically encoded message is produced as the stream of sounds, by executing acoustic gestures specific to L2 phonetic categories. Both formulation and articulation draw on L2-specific knowledge, and thus the efficiency in these two phases of speech production is largely reflective of the degree of automatization of L2 knowledge. In addition to three major processes, the function of *self-monitoring* plays an essential role in successful communication. However, self-monitoring is mostly a conscious process, and the concept of automaticity (i.e., automatic processing) may not be directly relevant to the self-monitoring function (see also Albarqi & Tavakoli, in press). Therefore, the speaking part of this chapter mainly discusses the validity of utterance fluency measures as an indicator of L2 automaticity in terms of the extent to which the given fluency measures are reflective of formulation and articulation processes.

Like models of L2 speech production, cognitive models of written production (e.g., Hayes, 2012; Kellogg, 1996) generally see writing as an interactive process, entailing four subprocesses: planning, translation, execution, and monitoring. *Planning* entails setting the communicative goals of writing, retrieving ideas from long-term memory or the task input, and organising these ideas into a coherent plan. As with the conceptualisation stage in speech production, planning in writing is assumed to be pre-verbal, not requiring the deployment of L2 knowledge. The *translation* stage, like formulation in speaking, involves the writer in turning the content they have planned into the corresponding linguistic form through linguistic encoding processes. Thus, L2-specific knowledge and skills are the key to the successful translation of the writer's plan. During *execution*, writers employ motor movements to create a hand-written or typed text in a similar manner to articulation in L2 speech production. Finally, *self-monitoring* entails checking whether the evolving text expresses the writers' intended content. The resulting revisions frequently entail greater conscious effort and elaboration in writing than speaking, involving various levels of the text and made some time succeeding initial text production (Schoonen et al., 2009). Similar to speaking, the writing part of this chapter will evaluate measures of writing fluency with regard to how well they capture the efficiency of translation processes. Planning and self-monitoring processes seemed less relevant to our focus, given that planning largely relies on the use of non-L2 specific knowledge and skills, and self-monitoring is a conscious process that primarily draws on learners' metacognitive knowledge (Schoonen et al., 2009).

In sum, theoretical models of speaking and writing assume the involvement of similar mechanisms, although the terms used to describe various subprocesses differ. The major components of speech and written production are summarised in Table 11a, highlighting parallels between the speaking and writing models. The focus in this chapter is linguistic encoding processes (formulation and articulation in speaking and translation in writing), as these stages are expected to best capture facets of L2 automaticity.

Table 11a. Major processes of speaking and writing production.

Speaking	Writing	Function
Conceptualization	Planning	Language-general processes that establish the conceptual message to speak or write
Formulation	Translation	Language-specific processes that convert the conceptual message into the corresponding linguistic form
Articulation	Execution	Gestural movement that transmits the linguistic representation of the message in the form of speech sounds or letters
Self-monitoring	Self-monitoring	Both language-general and language-specific conscious processes that check the interim and eventual outcome of oral and written processes

Methodological Synthesis

This section thereby reports on a methodological synthesis, which involved a systematic review of previous studies offering validity evidence regarding which utterance/writing fluency measures are reflective of CF. The section also reviews methodological trends regarding how speech and written production is elicited, focusing on elicitation task characteristics and task implementation conditions as well as software to obtain fluency indices. The results of the methodological synthesis are then discussed in relation to relevant theoretical issues that need considering when designing research on L2 speaking and writing fluency.

Research Questions

As the chapter explores valid measures of fluency as an indicator of automatised L2 knowledge, the current systematic review aims to summarise primary studies that offer concurrent validity evidence regarding the association between CF and utterance/writing fluency. Another objective of the chapter is to introduce essential methodological variables that can affect the validity of fluency measures as an indicator of automatised L2 knowledge, as well as methodological practises and tools. L2 fluency research typically takes three methodological phases—speech/writing elicitation, selection of measures and analysis of data. To give methodological insights into these major phases in conducting research, the systematic reviews for speaking and writing fluency measures is guided by the following RQs:

1. How and what kind of tasks were used to elicit speech/writing data?
2. What utterance/writing fluency measures were examined in relation to CF measures as a proxy for automatized L2 knowledge?
3. What methodological tools are commonly employed to obtain utterance/writing fluency measures?

Literature search

Following previous meta-analyses (Suzuki et al., 2021; Uchihara et al., 2019), the current systematic review began with the definition of target domain. For utterance fluency, we regarded the relationship between utterance fluency and CF measures (i.e., correlation coefficients) as the target domain. The effect sizes can thus serve as the evidence regarding the extent to which utterance fluency measures can reflect the variability in efficiency in L2-specific processing. Following a broader definition of CF (Segalowitz, 2010; Suzuki & Kormos, in press), we operationalised CF as linguistic resources and processing skills (not limited to speed dimension of linguistic knowledge) and utterance fluency as objective temporal measures (e.g., articulation rate).

We operationalised writing fluency in terms of process-based objective temporal measures (Latif, 2009). Given the smaller amount of research available on writing fluency, however, we kept the domain definition broader than the one for utterance fluency. We operationalized CF as the writers' linguistic knowledge, processing, and/or experience. This broader definition allowed for the inclusion of within-subject longitudinal studies and cross-sectional studies of writers with varied linguistic experience (e.g., placed in different-level language courses). We also interpreted linguistic knowledge broadly to include general proficiency measures. We regarded any positive statistically significant relationship between writing fluency and CF, as well as significant within-subject change, as evidence supporting the validity of a certain writing fluency to measure automatized L2 knowledge.

We used the same four databases for speaking and writing fluency research: *the Education Resources Information Center (ERIC), the Linguistics and Language Behavior Abstracts (LLBA), the ProQuest Dissertations and Theses and the PsycINFO*. We also searched previous studies in the following twelve journals: *Studies in Second Language Acquisition, Language Learning, Second Language Research, Language Teaching Research, The Modern Language Journal, System, TESOL Quarterly, International Journal of Applied Linguistics* (Wiley), *International Review of Applied Linguistics in Language Teaching* (DeGruter), *Annual Review of Applied Linguistics, Language Teaching,* and *Applied Psycholinguistics*. The following keywords were used to identify utterance and writing fluency research:

Utterance Fluency:
second language, foreign language, L2, FL, oral fluency, speech fluency, speaking fluency, utterance fluency, L2 fluency, speed, articulation rate, speech rate, syllable duration, pause frequency, pause ratio, pause duration, silent pause, linguistic knowledge, cognitive fluency, processing, lexical, syntactic, retrieval.

Writing Fluency:
second language, foreign language, L2, FL, writing fluency, fluency in writing, L2 fluency, written production, written language production, writing behaviour, textual production, production fluency, text

production, speed, burst, P-burst, R-burst, pause, pausing, revision, revising, composing rate, production rate, composition rate, productivity, process time, words per minute, characters per minute, product/process ratio.

Note that we also used a NOT function of Boolean operators with *"reading fluency*, *writing fluency*" to exclude studies about fluency in reading and writing from the pool for utterance fluency and with *"reading fluency, speech fluency, oral fluency, spoken fluency, utterance fluency, speaking fluency"* to exclude studies about fluency in speaking and reading from the pool for writing fluency.

Eligibility criteria

In the course of the literature search, we retrieved 208 primary studies for L2 speaking and 662 for L2 writing research, after removing duplicates. Those studies were then screened with their titles and abstracts and with the following eligibility criteria:

1. The study employed at least one CF measure separately measured from spontaneous production.
2. The study adopted spontaneous speaking/writing tasks to calculate utterance/writing fluency measures.
3. The study on speaking skills used at least one objective measure of utterance fluency, as opposed to subjective ratings of fluency (i.e., perceived fluency; Segalowitz, 2010). We only included the effect sizes based on objective measures which tap into any of speed, breakdown, repair fluency or composite measures. The study on writing skills used at least one process-based fluency measure, which assessed speed fluency or production rate, pausing behaviours, or revisions behaviours. Studies using traditional product-based fluency (e.g., number of words in final product per min) indices were removed.
4. The study reported effect sizes such as correlation coefficients representing the relationship between CF and utterance/writing fluency variables. Although studies that employed advanced regression analyses without reporting correlation coefficients were removed to count target effect sizes, they were kept in the pool of primary studies for the current purpose of methodological synthesis of the target domain.

As a result, we obtained 10 primary studies with 136 effect sizes for utterance fluency and 9 studies with 32 effect sizes for writing fluency. Given the limited number of primary studies, the current chapter decided to only narratively synthesise the primary studies that provide the concurrent validity evidence for utterance/writing fluency measures as a proxy for automatized L2 knowledge (for a similar methodological decision, see Suzuki et al., 2021).

Findings and Discussion from the Methodological Synthesis on Utterance Fluency

RQ1: Speech Data Collection

Previous research has shown that speaking performance is subject to how the speech is elicited (see Tavakoli & Wright, 2020). It is even possible to assume that L2 speaking performance is the result of interaction between learners' own L2 system including CF and task (Skehan, 1998). However, there were no studies examining the impact of task conditions on the CF-utterance fluency link. In order to discuss how tasks can affect the validity of utterance fluency measures as an indicator of L2 automatization, we thus present the summary of previous studies in terms of task characteristics and only introduce theoretical issues around the task conditions in relation to the CF-utterance fluency link.

Task characteristics. Table 11b summarises speaking tasks that were used to elicit speaking performance in primary studies. We applied two coding criteria: *Mode of speaking* (Monologue vs. Dialogue) and *Demands on content planning* (Open vs. Closed task). The former variable has been found to affect the construct of fluency (see Suzuki et al., 2021), and the latter has been examined to give insights into learner's speech production processes (Préfontaine & Kormos, 2015; Skehan, 2009).

Table 11b. Summary of speaking task types

Task type			No. of studies
Monologue	Open task	Argumentative task	2
		Personal narrative	2
	Closed task	Role-play	3
		Picture narrative	3
		Video narrative	2
Dialogue		Interview task	2

Note. Some studies employed multiple speaking tasks.

Regarding the mode of speaking, given the discourse and even speaking performance are co-constructed with the interlocutor(s) in dialogic tasks, the speaker's utterance fluency is inevitably affected by how the interlocutor behaves in the interaction. Possibly due to the co-constructive nature of dialogic fluency, previous studies may have tended to adopt monologic speaking tasks to examine the relationship between CF and utterance fluency variables.

Similarly, demands on content planning can also affect the degree of reflection of CF in utterance fluency. Due to the limited capacity of attentional resources (Skehan, 2014), L2 speaking performance is subject to how learners assign their limited attentional resources to different speech production processes. For instance, when the task is designed to draw students' attention to content planning, the overall efficiency of speech production can thus be largely dependent on students' efficiency in content planning. The speaking tasks that impose relatively high content planning demands can be categorised as an open task where speakers are required to decide most parts of the content and organisation of speech (Pallotti, 2009; e.g., argumentative tasks, personal narratives). Meanwhile, when the content of speech is largely predefined (i.e., closed task; e.g., picture narratives), a relatively large amount of attentional resources can be saved for linguistic processing—formulation, articulation and self-monitoring. As a result, the extent to which the efficiency in automatized L2 knowledge is reflected in the temporal characteristics of speech can increase in closed tasks (Suzuki & Kormos, in press).

Task conditions. Another important methodological factor that can affect students' utterance fluency is how the speaking tasks are implemented, that is, task condition. Two major variables of task condition are typically operationalized as *pre-task planning* (i.e., how long speakers can prepare before performing the task) and *within-task planning* (i.e., whether the time for speaking is pressured or not). Prior research has suggested that longer pre-task planning time has beneficial effects on utterance fluency possibly due to the pre-emptive activation of linguistic knowledge, while the impact of within-task planning time on utterance fluency can be either positive or negative, possibly due to individual variability in which aspects of speaking performance students devote their attentional resources to (Awwad & Alhamad, 2021; Tavakoli & Wright, 2020). Those previous studies, however, have only examined the variability of utterance fluency across different task implementation conditions. Therefore, to date, it has not yet been established how task implementation conditions can affect the validity of the use of utterance fluency measures as an index of automated L2 knowledge. However, in order to reduce the contribution of controlled processing and declarative knowledge, as opposed to automatized L2 knowledge, shorter pre-task planning with a timed within-task condition may be an optimal methodological option. As a result, the association between the variability in utterance fluency measures and the degree of automatization of L2 knowledge may be relatively strengthened (for a similar assumption in grammaticality judgement tasks, see Godfroid et al., 2015; Maie & Godfroid, 2022).

RQ2: Representative Measurements of Automatization in L2 Speaking

The frequency of utterance fluency measures used in the primary studies is summarised in Table 11c. It can be indicated that given the total number of primary studies (*N* = 10), the comparability of utterance fluency measure selection is not sufficient among the studies. Even at maximum, the most common utterance fluency measures for each dimension of utterance fluency to examine the association with CF included articulation rate (speed), silent pause frequency (breakdown), self-correction rate (repair), and speech rate (composite). For the detailed results of coding, see the supplementary information.

Following the domain-specific criteria for effect sizes (Plonsky & Oswald, 2014), utterance fluency measures that strongly correlated with CF measures ($r > .60$) were *mean length of run*, *silent pause frequency*, and *speech rate* in Hilton (2008) and articulation rate in De Jong et al. (2013). These results indicate that utterance fluency measures of composite, speed and breakdown fluency are likely to largely reflect students' CF (i.e., automatized L2 knowledge).

Due to the fact that composite measures tap into multiple dimensions of utterance fluency, composite measures tend to show strong associations with the measurements of L2 proficiency including CF measures. Conversely, the findings based on composite measures have been considered difficult to interpret (Bosker et al., 2013; Suzuki et al., 2021), and especially in the context of research on automatization of L2 knowledge, it is relatively unclear what cognitive and linguistic processes underlie the composite measures.

Meanwhile, the strong correlation coefficient in De Jong et al.'s (2013) study was found between articulation rate and sentence building speed measured via a sentence construction task, whereas articulation rate in their study only weakly or moderately correlated with lexical retrieval speed scores and other linguistic resource scores. The central role of efficiency in manipulating L2 knowledge in speed fluency was also reported in Suzuki and Kormos' (in press) study. It is thus plausible to argue that articulation rate captures some aspects of automatized L2 knowledge.

Table 11b. Summary of correlation coefficients of utterance fluency measures with cognitive fluency measures in the pooled studies

	Speed	Breakdown (frequency)		Breakdown (duration)	Repair			Composite		
Study	Articulation rate	Silent pauses	Filled pauses	Mean silent pause duration	Number of corrections	Number of repetitions	Disfluency rate	Speech rate	Mean length of run	Phonation time ratio
Aziz & Nicoladis (2018)								.06–.37		
De Jong & Mora (2019)	.03–.31	.14–.28		.12–.35						
De Jong et al. (2013)	.15–.66	.11–.41	.02–.33	.02–.16	.02–.43	.06–.24				
De Jong et al. (2015)	0.30	0.17	0.09	WAP: .01 BAP: .02	0.10	0.13				
Hilton (2008)		.66–.73		.39–.47	.52–.57			.58–.68	.67–.73	.55–.59
Kahng (2020)	.02–.50	MCP: .04–.33 ECP: .13–.47	MCP: .02–.36 ECP: .03–.18	.02–.22	.05–.54	.02–.33				
Koizumi & In'nami (2013)							NA	NA		
Leonard & Shea (2017)										
Segalowitz & Barbara (2004)									.38–.38	
Uchihara et al. (2020)	.31–.48	.23–.43	.03–.07							
No. of studies	5	5 MCP: 1 ECP: 1	3 MCP: 1 ECP: 1	4 WAP: 1 BAP: 1	4	3	1	3	2	1

Note. Koizumi and In'nami (2013) used structural equation modelling instead of correlation coefficients, and thus the associations that the study examined are indicated by NA; although Leonard & Shea (2017) adopted several utterance fluency measures, they merge them into a composite score and thus no correlation coefficients are reported; MCP = Mid-clause pause; ECP = End-clause pause; WAP = Within-AS-unit pause; BAP = Between-AS-Unit pause. For the calculation method for each measure, see the original primary studies.

Moreover, strong correlations were also found in Hilton's (2008) study between the DIALANG scores in both vocabulary and structure sections (Chapelle, 2006) and silent pause frequency. In general, pauses have been regarded as phenomena indicating breakdowns in speech production processes. It is thus assumed that the frequency of pauses (breakdown fluency) could be an indicator of lack of linguistic resources and slow speed of manipulating them (see Suzuki & Kormos, in press). Fluency research has shed light on the importance of the distinction of pause location (De Jong, 2016; Kahng, 2018; Suzuki & Kormos, 2020, in press). More specifically, pauses in the middle of clauses are reflective of the breakdowns in L2-specific processing such as lexical retrieval and sentence construction, while pauses at clausal boundaries are likely to be caused by breakdowns in content planning. Although only Kahng (2020) and De Jong et al. (2015) among the current primary studies computed pause measures separately for pause location, these studies failed to detect any differences in the strength of correlation between mid-clause and end-clause pauses. However, other studies (e.g., Suzuki & Kormos, in press) tend to show clearer differences in the effect sizes between mid- and end-clause silent pauses. Another issue around the selection of breakdown fluency measures is the distinction between silent and filled pauses. As indicated by the fact that most primary studies reported negligible strength of correlation between filled pause frequency and CF measures, L2 fluency research has shown that filled pauses can be reflective of personal speaking style (De Jong et al., 2015; Peltonen, 2018) or demands on content planning as opposed to linguistic processing (Fraundorf & Watson, 2014).

Regarding the duration measure of breakdown fluency, Hilton's (2008) study showed moderate correlation between mean duration of silent pauses with the DIALANG scores, while other studies commonly reported negligible or weak correlations with utterance fluency measures. From a theoretical perspective, the duration of pauses can be regarded as the time required for solving the breakdown, which can be achieved by either modifying the content of the message or retrieving alternative linguistic items. In other words, the duration of pauses can tap into a wide range of self-monitoring processes and thus may not be a valid measure of automatized L2 knowledge (see also Suzuki & Kormos, in press).

Finally, the measures of repair fluency in general showed negligible or weak correlations, with a few exceptional cases in Hilton’s (2008) and De Jong et al.’s (2013) studies. However, these relatively stronger correlations were only found between the frequency of self-correction and lexico-grammatical resources, meaning that repair fluency measures might primarily reflect the availability of linguistic resources when consciously modifying the message rather than the automatization of L2 knowledge (Suzuki & Kormos, in press). Taken together, mean length of run, articulation rate and the frequency of (mid-clause) silent pauses might thus be relatively robust indicators of automatized L2 knowledge.

RQ3: Analysis tools and software

Although some primary studies failed to fully describe the procedures of speech transcription and annotation for hesitations and disfluency phenomena, other studies employed either manual transcription and annotation (e.g., Koizumi & In’nami, 2013) or some software for automated annotations, such as Praat (Boersma & Weenink, 2012) and CLAN (MacWhinney, 2000). In particular, the Praat software has the function of annotating silences automatically, and there are scripts that can compute utterance fluency measures (e.g., De Jong & Wempe, 2009). The latest version of the Praat script, developed by De Jong et al. (2021), and other scripts available to date have been validated with certain groups of L2 learners. In other words, the detection of silent pauses and filled pauses by those scripts may not work for other groups of L2 learners. Therefore, the validity of the use of those scripts for automated annotation of pauses is recommended to be checked by manually annotating a small portion of speech data, analogous to intercoder reliability for manual annotation.

Findings and Discussion from the Methodological Synthesis on Writing Fluency

RQ1: Writing Data Collection

Similar to speaking, writing behaviours are likely to be influenced by task characteristics and conditions. While studies exploring the effects of task design characteristics (e.g., Lee, 2019; Lu, 2022; Révész et al., 2017; Spelman Miller, 2000) and task conditions (e.g., Lee, 2019) on L2 writing fluency are on the rise, only few have considered whether and how task features may moderate the link between CF and writing behaviours.

Task characteristics. In terms of task characteristics, only one coding category *Genre* emerged from our synthesis. As shown in Table 11c, most studies employed argumentative tasks, followed by narrative and descriptive tasks in this order. It appears reasonable to assume that, depending on genre, the strength of the link between aspects of cognitive fluency and writing fluency may vary. For example, argumentative writing may induce greater subordination complexity and syntactic sophistication, given that argumentation typically requires taking a position and supporting that position through generating and evaluating evidence. Narratives and descriptive tasks, on the other hand, may elicit increased phrasal complexity, involving the description of characters, events, and/or places. Indeed, Michel et al.'s (2019) learner corpus-based study provided evidence in support of genre effects on the outcomes of L2 writing, observing greater type frequency of base verb forms and modals on argumentative tasks than narrative and descriptive tasks. However, the researchers found that narration was associated with a larger number and variety of past tense forms. Nevertheless, L2 writers may probably face lower difficulty when they convert their ideas into linguistic form during narrative than argumentative writing in terms of verb use, as past tense is usually acquired earlier than modals (Bardovi-Harlig, 2000). In other words, the linguistic encoding processes required to use appropriate verb form-meaning mappings during argumentative writing might be less automatised for L2 writers than those utilised when writing narratives. From this follows that CF might be less predictive of L2 writers' fluency in producing verbs during narrative and descriptive tasks as compared to argumentative writing. When it comes to other type of constructions such as complex nominals, the nature of this relationship might differ, with descriptions and narratives eliciting lower fluency by less proficient writers.

Table 11c. Summary of writing task genres in the pooled studies

Genre	Example	Number of studies
Argumentative	People who are fleeing from situations of political and/or civil unrest should be able to freely enter another country without any requirements or limitations, such as a passport. (Vallejos, 2020)	6
Descriptive	Describe your town and surroundings. (Spelman Miller et al., 2008)	2
Narrative	The first time we try something new can be both exciting and nervous. Write a story about a memorable 'first' in your life. You may include what you did, when and where you did it, who else	3

	was involved, and explain how you felt about the experience. (Lu, 2022)	

Note. Some studies employed multiple writing tasks.

Thus far, only two studies have directly examined the extent to which genre may impact on the relationship between CF and writing fluency. In Lee's (2019) study, L2 English writers at intermediate and advanced proficiency produced one argumentative and one narrative text. Proficiency level was determined by the means of a cloze test, and the keystroke-logging software Inputlog was used to record writing behaviours. Speed fluency was assessed in terms of several measures, including words per minute and number and length of pause-bursts (P-bursts; i.e., text production units between pauses). Pausing patterns were studied through comparing frequency of pauses at various textual units (within and between words, sentences and paragraphs). Revision was expressed as indices of number and length of revision-bursts (R-bursts; i.e., text production units ending with a revision) and revision ratio (i.e., ratio of process and product measures; e.g., number of characters produced during writing divided by number of characters in final text). The study, contrary to Lee's prediction, yielded no effects for genre, that is, CF had a comparable relationship to writing fluency regardless of whether participants engaged in argumentative or narrative writing.

Following Lee's (2019) investigation, Lu's (2022) study compared writing fluency on argumentative and narrative tasks as a function of proficiency. Lu focused on L2 Chinese writing and utilised a cloze test to establish L2 Chinese proficiency. Over two sessions, L2 Chinese learners wrote two argumentative and two narrative texts using the Pinyin method while their keystrokes were logged. The Pinyin method involves entering characters from the Roman alphabet to identify Chinese characters and then selecting the intended character from among homophones. The resulting keystroke logs were analysed in terms of speed fluency, pausing, and revision measures (for details, see Supplementary information). As in Lee (2019), genre did not emerge as a significant moderator of the relationship between CF and writing fluency.

Although previous research has identified no moderating effects of genre, further studies are needed to explore whether this and other task features may influence the link of CF to writing fluency. For example, increased demands on content planning, as in speaking, may decrease the attentional resources available for linguistic encoding. This, in turn, may adversely affect writers with lower CF. Several task characteristics, such as increased reasoning demands (Kuiken & Vedder, 2008; Ruiz-Funes, 2014), greater story complexity (Tavakoli, 2014), and more task elements (Kuiken & Vedder, 2008), are likely to put greater pressure on planning processes. In future research, it would be worthwhile to examine the potential moderating influence of such task characteristics. It would also be interesting to investigate how task characteristics may affect writing fluency as a function of specific linguistic features produced. As discussed above, various genres (and other task features) may elicit different linguistic constructions to varied degrees, and the writers' fluency in producing these constructions may differ, ultimately influencing the relationship between CF and writing fluency.

Task conditions. Parallel to task characteristics, we identified a single coding category, *Time limit*, that distinguished how tasks were implemented in studies of CF and writing fluency. There is an even distribution of studies including or excluding a time limit (n =5 each). Although a time limit does not necessarily imply time pressure (in fact, in some studies piloting determined a sufficient time limit), it is still expected that some writers will display differential behaviours under timed conditions. A time constraint might put greater pressure on all writing stages including linguistic encoding, thereby affecting the fluency of writers with low CF to a greater degree.

To date, only Lee’s (2019) study examined directly how time pressure may influence the link between CF and L2 writing fluency. Under one condition, participants were provided with 30 minutes for writing, whereas the other condition allowed writers 60 minutes for the same tasks. Contrary to the researcher’s expectation, the study yielded no interaction between time limit and proficiency. As Lee speculated, this might have been due to participants' previous experience taking standardised writing tests, which typically involve short time limits.

Clearly, further research is needed to clarify the moderating effect of time limit and other task implementation factors on the relationship between CF and L2 writing fluency. Two task conditions that appear to deserve attention are availability of pre-task planning time and content support. These factors may help learners free up attentional resources due to less pressure on planning content and organisation, thereby allowing writers to dedicate more attention to linguistic encoding. This, in turn, may result in decreased impact of CF on writing fluency. While some studies have confirmed a positive link between these factors and writing outcomes (Johnson, 2017) and fluency behaviours (e.g., Révész et al, 2017), it is yet to be established empirically how CF may moderate the nature of this relationship.

RQ2: Representative Measurements of Automatization in L2 Writing

Table 11d provides a summary of the writing fluency measures utilised in primary studies of CF and writing fluency. As in the area of speaking, previous studies have employed a large variety of measures. In writing, the situation is further complicated by the fact that only few indices have been used in more than one study, and studies utilised varied research designs making it difficult to compare effect sizes across studies. Thus, it is difficult to reach firm conclusions about which measures constitute valid automaticity indices. Nevertheless, it appears worthwhile to discuss which fluency measures have shown a significant association with CF measurements to inform further validation work. We have restricted our discussion to measures that have been used in more than one study, thereby increasing the likelihood that any patterns observed generalise to different contexts and populations.

Table 11c. Summary of significance of relationship between writing fluency measures and proficiency and longitudinal development in the pooled studies

Study	Speed fluency/ Production rate	Pause frequency	Pausing duration	Revision	Composite
Cross-sectional design					
Chenoweth & Hayes (2001)	*			*	
Xu & Ding (2014)	*	*			
Xu & Xia (2021)	*			*	
Longitudinal design					
Kowal (2014)	*				
Spelman Miller et al. (2008)	*	*	*		*
Correlational design					
Lee (2019)	*				
Lu (in press)	*	*	*	*	
Vallejos (2020)		*	*		
Latif (2009)					*

Note. Speed fluency/ production rate includes words/characters per minute, Chinese characters/pinyins over total writing time (excl. pauses), words/characters per P-burst; Pause frequency includes the total number of all pauses or pauses within words or between clauses/sentences/paragraphs pauses (pause > 200s/2000s); Pausing duration includes the total duration of all pauses, pause time over writing time, within words, between Pinyin and Chinese character/words/clauses/sentences/paragraphs (pause > 200s/2000s); Revision includes the number of words/letters produced/in final text, total number, number below word/at word/below clause/transcription levels, number/percentage of R-bursts, characters per R-burst, number of letters deleted/added, number of letters between revision and production point, duration; Composite includes characters/words over P-bursts and R-bursts, number of pauses between revisions, pause length between revisions, writing time (excl. pauses)/number of revisions and pauses

Thus far, *words per minute* and *characters per minute*, measures of speed fluency or production rate, have been the most frequently used indices in studies of CF and writing fluency (4 and 2 studies, respectively). Out of the six studies including this index type (Chenoweth & Hayes, 2001; Latif, 2009; Lee, 2019; Xu & Ding, 2014; Spelmann Miller et al., 2008; Xi & Xia, 2021), all have identified a significant positive link between CF and writing fluency, except for Latif's (2009) study. In the three studies where effect sizes were reported, the strength of these relationships ranged from small (*partial* η^2 = .11; Xu & Xia, 2021,) through medium (r = .41; Xu & Ding, 2014) to large (d = 1.00; Lee, 2019).

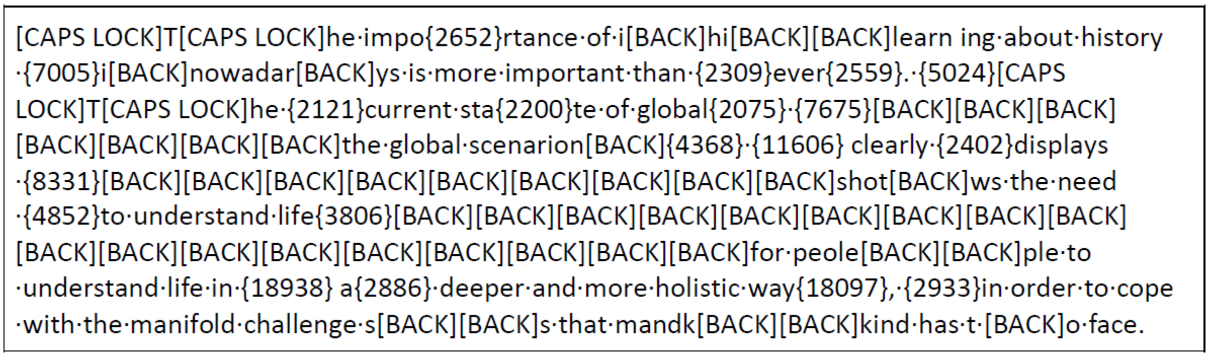
Among measures of pausing, two have been identified to have a significant relationship with linguistic knowledge in more than one study. Both Lu (2022) and Vallejos (2020) identified negative associations of L2 proficiency to *number of pauses between words* and *pause length between sentences*. Writers with higher L2 proficiency produced fewer pauses between words and shorter pauses between sentences. The effect sizes were considerable for both pause frequency between words (R^2 = .17, Vallejos, 2020; b = .56, Lu, 2021) and pause length between sentences (R^2 = .10, Vallejos, 2020; b = .28, Lu, 2021)

Although none of the revision or composite measures were found to have a significant link to linguistic knowledge in more than one study, it is worth noting that, in some studies, some of the revision and composite measures appeared to be related to CF. This was somewhat unexpected, as revision is assumed to draw on writers' metacognitive knowledge. Rather than indicators of automaticity, these associations were probably an artifact of the broad operationalisation of CF in the current study. This broader CF definition included performance on general proficiency tests and extent of linguistic experience, which likely captured differences in both declarative and automatised knowledge across participants.

In summary, based on the little research available, three measures appear to have potential to tap automaticity in L2 writing: words or characters per minute, pause frequency between words, and pause length between sentences. Clearly, more research is needed to explore further writing fluency indices in relation to L2 proficiency.

RQ3: Analysis tools and software

Most primary studies have utilised keystroke-logging software to record and analyse writing fluency behaviours. Various programs have been employed (see Lindgren et al., 2019 for a review of logging tools); among them, Inputlog (Leijten & Van Waes, 2013) has been the most popular among researchers exploring the relationship between proficiency and fluency in L2 writing. Keystroke-logging programs record all the writers' keystrokes while they are composing, and the resulting logs can be used to extract indices of speed fluency, pausing, and revision (Van Waes et al., 2015). These measures can be obtained manually (Lu, 2022), but some software (e.g., Inputlog) allows for obtaining writing fluency indices automatically. Figure 11a provides an example Inputlog output taken from Révész et al. (2017).



```
[CAPS LOCK]T[CAPS LOCK]he·impo{2652}rtance·of·i[BACK]hi[BACK][BACK]learn ing·about·history
·{7005}i[BACK]nowadar[BACK]ys·is·more·important·than·{2309}ever{2559}.·{5024}[CAPS
LOCK]T[CAPS LOCK]he·{2121}current·sta{2200}te·of·global{2075}·{7675}[BACK][BACK][BACK]
[BACK][BACK][BACK][BACK]the·global·scenarion[BACK]{4368}·{11606} clearly·{2402}displays
·{8331}[BACK][BACK][BACK][BACK][BACK][BACK][BACK][BACK][BACK]shot[BACK]ws·the·need
·{4852}to·understand·life{3806}[BACK][BACK][BACK][BACK][BACK][BACK][BACK][BACK][BACK]
[BACK][BACK][BACK][BACK][BACK][BACK][BACK][BACK][BACK]for·peole[BACK][BACK]ple·to
·understand·life·in·{18938} a{2886}·deeper·and·more·holistic·way{18097},·{2933}in·order·to·cope
·with·the·manifold·challenge·s[BACK][BACK]s·that·mandk[BACK][BACK]kind·has·t·[BACK]o·face.
```

Figure 11a. Excerpt from an Inputlog output file. The labels in square brackets "[ ]" refer to function keys. Curly brackets "{ }" indicate pauses and the number between them provides pause duration in ms. The symbol "·" indicates SPACE. Pause threshold was 2s.

Methodological Guidelines

Utterance Fluency

In this section, the step-by-step methodological guidelines to measure learners' automaticity in spontaneous L2 speaking are introduced, consisting of three major phases—preparing speech elicitation tasks, speech data elicitation, and fluency measure calculation (for a similar overview, see Suzuki et al., 2021). As the first step, researchers choose the prompts for speech elicitation with regard to task characteristics, such as the mode of speaking (monologue vs. dialogue) and the demands on content planning (continuum between open and closed tasks), as well as task conditions including the amount of time for pre-task planning and the time pressure while speaking (see also Tavakoli & Wright, 2020).

Using the speech samples collected, researchers calculate utterance fluency measures to capture the speakers' automaticity. Ideally, the selection of utterance fluency measures should be done prior to speech data collection. As noted earlier, the validity of utterance fluency or what utterance fluency measures are reflective of can be affected by how the speech is elicited (i.e., task characteristics and task condition), meaning that the target fluency measures should be considered in designing speech elicitation tasks. However, thanks to the development of open learner corpus data (e.g., ICNALE http://language.sakura.ne.jp/icnale/symposium.html; the Rated L2 speech corpus http://isle.illinois.edu/sst/data/RatedL2Speech/), target utterance fluency measures can be chosen after exploring the speech data available to researchers with respect to some potential effects of speech elicitation methods.

When it comes to the analysis of speech samples, utterance fluency measures are calculated through three major phases: speech transcription, disfluency feature annotation, and measure computation. First, traditionally, researchers manually transcribe speech data into written texts so that the number of words and syllables can be calculated. The number of syllables or words in speech is then used to standardize disfluency features in speech samples to calculate utterance fluency measures such as articulation rate and speech rate. Due to the recent technological advances, there have been several options to compute the written transcription by means of automated speech-to-text services, such as Google Cloud (https://cloud.google.com/speech-to-text) and Rev.ai (https://www.rev.ai/), both of which are commercial services. As the scope of transcription can differ across services and transcribers (e.g., whether to include fillers; see also Foster et al., 2000) and the accuracy of transcription can depend on the accentedness of speaker's speech, it would be recommended to initially use those services with a few samples to check the accuracy of transcription. In research articles, the reliability of those automated speech recognition services should be reported, for instance, using the index of word error rate (WER).

Second, the frequency and duration of disfluency features, such as silent and filled pauses, self-corrections and self-repetitions, need to be annotated to speech samples wit §h assistance of software such as Praat. In previous studies, disfluency features are manually annotated (e.g., Saito et al., 2018), considering the fact that the identification of some disfluency phenomena requires the annotator's interpretation (for automated annotation, see Matsuura et al., in press). In other words, it is important to consider how those disfluency features are operationalized in the study prior to the actual annotation of those features. Another essential issue around the annotation of temporal features is the threshold for silent pauses. A meta-analysis on the relationship between utterance and perceived fluency measures demonstrated that from the perspective of predictive power for listener-based judgements, the optimal threshold for silent pauses can be 250 ms (Suzuki et al., 2021), meaning that silences longer than 250ms can be regarded as silent pauses and may tend to be reflective of breakdowns in L2 speech production processes. In other words, shorter silences have been regarded as micropauses (Riggenbach, 1991), which are unlikely to reflect such breakdowns. However, depending on the target disfluency phenomena, some automatic systems for annotating disfluency features have been developed and available to date (Chen et al., 2018; De Jong et al., 2021; Matsuura et al., in press).

Finally, with the annotation of disfluency features and the information about the number of syllables or words produced in the speech samples, the target utterance fluency features are computed. To check whether the computation process is successful, the descriptive statistics of the fluency measures as well as the shape of distributions need to be inspected. Since most fluency measures are based on the frequency of some phenomenon, their distributions tend to be positively skewed or similar to a zero-inflated distribution, particularly when most participants do not produce the target phenomenon such as filled pauses. Figure 11b below is a set of the density plots describing the distribution of mid-clause pause ratio in Suzuki and Kormos' (in press; *N* = 128) study across four different speaking tasks. As the measure is based on the ratio of mid-clause pauses to the number of syllables produced, the range of the measure tend be between 0 and 1. In addition, the frequency of pauses is generally not quite high in speech and thus the large portion of the distribution is located under 0.5. If the annotation of disfluency features is conducted with *Praat*, the TextGrid files need to be transformed into some table-formatted files such as the csv file format. The example TextGrid file and R code to convert TextGrid files to a .csv file with the phonfieldwork package (Moroz, 2020) is available via OSF (https://osf.io/d2w4u/).

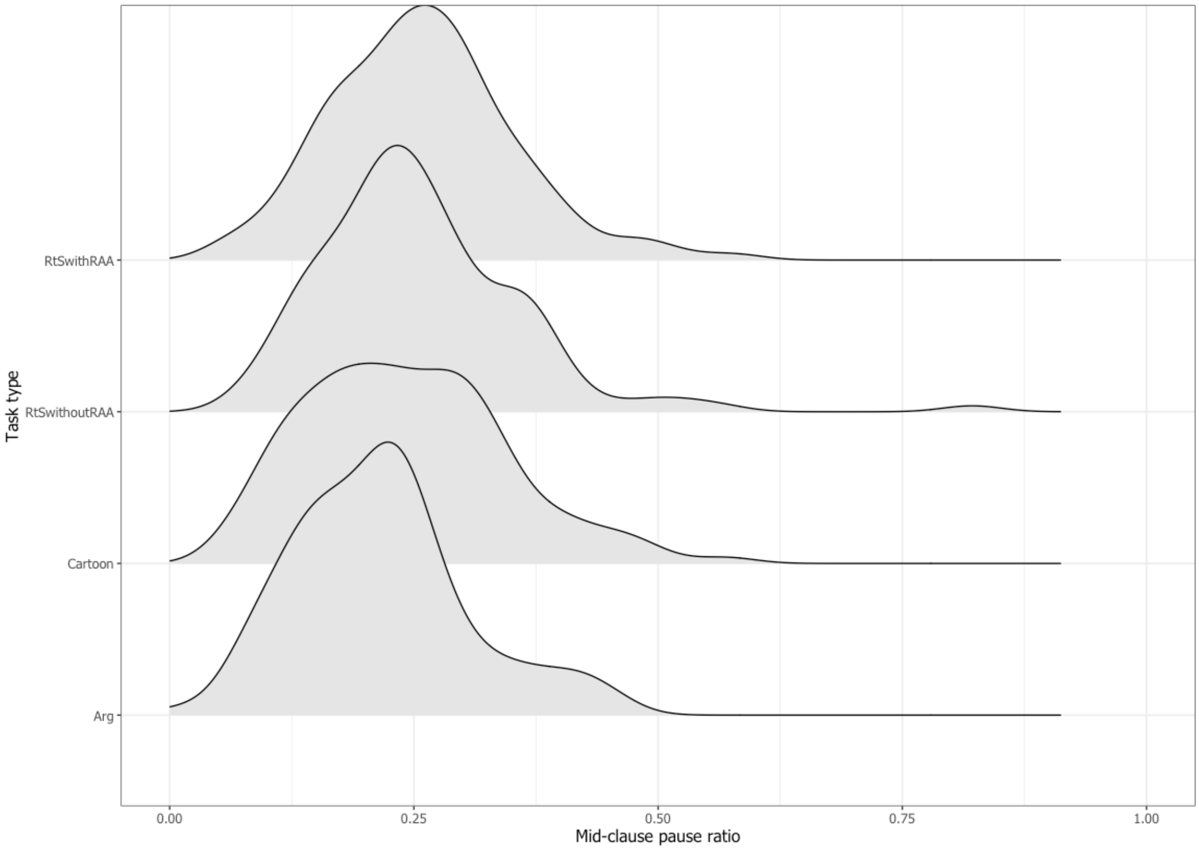


Figure 11b. The density plot of mid-clause pause ratio across four speaking tasks in Suzuki and Kormos' (in press) study.

Exemplary Study (Speaking)

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.

Building on Segalowitz's (2010) framework of cognitive and utterance 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.

Writing Fluency

When investigating automaticity in L2 writing, one of the first decisions to make is what elicitation instrument to use. It is important to select a writing task that is appropriate for the participants' writing ability while being challenging enough to detect differential levels of automaticity among them. As discussed above, task difficulty can be altered through manipulating task design characteristics and implementation factors.

Prior to data collection, it is also important to identify suitable logging software to record the writing process. Among the factors to consider are whether handwritten or typed production the researcher intends to capture and whether the researcher intends to align the software with other tools such as eye-tracking, speech recognition, linguistic analysis, or automated feedback software (for a review of logging tools, see Lindgren et al., 2019). Another essential consideration is the type of analyses planned. For example, in studies of automaticity a key question is whether the software allows for automated fluency analyses.

Once the data have been collected, the process of analysis can begin. First, data quality needs to be checked. Some data loss is expected due to technical errors and/or incomplete participant data. The remaining logs need to be filtered to remove noise to ensure that subsequent analyses are limited to actual text production. For example, it is important to remove the beginning and end of the writing process which involve the start and closing of the logging program. Some logging programs perform this pre-processing stage automatically.

The next step entails deciding on the type of text production the researcher wants to study. For example, Baaijen et al. (2012) made a distinction between the initial drafting stage and postdraft revision, between linear and nonlinear transitions during initial drafting, and between nonlinear revision events at the leading edge or elsewhere.

Finally, the researcher needs to identify measures of writing fluency that they will extract from the dataset. Given the scarce research on the link between CF and writing fluency, it is difficult to make recommendations about what specific measures to adopt. Nevertheless, a few considerations are worth highlighting. When investigating pausing, a key issue is what pause threshold to define. The majority of L2 studies have employed a 2s threshold, but a limitation of this approach is that pauses longer than 2s likely reflect higher-order writing processes. Thus, they do not capture lower-level linguistic encoding processes that are particularly relevant to studying automaticity. Possible ways to resolve this issue are to employ a lower pause threshold (e.g., Michel et al., 2020; Vallejos, 2020; Van Waes & Leijten, 2013) or use raw length of intervals between keystroke events rather than pause thresholds as a basis for studying pause patterns (Baaijen et al., 2012).

Another decision concerns whether to compute separate or combined indices for P-bursts and R-bursts (Galbraith & Baaijen, 2019). P-bursts are assumed to indicate complete text production units, thus mirroring the capacity of the translator (Kaufer et al., 1986). On the other hand, R-bursts are presumed to represent incomplete units of production, where the writer decided to terminate the burst before it was completed. Hence, R-bursts are less likely to capture linguistic encoding capacity than P-bursts (Kaufer et al., 1986). Indeed, Baaijen et al. (2012) found that writers in their dataset produced longer P-bursts than R-bursts, confirming that P-bursts are more relevant to measuring automaticity than R-bursts. Before this assumption is empirically confirmed, studies of writing fluency should calculate separate indices for P-bursts and R-bursts rather than combine them with a view to identifying valid indices of automaticity in writing.

When researchers compute measures of pausing and revision with automaticity in mind, it is also necessary to decide whether to obtain a total score for various indices or distinguish them according to textual location. There is growing evidence in L2 writing research that pausing and revision at different textual locations are associated with different underlying processes, with pausing and revision behaviours at lower textual units more likely to be related to linguistic encoding (e.g., Révész et al., 2017; Révész et al., 2019; Spelman Miller, 2000). Thus, we would expect fluency indices involving pausing at lower textual locations to be more suitable for measuring automaticity. Although the current evidence is mixed regarding this prediction as discussed above (Lee, 2019; Lu, 2022; Vallejos, 2020), more studies including measures of pausing and revision at different textual boundaries are needed to help identify valid measures of automatization in writing.

Exemplary Study (Writing)

Lu, X. (2022) Second language Chinese computer-based writing by learners with alphabetic first languages: Writing behaviours, second language proficiency, genre, and text quality. *Language Learning, 72*(1), 45–86.

This study investigated the link between L2 Chinese proficiency and writing behaviours and whether this relationship was influenced by genre. In addition, it examined the extent to which writing behaviours relate to text quality.

Thirty-two L2 writers of Chinese carried out two argumentative and two narrative tasks using the Pinyin method to type their texts. Participants wrote one narrative and one argumentative essay in each one of two sessions, with task versions counterbalanced across sessions. A 30-minute time limit was specified for each writing task. While composing, participants' keystrokes were logged. A cloze test was employed to establish L2 proficiency, and holistic ratings were used to assess text quality.

Writing behaviours were evaluated in terms of speed fluency, pausing, and revision. Speed fluency was computed by dividing the number of Pinyin/Chinese characters with the total time spent writing excluding pauses. Pausing patterns were evaluated in terms of pause frequency and duration by location (within words; between words, clauses, sentences; between Pinyin and Chinese characters; between revision pauses), with a pause threshold of 2 s. Revision was studied through calculating the number of revisions according to location (below word, word level, below clause, clause level and above) and transcription (change to Pinyin or Chinese character).

Series of mixed effects regression models found that proficiency had a positive relationship with speed fluency, between-word pause duration, and below clause revision frequency, and it negatively correlated with between-word pause frequency and between-sentence pause duration. The study yielded no moderating effects for genre, but texts with higher ratings were associated with fewer between-word pauses.

Exercises on Measurement Development

Speech Data Analysis

Exercise 1

Using the utterances below, discuss the advantage of using syllables as the standardizing unit for calculating speed fluency measures over using words.

Note. Both utterances were produced in 5 seconds.

Speaker A (Upper intermediate)

The psychiatrist recommended the businessman to rest mentally and physically...

- No. of syllables: 24
- No. of words: 11

Speaker B (Beginner)

He asked the guy to have some rest in the...

- No. of syllables: 10
- No. of words: 11

Exercise 2

Using the utterances below, discuss the advantage of using a pruned transcription as the standardizing unit for calculating utterance fluency measures over an unpruned transcription.

Note. The square brackets in the utterances indicate disfluency words, which are either repeated or corrected.

<u>Speaker C (Upper intermediate)</u>
One of the lovely things that occurred to [the] the area

- No. of syllables: 15 (unpruned) / 14 (pruned)
- No. of words: 11 (unpruned) / 10 (pruned)

<u>Speaker D (Beginner)</u>
[Some some of the] some of the areas [were] were cultivated

- No. of syllables: 15 (unpruned) / 11 (pruned)
- No. of words: 11 (unpruned) / 7 (pruned)

Writing data analysis

Exercise 3

Consider the Inputlog output excerpt presented in Figure 11a. Based on the keystroke log, produce the resulting text.

Exercise 4

Based on the Inputlog output excerpt presented in Figure 11a, calculate the following indices:

1. Number of pauses within words, number of pauses between words, and number of pauses between sentences. Pauses separated by SPACE or punctuation marks count as one pause.
2. Number of revisions below word level, at word level (i.e., an entire word is revised), below clause (i.e., a unit larger than a word but smaller than a clause is revised), and clause and above (i.e., an entire clause or a larger unit is revised).

References

To view the list of all studies in this synthesis, visit the following link: https://osf.io/d2w4u/

Abdel Latif, M. M. (2013). What do we mean by writing fluency and how can it be validly measured? *Applied Linguistics, 34*, 99–105.

Albarqi, G., & Tavakoli, P. (in press). The effects of proficiency level and dual-task condition on L2 self-monitoring behavior. *Studies in Second Language Acquisition*.

Awwad, A., & Alhamad, R. (2021). Online task planning and L2 oral fluency: does manipulating time pressure affect fluency in L2 monologic oral narratives? *International Review of Applied Linguistics in Language Teaching, 59(4), 605–627.*

Baaijen, V. M., Galbraith. D., & de Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication, 29*, 246– 277.

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use.* Blackwell.

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.

Chapelle, C. A. (2006). DIALANG: A diagnostic language test in 14 European languages. *Language Testing*, *23*(4), 544–550.

Chen, L., Zechner, K., Yoon, S., Evanini, K., Wang, X., Loukina, A., Tao, J., Davis, L., Lee, C. M., Ma, M., Mundkowsky, R., Lu, C., Leong, C. W., & Gyawali, B. (2018). Automated Scoring of Nonnative Speech Using the SpeechRater SM v. 5.0 Engine. *ETS Research Report Series*, *1*, 1–31.

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.

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., Pacilly, J., & Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice*, *28*(4), 456–476.

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.

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

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.

Galbraith, D. & Baaijen, V. M. (2019). Aligning keystrokes with cognitive processes in writing. In Observing writing (pp. 306–325). Brill.

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.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*, 369–388.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Language Learning Journal*, *36*(2), 153–166.

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, lexical complexity, accuracy, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing,* *37*, 13-38.

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*(3), 569–591.

Kaufer, D., Hayes, J.R., & Flower, L.S. (1986). Composing written sentences. *Research in the Teaching of English, 20*(2), 121–140.

Kellogg, R. T. (1996). A model of working memory in writing. In M. C. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and application* (pp. 57–71). Lawrence Erlbaum Associates.

Koizumi, R., & In'nami, Y. (2013). Vocabulary Knowledge and Speaking Proficiency among Second Language Learners from Novice to Intermediate Levels. *Journal of Language Teaching and Research*, *4*(5), 900–913.

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

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17*, 48– 60.

Lee, J. (2019). *The effects of time constraints, genre, and proficiency on l2 writing fluency behaviors and linguistic outcomes*. [Doctoral dissertation, Michigan State University].

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing Research: Using Inputlog to analyze writing processes. *Written Communication 30*(3), 358-392.

Lindgren, E., Knospe, Y., & Sullivan, K. P. H. (2019). Researching writing with observational logging tools from 2006 to the present. In *Observing writing: insights from keystroke logging and handwriting* (pp. 1–29).

Lu, X. (2022). Second language Chinese computer-based writing by learners with alphabetic first languages: Writing behaviors, second language proficiency, genre, and text quality. *Language Learning, 72* (1), 45-86.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed. In *The CHILDES project: Tools for analyzing talk: Transcription format and programs, Vol. 1, 3rd ed.* Lawrence Erlbaum Associates Publishers.

Maie, R., & Godfroid, A. (2022). Controlled and Automatic Processing in the Acceptability Judgment Task: An Eye-Tracking Study. *Language Learning*, *72*(1), 158–197.

Matsuura, R., Suzuki, S., Saeki, M., Ogawa, T., & Matsuyama, Y. (in press). Refinement of utterance fluency feature extraction and automated scoring of L2 oral fluency with dialogic features. *Proc. The 14th Asia Pacific Signal and Information Processing Association Annual Summit and Conference* (APSIPA ASC).

Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: evidence from a large learner corpus of A1 to C2 writings. *Instructed Second Language Acquisition*, *3*(2), 124–152.

Michel, M., Révész, A., Lu, X., Kourtali, N.-E., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study. *Second Language Research, 36*, 307–33.

Moroz G (2020). Phonetic fieldwork and experiments with phonfieldwork package. https://CRAN.R-project.org/package=phonfieldwork.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601.

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

Plonsky, L., & Oswald, F. L. (2014). How Big Is "Big"? Interpreting Effect Sizes in L2 Research. *Language Learning*, *64*(4), 878–912.

Révész, A., Kourtali, N., & Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning, 67*, 208–241.

Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviours: A mixed-methods study. *Studies in Second Language Acquisition, 41*, 605–631.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423.

Ruiz-Funes, M. (2014). Task complexity and linguistic performance in advanced college-level foreign language writing. In H. Byrnes & R. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 163–191). Amsterdam: John Benjamins.

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.

Schoonen, R., Snellings, P., Stevenson, M., & van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchón (Ed.), *Learning, teaching, and researching writing in foreign language contexts* (77-101). Multilingual Matters.

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.

Spelman Miller, K. (2000). Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research, 4*, 123–148.

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. (in press). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*.

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.

Tavakoli, P. (2014). Storyline complexity and syntactic complexity in writing and speaking tasks. In H. Byrnes & R. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 163– 191). Amsterdam: John Benjamins.

Tavakoli, P. (2019). Automaticity, fluency and second language task performance. In Z. Wen & M. J. Ahmadian (Eds.), *Researching L2 task performance and pedagogy: In honour of Peter Skehan* (pp. 39–52). John Benjamins Publishing Company.

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

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.

Vallejos, C. (2020). *Fluency, working memory and second language proficiency in multicompetent writers*. [Doctoral dissertation, Georgetown University].

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325–344.

Van Waes, L., & Leijten, M. (2015). Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers & Composition, 38*, 79–95.

Van Waes, L., Leijten, M, Lindgren, E, & Wengelin, Å. (2015). Keystroke logging in writing research: Analyzing online writing processes. In C. MacArthur, A. S. Graham, & J. Fitzgerald (Eds.), Handbook of writing research 2 (pp. 410-426). Guilford Press.

Wolf-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center.