To be published in *The Modern Language Journal* (Wiley Blackwell)

To cite: 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*.

The Relationship Between Utterance and Perceived Fluency: A Meta-analysis of Correlational Studies

Shungo Suzuki[1]
Lancaster University

Judit Kormos
Lancaster University

Takumi Uchihara
Waseda University

Abstract

Listener-based judgements of fluency play an important role in second language (L2) communication contexts and in L2 assessment. Accordingly, our meta-analysis examined the relationship between different aspects of utterance fluency and listener-based judgements of perceived fluency by analyzing primary studies reporting correlation coefficients between objective measures of temporal features and subjective ratings of fluency. We analyzed 263 effect sizes from 22 studies (N = 335–746) to calculate the mean effect sizes of the links between utterance and perceived fluency. We also investigated the moderator effects of 11 methodological factors, such as speech stimuli, listeners' background, rating procedure, and computation of utterance fluency measures, on the relationship between utterance and perceived fluency. Perceived fluency was strongly associated with speed and pause frequency (r = |.59–62|), moderately with pause duration (r = |.46|), and weakly with repair fluency (r = |.20|). Composite measures showed the strongest effect sizes (r = |.72–.76|). Moderator analyses revealed that the utterance-perceived fluency link is influenced by methodological variables related to how speech samples are prepared for listeners' judgements and how listeners' attention is directed in evaluations of fluency. These findings suggest future directions for L2 fluency research and implications for language assessment.

Keywords: perceived fluency; utterance fluency; meta-analysis; speech perception; second language speaking

[1] We are grateful to *The Modern Language Journal* reviewers as well as the journal editor, Marta Antón, and the handling editor, Shawn Loewen, for their constructive feedback on earlier versions of the manuscript. We would also like to thank Pavel Trofimovich, Hans Rutger Bosker, Michael McGuire; Kazuya Saito, Fumiyo Nakatsuhara, Noriko Iwashita for providing information needed to complete our dataset for the current meta-analysis.

Second language (L2) oral fluency has been recognized as an essential characteristic for successful L2 communication. An optimal level of oral fluency is necessary for speakers to maintain listeners' attention and to be able to save face (Lennon, 2000). It is thus useful for L2 speakers and language teachers to know the extent to which different speech characteristics, such as speed of delivery and hesitations, contribute to listeners' perceptions. From a pedagogical perspective, understanding the role of temporal features in L2 speech perception can yield valuable information for setting curricular objectives and enhancing L2 learners' fluency in classroom language teaching. As oral fluency is a robust indicator of L2 oral proficiency (Baker-Smemoe et al., 2014; De Jong et al., 2013; Tavakoli et al., 2020), listener-based judgements of fluency also play a crucial role in language assessment contexts. Therefore, a better understanding of the association between speech characteristics and listener-based judgements of fluency is of great importance for the development of research-informed assessment rubrics, rater training, and automated scoring systems (see De Jong, 2018; Duijm et al., 2018; Ginther et al., 2010), which in turn has a substantial impact on high-stake proficiency tests. L2 fluency research has thus examined how listeners' perceptions of fluency are associated with temporal features of the speech (Bosker et al., 2013; Kormos & Dénes, 2004; Rossiter, 2009; Saito et al., 2018; Suzuki & Kormos, 2020).

In the literature on L2 fluency, listener-based judgements of fluency and temporal features of speech have been termed *perceived fluency* and *utterance fluency*, respectively (Segalowitz, 2010, 2016). Previous studies have shown that perceived fluency is primarily associated with speed of delivery and pausing behaviour (Saito et al., 2018; Suzuki & Kormos, 2020). However, the findings regarding the contribution of disfluency phenomena, such as self-repetition and false starts, to perceived fluency are contradictory. In addition, the extent to which utterance fluency measures explain the variance of in perceived fluency scores has been found to vary considerably across studies. From a methodological perspective, research on the utterance-perceived fluency link entails a range of methodological choices, such as the selection of the group of listeners, target language to be investigated, and speaking tasks to be used to elicit speech samples. Therefore, the contradictory findings might have been due to methodological differences across studies.

The lack of a thorough understanding of how methodological choices affect the utterance-perceived fluency link may reduce the transferability of findings to the domain of L2 language assessment. Due to a large number of methodological factors, it is arguably unrealistic for individual studies to address those concerns. However, meta-analyses, albeit restricted to the existing methodological trends, can draw relatively robust conclusions regarding the target research domain with a higher statistical power (Hunter & Schmidt, 2015). To the best of our knowledge, no study to date has systematically meta-analysed the utterance-perceived fluency link (for a meta-analysis including perceived and utterance fluency as pronunciation measurements for instructional outcome, see Saito & Plonsky, 2019). Therefore, the current study aims to synthesise and meta-analyse prior work on the utterance-perceived fluency link with regard to a comprehensive set of methodological factors as moderator variables. Based on the findings, we also suggest methodological improvements for the assessment and measurement of L2 fluency.

In this paper we first provide a theoretical and methodological overview of previous research on utterance fluency and perceived fluency. This is followed by a description of our research procedure and a presentation of the findings. Next, we discuss the results of our research with reference to the moderator effects of methodological factors. We conclude our paper by outlining the implications for L2 fluency research and language testing.

DEFINITIONS OF PERCEIVED FLUENCY AND UTTERANCE FLUENCY

When making judgements about perceived fluency, listeners can either exclusively focus on the temporal features of speech or subjectively evaluate the speaker's capability to mobilize their linguistic resources. Previous research findings suggest that even while having received an instruction to focus on temporal features, raters' perceptions of fluency tend to be influenced by non-temporal features as well (e.g., grammatical errors; Kormos & Dénes, 2004; Rossiter, 2009; Suzuki & Kormos, 2020). Therefore, perceived fluency is closely associated with cognitive fluency, which involves a range of linguistic knowledge and processing skills (De Jong et al., 2013). Listeners inherently make inferences about how efficiently the speaker encodes their intended message by paying selective attention to utterance features that they believe reflect the speaker's efficiency of mobilizing L2 knowledge for speech production (i.e., cognitive fluency; Segalowitz, 2010).

Within Segalowitz's (2010, 2016) framework, utterance fluency refers to observable temporal features, such as pauses and hesitations, that reflect the operation of L2 speech production mechanisms (i.e., cognitive fluency). Utterance fluency is generally divided into a triad of utterance fluency subcomponents—speed, breakdown, and repair fluency (Tavakoli & Skehan, 2005). *Speed fluency* is concerned with the density of information or the speed of delivery, and it is typically measured by articulation rate (i.e., the mean number of syllables produced per minute, excluding pauses). *Breakdown fluency* refers to pausing behaviours including silent and filled pauses. Breakdown fluency is traditionally operationalized in terms of the length and frequency of pauses. There has been an ongoing debate about the minimum length of pauses that reflect breakdowns attributed to disruptions in linguistic processes in speech production, such as lexical retrieval and syntactic procedures (De Jong & Bosker, 2013; De Jong et al., 2013), because short pauses are less likely to reflect such breakdowns in speech production processes (i.e., so-called *micropauses*; Riggenbach, 1991). Thus, scholars have attempted to identify an optimal threshold for silent pause length and generally define pauses as silence longer than 250 ms (De Jong, 2016b; De Jong & Bosker, 2013). Recent studies have also recognized the importance of pause location in predicting perceived fluency (Kahng, 2018; Saito et al., 2018; Suzuki & Kormos, 2020). Pauses in the middle of clauses have been found to be more strongly associated with perceived fluency than pauses at clause boundaries, because pauses within clauses are hypothesized to reflect disruptions in L2-specific linguistic processing (De Jong, 2016b; Suzuki & Kormos, 2020). Finally, repair fluency covers a range of disfluency phenomena including self-corrections, false starts, and verbatim repetitions. Repair fluency is in a supplementary relationship with breakdown fluency (Williams & Korko, 2019), as repairs can reflect the operation of self-monitoring processes (Kormos, 2006) and offer an opportunity for speakers to buy time to deal with disruptions in speech production processes (Bui et al., 2019). Repair fluency has been found to be consistent across first language (L1) and L2 production (Peltonen & Lintunen, 2016) and across L2 proficiency levels (Tavakoli et al., 2020), suggesting that it is more strongly associated with individual speaking style than L2 competence.

LINK BETWEEN UTTERANCE AND PERCEIVED FLUENCY

L2 fluency research has extensively investigated which temporal features of utterances can explain listeners' perceptions of fluency. Previous studies have shown that perceived fluency is primarily associated with speed and breakdown fluency and, to a lesser degree, with repair fluency (for a similar review, see Saito et al., 2018; Suzuki & Kormos, 2020). Despite the findings of large variances in perceived fluency scores explained by a set of utterance fluency measures, there is still quite a large variability in the amount of variance explained across studies (e.g., $R^2 = 0.84$ in Bosker et al., 2013 vs. 0.57 in Saito et al., 2018).

It may thus be plausible that the connection between utterance and perceived fluency is affected by methodological differences across studies.

In addition to the amount of explained variance of perceived fluency scores, there are several inconsistent findings regarding the utterance-perceived fluency link. First, some studies have shown that speed fluency measures have higher correlation coefficients with perceived fluency scores than breakdown fluency measures (Bosker, et al., 2013; Kormos & Dénes, 2004). However, other studies, especially those considering pause location, have reported that breakdown fluency measures correlate with perceived fluency scores more strongly than speed fluency measures (Cucchiarini et al., 2002; Suzuki & Kormos, 2020). These contradictory findings may indicate that mid-clause pause frequency measures tend to have a strong predictive power for perceived fluency. The relationship between breakdown fluency and perceived fluency is also influenced by the type of pause—silent versus filled pauses. Measures based on silent pauses tend to correlate with perceived fluency scores more strongly than those based on filled pauses (Bosker et al., 2013; Cucchiarini et al., 2002; Suzuki & Kormos, 2020).

Second, another inconsistent finding observed in L2 fluency research is the role of repair fluency in perceived fluency. Since repair fluency entails a range of disfluency phenomena, the selection of targeted disfluency phenomena has varied across previous studies. Studies that did not distinguish different types of disfluency phenomena and used a composite measure such as disfluency rate, tended to find no significant correlation between repair fluency measures and perceived fluency scores (e.g., Cucchiarini et al., 2002; Kormos & Dénes, 2004; Suzuki & Kormos, 2020). On the other hand, repair fluency measures with a particular focus on specific disfluency phenomena, such as self-repetitions and self-corrections, were found to correlate significantly with perceived fluency scores in some studies (e.g., Bosker et al., 2013), but not in others (Magne et al., 2019; Saito et al., 2018).

Third, composite measures, such as speech rate and mean length of run (MLR), can capture multiple dimensions of utterance fluency and thus tend to correlate strongly with perceived fluency scores (Derwing et al., 2009; Kormos & Dénes, 2004; Préfontaine et al., 2016; Rossiter, 2009). Despite having such strong predictive power for perceived fluency, it is not always appropriate to select these composite measures, especially when researchers aim to use multiple utterance fluency measures to predict perceived fluency scores (Bosker et al., 2013). These composite measures make it difficult to interpret the findings, because it is unclear which temporal features a given composite measure represents (e.g., speed vs. pause frequency for MLR).

MODERATOR VARIABLES IN THE UTTERANCE-PERCEIVED FLUENCY LINK

The preceding literature review suggests that methodological differences across studies may contribute to inconsistent results regarding the relationship between utterance and perceived fluency. As illustrated in Figure 1, research into the utterance-perceived fluency link involves five major methodological phases, each of which entails a set of methodological decisions.

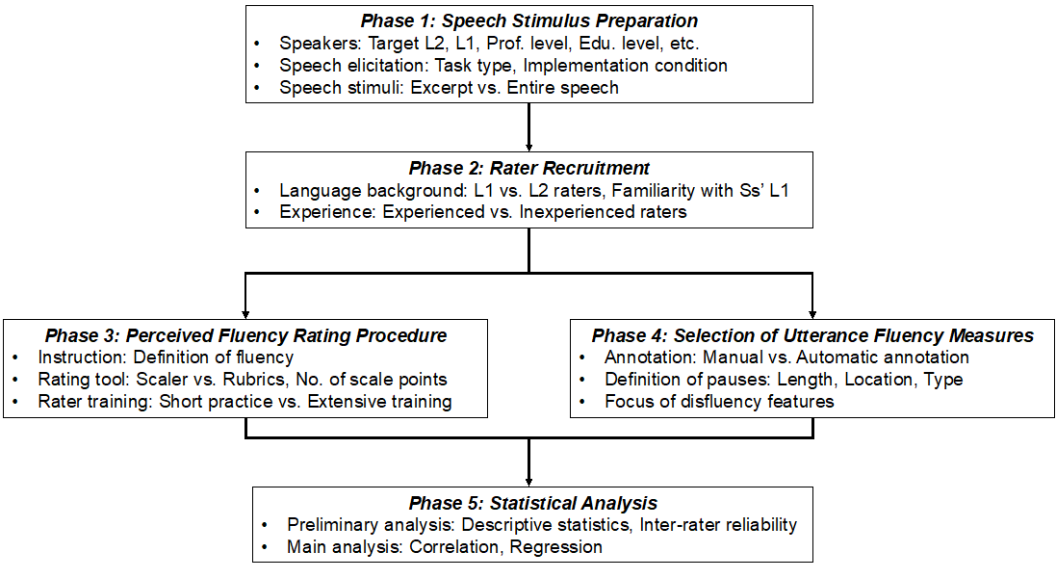


FIGURE 1. Five Major Phases in L2 Research into Utterance-perceived Fluency link

Speech Stimulus Preparation

The first phase of L2 perceived fluency research is the preparation of speech stimuli for fluency judgements. First, researchers specify the target population of speakers in terms of L1, L2, proficiency level, age, etc. Second, researchers determine speech elicitation methods, such as speaking task type and implementation condition. Cucchiarini et al. (2002) found that the correlation coefficients between perceived fluency scores and various utterance fluency measures were higher in controlled speech (read-aloud speech) than in spontaneous speech (opinion-giving speech). Similarly, predicting perceived fluency scores using mixed-effects modelling, Préfontaine et al. (2016) reported that the relative magnitudes of regression coefficients among utterance fluency measures varied across tasks. In addition, L2 fluency research has recently been extended to dialogic speaking tasks and has argued that dialogic fluency is theoretically distinctive from monologic fluency (Tavakoli, 2016; Tavakoli & Wright, 2020). After collecting speech data, researchers need to decide whether speech stimuli are presented to their raters either as entire speech samples or as short excerpts from those samples. Some scholars claim that short excerpts of speech (e.g., initial 30 seconds) are sufficient to elicit listener perception data in research contexts (Derwing et al., 2006, 2009), whereas some studies have presented entire speech as stimuli, emphasising the ecological validity of findings for language assessment contexts (e.g., Préfontaine et al., 2016; Suzuki & Kormos, 2020). However, it is still unclear how the length of speech stimuli affects the connection between utterance fluency features and listeners' perceptions of L2 fluency.

Rater Recruitment

The second phase of L2 perceived fluency research is the recruitment of listeners for perceived fluency judgements. One of the relevant listener characteristics is language background, namely whether raters are speakers of the target language as their L1 or L2. Previous studies examining the effects of language background (Magne et al., 2019; Rossiter, 2009; Saito et al., 2018) have reported that L1 and L2 listeners' perceptions of fluency tended to be similar. Moreover, listener-based judgements of speech can potentially be influenced by raters' experience, such as examination experience for high-stakes tests, teaching experience,

and expertise in linguistics (Isaacs & Thomson, 2013). However, these potential mediating factors have not yet been systematically examined (for a rare exception, see Rossiter, 2009).

Perceived Fluency Rating Procedure

The third phase of L2 perceived fluency research is the actual implementation of rating procedures. Previous studies either instructed their listeners to focus narrowly on temporal aspects of speech (i.e., lower-order fluency; e.g., Bosker et al., 2013) or provided no definition to allow for their intuitive judgements of fluency, typically interpreted as overall command of language (i.e., higher-order fluency; e.g., Suzuki & Kormos, 2020). In the former case, most studies presented a narrow definition of fluency based on research findings, while some studies employed existing assessment tools, such as the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) assessment scales (Préfontaine et al., 2016) or even created rubrics for their research purposes (Sato, 2014). Although different presentations of the definition of fluency to listeners may impact on listeners' judgements, this issue has rarely been investigated (cf. Dressler & O'Brien, 2019).

When it comes to rating scales, there is great variation in the number of scale points. According to Isaacs and Thomson (2013), the use of five- and nine-point scales did not result in significant differences in rater severity, but their Rasch probability plots revealed that the distinguishability of adjacent levels on scales was more meaningful on a five-point scale rather than on a nine-point scale.

Researchers also need to decide the amount of practice and training that raters will have before the actual rating task. Most previous studies have asked their raters to judge a few speech samples to familiarize themselves with the use of rating scales. In other cases, especially when rubrics were used for fluency judgements, raters were given in-depth training. They received feedback from researchers, and there was a discussion among raters so that they would reach agreement on the scores awarded (e.g., Sato, 2014). However, the extent to which such training affects the strength of association between utterance fluency features and fluency judgements remains unclear.

Selection of Utterance Fluency Measures

The fourth phase of L2 perceived fluency research is the computation of utterance fluency measures by annotating temporal features of speech samples. First and foremost, researchers need to select utterance fluency measures to predict perceived fluency ratings. Although the selection of utterance fluency measures is dependent on the focus of the research, such as the scope of perceived fluency (e.g., higher- vs. lower-order fluency), researchers are advised to ensure the construct validity of the measures selected (Lambert & Kormos, 2014), comparability with previous studies (Michel, 2017), and a lack of intercollinearity among the measures (Bosker et al., 2013). Research focusing on higher-order fluency tends to employ linguistic measures in addition to utterance fluency, such as grammatical errors and lexical repertoire (e.g., Kormos & Dénes, 2004; Suzuki & Kormos, 2020).

Subsequent to the selection of utterance fluency measures, researchers must decide whether to annotate speech samples manually or automatically. Either way, researchers need to specify temporal features relevant to the utterance fluency measures selected, such as pauses and hesitation phenomena. Although based on their simulation data, De Jong and Bosker (2013) suggested that the optimal minimum length of silent pauses is 250 ms, a body of research employed different cut-off lengths for silent pauses (e.g., 200 ms, Cucchiarini et al., 2002; 400 ms, Derwing et al., 2009). In addition, some studies set a maximum length of pauses (e.g., 3,000 ms, Kormos & Dénes, 2004) to avoid counting breakdowns due to non-

linguistic processing. Based on the assumption that different speech production processes and types of breakdowns might explain the occurrence and length of pauses within and between clauses (Kormos, 2006; Lambert et al., 2017), L2 fluency research has also shed light on the differential role of pause location when calculating pause-related measures (e.g., De Jong, 2016b). Mid-clause pause measures (frequency and duration) have been found to show a stronger association with perceived fluency judgements than end-clause pause measures (Kahng, 2018; Suzuki & Kormos, 2020). Another methodological issue around breakdown fluency measures is the distinction of silent and filled pauses. Most studies have counted silent and filled pauses separately (e.g., Bosker et al., 2013; Suzuki & Kormos, 2020), but others did not make a distinction between them (e.g., Trofimovich et al., 2017). However, it is still unclear to what extent pause type (silent vs. filled pauses) differentiates the predictive validity of pause-related measures in perceived fluency. We also lack insights into how the association between repair fluency and perceived fluency varies, depending on the range and focus of disfluency features.

Statistical Analysis

The final phase of L2 perceived fluency is the actual implementation of statistical analysis. Although the current study focuses on correlation coefficients, prior research has usually conducted regression analyses to predict the scores of perceived fluency from a set of utterance fluency measures. Conventionally, multiple regression with stepwise procedure has been used to control for the intercollinearity among predictor variables (i.e., utterance fluency measures). In traditional multiple regression, previous studies commonly averaged the perceived fluency scores for each speaker, once the inter-rater reliability was established. However, even with high inter-retar reliability, the average score may lose the information about the variability of score assignments among raters, subsequently lowering the accuracy of prediction. In response to this problem, recent studies tend to employ mixed-effects modelling, using individual raters as a random-effects predictor (e.g., Bosker et al., 2013; Préfontaine et al., 2016). Mixed-effects modelling allows for multiple observations for the same item (here, ratings from multiple raters to one speech sample), as opposed to multiple regression (Barr et al., 2013). Scholars can thus build regression models with the raw scores of perceived fluency judgements, while maintaining the variability in rating among listeners.

THE CURRENT STUDY

Despite an extensive investigation into L2 perceived fluency, a closer look at L2 fluency research reveals that predictors of perceived fluency have varied across studies and that methodological factors, such as rating procedure and listener's background, may affect the relationship between utterance fluency and perceived fluency. Therefore, in the current study, we conducted a meta-analysis of the correlation coefficients between perceived and utterance fluency measures, and we also examined the moderator effects of methodological factors on the utterance-perceived fluency connection. The current study was thus guided by two research questions (RQs):

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

RQ2. To what extent does the relationship between perceived fluency and utterance fluency vary, according to methodological factors in different phases of L2 perceived fluency research—*speech stimuli preparation*, *rater recruitment, rating procedure*, and *selection of utterance fluency measures*?

METHOD

Literature Search

In order to identify a comprehensive pool of previous studies, we conducted three different literature searches: database search, journal search, and ancestry search from review papers. Following the guidelines on literature search for a meta-analysis (In'nami & Koizumi, 2010; Plonsky, 2015; Plonsky & Brown, 2015), five databases were selected: *Linguistics and Language Behaviour Abstract* (LLBA), *the Educational Resources Information Center* (ERIC), *ProQuest Dissertations and Theses*, *PsycINFO*, and *Academic Search Ultimate*. In order to reduce the effects of publication bias (i.e., the tendency of published studies to report larger or significant effect sizes and, subsequently, the potential suppression of small or non-significant effect sizes in published articles; Pigott & Polanin, 2019), we included dissertations and conference proceedings. Keywords were collected covering target variables (e.g., *assessment, perception, rating* for perceived fluency) and relevant methodologies, including statistical analyses. We also conducted a journal search, using the same keywords, on 23 journals of applied linguistics and speech-related phenomena (for the entire list of keywords and journals, see Supplementary Material 1). In addition, we conducted an ancestry search on recent review papers of L2 fluency (De Jong, 2016a, 2018; Segalowitz, 2016).

The literature search identified 5,061 papers, published from 1943 to 2019. Following methodological guidelines for meta-analysis (Boers et al., 2020; Moher et al., 2009; Plonsky & Oswald, 2015), their titles, abstracts and study descriptors (e.g., keywords, subject categories) were then inspected to see if (a) the study measured any aspects of oral fluency in any form and (b) speech data were produced by L2 learners. A sample of approximately 2% (k = 100) of the 5,061 studies was independently examined by the first and third authors. As a result, we reached 93% agreement at this screening stage, and disagreements were resolved through discussion. This screening process identified 318 studies (for the process of retrieving studies, see Figure 2).

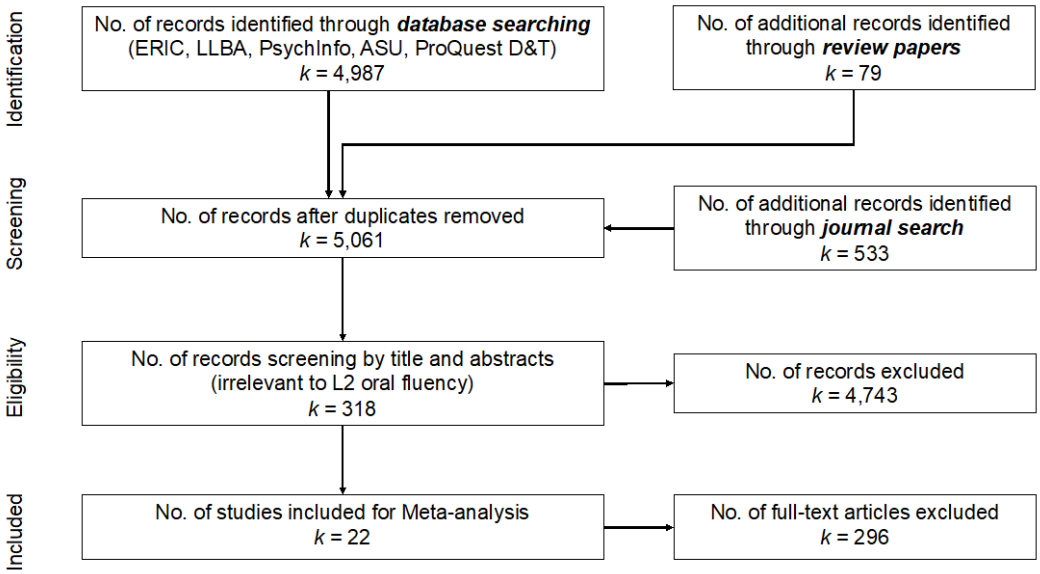


FIGURE 2. The Process of Retrieving Studies for the Current Study

Criteria for Eligibility

We set the following eight criteria for the eligibility of retrieved studies:

1. The study explicitly mentioned that speech stimuli were collected from L2 learners. We excluded studies that employed speech data for the purpose of clinical assessment (e.g., speech disorders).
2. The study may include speech samples elicited from L1 speakers of the target language. However, L1 speakers' speech samples must only be used as the reference point for perceived fluency judgements (e.g., Bosker et al., 2013).
3. The study may employ different speaking tasks for speech elicitation. However, since our target research domain exclusively focuses on L2 speech, we excluded interpreting speech which required speakers to process two languages simultaneously.
4. The study evaluated L2 oral fluency using listener-based scaler ratings.
5. The study may have operationalized perceived fluency as component scores of oral proficiency tests (e.g., TOEFL iBT), if the fluency scores were determined by listener-based judgements. However, we excluded studies if such fluency scores were combined with other constructs (e.g., *fluency and coherence* in IELTS band descriptor; Korko & Williams, 2017).
6. The study employed at least one objective measure of utterance fluency (e.g., speech rate, pause frequency).
7. The study must have either reported correlation coefficients between listener-based and objective measures of fluency or provided information needed to calculate correlation coefficients, such as raw data.
8. The article reporting on the study must have been written in English.

Using 30 studies randomly selected out of 318 studies, we established the reliability of inclusion by 96.7% agreement between the first and third authors. After disagreements were solved through discussion, the first author coded the remaining studies and identified 28 studies that met all eight criteria in our meta-analysis. However, some studies used identical data sets across studies. Accordingly, six studies were excluded, and thus 22 studies were included for the current meta-analysis, which provided in total 263 effect sizes. These 22 studies comprised 17 journal articles, one book chapter, one conference proceeding, two PhD theses, and one MA thesis. Information about these studies is presented in Table 1 (for the list of primary studies, see Supplementary Material 3).

TABLE 1
Basic Information about the 22 Primary Studies

Study	*n*	L1–L2 (Speaker)	L1 vs. L2 raters	Task type
Ahmadi & Sadeghi (2016)	23	Persian–English	L2	Personal narrative, Interview, Group conversation
Bosker et al. (2013)	90	English/Turkish–Dutch	L1	Different types of role play
Cucchiarini et al. (2000)	60	Varied–Dutch	L1	Controlled speech
Cucchiarini et al. (2002)	28, 29, 60	Varied–Dutch	L1	Role play, Argumentative speech, Controlled speech
Derwing et al. (2009)	32	Slavic–English, Slavic–English	L1	Picture narrative
Doe (2017)	32	Japanese–English	L1	Personal narrative
Dubiner et al. (2007)	46	English–Spanish	L1	Interview
Kahng (2018)	74	Korean–English	L1	Personal narrative
Kormos & Dénes (2004)	16	Hungarian–English	L1, L2	Picture narrative
Lam (1994)	15	Chinese–English	L1	Story retelling, Personal narrative
Magne et al. (2019)	90	Japanese–English	L2	Picture description
McGuire (2009)	19	Chinese/Japanese/Thai–English	L1	Paired conversation
Negishi (2012)	135	Japanese–English	L2	Group conversation
Préfontaine et al. (2016)	40	English–French	L1	Picture narrative, Story retelling
Rossiter (2009)	24	Varied–English	L1, L2	Picture narrative
Saito et al. (2017)	40	French–English	L1	Picture narrative
Saito et al. (2018)	90	Japanese–English	L1	Picture description
Sato (2014)	112	Japanese–English	L1	Picture description, Paired decision-making task
Smyk et al. (2013)	76	Spanish–English	L1	Story retelling
Suzuki & Kormos (2020)	40	Japanese–English	L1	Argumentative speech
Tajima (2003)	61	Korean–Japanese	L1	Role play
Trofimovich et al. (2017)	30	Varied–French	L1	Controlled speech, Picture narrative

Note. *n* = Number of speech samples

Selection of Utterance Fluency Measures

Due to a large number of different utterance fluency measures across studies, we decided to reduce the number of utterance fluency measures for the current meta-analysis. In order to select appropriate ones, we combined an a priori theoretically driven approach with methodological trends in our pooled previous studies. To this end, we first decided to include one or two representative measures for each subdimension of utterance fluency—speed, breakdown, and repair fluency (Tavakoli & Skehan, 2005). Also, we decided to include some composite measures, considering their prevalent use in previous studies. In order to reflect methodological trends, we then summarized the frequency of different utterance fluency measures in our pooled studies (see Table 2).

TABLE 2.
Summary of Utterance Fluency Measures in the Pooled Studies

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

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

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

As for breakdown fluency, motivated by the multidimensional nature of pausing behaviours, we first divided breakdown fluency measures into frequency and duration measures. Then, both measures were further coded in terms of pause location (mid-clause, end-clause or both) so that we could use pause location labels for moderator analysis. Considering the number of effect sizes available, we focused on

silent pause frequency and mean duration of silent pauses for effect size aggregation (RQ1), and we also decided to use pause location, pause length, and pause type (silent vs. filled pauses) for moderator analyses (RQ2).

As can be seen in Table 2, the focus of disfluency phenomena varied across studies. We selected disfluency rate (mean number of all types of disfluency features) as a representative measure for repair fluency. However, to obtain a relatively large number of effect sizes for our effect size aggregation, we decided to include any repair fluency measure capturing the frequency of any type of disfluency feature. For the sake of the independence of observations, we averaged effect sizes across frequency-based repair fluency measures for effect size aggregation if a study reported multiple repair measures. For instance, Saito et al. (2018) reported the effect sizes for self-repetition and self-correction as repair fluency measures, and thus the averaged effect size of these measures was entered into effect size aggregation (RQ1). Finally, for composite measures of utterance fluency, we selected speech rate and mean length of run for our meta-analysis, considering their relative prevalence in prior research.

Moderator Variables

Motivated by methodological differences among previous studies, we initially intended to code a total of 16 moderator variables. However, due to an unsatisfactory level of comparability across studies (e.g., different criteria for proficiency levels across studies), we eventually included 11 out of 16 moderator variables for our moderator analysis, using the following criteria. Regarding the excluded moderator variables (speakers' L1, L1-L2 pair, proficiency level, and education level, and listeners' familiarity with speakers' L1), we descriptively synthesized previous studies for the purpose of providing some insights into future directions in our Supplementary Material 2.

Speakers' Target L2. We first classified studies by speakers' L1 and L2. If studies included speakers of multiple L1 backgrounds, we coded them as "Varied" (see Table 3). Despite the huge variability of L1s among studies, we decided to submit only the variable of target L2 to our moderator analysis.

TABLE 3.
Summary of Frequency of Researched L1 and L2 of Speakers

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

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

Task Type. We first categorised studies into monologic and dialogic speech according to the speech elicitation tasks. Furthermore, studies using monologic tasks were further categorised into the following sub-levels, in accordance with the extent to which the content of speech was pre-determined by the task: Controlled production (e.g., reading a text out loud), Closed task (e.g., picture narrative task in which participants have to narrate a given set of events) and Open task (e.g., an argumentative task in which students are free to produce their own arguments).

Excerpts vs. Entire Speech. Motivated by two approaches to presenting speech stimuli to listeners, we coded studies as either entire speech or excerpt (see Table 4). However, excerpts also varied in their exact length (20 to 300 seconds); thus, this moderator variable reflects the completeness of discourse of speech, rather than the absolute length of speech stimuli.

TABLE 4
Frequency of Different Task Types and Stimulus Type

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

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

Listeners' Language Background. This variable simply consisted of L1 raters and L2 raters (see Table 5). L1 raters refer to listeners whose first language is the target language of the speakers, while L2 raters are those who speak the target language of the speakers as a second language. Note that the group of L2 raters can either share the same L1 as the speakers (k = 3) or have an L1 background other than the speakers' L1 (k = 2). The group of raters consisting of both L1 and L2 raters was labelled as "Mixed".

Listeners' Experience. As shown in Table 5, there were different types of experience relevant to L2 perceived fluency judgements, such as teaching experience and linguistic expertise. However, due to the potential overlap between different types of raters' experience, we dichotomously coded experiences as to whether raters had any relevant experience or not. As a result, this variable consisted of Inexperienced raters, Experienced raters and Mixed, which included both inexperienced and experienced raters.

TABLE 5.
Summary of Listeners' Background

Listeners' background	The number of subgroups
Language background	
L1 rater	17
L2 rater	5
Mixed	1
Experience	
Inexperienced raters	11
Experienced	17
Language teaching experience	8
Expertise in linguistics	5
Expertise in language assessment	1
Mixed	2

Note. The total number of studies = 22. Some studies employed multiple groups of listeners. There were eight subgroups without information about listeners' language background and one subgroup without information about listeners' experience.

Definition of Perceived Fluency for Raters. We first categorized pooled studies with semantic scales based on whether researchers provided a definition of fluency to raters: (a) No definition and (b) Researcher's definition. Furthermore, some studies provided rubrics of existing assessment tools (e.g., CEFR; Préfontaine et al., 2016) or created their own tool for research purposes (e.g., Sato, 2014). Therefore, we added two categories: (c) Rubrics of existing assessment tools and (d) Research-based rubrics (see Table 6).

Number of Scale Points. We coded this moderator variable as a categorical variable (5 to 1000, see Table 6). One study, using a sliding bar scale without numerical values on it (Saito et al., 2017), was not included for the moderator analysis for the number of scale points.[1]

Amount of Practice. This variable consisted of two categories: Short practice and Extensive training. In our study, studies were labelled as short practice when researchers asked their raters to use the rating scale to judge several speech samples (e.g., three samples; see Table 6), immediately before the rating session. Studies were categorized as extensive training when researchers provided more extensive training, such as feedback and discussion among raters.

TABLE 6.
Summary of Perceived Fluency Rating Procedure

Rating procedure	*k*
Definition of fluency	
No definition (raters' intuition)	9
Researchers' definition	11
Existing assessment tools	6
Researcher-based rubrics	3
No. of scale points	
5	4
6	6
7	5
9	10
10	4
1000 (with no numerical points)	1
Pre-rating training	
Short practice	15
3 samples	5
4 samples	3
5 samples	6
6 samples	1
Extensive training	4

Note. k = number of studies. The total number of studies = 22. There were eight studies without information about the amount of rating practice.

Speech Annotation Method. This variable has two categories: Manual coding and Automatic annotation (see Table 7). Manual coding refers to studies where researchers manually transcribed and annotated temporal features with some assistance from acoustic analysis software, such as Praat (Boersma & Weenink, 2012). Studies were coded as automatic annotation when researchers annotated speech or computed utterance fluency measures only with the help of a computer program. In our pooled studies, studies coded as automatic annotation used either De Jong and Wempe's (2009) script in Praat (Praat Script Syllable Nuclei v2) or a continuous speech recognizer (Strik et al., 1997).

TABLE 7.
Summary of Speech Annotation Methods

Speech annotation method	*k*
Manual coding	17
Automatic annotation	3

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

Definition of Pauses. Considering the fact that studies can specify pauses differently according to measures, coding specific to pause measures was conducted at the level of effect sizes rather than the level of studies. As reviewed previously, some studies specified the threshold for silent pauses in terms of not only the minimum, but also the maximum length of pauses. However, due to the limited number of studies specifying an upper bound for silent pauses ($k = 2$), the current study focused only on the minimum length of silent pauses. As a result, this moderator variable consisted of the following categories: 200 ms, 250 ms, and 400 ms, with 300 ms excluded due to limited sample size ($k = 2$). We also focused on pause location as another moderator variable. Pause measures were thus classified by three categories: Within clause (pauses in the middle of clauses), Between clause (pauses at clause boundaries) and, Both (counting pauses regardless of location). Finally, we examined the moderator effects of pause type—silent and filled pauses. Since some studies counted silent and filled pauses together (e.g., Trofimovich et al., 2017), we classified pause measures into the following categories: Filled pauses, Silent pauses, and Mixed (counting pauses regardless of type).

TABLE 8.
Summary of Definition and Scope of Pauses and Disfluency Features

Temporal Features	*n*
Pause length	
200	39
250	79
400	47
Pause location	
Both	62
Within clause	14
Between clause	7
Pause type	
Filled pauses	17
Silent pauses	79
Mixed	4
Disfluency features	
Mixed	10
Repetition	6
Self-correction	6

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

Selection of Disfluency Features. As with pause measures, repair fluency measures were also classified according to their target of disfluency phenomena. We labelled effect sizes by targeted disfluency features, while repair fluency measures based on multiple phenomena were labelled as mixed. The frequency of each category is summarized in Table 8. Due to the limited number of effect sizes, we excluded

false starts measures (n = 1), and this resulted in three subgroups: Mixed, Repetition, and Self-correction.

Reporting Practice of Statistics. Following previous meta-analyses, the reporting practice of statistics in primary studies was examined for descriptive statistics, reliability estimates, and type of regression analyses. Among 22 primary studies, 16 studies reported descriptive statistics for perceived fluency scores, and 15 studies included descriptive statistics for utterance fluency measures. As shown in Table 9, we found a range of inter-rater reliability indices for perceived fluency scores, while only few studies reported inter-coder reliability for utterance fluency measures. The trend in regression analysis is summarized in Table 10, showing that many of primary studies relied only on correlation analyses. Meanwhile, as mentioned previously, the recent use of linear mixed-effects modelling is notable (Bosker et al., 2013; Préfontaine et al., 2016).

TABLE 9
Summary of Reliability Indices for Measures of Perceived Fluency and Utterance Fluency

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

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

TABLE 10
Summary of Types of Regression Analysis for the Utterance-perceived Fluency Link

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

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

Coding

To establish the reliability of coding effect sizes and relevant moderator variables, the first and third authors blind-coded a randomly selected sample of 10 studies out of 22. The overall agreement between authors reached 95.8%, and disagreements (see Appendix G) were resolved through discussion. Accordingly, the coding scheme was revised multiple times based on discussion. Then, the first author coded the remaining studies. Our coding scheme and raw data will be available on the IRIS Database (https://www.iris-database.org).

Statistical analysis

We performed all the statistical analyses using the *meta* package (version 4.11-0; Schwarzer, 2007) in R (version 3.6.2; R Development Core Team, 2019). We first examined the extent to which the current data set was influenced by publication bias, using funnel plots and Egger's tests. Visual inspection of funnel plot (see Appendix H) as well as the results of Egger's tests (see Table 11 below) indicated no substantial influences from publication bias on the findings. In addition, we examined the independence of observations in our pooled effect sizes (Plonsky & Oswald, 2015) and then averaged multiple effect sizes across studies to calculate overall effect sizes (RQ1; for details of the averaging process, see Supplementary Material 1).

Prior to the analysis answering our RQs, we tested the moderator effects of interactivity (monologue vs. dialogue) to decide whether to include dialogic fluency in our meta-analysis. A heterogeneity test showed that the effect of interactivity was significant ($Q(1) = 29.14$, $p < .0001$). Moreover, the correlation coefficients in dialogic speech were not significant ($r = .08$, CI[–.10, .25], $p = .389$), indicating the possibility that utterance fluency in dialogic speech contributes differently to perceived fluency. We thus decided to meta-analyse effect sizes based on monologic speech data (for pooled results based on both monologic and dialogic speed data, see Supplementary Material 1).

In order to answer our first research question, inverse-variance weighted overall effect sizes were computed separately for six utterance fluency measures, using a random-effects model with the Restricted Maximum Likelihood estimation method (Novianti et al., 2014). We decided to exclude influential cases for the sake of robust estimates of aggregated effect sizes (Viechtbauer & Cheung, 2010). The exclusion criteria were set based on the prediction intervals of target measures, which suggest the possible range of correlation coefficients in future studies (Nagashima et al., 2019). We employed a within-group Q statistic to detect the potential heterogeneity of effect sizes across the studies included in our analyses.

As regards our second research question, we conducted subgroup analyses for moderator variables. As with the first research question, we also used random-effects modelling to pool the effects within each subgroup. Furthermore, considering the possibility that our categorization of subgroups might introduce a new sampling error at the subgroup level, we decided to use random-effects modelling for between-subgroup comparisons while controlling for such sampling errors (Harrer et al., 2019; Plonsky & Oswald, 2015). We set the minimum number of studies for each category of moderator variables to $k = 3$, following previous meta-analyses in L2 research (e.g., Uchihara, Webb, & Yanagisawa, 2019). Since all of our moderator variables were categorical variables, we calculated a between-group Q statistic to examine the impact of moderator variables on effect sizes. Given that heterogeneity analysis is sensitive to sample size (Borenstein et al., 2009), results of the analysis were further examined and interpreted along with the results of confidence intervals and magnitude of

correlation coefficients according to Plonsky and Oswald's (2014) effect-size benchmarks: Small = |.25–.40|, Medium = |.40–.60|, Strong = |.60–1.00|.

Based on an inspection of the initial results of forest plots of six utterance fluency measures, we identified silent pause duration measures in Préfontaine et al. (2016) as an influential case (for an initial forest plot of pause duration, see Supplementary Material 1). Accordingly, we excluded their averaged effect size of pause duration (k = 1, from three different tasks) from the effect size aggregation (RQ1) and raw effect sizes (k = 3) from relevant moderator analyses (RQ2).

RESULTS

Effect Size Aggregation

To answer our first research question, about the overall relationship between perceived fluency and six selected utterance fluency measures, we conducted a set of effect size aggregations to determine overall effect sizes. As summarized in Table 11, results suggested that all utterance fluency measures were significantly associated with perceived fluency ratings (for forest plots, see Appendices A–F). Both composite measures (mean length of run, speech rate) can be considered as showing strong effects (r = .72, .76, respectively), while the effect size for speed fluency measures (articulation rate) was slightly smaller than that of composite measures, but it still indicated a strong effect (r = .62). Interestingly, within breakdown fluency measures, pause frequency measures (r = –.59) showed a stronger association with perceived fluency than pause duration measures (r = –.46), highlighting the importance of multidimensionality of pausing behaviour in perceived fluency judgements (Kahng, 2018; Saito et al., 2018; Suzuki & Kormos, 2020). Moreover, the 95% confidence interval of repair fluency measure (r = –.20, CI[–.30, –.09]) did not overlap with the confidence intervals of other utterance fluency measures, suggesting that the correlation between repair fluency and perceived fluency was significantly smaller than the correlations between perceived fluency and speed or breakdown fluency. Finally, the aggregated effect sizes for all utterance fluency measures, except for repair fluency, showed significant heterogeneity among the studies, confirming the possibility that moderator variables may affect the association between perceived fluency scores and different utterance fluency measures.

TABLE 8.
Results of Effect Size Aggregations for Six Utterance Fluency Measures

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

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

Moderator Analysis

Speech Stimulus Preparation. Three moderator variables related to speech stimulus preparation were examined. First, although the difference in effect sizes between the subgroups did not reach statistical significance ($Q(1) = 3.15$, $p = .076$), studies using entire speech as speech stimuli ($r = .59$) tended to demonstrate slightly higher correlation coefficients than those using excerpts of speech ($r = .50$). Second, we found significant effects of speaking task type on the correlation coefficients between utterance and perceived fluency measures ($Q(3) = 7.91$, $p = .019$). A set of post-hoc Q tests revealed that effect sizes based on controlled production ($r = .74$; e.g., read-aloud speech) showed higher correlation coefficients than the other two types of monologic speech (both $ps < .01$). In addition, there was no significant difference between closed tasks ($r = .53$) and open tasks ($r = .51$) in the size of the correlation coefficients ($Q(1) < 0.01$, $p = .983$). Third, we also found a significant effect of target L2 on the utterance-perceived fluency connection ($Q(3) = 28.58$, $p < .0001$). A series of post-hoc Q tests revealed that there were no significant differences among the subgroups of L2 Dutch, English, and French ($r = .52$–$.61$), while studies investigating fluency in L2 Japanese ($r = .77$) showed higher correlation coefficients than these three L2 subgroups (all $ps < .001$).

Listeners' Background. Regarding the moderator variables related to listeners' background, we examined the effects of listeners' experience (Experienced vs. Inexperienced raters) and language background (L1 vs. L2 speakers) on the utterance-perceived fluency link. We found no significant effects of listener experience ($Q(1) = 1.96$, $p = .162$). Similarly, a heterogeneity test revealed that listeners' language background did not differentiate the strength of the association between perceived and utterance fluency ($Q(1) = 0.86$, $p = .355$). However, comparing their ranges of 95% confidence intervals, it should be noted that L1 raters ($r = .56$, CI[.51, .61]) indicated a narrower range of confidence intervals than L2 raters ($r = .48$, CI[.29, .64]).

Rating Procedure. None of the moderator variables of the rating procedure showed significant effects on the correlation between perceived fluency scores and utterance fluency measures. As regards the definition of fluency presented to listeners, the category of research-based rubrics suggested a strong effect size ($r = .67$), while the other three categories indicated medium-to-strong effect sizes ($r = .51$–$.59$). Post-hoc Q tests found that a significant difference only between research-based rubrics and researcher's definition ($Q(1) = 5.38$, $p = .020$).

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

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

TABLE 9.
Results of Moderator Analysis of Methodological Variables

Moderator variable	*n*	*r*	*95%*	*Q(df)*	*p*
Speech stimulus preparation					
Target L2				28.58(3)	< .0001
Dutch	44	0.52	[.42, .62]		
English	65	0.54	[.47, .61]		
French	12	0.61	[.55, .67]		
Japanese	4	0.77	[.71, .81]		
Speaking task type				7.91(2)	0.019
Monologue (Controlled)	14	0.74	[.60, .83]		
Monologue (Closed task)	61	0.53	[.46, .59]		
Monologue (Open task)	50	0.51	[.43, .58]		
Speech sample				3.15(1)	0.076
Entire speech	65	0.59	[.52, .66]		
Excerpt	60	0.50	[.44, .57]		
Listeners' background					
Experience				1.96(1)	0.162
Experienced	70	0.58	[.51, .65]		
Inexperienced	52	0.51	[.44, .58]		
Language background				0.86(1)	0.355
L1 raters	109	0.56	[.51, .61]		
L2 raters	16	0.48	[.29, .64]		

Rating procedure					
Definition of fluency for raters				6.52(3)	0.089
Researcher's definition	59	0.51	[.44, .57]		
No definition (Intuitive judgements)	47	0.57	[.47, .66]		
Existing assessment tools	10	0.59	[.49, .68]		
Research-based rubrics	8	0.67	[.55, .77]		
No of scale points				3.41(3)	0.333
5-point	23	0.58	[.44, .69]		
6-point	9	0.63	[.55, .70]		
9-point	58	0.53	[.46, .60]		
10-point	26	0.57	[.42, .69]		
Rater training				1.43(1)	0.232
Short practice	90	0.54	[.48, .60]		
Extensive training	6	0.66	[.47, .79]		
Utterance fluency measure					
Speech annotation				0.58(1)	0.448
Manual coding	89	0.54	[.48, .59]		
Automatic annotation	34	0.59	[.48, .68]		

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

As for pause frequency measures, we conducted a set of moderator analyses of pause location, pause length, and pause type. As summarised in Table 13, despite the non-significant effect of pause location on the whole ($Q(2) = 4.25$, $p = .119$), the effect size of pauses within clauses was considered strong, while both categories of pauses between clauses and pauses including both locations were regarded as medium effect sizes. Post-hoc Q tests revealed that frequency of pauses within clauses ($r = –.72$) tended to show higher correlation coefficients than that of pauses between clauses ($r = –.48$; $Q(1) = 4.01$, $p = .045$). Despite the non-significant result of post-hoc Q test ($Q(1) = 3.47$, $p = .062$), the difference in effect sizes between pauses within clauses and those with both locations ($r = –.55$) appeared to be substantial.

With regard to pause length, we did not find any significant effects on the correlation with perceived fluency scores ($Q(2) < .01$, $p = .999$). However, the range of confidence intervals of the subgroups suggested that the longer threshold of silent pauses tended to have a narrow confidence interval (e.g., 400 ms, $r = –.57$, CI[–.64, –.50] vs. 200 ms, $r = –.56$, CI[–.80, –.19]). In other words, pause length did not affect the predictive power of the measure in listener-based judgements of fluency, while the longer cut-off duration of silent pauses may enhance its stability.

Furthermore, we found significant effects of pause type on the correlation coefficients between perceived and utterance fluency scores ($Q(2) = 32.57$, $p < .0001$). A set of post-hoc Q tests demonstrated that the difference between silent pauses ($r = –.57$) and a combination of both silent and filled pauses ($r = –.47$) did not reach statistical significance ($Q(1) = 3.14$, $p = .076$), while filled pause frequency measures ($r = –.24$) showed significantly lower correlational coefficients than the other two subgroups (both $ps < .01$).

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

TABLE 10.
Results of Moderator Analysis of Utterance Fluency Measure-specific Variables

Moderator variable	*n*	*r*	*95%*	*Q(df)*	*p*
Mean pause duration					
Pause location				–	–
Both	8	–0.42	[–.55, –.27]		
Within clauses	2	–0.71	[–.90, –.27]		
Between clauses	1	–0.63	[–.79, –.39]		
Pause length				1.93(1)	0.165
200ms	5	–0.41	[–.59, –.19]		
250ms	6	–0.60	[–.75, –.39]		
400ms	0	–			
Pause frequency					
Pause location				4.25(2)	0.119
Both	23	–0.55	[–.62, –.47]		
Within clauses	6	–0.72	[–.84, –.55]		
Between clauses	4	–0.48	[–.64, –.27]		
Pause length				0.00(2)	0.999
200ms	6	–0.56	[–.80, –.19]		
250ms	13	–0.57	[–.67, –.46]		
400ms	14	–0.57	[–.64, –.50]		
Pause type				32.57(2)	< .0001
Both	5	–0.47	[–.59, –.33]		

Filled	10	–0.24	[–.34, –.14]		
Silent	29	–0.57	[–.67, –.52]		
Disfluency rate					
Type of repair features				1.29(2)	0.524
Mixed	8	–0.22	[–.33, –.10]		
Repetition	3	–0.13	[–.45, .22]		
Self-correction	3	–0.08	[–.30, .13]		

Note. *n* = number of effect sizes.

DISCUSSION

Overall Link Between Utterance Fluency and Perceived Fluency

With the primary goal of quantifying the overall strengths of association of different dimensions of utterance fluency with listener-based judgements of fluency (RQ1), we meta-analysed the correlation coefficients between six representative measures of utterance fluency and perceived fluency scores. The results demonstrated strong effect sizes for speed fluency (r = .62) and composite measures (r = .72, .76). The strong predictive power of speed fluency and composite measures in fluency judgements align with findings that have demonstrated that these two measures distinguish performance at different levels of proficiency (e.g., Tavakoli et al., 2020). The results indicate that perceived fluency judgements in previous research tend to have been based on what Tavakoli and Hunter (2018) call narrow definitions of fluency. The finding that these composite measures explain a large variance in listeners' judgements suggests that they mostly regard fluency as "ease, flow and continuity of speech" (Tavakoli & Hunter, 2018, p. 343). However, a considerable proportion of variance in fluency judgements still remains unexplained after utterance fluency measures are accounted for (i.e., leftover variance ranges from 38.4 to 57.8%). The results of our meta-analyses suggest that listeners do not rely on 'very narrow' conceptualizations of fluency or take only speed, breakdown and repair features into account (cf. Tavakoli & Hunter, 2018). To some extent, listeners might also attend to linguistic aspects, such as lexis, grammar, and pronunciation.

As regards breakdown fluency measures, the effect sizes were stronger for pause frequency measures (r = –.59) than pause duration measures (r = –.46), indicating that listeners might pay more attention to the frequency of pauses than their duration. This finding aligns with De Jong et al.'s (2013) results showing that pause frequency is associated with a wider range of cognitive fluency measures than pause duration. Regarding the relationship between repair fluency and perceived fluency, the aggregated effect sizes in the current study demonstrated a small but significant correlation (r = –.20, CI[–.30, –.09]). These findings are in line with those of Tavakoli et al. (2020), who investigated the discriminatory role of breakdown fluency measures in the assessment of oral language proficiency. They found that the frequency of repairs did not differ across levels of proficiency. Repair phenomena also tend to be associated with speakers' L1 speaking style (Peltonen & Lintunen, 2016), and consequently they might serve as less reliable cues for listeners than speed, breakdown and composite measures.

Moderator Effects of Methodological Variables

Motivated by the results of heterogeneity tests as well as our review of the literature, we conducted moderator analyses to identify which methodological variables moderate the association between utterance fluency and perceived fluency (RQ2).

Target L2. We observed medium-to-strong effect sizes in Dutch, English, and French (r = .52–.61), while Japanese showed a stronger effect size than these three languages (r = .77). One possible explanation for this difference may lie in cross-linguistic differences in phonological units. Dutch, English, and French are syllable-based, whereas Japanese is mora-based. The basic form of mora consists of one consonant and one vowel, and typically ends with vowel sounds. Accordingly, consonant clusters between units are unlikely to occur in mora-based languages, and the length of basic units tends to be shorter in morae than in syllables (see Collins &

Mees, 2003; Vance, 2008). Therefore, there might be less rhythmic variation in mora-based languages if temporal features (e.g., speed, pauses) are constant, compared to syllable-based languages. Building on these assumptions, fluency judgements of L2 Japanese might be less susceptible to suprasegmental features (stress, rhythm), and thus might be more closely aligned with objective measures of utterance fluency than those of the other three languages. Conversely, particularly in the context of syllable-based languages, such rhythmic/suprasegmental aspects might affect listener-based judgements of fluency (Kormos & Dénes, 2004; Suzuki & Kormos, 2020). These crosslinguistic differences are particularly important if the scoring of fluency is automated and relies on utterance fluency measures alone. For some languages, such as L2 Japanese, utterance fluency measures might be more reliable indicators of fluency judgements than for other L2s, such as English, Dutch, and French.

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

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

Length of Speech Stimuli. Despite no significant difference between short excerpts and entire speech ($Q(1) = 3.15$, p = .076), the results suggested that the effect sizes of entire speech were virtually large (r = .59), while those of excerpts were medium (r = .50). Entire speech samples might provide listeners with more information for judgements than excerpts. As raters can listen to the complete discourse and are exposed to longer input, their subjective perceptions might align better with the objective temporal features of speech. In sum, either type of stimulus might be used, but for the sake of more valid assessment (e.g., language assessment contexts), entire speech may be a better choice for fluency judgements (Isaacs & Thomson, 2013).

Listeners' Background. We examined the moderator effects of two major variables of listeners' background—experience (Experienced vs. Inexperienced raters) and language background (L1 vs. L2 raters). Although the *Q*-tests revealed that neither of the moderator variables differentiated the effect sizes, the aggregated effect sizes were substantially different between the subgroups. Regarding experience, the effect sizes of experienced raters were virtually large (r = .58), while those of inexperienced raters were medium (r = .51). The slightly closer alignments of fluency judgements with temporal features in experienced raters may be in line with Rossiter's study (2009), in which novice and expert raters tended to pay attention to different temporal features, despite similarities in the severity of judgements. Moreover, in the context of holistic assessment of speaking, professional raters tend to be more sensitive to variability in temporal features when it comes to less fluent speech (Duijm et al., 2018). For a better understanding of the role of experience in perceived fluency judgements, the effects of rater experience should be more carefully examined with reference to the overall level of utterance fluency. As for language background, a relatively wide range of 95% confidence intervals in the group of L2 raters (r = .48, 95%CI[.29, .64]), compared to L1 raters (r = .56, 95%CI[.51, .61]) indicated that correlation coefficients tend to be more stable for L1 raters. However, a variety of factors may underlie the distinction between L1 and L2 raters. Therefore, it is still unclear what individual difference variables, such as proficiency and L2 learning experience, contribute to L2 raters' variability in the utterance-perceived fluency link (for the dynamicity of L2 listeners, see Magne et al., 2019; Saito et al., 2019).

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

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

Rater Training. Despite the non-significant difference between the two subgroups of rater training, the effect size of the subgroup of extensive training (r = .66) can be considered large, while that of the subgroup of short practice was medium (r = .54). Considering the possibility that the non-significant difference may have derived from the small number of effect sizes in the subgroup of extensive training (n = 6), the difference in the effect sizes between extensive training and short practice can be considered meaningful. This finding suggests that the length/amount of rater training may enhance the influence of temporal correlates on fluency judgements. Due to the broad category of extensive training in the current study,

however, it is still unclear what type of rater training would significantly increase the association between utterance and perceived fluency measures.

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

Location of Pauses. Due to the limited number of effect sizes in pause duration measures, we conducted moderator analysis of pause location only for pause frequency measures. There were no significant effects of pause location, possibly due to the small number of subgroups (e.g., $n = 4$ for pauses between clauses). However, a similar pattern of pause location effects was demonstrated in both pause measures, showing the highly strong effect sizes for the category of pauses within clauses ($r = –.71$ for pause duration, $r = –.72$ for pause frequency). Meanwhile, the remaining categories were regarded as showing medium-to-large effect sizes ($r = –.42$—$.63$). From the perspective of L2 speech production, pauses within clauses tend to reflect disruptions in linguistic encoding processes, such as lexical retrieval and syntactic procedures (De Jong, 2016b; Kormos, 2006). Therefore, the findings suggest that listeners' perceptions of fluency are established using pause location as an important cue for speakers' efficiency in L2 speech production (i.e., cognitive fluency).

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

Type of Pauses. The effect size of silent pauses approached large ($r = –.57$), while that of filled pauses was regarded as being small ($r = –.24$). Possibly due to the weak predictive power of filled pauses, we found a medium effect size when combining both filled and silent pauses ($r = –.47$). From the perspective of speech production mechanisms, both filled and silent pauses are assumed to reflect breakdowns in speech production processes (Kormos, 2006; Segalowitz, 2010) and the time needed to handle such disruptions (Bui et al., 2019). However, the current findings suggest that listeners may not always perceive filled pauses as an indication of disruption to speech production. The weak role of filled pauses in perceived fluency may be due to the fact that filled pauses can provide listeners with the impression of continuation of speech rather than breakdowns (Clark & Fox Tree,

2002). It is thus recommended to calculate pause frequency measures separately for silent and filled pauses in L2 speech perception research.

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

LIMITATIONS AND FUTURE DIRECTIONS

Several methodological limitations of our meta-analysis need to be acknowledged to avoid overinterpretation of our findings. First, the total number of primary studies was relatively small, because of our strict screening procedure, which is crucial for the robustness of findings from meta-analyses (Boers et al., 2020). Meanwhile, the number of effect sizes in some subgroups in moderator analyses was too small to perform some subgroup analyses. Therefore, the limited number of studies included highlights the need for more studies that directly examine the utterance-perceived fluency link. Second, the significant moderator effects of target L2 in the current study might be subsumed under the effects of L1-L2 combination, because we could not control for the L1 background of speakers due to the huge variability in L1s across studies. Third, due to the variability in methodological practice, we could not include some empirically motivated methodological variables, such as speakers' L2 proficiency levels and listeners' familiarity with the speakers' L1, in our moderator analyses (for the descriptive synthesis, see Supplementary Material 2). Similarly, we acknowledge that our categories of some moderator variables were broad (e.g., listeners' experience, rater training, task type), calling for future studies carefully manipulating specific variables. Finally, due to the limited number of studies reporting reliability estimates for utterance fluency measures, we could not correct the aggregated correlation coefficients for reliability estimates (i.e., measurement errors), indicating that our calculated effect sizes might have been slightly attenuated (cf. Saito & Plonsky, 2019).

The current meta-analysis revealed several methodological factors in need of further investigation. First, relating to the abovementioned methodological incomparability across L2 fluency studies, it would be useful to develop a comprehensive background questionnaire for listeners (cf. Saito et al., 2019). Scholars should also report speakers' proficiency levels in relation to established benchmarks such as CEFR (for a similar suggestion, see Webb et al., 2020), with some justification for their assessment of proficiency (Plonsky & Kim, 2016). Second, our comprehensive library search did not find studies correcting utterance fluency measures by the speakers' L1 utterance fluency. Comparing L1-corrected measures with the raw L2 fluency counterparts, future studies can explore listeners' sensitivity to the influence of speakers' personal speaking style on perceived fluency judgements. Third, we encourage researchers to report the reliability estimates for both perceived and utterance fluency measures, unless automated annotation of

temporal features is used. This practice would allow future meta-analyses to calculate the effect sizes more precisely by correcting for reliability estimates. Finally, following the recommended practice in L2 speech perception research (Isaacs & Thomson, 2020), supplementary qualitative data may also provide some insights into how listeners selectively pay attention to specific speech characteristics (e.g., Magne et al., 2019; Suzuki & Kormos, 2020).

IMPLICATIONS AND CONCLUSIONS

Despite an extensive investigation into listener-based judgements of fluency, prior research provided inconsistent findings regarding the strengths of association between each subdimension of utterance fluency and perceived fluency, possibly due to methodological differences across studies. Our aggregated effect sizes confirmed that perceived fluency is strongly associated with speed fluency and pause frequency, moderately with pause duration, and weakly with repair fluency. In addition, a series of moderator analyses revealed that the utterance-perceived fluency link may be influenced by methodological variables particularly related to how speech samples are prepared for listeners' judgements (target L2, task type, and length of speech stimuli) and how listeners' attention is directed (listeners' experience, rater training, and the definition of fluency presented to raters). As regards the specification of temporal features, the current study also confirmed the importance of distinguishing pause location (pauses within vs. between clauses) and type (filled vs. silent pauses) as well as 250 ms as an optimal threshold for silent pauses.

The current meta-analysis has several implications for language assessment. First and foremost, our findings suggest that assessment tools for fluency rating should place less emphasis on repair phenomena and the frequency of filled pauses. Furthermore, our meta-analysis revealed that automated annotations of speech and manual coding of fluency had similar associations with listener-based judgements of fluency. In combination with the importance of distinguishing the location of pauses, automated scoring of fluency could thus be further improved if pause location was identified with the assistance of speech recognition software and natural language processing techniques. However, considering the fact that there is still substantial variance in listeners' perceptions of fluency that utterance fluency features do not account for, care needs to be taken when using fully automated assessment. In addition, we found a potential cross-linguistic difference in the temporal correlates of fluency judgements, as well as benefits in adjusting the number of scale points according to speakers' proficiency levels. Moreover, Tavakoli et al.'s (2020) research indicates that there is a linear relationship between speed fluency and oral language competence only up to B2 level on the CEFR, beyond which speed measures do not distinguish L2 speakers. Therefore, care needs to be taken if speed and relevant composite measures of fluency are used in automated assessments of fluency, because ratings generated in this way might not fully align with the perceptions of human judges. Based on these findings, we also recommend that in order to enhance the validity of fluency assessment, rubrics and rating scales need to be adjusted to the target population, especially with regard to the range of their proficiency levels and cross-linguistic characteristics of the target L2.

As regards language teaching pedagogy, our results suggest that L2 learners' awareness should be raised of the importance of delivering their speech at an appropriate speed and with relatively low frequency of pauses, particularly in mid-clause locations. Strategy-training activities can be used in the classroom to assist

students in exploiting lexical fillers or pauses for planning ahead at the end of clauses. Besides awareness raising and strategy training activities, pre-task planning time and rehearsal can also be beneficial for increasing speed of delivery and reduction in pausing and hesitations (e.g., Lambert et al., 2020; Tavakoli et al., 2016). Teaching chunks (e.g., collocations and fixed expressions) might have a central place in vocabulary instruction as phraseologically proficient speakers are less likely to pause in the middle of clauses (Tavakoli & Uchihara, 2020). In addition, repeated task performance has also been shown to result in gains in speed fluency and decrease of mid-clause pause frequency (Lambert et al., 2017).

REFERENCES

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly*, *13*(4), 341–358.

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, *47*(4), 707–728.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Boers, F., Bryfonski, L., Faez, F., & McKay, T. (2020). A call for cautious interpretation of meta-analytic reviews. *Studies in Second Language Acquisition*, *15*(1), 1–23.

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.

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.

Bui, G., Ahmadian, M. J., & Hunter, A.-M. (2019). Spacing effects on repeated L2 task performance. *System*, *81*, 1–13.

Clark, H., & Fox Tree, J. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111.

Collins, B., & Mees, I. M. (2003). *The phonetics of English and Dutch* (5th ed.). Brill.

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

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, *107*(2), 989–999.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862–2873.

De Jong, N. H. (2016a). Fluency in second language assessment. *Handbook of Second Language Assessment*, 203–218.

De Jong, N. H. (2016b). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113–132.

De Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, *15*(3), 237–254.

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. *DiSS 2013. Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech*, *January 2013*, 17–20.

De Jong, N. H., 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.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The

relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(4), 533–557.
Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgements on different tasks. *Language Learning*, *54*(4), 655–679.
Derwing, T. M., Thomson, R. I., & Munro, M. J. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System*, *34*(2), 183–193.
Doe, T. (2017). *Oral fluency development activities: A one-semester study of EFL students*. Temple University.
Dressler, A. M., & O'Brien, M. G. (2019). Rethinking perceptions of fluency. *Applied Linguistics Review*, *10*(2), 259–280.
Dubiner, D., Freed, B. F., & Segalowitz, N. (2007). Native speakers' perceptions of fluency acquired by study abroad students and their implications for the classroom at home. In S. Wilkinson (Ed.), *Insights from study abroad for language programs* (pp. 2–21). Thomson Heinle.
Duijm, K., Schoonen, R., & Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing*, *35*(4), 501–527.
Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–399.
Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2019). *Doing meta-Analysis in R: A hands-on Guide*. PROTECT Lab.
Hunter, J. E., & Schmidt, F. L. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd edit).
In'nami, Y., & Koizumi, R. (2010). Database selection guidelines for meta-analysis in applied linguistics. *TESOL Quarterly*, *44*(1), 169–184.
Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159.
Isaacs, T., & Thomson, R. I. (2020). Reactions to second language speech. *Journal of Second Language Pronunciation*, *6*(3), 402–429.
Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*(3), 569–591.
Korko, M., & Williams, S. A. (2017). Inhibitory control and the speech patterns of second language users. *British Journal of Psychology*, *108*(1), 43–72.
Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum Associates.
Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145–164.
Lam, W. (1994). Investigating the oral fluency of 15 EFL teachers: A quantitative approach revisited. *Paper Presented at the International Language in Education Conference in Hong Kong. (ERIC Document Reproduction Service No. ED389168)*.
Lambert, C., Aubrey, S., & Leeming, P. (2020). Task Preparation and Second Language Speech Production. *TESOL Quarterly*, *0*(0), tesq.598.
Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, *35*(5), 607–614.
Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language

speech processing. *Studies in Second Language Acquisition*, *39*(1), 167–196.
Lennon, P. (2000). The lexical element in spoken second language fluency. In H Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). University of Michigan Press.
Magne, V., Suzuki, S., Suzukida, Y., Ilkan, M., Tran, M., & Saito, K. (2019). Exploring the dynamic nature of second language listeners' perceived fluency: A mixed-methods approach. *TESOL Quarterly*, *53*(4), 1139–1150.
McGuire, M. (2009). *Formulaic sequences in English conversation: Improving spoken flunecy in non-native speakers*. Unpublished MA dissertation, University of North Texas.
Michel, M. C. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 50–68). Taylor & Francis.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., Atkins, D., Barbour, V., Barrowman, N., Berlin, J. A., Clark, J., Clarke, M., Cook, D., D'Amico, R., Deeks, J. J., Devereaux, P. J., Dickersin, K., Egger, M., Ernst, E., … Tugwell, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7).
Nagashima, K., Noma, H., & Furukawa, T. A. (2019). Prediction intervals for random-effects meta-analysis: A confidence distribution approach. *Statistical Methods in Medical Research*, *28*(6), 1689–1702.
Negishi, J. (2012). Relationships between L2 speakers' development and raters' perception on fluency in group oral interaction. *Jounral of Pan-Pacific Association of Applied Linguistics*, *15*(2), 1–26.
Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing*, *31*(2), 147–175.
Novianti, P. W., Roes, K. C. B., & van der Tweel, I. (2014). Estimation of between-trial variance in sequential meta-analyses: A simulation study. *Contemporary Clinical Trials*, *37*(1), 129–138.
Peltonen, P., & Lintunen, P. (2016). Integrating quantitative and qualitative approaches in L2 fluency analysis: A study of Finnish-speaking and Swedish-speaking learners of English at two school levels. *European Journal of Applied Linguistics*, *4*(2), 209–238.
Pigott, T. D., & Polanin, J. R. (2019). Methodological guidance papers: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46.
Plonsky, L. (2015). *Advancing quantitative methods in second language research*.
Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). *Second Language Research*, *31*(2), 267–278.
Plonsky, L., & Kim, Y. (2016). Task-Based Learner Production: A Substantive and Methodological Review. *Annual Review of Applied Linguistics*, *36*, 73–97.
Plonsky, L., & Oswald, F. L. (2014). How Big Is "Big"? Interpreting Effect Sizes in L2 Research. *Language Learning*, *64*(4), 878–912.
Plonsky, L., & Oswald, F. L. (2015). Meta-anlyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). Routledge.
Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language*

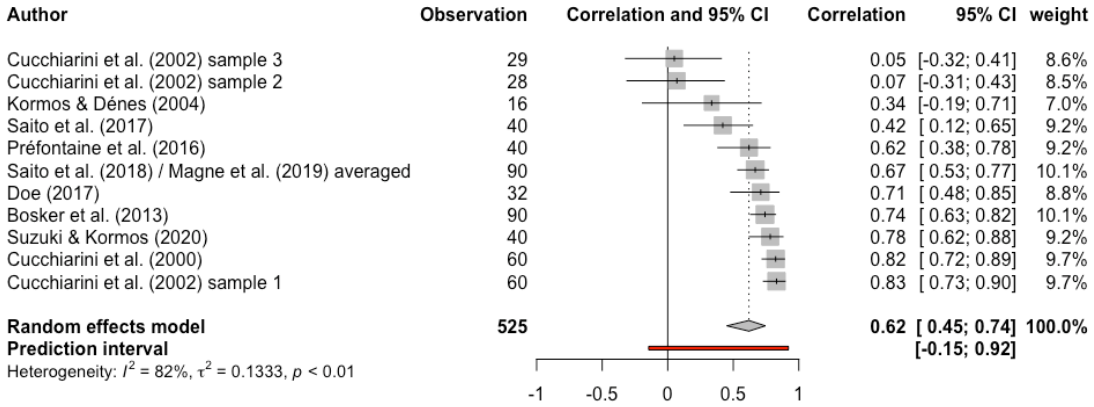
Testing, *33*(1), 53–73.
R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org
Riggenbach, Heidi. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, *14*(4), 423.
Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review*, *65*(3), 395–412.
Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics*, *39*(3), 593–617.
Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708.
Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? *Studies in Second Language Acquisition*, *41*(5), 1133–1149.
Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439–462.
Sato, M. (2014). Exploring the construct of interactional oral fluency: Second Language Acquisition and Language Testing approaches. *System*, *45*, 79–91.
Schwarzer, G. (2007). meta: An R package for meta-Analysis. *R News*, *7*, 40–45.
Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.
Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 79–95.
Smyk, E., Restrepo, M. A., Gorin, J. S., & Gray, S. (2013). Development and validation of the Spanish–English Language Proficiency Scale (SELPS). *Language, Speech, and Hearing Services in Schools*, *44*(3), 252–265.
Strik, H., Russel, A., Van Den Heuvel, H., Cucchiarini, C., & Boves, L. (1997). A spoken dialog system for the Dutch public transport information service. *International Journal of Speech Technology*, *2*(2), 121–131.
Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167.
Tajima, M. (2003). *The effects of planning on oral performance of Japanese as a foreign language*. Unpublished PhD dissertation, Purdue University.
Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 133–150.
Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*, *50*(2), 447–471.
Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of Fluency Across Assessed Levels of Speaking Proficiency. *The Modern Language Journal*, *104*(1), 169–191.
Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language*

(pp. 239–273). John Benjamins.
Tavakoli, P., & Uchihara, T. (2020). To what extent are multiword sequences associated with oral fluency? *Language Learning*, *70*(2), 506–547.
Tavakoli, P., & Wright, C. (2020). *Second language speech fluency: From research to practice*. Cambridge University Press.
Trofimovich, P., Kennedy, S., & Blanchet, J. (2017). Development of second language French oral skills in an instructed setting: A focus on speech ratings. *Canadian Journal of Applied Linguistics*, *20*(2), 32–50.
Uchihara, T., 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.
Vance, T. J. (2008). *The sounds of Japanese*. Cambridge University Press.
Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125.
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.

APPENDICES

Appendix A

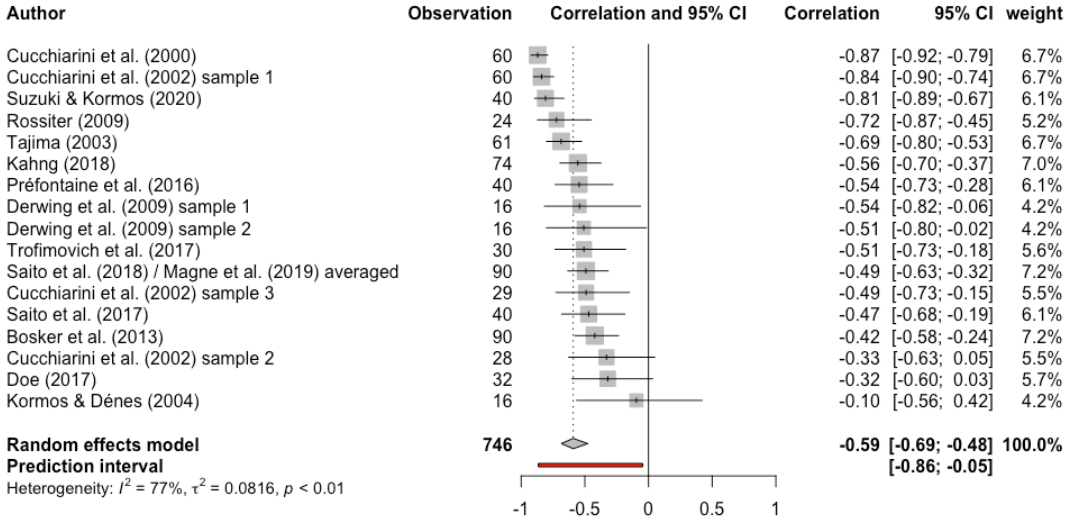
Forest Plot of the Relationship Between Perceived Fluency and Articulation Rate



Note. A diamond indicates the overall average correlation; and a red line showed a prediction interval.

Appendix B

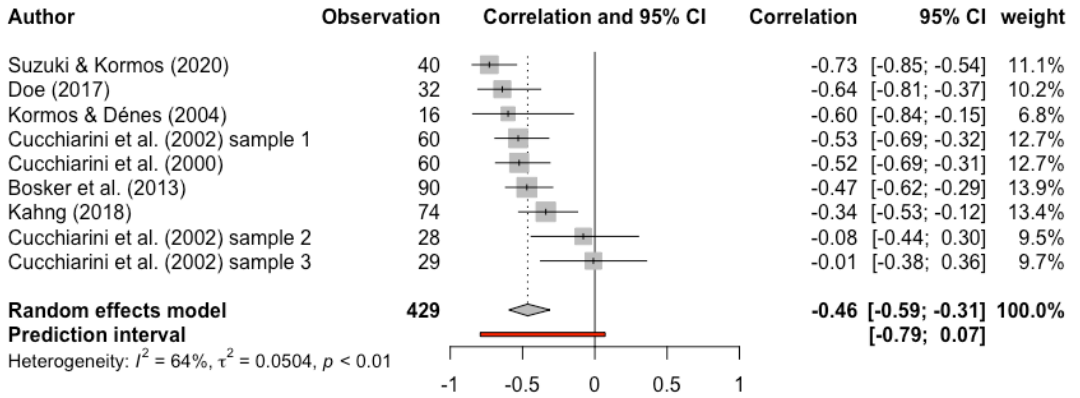
Forest Plot of the Relationship Between Perceived Fluency and Silent Pause Frequency



Note. A diamond indicates the overall average correlation; and a red line showed a prediction interval.

Appendix C

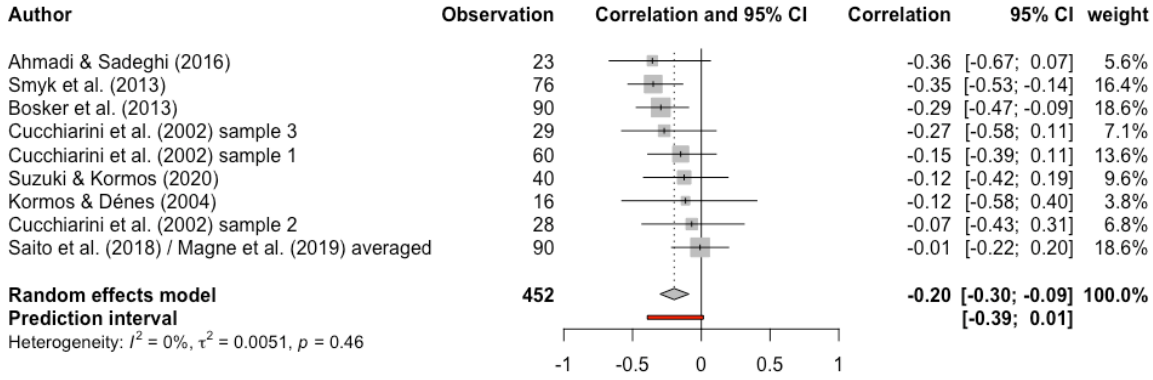
Forest Plot of the Relationship Between Perceived Fluency and Silent Pause Duration



Note. A diamond indicates the overall average correlation; and a red line showed a prediction interval.

Appendix D

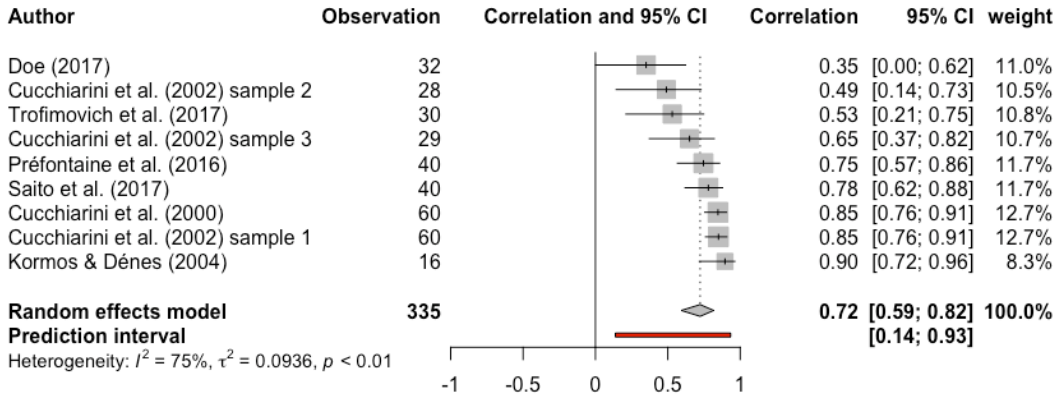
Forest Plot of the Relationship Between Perceived Fluency and Disfluency Rate



Note. A diamond indicates the overall average correlation; and a red line showed a prediction interval.

Appendix E

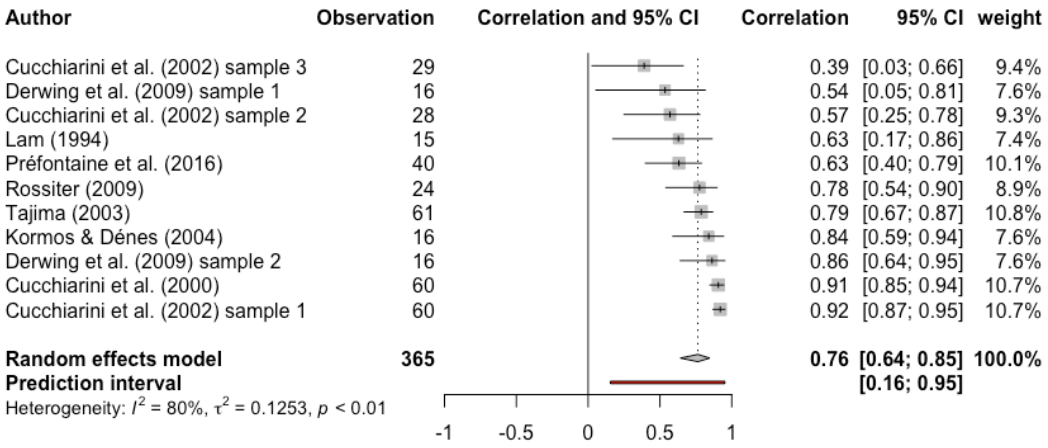
Forest Plot of the Relationship Between Perceived Fluency and Mean Length of Run



Note. A diamond indicates the overall average correlation; and a red line showed a prediction interval.

Appendix F

Forest Plot of the Relationship Between Perceived Fluency and Speech Rate



Note. A diamond indicates the overall average correlation; and a red line showed a prediction interval.

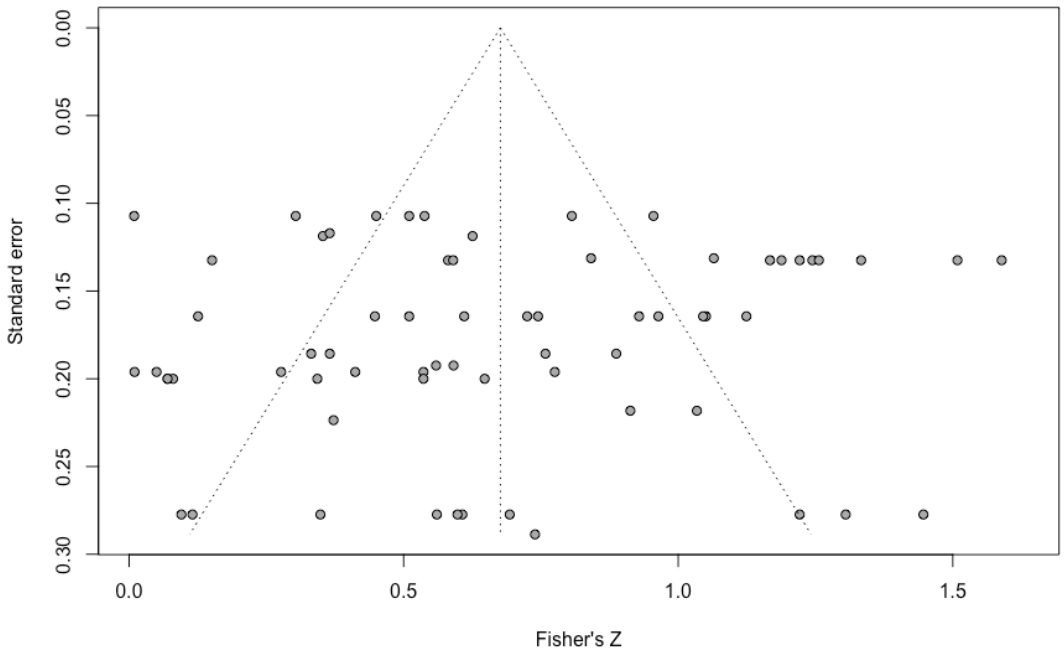
Appendix G

Summary of inter-coder agreements for selected variables

Coded variables	Total number of coding	Raw frequency of agreements	% agreement
Speaker variable			
Sample size	11	11	100
L1	11	11	100
L2	11	11	100
L2 proficiency	11	9	81.8
Education level	11	8	72.7
Listener variable			
Sample size	12	12	100.0
L1 or L2 raters	12	12	100.0
Experience	12	11	91.7
Familiarity with speakers' L1	12	12	100.0
Speech stimulus			
Task type	13	10	76.9
Entire vs. experpt	13	12	92.3
Perceived fluency			
Source of definition of fluency	13	11	84.6
Amount of practice	13	12	92.3
No of scale points	13	11	84.6
Reliability index	12	12	100.0
Reliability estimates	12	12	100.0
Descriptive statistics	12	12	100.0
Utterance fluency			
Measure	41	39	95.1
Construct	41	39	95.1
Length of pause	41	41	100.0
Pause type	20	20	100.0
Pause location	20	20	100.0
Disfluency features	8	7	87.5
Reliability index	41	41	100.0
Reliability estimates	41	41	100.0
Descriptive statistics	57	57	100.0
Statistics			
Effect size	69	65	94.2
Standard error	65	69	106.2
Regression type	12	11	91.7
Type of R2 index	12	11	91.7
R2 value	12	11	91.7

Appendix H

Funnel plot for six selected utterance fluency measures in RQ1



[1] The reason for the exclusion of Saito et al. (2017) from the moderator analysis for the number of scale point is ultimately due to the statistical constraints; we decided that one primary study is not appropriate to create the subgroup for the moderator variable.