Article:
Wright, C (2021) Effects of Task Type on L2 Mandarin Fluency Development. Journal of Second Language Studies, 3 (2). pp. 157-179. ISSN: 2542-3835

https://doi.org/10.1075/jsls.00010.wri

Effects of Task Type on L2 Mandarin Fluency Development

Clare Wright

Department of Linguistics, School of Languages, Cultures and Societies, University of Leeds, UK

Author Note

Clare Wright orcid.org/0000-0003-3962-7903

No known conflict of interest to disclose.

Correspondence concerning this article should be addressed to Clare Wright, School of Languages, Cultures and Societies, Michael Sadler Building, University of Leeds, Leeds LS2 9JT, UK. Email: c.e.m.wright@leeds.ac.uk

ABSTRACT

This study explores task effects on fluency development in second language (L2) Mandarin during study abroad (SA) in China, given linguistic and pedagogic challenges facing western learners of Mandarin (Zhao, 2011). Data from 10 adult English learners of Mandarin were compared pre/post 10 months' SA in China. We measured performance in 4 tasks with different task loads (rehearsed vs. spontaneous speech, in monologic and dialogic mode). Significant differences between the rehearsed monologue and other tasks found pre-SA were generally not found after SA. Some differences remained between monologues and dialogues, suggesting task load effects may override SA impact. Claims about the impact of SA on L2 oral development should take more account of different task demands, to help further illuminate our understanding of how SA may benefit L2 fluency.

Keywords: L2 Mandarin fluency; task effects; Study Abroad; prepared speech; spontaneous speech

Effects of Task Type on L2 Mandarin Fluency Development

Studies on oral proficiency among foreign language learners during Study Abroad generally concur over benefits in fluency gained from exposure during immersion in the target language country, largely assumed to be due to the effects of immersion and opportunities for authentic meaningful interactions with local speakers (Collentine & Freed, 2004; Derwing, Munro, & Thomson, 2008; Sanz, 2014; Tullock & Ortega, 2017). However, immersion is not always found to be a unique "magic formula" (Kinginger, 2011, p. 58) for improvement; fluency can also significantly improve in non-immersed settings (Segalowitz & Freed, 2004). Understanding precisely what to expect from effects of SA on fluency can also be complicated by variability in methodologies used, e.g. different speech tasks, making it difficult to compare and replicate studies of SA fluency, particularly longitudinally (Tullock & Ortega, 2017). Finally, current assumptions of SA effects on fluency have not yet been widely tested in languages other than English or western European languages, particularly for languages which are seen as typologically and socioculturally distant from western European languages wsuch as Mandarin (Duff et al., 2013; Zhao, 2011). In such settings, linguistic challenges and limited contexts for engagement out of the classroom can reduce opportunities to build communicative fluency across a wide range of tasks, including spontaneous dialogic interaction (Mitchell, Tracy-Ventura, & McManus, 2017; Sanz, 2014; Wright, 2018; Zhao, 2011). This article therefore seeks to add to current discussions of fluency development during SA, by providing a small scale but in-depth exploration of task effects on longitudinal development of fluency in L2 Mandarin.

Literature Review

Task effects on fluency development

One challenge for SA researchers is how to identify an appropriate operationalization of fluency, e.g. whether to look at fluency in a holistic broad sense or in a more narrowly defined sense (Lennon, 1990; Tavakoli & Hunter, 2018). While there are good arguments for considering both perspectives, many current studies focus on a narrow temporal definition of fluency, aiming to identify how SA affects aspects of L2 utterance fluency in relation to speed, breakdown and repair (Skehan, 2003). Such research is based on models of speech production (e.g., Levelt, 1989) in which speech consists of a set of processes in stages taking an idea from conceptualization, through formulation in grammatical and lexical form, to overt articulation as the intended utterance. In fluent L1 speakers, these processes are effortless, automatic, and operate at high speed and in parallel. Accordingly, speech can be formulated and articulated as the next idea is generated (Levelt, 1989).

In L2 learners (Kormos, 2006, 2011), these processes are not yet automatized, requiring serial processing stage by stage, typically making L2 speech more effortful, with more silent pauses. Gaps in linguistic knowledge or slow processing in accessing knowledge can affect the formulation of accurate or complex grammar and lexis, creating slower speech speed, hesitations, filled pauses and repairs (Segalowitz, 2010, 2016; Skehan, 2003; Tavakoli, 2011). Unfamiliarity with target sounds can lead to problems in articulation, impacting intelligibility and comprehensibility (Magne et al., 2019; Préfontaine & Kormos, 2016; Révész, Ekiert, & Torgersen, 2016; Saito, Trofimovich, & Isaacs, 2016).

Fluency can be taught and practised though activities in which familiarity and repetition can reduce the processing load on planning or formulation (Gatbonton & Segalowitz, 2005; Tavakoli, Campbell, & McCormack, 2016). Familiar routines allow speakers to use pre-learned chunks (Wood, 2010; Wray 2002); automaticity in formulation and articulation can be built up by activities such as the 4-3-2 task (de Jong & Perfetti, 2011; Nation, 1989), in

which a speaker talks about the same topic first for 4 minutes, then 3, and then 2 minutes. However, building up fluency through repetition in one type of task may not always have a broad effect on fluency across all types of speech (Wright & Tavakoli, 2016).

Within L2 fluency research it is clearly established that there can be significant task effects on oral performance, particularly at the formulation or articulation stage (see, e.g., de Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012; Pallotti, 2009, 2017). Tasks carrying different cognitive loads, depending on task design and levels of complexity, can thus differently affect fluency in speech performance (Awwad, Tavakoli, & Wright, 2017; Ellis, 2003; Robinson, 2003, 2011; Robinson & Gilabert, 2013; Skehan & Foster, 2001).

One factor affecting task load is the degree of preparation or rehearsal time. Reduced preparation time increases task complexity by dispersing learners' attentional resources (Gilabert, Baron, & Levkina, 2011; Ortega, 2005). Prepared speech, particularly on a familiar topic, allowing speakers to use pre-learned chunks is thus seen as having a lower task load than unprepared spontaneous speech (Wood, 2011; Yuan & Ellis, 2003). Typical classroom tasks for speaking practice, such as short pre-rehearsed talks, can reduce load and improve performance as measured at utterance level in terms of speed, breakdown and repair (Skehan, 2003). Such rehearsed speech may well be fluent in terms of temporal speed and lack of breakdown; however, such fluent delivery may be based on good use of memorised phrases and efficient articulatory skills at utterance level, rather than well-developed automatic processes at the deeper cognitive level (Segalowitz, 2010).

Task load can also differ based on discourse type, when comparing monologic vs. dialogic speech. Interactions in dialogic mode, especially spontaneous creative interaction, are usually seen as attention-dispersing tasks which may create heavier cognitive load, by requiring layers of linguistic, cognitive and communicative abilities to maintain fluency (Faerch, Haastrup, & Phillipson, 1984; Segalowitz, 2010; Tavakoli, 2016; Witton-Davies, 2014).

Utterance fluency found in monologic form may therefore not always be sustained in dialogues, though it depends on dialogic task requirements (Tavakoli, 2016), and it has been noted that some dialogic tasks, particularly on familiar topics and using conventional routines, may not be particularly taxing (Michel, 2011; Witton-Davies, 2014).

From the literature, it seems plausible to hypothesise a task trade-off effect on fluency, based on the two dimensions of level of rehearsal, on the one hand, and discourse mode, on the other. In rehearsed or prepared speech, on familiar topics, the lower task load and monologic discourse mode may facilitate better utterance-level or performative fluency ("performative competence", Wright, 2018, p.176). By comparison, in spontaneous speech (requiring "creative competence", Wright, 2018, p. 176), particularly in interaction, heavier task load and discourse mode demands could reduce utterance-level fluency. Alternatively, it could be hypothesized that utterance fluency in one task, e.g. a prepared monologue, could facilitate fluency development in other tasks, e.g. in unprepared or interactive tasks. Relatively little research has explored this aspect of task effects on fluency, particularly in SA contexts. It remains therefore an open question as to whether or how such task trade-off or facilitation effects might be impacted by immersion during SA.

Fluency during SA

Generally, it has been assumed that SA should be an effective context for triggering fluency (see, e.g., Diao, Donovan, & Malone, 2018; Freed, 1995; Mora & Valls-Ferrer, 2012; Sanz, 2014; Segalowitz & Freed, 2003). However, there are contradictory findings in SA research over precise SA benefits for fluency, and task-based effects have not yet, to our knowledge, been extensively studied in SA contexts. Thus examining fluency development during SA, using a set of clearly operationalised measures across different tasks based on a performative/creative distinction, could be valuable in testing predicted SA benefits, to see if

fluency develops in similar or different ways, e.g. on carefully rehearsed performative competence vs. unprepared, more pressured spontaneous communicative competence. Such an approach is particularly useful for SA research on a language such as Mandarin which may present particular challenges for evaluating fluency development.

L2 Mandarin fluency research in SA

In non-European languages such as Mandarin, typological distance and instructional context may impact current norms for measuring L2 development and oral fluency in particular (Han & Finneran, 2014; Lu, 2017; Wright, 2018, 2019). While Mandarin can be seen as "easy" for western learners in terms of grammar formulation (Hu, 2010), there are challenges in learning specific word order rules, information structure affecting phrase construction, as well as variability in what may or may not be required due to discourse and pragmatic constraints (Xing, 2006). For example, variability in *shi*-copula and pronouns, word order rules for adverbials, relative clause and *ba* structures, or topicalization, would require an emerging L2 speaker of Mandarin to plan the whole utterance from the start (Zhao, 2011). In addition, the sound system of Mandarin leads to challenges in semantic comprehensibility and avoidance of ambiguity; the limited number of syllables in Mandarin leads to multiple polysemous and polyphonous words, while the tone system creates difficulties in accurate articulation for speakers without tone in their L1 (Wright, 2019). These problems can affect speech production at all stages from conceptualization through formulation to articulation.

There is therefore the potential for L2 Mandarin learners even at early stages to be relatively fluent in rehearsed tasks, using practised routines and familiar lexical and grammatical structures (Wright & Zhang, 2014). However, their ability to manage creative spontaneous speech, e.g. in dialogic tasks, can be more limited, particularly if their

instructional experience has been of teachers who value accuracy over fluency using traditional memorization and choral drilling techniques (Perez-Milans 2015; Everson & Xiao, 2009; Wray, 2002). Even speakers at mid-level proficiency report relying on classroom-drilled chunks, and finding frustration with the capacity to express themselves creatively (Wright, 2019); likewise, L2 learners completing language degrees which require time spent abroad, may also express anxiety over their capacity to handle interaction with target-language speakers in China even after intensive study (Peng & Wright, 2020). Gaining greater insight into the challenges and realities of fluency development in China is therefore valuable for exploring the issues highlighted here.

There has been a recent explosion of interest in studying Mandarin (e.g. Han, 2014; Lu, 2017; Tao, 2016), but only a handful of empirical studies so far have emerged investigating L2 Mandarin speech development during SA (author). Research into task effects comparing monologic and dialogic fluency also seems to be rare in the L2 Mandarin literature, at least in English-language publications.

Liu's (2009) study found evidence of improvements in a US study abroad program (n=12), based on data taken from a wide range of different measurements and tests, including an Oral Proficiency Interview (OPI). After 1 year of at-home short-term immersion and class instruction, followed by an intensive 4-week SA period in China living with a family, students' test scores improved, at least descriptively, if not significantly. However, no specific details were provided, e.g. on sub-measures of oral improvements within the different segments of the OPI, to indicate potential task differences. Kim et al. (2015) provided rich data on fluency among other aspects of linguistic development, based on computerised OPI tests, finding faster speech rate and shorter pauses among a cohort of 22 US students after a 16-week semester stay in China. However, again the OPI data were not broken down for possible within-task variation. Diao, Donovan and Malone (2018) also

found clear evidence of oral fluency development, among home-stay students during one-semester SA visits, noting particularly the effects of different levels of engagement and interaction with the home-stay hosts. However, fluency was evaluated using broad holistic measures of oral proficiency, and not easily related to specific task-load effects.

Du's (2013) study of 29 US students during one semester in China gathered speech data in different contexts and tasks on and off campus: for example, formal prepared talks were recorded in Chinese speaking classes; spontaneous speech was recorded in individual informal interviews off campus. This study showed significant fluency development over time in both types of speech, but the results were taken from specific segments chosen to highlight the most productive segments of fluent speech, which in our view limits the reliability and generalizability of the findings. We therefore believe that there is a gap in consistent and replicable task-focused SA research in Mandarin, which this study aims to address.

Study rationale and research questions

The study reported here is one of very few, as far as the authors are aware, that tracks longitudinal development of L2 Mandarin fluency across different tasks, aiming to inform task-based and SA-based paradigms of fluency development. The data formed part of a wider project tracking development across written and oral modes (see Wright, 2019; Wright & Zhang, 2014); we report here on novel analyses of the oral production data across four different speaking tasks, to investigate specific questions of task effects identified above. We acknowledge that quantitative SA research should also, where possible, be supported by contextual qualitative data collection (Kinginger, 2011), and that perceptions of fluency can also be added to get a fuller picture of a speaker's performance (Segalowitz, 2010). However, the research questions here focus on the quantitative measures of speech production, as a

baseline for further research into the linguistic experiences of learners of Mandarin during SA.

Using the assumptions of task load explained above for rehearsed and spontaneous speech, and of discourse effects for monologic and dialogic mode, we hypothesized that there will be differences found pre-SA between monologic and dialogic speech but that these will reduce after SA; we also hypothesized that the gap would reduce most clearly for spontaneous dialogic speech as speakers would become more able to handle the task load and discourse demands, through encountering more opportunities for interaction during SA. We aimed to identify to what extent changes occur in parallel across all tasks, or if there is some kind of trade-off effect, seen in differences between specific variables measuring speed, repair and breakdown.

Our research questions were:

RQ1: Does fluency in L2 Mandarin change significantly over time during SA, tested through a battery of fluency measures?

RQ2: Are there task differences when measuring speech fluency development between task load in rehearsed vs. spontaneous speaking tasks, and between discourse type in monologic and dialogic mode? Is any kind of trade-off or facilitation effect found?

Method

Participants

We recruited a volunteer set of ten adult English university learners of Mandarin from a UK university, after 2 years' instruction *ab initio*, with no contextual or long-term exposure to Mandarin. Our original pool was 22 students, but only ten completed all tasks at both times of testing pre/post SA. During their time in China, the participants were all expected to have comparable experiences, to fit with university curriculum requirements; all attended classes at

the host institution given in Mandarin Chinese (typically 15 hours per week) and were able to communicate with Chinese people on a daily basis, based on self-reports of language usage collected at three points during the period. The usual SA duration was an academic year, approximately ten months (2 participants stayed for an additional month's travel after the end of their final semester). Data was collected at the end of their second year on the programme, and again on their return to the UK at the start of their final year of study (for further details, see Wright, 2019).

Task Design

We used two monologic speaking tasks and two dialogic tasks, to track changes in fluency, recorded as part of the students' end-of-second-year summer assessments, and repeated after the period of SA, on the participants' return to the UK university at the start of their Year 4. All four tasks were based on the theme of daily routines, commonly used in class discussions and for practice oral assessments before the first time of recording. All tasks were related to daily routines, family and friends, or typical social activities, and aimed to tap related familiar vocabulary and grammatical structures. The data were taken from the university's standard year-end assessment tests, in which the teacher also played the role of interlocutor in the dialogue tasks. We suggest the tests provide useful ecological validity for analysing fluency development within the context of a typical university degree programme experience.

The rehearsed oral test (Task 1) was a pre-prepared talk on one of three topics – one of which was then chosen by the teacher for all participants, which was about some aspect of participants' daily life in China (thinking ahead to what they expected before their study abroad at Time 1, or reflecting back after their return at Time 2). Participants were given at least 48 hours before the test to prepare the three possible topics outside class and were

instructed to practice their talks. Performance on this task was taken primarily as a baseline measure for fluency prior to SA in demonstrating fluency in a rehearsed task with minimal cognitive load; we could then see how this baseline might change after SA, and test if performance on this task might bootstrap fluency development in other tasks.

The spontaneous oral test (Task 2) was an unprepared description of a picture, using a scene involving people doing typical daily-routine activities (people dining in a restaurant or having birthday parties). Task load was hypothesized to be higher than Task 1, due to lack of rehearsal.

Task 3 consisted of a prepared role-play from a list of 3 options, which the students knew in advance and had practiced in class (ordering in a restaurant, going to a friend's party, or asking a friend to go on a trip). Cognitive load was hypothesized to be greater in Task 3 compared to Task 1, due to dialogic mode; some rehearsal benefits may be seen, since the topics could be prepared in advance, but it would not be possible to prepare what the interlocutor would say.

Task 4 was an open conversation discussing what the students' expectations were of life in China (prior to SA) and what their life had been like (after SA), but they were not told of the topic beforehand, and questions could take any direction, so conversation had to be genuinely spontaneous; we assumed this task would have the highest load.

While we acknowledge there would often be some degree of linguistic structural differences between these four tasks, e.g. referring to first person (used in the rehearsed monologue), second person (dialogues) or third person activities (spontaneous monologue pictures), in Mandarin these linguistic differences are minimized as there is no overt person marking. We therefore considered that these four activities could be considered to be as similar as possible in task design terms, with appropriate ecological validity, apart from the two task-load conditions of rehearsal vs. spontaneous speech, and two modes of monologic

vs. dialogic speech. We therefore judged they were a fair test to reflect our assumptions of variation across task types and to track changes over time.

The participants were tested individually, using a digital voice recorder, and were allowed approximately 2 minutes for each task; additionally, some time at the start of each task (up to 30 seconds) allowed the teacher to introduce each task and ensure the students knew what to do. Sound files were transcribed using CHAT (MacWhinney, 2000), then analysed further using CLAN or PRAAT software (Boersma & Weenink, 2014). Given the complex nature of analysing Mandarin word boundaries (Du, 2013; Kim et al., 2015; Li & Yang, 2009), we follow Kim et al. (2015) in calculating single characters as equal to a syllable, and report all our measures as characters per second. In view of the limited sample size, non-parametric inferential analysis was used; all resulting measures were analysed using SPSS. The data were also analysed for individual outlier performances to compare against the trends shown in group means.

Variable measurement

The battery of measures used here were intended to demonstrate an in-depth evaluation of fluency, including both general measures of overall productivity and temporal analyses (de Jong et al., 2012; Freed, Segalowitz, & Dewey, 2004; Skehan, 2003), and combining both broad and narrow senses of fluency (Tavakoli & Hunter, 2018). We also included a measure of lexical diversity to test for speakers' ability to move away from formulaic routines, given the specific issues identified earlier in typical Mandarin classroom settings. The measures covered amount spoken (total output, lexical diversity), speech speed, hesitations, length of run, and silence. We measured output as total characters, and lexical diversity was measured as G (Guiraud's index, see Richards & Malvern, 2002). Speed was measured as articulation rate (calculated as characters per second of phonation time, i.e., total time on task minus

silent pause time). Hesitation rate (Wright, 2013) was calculated as a ratio of hesitations, a composite total of number of repetitions, retracings, repairs and filled pauses, divided by the total number of characters. Mean length of run was calculated as characters bounded by a silent pause.

We measured silent pausing patterns (Tavakoli, 2011) as number of clause-internal pauses, and mean length of pauses, adopting 250 ms. as the cut-off in line with other L2 studies (e.g., Kahng, 2014; Kormos & Denés, 2004; Towell, Hawkins, & Bazergui, 1996)[1]. Clauses were defined in line with other studies as meaningful units of an utterance (Foster, Tonkyn, & Wigglesworth, 2000); we took the start of each utterance as initiated with some kind of lexically meaningful word, and where any following pause was less than 3 seconds. We did not include clause-external silence since it can be very unclear how to assign such pausing in dialogic speech (Tavakoli, 2016). Future research may need to investigate silence at discourse turn-taking level in more detail, but this issue is beyond the scope of this study.

To assure reliability and validity, the second author trained a transcription team of three native Mandarin speakers. All transcripts and variable coding (produced in the CLAN and PRAAT analyses) were double-checked by a fourth native speaker for inter-rater reliability; any discrepancies were discussed and finalised at 100% agreement.

Results

The study investigated changes in fluency among ten L2 Mandarin adult university learners after their study year abroad immersed in China, using a battery of seven speech performance measures combining CLAN and PRAAT methodologies; the study was part of a

[1] We note that pausing, silent or filled, can be complex as measures for fluency research – they can be affected by a speaker's individual patterns of speech (i.e. a speaker may tend to pause or repair more than others in their L1, not because of any problems in creating L2 speech); we also acknowledge that silent pausing may be an indication of using time for speech planning processes, rather than utterance planning, and also that filled pauses may indicate successful strategies for holding a turn, particularly in dialogues, and may not always be a very clear indication of lack of articulatory fluidity (de Jong, 2016; Tavakoli, 2011). Nevertheless they are included here for comparability with other studies.

wider project reporting on both oral and written language development (Wright 2019, Wright & Zhang, 2014). We checked for effects of time across all measures (RQ1) and compared changes over time between four speaking tasks – one rehearsed monologue (Topic), one spontaneous monologue (Description), one rehearsed dialogue (Role-Play) and one spontaneous dialogue (Chat) – to see if there were clear differences based on rehearsal level or discourse type (RQ2).

Time and task effects

Fluency scores for the seven measures across the four tasks are shown in Table 1 below, in two columns for Time 1 (prior to SA) and Time 2 (post-SA); results showed a clear trend towards better performance across time on all tasks (RQ1), but also evident task differences, particularly at Time 1, although these tended to reduce by Time 2 (RQ2). To address RQ1 in more detail, we compared each variable at Time 1 and Time 2, using Wilcoxon signed rank tests to see if these improvements were significant, as shown in Table 2 below.

Table 1

Mean (SD) fluency scores by task at Time 1 (T1) and Time 2 (T2)

Variable		Topic (rehearsed monologue)		Description (spontaneous monologue)		Role-Play (rehearsed dialogue)		Chat (spontaneous dialogue)	
		Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Output	T1	169.90	(44.20)	120.70	(39.71)	123.86	(55.24)	140.43	(45.24)
	T2	290.60	(101.28)	174.40	(78.29)	162.00	(47.113)	305.00	(100.02)
G	T1	5.56	(.955)	5.05	(.486)	5.23	(.906)	5.038	(.330)
	T2	6.07	(.410)	5.67	(.664)	5.79	(.525)	6.65	(.678)
Articulation Rate	T1	2.85	(0.48)	2.15	(0.27)	1.67	(0.26)	1.90	(0.332)
	T2	3.09	(0.27)	2.70	(0.92)	2.45	(0.27)	2.77	(0.324)
Hesitation Rate	T1	0.144	(.059)	0.234	(0.11)	0.306	(0.13)	0.236	(0.06)
	T2	0.132	(.048)	0.148	(.064)	0.137	(0.04)	0.132	(0.05)
Mean Length of Run	T1	1.52	(0.65)	2.20	(0.44)	1.27	(0.54)	1.49	(0.70)
	T2	1.87	(.46)	2.69	(0.28)	1.61	(0.41)	1.69	(0.33)
Mean Length of Pause	T1	0.672	(0.15)	0.887	(0.23)	0.709	(0.25)	0.628	(0.16)
	T2	0.612	(0.14)	0.803	(0.31)	0.623	(0.19)	0.516	(0.08)
Number of Pauses	T1	38.9	(9.69)	44.80	(14.79)	25.0	(11.63)	22.14	(13.13)
	T2	51.2	(21.87)	35.40	(9.85)	12.14	(3.44)	24.29	(8.14)

Table 2

Significant differences between Time 1 and Time 2 in variables across four tasks

Variable	Topic (rehearsed monologue)		Description (spontaneous monologue)		Role-Play (rehearsed dialogue)		Chat (spontaneous dialogue)	
	p - value	Z - value	*p* - value	Z - value	*p* - value	Z - value	*p* - value	Z - value
Output	.009**	-2.599	.005**	-2.803	NS		.018*	-2.366
G	NS		NS		NS		.018*	-2.366
Articulation Rate	NS		.037*	-2.09	.018*	-2.366	.018*	-2.366
Hesitation Rate	NS		.028*	-2.193	.018*	-2.366	.028*	-2.197
Mean Length of Run	NS		.025*	-2.243	NS		NS	
Mean Length of Pause	NS		NS		NS		.028*	-2.197
Number of Pauses	NS		NS		.018*	-2.366	NS	

* p <.05, ** p < .01

The results in Tables 1 and 2 showed that performance on the rehearsed monologue (Topic) was generally the best, even at Time 1, though fluency did not significantly improve during SA other than in total output. The spontaneous monologue (Description) showed significant differences over time on all measures apart from length or number of pauses; it was also the only task to show significant improvement on mean length of run. The dialogues also showed some significant improvements over time; in both dialogue tasks, improvements were found on articulation rate, hesitation rate. Improvement was also found on mean length of pause and G (though only for the spontaneous Chat task, out of all 4 tasks) and number of

pauses (though only for the rehearsed Role Play task out of all 4 tasks); neither dialogue task improved on mean length of run.

Differences in performance by task

To check if the differences observed above were consistent with our hypothesized task-load effects, we compared tasks by load (prepared/spontaneous) and by mode (monologue/dialogue). Using Kruskal-Wallis tests, with Task as the group factor, at Time 1 all variables other than output and G were significantly different across the four tasks (articulation rate $\chi2(3) = 21.060$, $p = .000$; hesitation rate $\chi2(3) = 10.542$, $p = .014$; mean length of run $\chi2(3) = 12.451$, $p = .006$; mean length of pause $\chi2(3) = 8.120$, $p = .044$; number of pauses $\chi2(3) = 13.168$, $p = .004$). At Time 2, hesitation rate and mean length of pause were not significantly different across tasks; differences were found for output ($\chi2(3) = 15.853$, $p = .001$), G ($\chi2(3) = 8.758$, $p = .033$), articulation rate ($\chi2(3) = 10.937$, $p = .012$), mean length of run ($\chi2(3) = 19.085$, $p = .000$) and number of pauses ($\chi2(3) = 23.036$, $p = .000$).

However, some task effects remained, either in terms of rehearsal or discourse mode. Using Mann-Whitney U tests to analyse Time 2 measures, we compared the rehearsed tasks (Topic and Role Play). Significantly higher scores were found in the Topic task on output ($U = 4$, $p = .002$), articulation rate ($U = 3$, $p = .001$), and number of pauses ($U = 0$, $p = .000$). Comparing the spontaneous tasks (Description and Chat), significant differences at Time 2 remained, seen in significantly better scores in the Chat dialogue task on output, ($U = 8$, $p = .007$), on mean length of pause ($U = 11$, $p = .019$) and on number of pauses ($U = 13$, $p = .033$). The description task was better only on mean length of run ($U = 0$, $p = .000$).

For discourse-mode effects, Mann-Whitney U tests on the monologues at Time 2 showed that significant differences remained, generally seen in better scores on the Topic task on output ($U = 11$, $p = .002$), articulation rate ($U = 20.5$, $p = .023$), though the description

showed higher mean length of run (U = 7, p = .000). The dialogues showed fewer significant differences at Time 2; output was significantly higher in the Chat task (U = 6, p = .017; number of pauses was significantly lower in the Role-Play (U = 1, p = .001).

G showed few significant task or mode differences, other than in the spontaneous dialogic Chat task, but we noted the total number of characters across all tasks had significantly increased from 4756 to 7919 characters. We therefore conducted further lexical analysis on the 10 most frequent words to see if there were qualitative changes in the type of words used, e.g. in range of grammatical function words. We also compared the tokens to the top 10 frequent words in a very large learner corpus, the Guangwai-Lancaster Chinese Learner Corpus (GLCLC) of 1.2 million words, to see if our data would reflect any general learner similarities.[2]

Results are shown in Table 3 below, in descending order of frequency; each token's grammatical category is shown, with its nearest English translation where available, and total token counts; the numbers of tokens from the GLCLC are rounded to the nearest thousand (k).

The results at Time 1 and Time 2 showed some similarities though the order of frequency is slightly different; notably they do not include more complex grammatical particles found in the GLCLC, such as the perfective aspect marker (*le* particle).

[2] The GLCLC can be accessed at https://www.sketchengine.eu/guangwai-lancaster-chinese-learner-corpus/.

Table 3

Top 10 most frequent words pre/post SA, compared to GLCLC

Order of frequency	Before SA	After SA	GLCLC
1	我 pronoun (I, 261)	的 *de* particle (437)	的 *de* particle (700k)
2	的 *de* particle (146)	是 copula (be, 204)	是 verb (be, 161k)
3	很 adverb (very, 115)	有 verb (have, 168)	了 *le* particle (143k)
4	好 adjective (good, 63)	好 adjective (good/okay, 107)	一 numeral (one, 142k)
5	喜欢 verb (like, 41)	个 *ge* classifier (95)	在 preposition (at, 137k)
6	是 copula (be, 51)	对 adjective/ preposition (yes/to, 66)	和 conjunction (and, 105k)
7	有 verb (have, 47)	不 *bu* negation marker (60)	不 *bu* negation marker (80k)
8	不 *bu* negation marker (40)	很 adverb (very, 41)	有 verb (have, 74k)
9	在 preposition (at, 40)	你 pronoun (you, 37)	个 *ge* classifier (69k)
10	你 pronoun (you, 26)	一 numeral (one, 22)	我 pronoun (I, 58k)

Individual variation

We also investigated the scores for individual variation, as we found some individuals demonstrating very different performances at both times, with one individual particularly producing markedly higher output at Time 2. This individual was not markedly different on other fluency variables, and did not significantly affect the group mean patterns already discussed. However, for the sake of completeness, we checked other individuals' scores to see how far individual cases might affect the mix of task-general and task-specific trends observed in the group means. The highest performer at Time 1 for output and articulation rate remained the highest performer on those variables at Time 2, but produced more pauses at Time 2; the weakest performer at Time 1 for output and articulation rate improved, but remained in the lower half of the group at Time 2, suggesting there may be some kind of broad interim-level threshold effect below which or above which SA has less general effect on all tasks than may be expected (Wright & Zhang, 2014). Future research with more individuals would allow us to investigate this question more systematically.

Does baseline performative fluency bootstrap fluency development?

To further investigate whether performative fluency as displayed in the Topic task might bootstrap (or constrain) fluency across other tasks (RQ2), we ran a one-tailed Spearman correlation between articulation rate and hesitation rate scores from the Topic task (i.e. the two best-performing measures of fluency across all tasks at Time 1), with mean length of run, mean length of pause and number of pauses from the other 3 tasks at Time 2. No significant correlations were found, suggesting that fluency in one task does not seem to be associated with fluency in other tasks. In other words, how fast and fluidly a speaker performed on

rehearsed speech at Time 1 implied no particular influence on changes in fluency development in other tasks after SA.

Discussion

This study tracked development of L2 Mandarin fluency in adult learners after a period of SA, to evaluate potential patterns based on task characteristics. Specific differences in task load due to rehearsal and/or dialogic mode were hypothesized to impact on speech performance pre and post-SA; any benefits arising from immersion during SA was hypothesized to show clearest effect on spontaneous dialogic speech.

Our findings showed clear improvement on many measures of fluency in four speaking tasks (RQ1) comparing rehearsed vs. spontaneous monologues, and rehearsed vs. spontaneous dialogues, in terms of general fluidity (i.e. greater output, fewer hesitations, less silence). Changes in fluency over time seemed to retain some task effects, at least in relation to length of run or (to some extent) speed of articulation, and differences in overall performance between monologic and dialogic speech. However, the reduced number of significant task-based differences by Time 2 supported to some extent the hypothesis that overall fluency performance would be different in more taxing spontaneous tasks, but that SA would reduce the difference.

We found some support for our trade-off hypothesis, based on assumed differences of task load and discourse mode, that the rehearsed Topic would be the best performing task. In general, all scores were better on the rehearsed monologue (Topic) than on the spontaneous tasks at both times of testing, though no changes in the rehearsed task after SA were significant, other than total output. SA had significantly more effect on the spontaneous description as predicted, in line with the expectation of the effects of immersion for spontaneous creative speech (Segalowitz, 2010). However, the effect of rehearsal on the

monologues did not appear to have the same clear effect in the dialogues, where the spontaneous dialogue (Chat) often outperformed the rehearsed Role Play (though not always significantly) at both Time 1 and Time 2, except reduction in pauses. Our assumption that performance in the least taxing task (Topic) would bootstrap fluency in other tasks, was not found, suggesting that fluency in one task cannot predict general improvements in other tasks.

We also noted that the number of pauses found by Time 2 showed an intriguing pattern in that there were many more at both times in the rehearsed Topic task, even fewer at both times in the rehearsed dialogue (Role Play), and about the same in the spontaneous dialogue (Chat), suggesting that pausing may be one aspect of fluency performance that is particularly dependent on task load and mode.

Inevitably, in a small-size and exploratory study such as this, group means may not always be the most robust evidence of how individuals changed. However, our detailed focus on task effects across multiple fluency variables go beyond many other studies in tapping into and finding clear evidence of task differences. We particularly found evidence to support our prediction that the demands of spontaneous speech as well as dialogic demands can have a significant effect on speaking fluency. We therefore add to calls for further research into task load on more types of speech, and particularly for comparing monologues and dialogues (Awwad, Tavakoli, & Wright, 2017; Tavakoli, 2016).

We found specific task effects which impacted on the various measures we used here and which retained task differences at Time 2. Mean length of run and mean length of pause on the spontaneous monologue (Description) task were markedly longer than on any other tasks at both times, while number of pauses was the highest in the rehearsed task at both times. This suggests that the requirements of creating a meaningful run was different across the tasks, and that rehearsal of one type of task does not necessarily provide a robust foundation

for fluency in other tasks, in line with other studies looking at fluency in terms of skill development or automaticity(e.g., de Jong & Perfetti, 2011).

To explain further, in the rehearsed monologue (Topic) task which used rehearsed pre-planned speech, participants generally uttered simple shorter phrases; when we listened to these more carefully, we judged these as easily memorised and probably recited. This task factor created a performance advantage on many measures at Time 1, and also yielded more output at Time 2. We believe this is due to rehearsal aiding automaticity in articulation, by reducing the need to construct meaning in real time. This conclusion fits with de Jong & Perfetti's (2011) findings, which found that repeating the same story again and again but under increasing time pressure resulted in greater temporal fluency in terms of smoother faster speech, but only on a similarly structured task, and did not carry over to other tasks. We assume this is because the use of pre-planned verbal material will reduce the cognitive load, by accessing pre-created lexemes or sequences at the formulation stage; these sequences are rehearsed prior to the required time of utterance, which further boosts utterance fluency. This equates to what we term "performative competence".

By contrast, the unprepared monologue (the Description task), requiring freely produced spontaneous speech, seemed to lead to more complex language, at least if measured as longer runs. It also revealed a predictable trade-off with utterance-level fluency, in terms of more hesitation and pausing, at both times. However, this disadvantage in level of performance compared to the rehearsed monologue at Time 1 significantly diminished by Time 2, in line with other research finding benefits for spontaneous speech after SA (Segalowitz, 2010). In other words, SA led to improvements in learners' capacity to better manage spontaneously constructed smoother online speech – or what we term here "creative competence".

The effect of SA on dialogic speech was less clear; we had hypothesized that the unrehearsed dialogue (the Chat task) would be the most taxing, and would therefore show the

worst fluency performance. However, this was not the case. On many measures the rehearsed dialogue (the Role Play) produced the worst performances both at Time 1 and Time 2, including the shortest runs and lowest articulation rate. Our analysis of lexical development (G) did not reveal significant increase in lexical diversity, borne out by a comparison of the most frequent lexemes used across all tasks, so we do not have evidence that the greater output in the Chat task was based on greater lexical or grammatical range, but participants were clearly able to manage the Chat task more easily than the Role Play. This suggests that familiarity with both topic and type of interaction required in the Chat task created less challenge than expected (in line with other findings by Michel, 2011; Witton-Davies, 2014). We also note that the capacity to access formulaic chunks on free discourse on familiar topics may be higher than within the limitations of a set role-play (Wray & Perkins, 2000). As yet, there does not seem to be any research available on lexical chunks in Mandarin fluency research, so this seems to be a clear area for further investigation.

We also noted shorter runs on both dialogues, and fewer pauses, which changed little over time. As we saw above in the Description task, longer length of run can be seen to indicate more fluency, but given the patterns of short runs at both times and no change in pausing over the period of SA, we suggest that these findings may be impacted by discourse mode rather than fluency *per se*. We suggest that the pragmatic discourse constraints of maintaining a dialogue can have a significant impact on patterns of speech – shorter runs may indicate effective turn-taking, rather than holding the floor inappropriately long, and reduced pausing may be a strategy to try to maintain the flow, and avoid communication breakdown (Kasper & Kellerman, 1997).

While time during SA can therefore been seen in this study as enhancing fluency in across many measures, it is also clear from this research that task demands may be more nuanced than expected. We acknowledge that the study has limitations in generalizability, due to the

small sample size, and use of specific university assessment procedures as the basis for the empirical data analysis; however, we suggest that the range of measures used here provide a useful starting point to understanding more clearly how task effects can impact on fluency development. We look forward to further research based on careful operationalization and choice of variables used for speech performance, to ensure fluency development in SA is measured appropriately.

Conclusion

This investigation of L2 fluency development for learners of L2 Mandarin focused on how tasks and time spent during SA may affect different facets of fluency. Our results reinforce the general finding that immersion during SA helps learners' fluency (Collentine & Freed, 2004), which has been widely shown in L2 English and many European languages, but has been under-unexplored hitherto for L2 Mandarin (Wright, 2018). We also showed how our data fitted predictions of task-related performance drawn from current models of L2 speech fluency and task load (Robinson & Gilabert, 2013; Segalowitz, 2010; Skehan 2003). We predicted that SA would favour improvements in more taxing tasks requiring spontaneous speech ("creative competence"), compared to fluency in rehearsed speech with less task-load ("performative competence"), across monologic and dialogic speech performance. We found that performance in the rehearsed speech tasks, particularly the monologue, was better than in the spontaneous tasks across most measures at both times, reflecting the lower cognitive load on speech fostered by preparation time, and also reflecting the successful fluency that instructed learners with no immersion can achieve, at least on the temporal measures used here. However, we did not find any indication that fluency on a rehearsed task, e.g. in terms of speed of articulation or pausing, might facilitate improvements over other tasks.

Task differences in the dialogues compared to the monologues also overrode the assumed effect of different task load. Questions are still needed to understand how task requirements affect speech fluency in wider discourse contexts than are commonly tested in fluency development research (Iwashita, McNamara, & Elder, 2001), which we argue will be essential to understanding in more depth how SA may impact on fluency in different ways. We also noted that the highest performer and lowest performer in the group failed to show consistent progress on all tasks, suggesting some kind of broad threshold effect, where immersion may not significantly change performance if it is already high, or boost improvement if it is below a certain level. These caveats notwithstanding, the degree of improvement in the spontaneous tasks after SA across the cohort found here meant that differences from the rehearsed tasks were no longer generally significant.

We conclude therefore that SA can indeed foster fluency in the broad sense of aiding communicative creative competence based on cognitive fluency (Segalowitz, 2010), though mitigated by task-load, and that existing fluency in a performative articulatory sense does not necessarily bootstrap further fluency development in interaction. We look forward to other longitudinal studies using a wide range of comparable tasks to provide support for this conclusion, which has implications for how learners and teachers can expect to display fluent speech in different tasks or contexts, particularly in being realistic about how far successful monologic speech may represent the complex range of cognitive and linguistic factors required to boost fluency development across the range of "real-life" interaction during SA. Given the relative lack of investigations into L2 Mandarin, we also hope that this study, despite our small sample size, can be helpful in bringing studies of L2 Mandarin fluency more into the mainstream of SLA and SA research.

References

Awwad, A., Tavakoli, P., & Wright, C. (2017). "I think that's what he's doing": Effects of intentional reasoning on second language (L2) speech performance. *System, 67*, 158–169.

Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer*. Version 5.3.51. Retrieved 3 September 2019, from http://www.praat.org.

Collentine, J., & Freed, B. (2004). Learning context and its effects on second language acquisition: Introduction. *Studies in Second Language Acquisition, 26*(2), 153–171.

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

de Jong, N. H., Steinel, M., Florijn, A., Schoonen, R., & Hulstijn, J. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5–34.

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning, 61*(2), 533–568.

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29*(3), 359–380.

Du, H. (2013). The development of Chinese fluency during Study Abroad in China. *The Modern Language Journal*, *97*(1), 131–143.

Duff, P., Anderson, T., Ilnyckyj, R., Van Gaya, E., Wang, R., & Yates, E. (2013). *Learning Chinese: Linguistic, sociocultural, and narrative perspectives*. Berlin/ Boston: DeGruyter.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Everson, M., & Xiao, Y. (Eds.). (2009). *Teaching Chinese as a foreign language*. Boston, MA: Cheng & Tsui Company.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21,* 354–375.

Freed, B. (Ed.). (1995). *Second language acquisition in a study abroad context*. Amsterdam: John Benjamins.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition, 26*(2), 275–301.

Gatbonton, E. & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *The Canadian Modern Language Review / La revue canadienne des langues vivantes, 61*(3), 325–353.

Gilabert, R., Barón, J., & Levkina, M. (2011). Manipulating task complexity across task types and modes. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 105–138). Amsterdam: John Benjamins.

Han, Z.-H., (Ed.). (2014). *Studies in second language acquisition of Chinese*. Clevedon: Multilingual Matters.

Han, Z.-H., & Finneran, R. (2014). Re-engaging the interface debate: Strong, weak, none, or all? *International Journal of Applied Linguistics, 24*(3), 370–389.

Hu, B. (2010) The challenges of Chinese: A preliminary study of UK learners' perceptions of difficulty. *The Language Learning Journal, 38*(1), 99–118.

Iwashita, N., McNamara, T., & Elder. C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning, 21*, 401–436.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*(4), 809–854.

Kasper, G., & Kellerman, E. (Eds.). (1997). *Communication strategies: Psycholinguistic and sociolinguistic perspectives*. Longman: London.

Kim, J., Dewey, D., Baker-Smemoe, W., Ring, S., Westover, A., & Eggett, D. (2015). L2 development during study abroad in China. *System, 55*, 123–133.

Kinginger, C. (2011). Enhancing language learning in study abroad. *Annual Review of Applied Linguistics, 31*, 58–73.

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

Kormos, J. (2011). Speech production and the Cognition Hypothesis. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 39–60). Amsterdam: John Benjamins.

Kormos J. & Denés M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32,* 145–164.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387–417.

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

Li, W., & Yang, Y. (2009). Perception of prosodic hierarchical boundaries in Mandarin Chinese sentences. *Neuroscience, 158*(4), 1416–1425.

Liu, J. (2009). Assessing students' language proficiency: A new model of study abroad program in China. *Journal of Studies in International Education*, *14*(5), 528–544.

Lu, Y. (Ed.). (2017). *Teaching and learning Chinese in higher education*. London: Routledge.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk (3rd ed.).* Mahwah, NJ: Lawrence Erlbaum Associates.

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*. https://doi.org/10.1002/tesq.528

Michel, M. (2011). Effects of task complexity and interaction on L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 141–174). Amsterdam: John Benjamins.

Mitchell, R., Tracy-Ventura, T., & McManus, K. (2017). *Anglophone students abroad: Identity, social relationships, and language learning*. London: Routledge.

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly, 46*(4), 610–641.

Nation, I.S.P. (1989). Improving speaking fluency. *System 17*(3), 377–384.

Ortega, L. (2005). What do learners plan? Learner-driven attention to form during pre-task planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 77–110). Amsterdam: John Benjamins.

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

Pallotti, G. (2017). Assessing tasks: The case of interactional difficulty. *Applied Linguistics, 40*(1), 176–197.

Peng, Y. & Wright, C. (2020). Minding the expectation gap – student expectations pre-study abroad in China. In S. Salin, D. Hall, & C. Hampton (Eds.), *Perspectives on the year abroad: A selection of papers from the 2018 Year Abroad Conference* (pp. 1–10). Forthcoming at https://research-publishing.net/projects

Préfontaine, Y. & Kormos, J. (2016). A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics in Language Teaching, 54*(2), 151–169.

Révész, A., Ekiert, M., & Torgersen, E. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*, 828–848.

Richards, B., & Malvern, D. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*(1), 85–104.

Robinson, P. (2003). The Cognition Hypothesis, task design, and adult task-based language learning. *Second Language Studies, 21*(2), 45–105.

Robinson, P. (Ed.) (2011). *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance*. Amsterdam: John Benjamins.

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics, 38*, 439–462.

Sanz, C. (2014). Contributions of Study Abroad research to our understanding of SLA processes and outcomes: The Sala project. In C. Perez Vidal (Ed.), *Language acquisition in study abroad and formal instruction contexts* (pp. 1–13). Amsterdam: John Benjamins.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.

Segalowitz, N. & Freed, B. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition, 26,* 173–199.

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36,* 1–14.

Skehan, P., & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183–205). Cambridge: Cambridge University Press.

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching 54*(2), 79–96.

Tao, H. (Ed.) (2016). *Integrating Chinese linguistic research and language teaching and learning*. Amsterdam: John Benjamins.

Tavakoli, P. (2011). Pausing patterns: differences between L2 learners and native speakers. *ELT Journal, 65*(1), 71–79.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 115–150.

Tavakoli, P. & Hunter, A.-M. (2018). Is fluency being 'neglected 'in the classroom? Teacher understanding of fluency and related classroom practices. *Language Teaching Research, 22*(3), 330–349.

Tavakoli, P., 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.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*, 84–119.

Tullock, B., & Ortega, L. (2017). Fluency and multilingualism in study abroad: Lessons from a scoping review. *System, 71*, 7–21.

Witton-Davies, G. (2014). *The study of fluency and its development in monologue and dialogue.* Unpublished doctoral dissertation, Lancaster University.

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications.* London: Bloomsbury Publishing.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication, 20*, 1–28.

Wright, C. (2013). An investigation of working memory effects on oral grammatical accuracy and fluency in producing questions in English. *TESOL Quarterly* 47(2), 352–374

Wright, C. (2018). Effects of time and task on L2 Mandarin Chinese language development during study abroad. In C. Sanz & A. Front-Morales (Eds.), *The Routledge Handbook of Study Abroad Research and Practice* (pp. 166–180). New York: Routledge.

Wright, C. (2019). Developing communicative competence in adult beginner learners of Chinese. In C. Shei, M. Zikpi, & D-L. Chao (Eds.), *The Routledge handbook of Chinese language teaching* (pp. 134–148). London: Routledge.

Wright, C. & Tavakoli P. (2016). New directions and developments in defining, analyzing and measuring L2 speech fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 73–77.

Wright, C. & Zhang C. (2014). Examining the effects of study abroad on L2 Chinese development among UK university learners. *Newcastle & Northumbria Working Papers in Linguistics, 20*, 67–83.

Xing, J. X. (2006) *Teaching and learning Chinese as a foreign Language: A pedagogic grammar*. Hong Kong: Hong Kong University Press.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, *24*(1), 1–27.

Zhao, Y. (2011). Review article: A tree in the wood: A review of research on L2 Chinese acquisition. *Second Language Research, 27,* 559–572.