EXAMINING THE EFFECTS OF STUDY ABROAD ON MANDARIN CHINESE LANGUAGE DEVELOPMENT AMONG UK UNIVERSITY LEARNERS[*]

CLARE WRIGHT
(University of Reading)

CONG ZHANG
(University of Oxford)

Abstract

This study tracked ten third-year English students learning Mandarin Chinese as a second language (L2) at a UK university, to examine changes in L2 Mandarin during an eight-month period spent studying abroad (SA). We used three writing tasks and four speaking tasks as measures of writing and speaking proficiency, to assess total output, grammatical accuracy, lexical development, pronunciation and fluency, repeated before and after SA in China. Overall mean oral proficiency scores improved significantly ($p < .05$), especially speech rate ($p < .01$), supporting the claim that SA favours fluency development (Freed et al. 2004), although the measures highlighted difficulties in clarifying precisely how to assess oral proficiency. Written proficiency showed fewer marked improvements: only one writing test (an untimed short essay) significantly improved in length ($p < .05$), and increased complex grammar (use of *de*-relative clause morphemes, $p < .001$). A sub-group (n=7) provided quantitative data on L2 Mandarin use at different times during SA, showing clear individual differences, highlighting the value of capturing details of students' experiences during SA (Regan et al. 2009). We also note the lack of standardised linguistically-informed measures for tracking development in L2 Mandarin (Freed et al. 2004; Pallotti 2009, De Jong et al. 2012). Further research is therefore much needed to identify systematic linguistic development in L2 Mandarin, and also to bridge theory and practice in L2 Mandarin language teaching to clarify the interconnecting factors that affect L2 Mandarin language development.

1. Introduction

This small-scale exploratory study focuses on learning L2 Mandarin, which is as yet a radically understudied area in language pedagogy and L2 acquisition research. The study compared changes in oral and written measures of L2 Mandarin before and after a period of Study Abroad (SA) study for ten undergraduates from a UK university, who had started learning Mandarin *ab initio* on arrival at university two years earlier.

Research on L2 acquisition, especially development of L2 speech, has become a richly informed but often fragmented area, involving different disciplines and sub-fields. There is still relatively little research integrating formal research into L2 acquisition of grammatical or lexical knowledge (linguistic competence) with research into changes in L2 oral interaction (communicative competence), despite this problem being noted many years ago by Dell Hymes (1972). In addition, models of longitudinal stages of L2 development,

[*] We gratefully acknowledge the support of Newcastle University (HASS Faculty Board) for funding this research, also Linda Cheng of the School of Modern Languages for collecting the data, and Dr Alex Ho-Cheong Leung of Northumbria University and Dr Chi-Wai Lee of Newcastle University for assistance with data analysis. We are also most grateful to colleagues from Newcastle University and to this journal's reviewers, for their support, advice and comments.

especially oral development, remain under-researched, especially beyond the standard English/European language focus (Pienemann 1998; Vainikka & Young-Scholten 2005).

During the last twenty years, research focusing on L2 development of language learners studying abroad (SA), i.e. spending time in the immersion setting, has become increasingly popular. SA research encompasses a broad agenda covering many aspects of linguistic and socio-cultural issues affecting language learners (Kinginger 2011). SA research can thus be seen as resolving some of the fragmentation noted above, offering a way of capturing valuable data of how language knowledge and language use change, particularly in interaction, when the type of exposure to input changes. There is an added pedagogic driver to the research, in terms of evaluating the specific value of SA programmes as study abroad itself attracts more and more language learners, and more language programmes are set up to promote SA (Wright & Schartner 2013). Yet to our knowledge, a bridge between research into these linguistic, contextual and pedagogic factors has not been widely established in L2 Mandarin development.

Given the scope of the current study, we focus here primarily on research on SA effects on language development, taking SA to be a potential trigger for significant change in language proficiency, given the assumed increase in exposure moving from a foreign-language classroom to immersion. However, the results of many current SA studies vary. Some research findings claim that SA is much more beneficial for students' language proficiency than other contexts such as classroom teaching in the home country (e.g. Brecht et al. 1995; Davidson 2010; Du 2013). Other studies (e.g. Collentine 2004; Freed et al. 2004; Isabelli-Garcia 2010) argue that study abroad does not guarantee language proficiency as commonly assumed, that immersion is necessarily not as deep and effective as expected (Kinginger 2011), but that even the notion of language proficiency itself is complex, and needs careful analysis (see Kinginger 2011 for an overview).

Research on SA effects remains inconclusive at the level of effects on different aspects of language development, such as grammatical accuracy, written proficiency or oral proficiency. Collentine (2004) concluded that study abroad is not necessarily the best way to improve students' grammatical accuracy. This study examined 17 morphological, syntactic and morpho-syntactic variables to compare the improvements in a group of study-abroad students with a comparable group of at-home students over a semester. The at-home students had more improvements on grammatical items such as present-tense and indicative accuracy on the verbal level and subordinate conjunction selection on the syntactic level. Isabelli-Garcia (2010) focused on one grammatical item, gender agreement, and did not find significant difference between at-home students and study-abroad students after a period of four months.

A few studies have focused on SA effects on written data in general assessments of language development (e.g. Sasaki 2011; Serreno, Tragant & Llanes 2012), although some discrepancy has been found in the length of SA time needed to show significant progress. Sasaki (2011) investigated improvements in Japanese students' L2 written English with or without SA, focusing on argumentative compositions. This study discovered that the students who studied abroad for more than four months showed significant improvements on a global measure of written ability assessed by EFL specialists than the other students. Serreno et al. (2012) elicited a 150-word written descriptive piece from participants over a three-semester period. After a semester abroad, the students did not show significant progress in fluency, syntactic complexity, lexical richness or accuracy in this writing task from before going abroad. However, by the third semester, progress in those measures became significant.

By comparison, studies on oral proficiency, especially fluency, find more robust support for the benefits of study abroad, even where time spent abroad is not the same (Brecht et al. 1995; Davidson 2010; Du 2013). Brecht et al. (1995) found marked progress in

oral proficiency in US students on a Russian study-abroad project, as measured by oral test scores. Davidson (2010) replicated the study with 1,881 US-based students, and also found improvements in oral proficiency. Specific aspects of oral proficiency, such as target-like sociolinguistic usage, has also been found to improve after SA: studies have found improvements in grammatical/discourse factors in L2 French, such as the use of formal/informal modes of address or informal forms of the negative (e.g. Regan 1995).

However, the claim that study abroad is the best way of improving oral proficiency is not completely supported. Freed et al. (2004) also tested oral proficiency specifically focusing on temporal measures of oral fluency. They examined the performance of three groups of students, at-home students, students who studied in America but in a summer immersion French camp, and study-abroad students who studied in France. The summer immersion group showed improvements on oral fluency after seven weeks, while the at-home group and study-abroad group did not show significant improvements even after 12 weeks.

The variance across the different studies about the effects of SA may therefore, to some extent, be ascribed to methodological and design complexities in the field, as findings may reflect different amounts of time spent abroad (not always controlled for in SA studies), to different tasks used in different studies, time that students spent speaking or writing in the target language, quality of input, or the starting proficiency levels of the students.

Studies which use generalised measures of proficiency such as an essay or the standardised Oral Proficiency Interview (OPI) are difficult to compare with studies which test more specific targets at various levels or different contexts (e.g. comparing Freed et al. 2004 with Brecht et al. 1995). Quantity and quality of input and output in the target language during SA may also differ greatly: Du (2013) claimed that oral fluency is most influenced by time-on-task, i.e. the amount of time that students use the target language every day. However, Moyer (2011) evaluated the effect of both the quality and quantity of L2 experience on accent as another aspect of oral proficiency, and found that quality of L2 experience is more important than quantity, as measured in terms of significant context-specific interaction. Meanwhile, Wright's (2013) longitudinal study of oral proficiency among 32 Mandarin learners of English in postgraduate study in the UK found no significant effect on improvement associated with qualitative or quantitative differences in target language use by the study participants.

In addition to quantity and quality of input and output, the starting proficiency level is believed to be a relevant factor too. Davidson (2010) found speaking gains among English L1 students only at advanced levels in a Russian study abroad program. Marqués-Pascual (2011) discovered that advanced learners produced subject-verb inversions after a semester abroad in a Spanish verbal morphology study, while the intermediate-level students did not.

There is clearly great variance across research questions, methodologies and results for SA research generally, and more consistency of research tools is needed to establish greater cross-comparison among different measures of language knowledge and language use. To date most of the research has focused on European languages, which are arguably typologically and sociolinguistically somewhat related. It therefore needs to be established how far standard assumptions about SA drawn from SLA and pedagogic research can be transferred to non-European languages such as Chinese, and L2 Mandarin specifically. In this article, we will use the terms Mandarin throughout to avoid confusion, though we note that the sources referred use either Mandarin or Chinese to some extent interchangeably.

It is clear that SLA research in L2 Mandarin is much needed, given recent rapid increases in numbers of students studying Mandarin, including in study abroad settings. According to official Chinese sources, such as the China Scholarship Council and CUCAS, China's English-language University and College Admission System, the number of international students in China is rapidly increasing. In 2012, over 320,000 students were

registered in China from over 180 countries to study at both degree and non-degree level (CUCAS 2013), with a projected target of 500,000 by 2020 (China Scholarship Council 2013).

Yet within SLA and pedagogic SA literature, this explosion remains relatively underexplored, and somewhat disparately reported, at least in English-language publications. Two recent papers, Shi and Wen (2009, in Mandarin) and Zhao (2011), provide overviews of linguistic research on L2 Mandarin acquisition, but these typically examined levels of success in acquisition of specific linguistic features, and did not include studies relating to longitudinal development or effects of spending time in the SA context. A brief report assessing SA programmes run by Georgetown University's Office of International Programs (VandeBerg, Connor-Linton & Paige 2009) included L2 Mandarin SA data, gathered using the standardised ACTFL Oral Proficiency Interview (OPI). Overall, participants across the whole program were found to have significantly improved oral proficiency; however, since this report incorporated all the Georgetown study abroad programs across eight countries, it was impossible to identify the exact improvements for L2 Mandarin learners.

Other studies which specifically examined SA in China include research focusing on sociolinguistic and attitudinal change, such as Jin (2012), Liu (2010) and Yu (2010). Yu (2010) collected questionnaires from 90 L2 Mandarin learners over a period of nine months. The result partially supported the claim that language-related attitudes and motivation would increase while language anxiety would decrease over the period, and concluded that students' self-ratings on their own language proficiency would increase over time. However, there was no objective measure of language proficiency reported in the study. Liu (2010) studied twelve students' improvements in a US study abroad program, by integrating different measurements: OPI, Mandarin language standardised assessment tests (SAT College Board 2014), a portfolio of general writing tasks and a survey asking for self-ratings on reading, listening, speaking, writing, cultural awareness, and personal career development. After an 8-week at-home immersion preparation program (living with native Mandarin speakers), an 11-month year of academic study before going to China, and then 4 weeks' residence in China, the students' tests scores all showed descriptive increases, and the majority reached advanced level on the OPI ratings after the period of residence in China. However, there was no report of what language features improved the most in each specific context, or any quantitative results to show how the oral and writing task performances compared. Jin (2012) investigated the effect of a study abroad programme; again this paper has a mainly pragmatic or sociolinguistic focus, discussing whether the students could successfully learn the use of compliment words, compared to native speakers. Whether the students' language fluency or accuracy was improved was not specifically reported.

Du (2013) is one of very few studies to take oral L2 Mandarin as its focus, investigating the development of Mandarin fluency over one SA semester during a 3-year study program. The researcher collected a range of speaking data every month from 29 students during their semester in China, in different contexts on and off campus, using both recorded Mandarin speaking classes for planned instructed output and using Labovian-style individual interviews[1] (Silver & Lwin 2013: 78), to elicit a range of spontaneous output. This study showed significant fluency progress over time but the results were taken from specific 2-minute segments chosen to highlight the students' most "productive" moments (Du 2013: 135) in terms of numbers of morphemes produced, from each of the four sessions across the semester. This may not therefore be as representative of changes in fluency as at first sight.

[1] [1] A Labovian sociolinguistic interview typically covers a range of topics and activities, including the interviewee retelling an emotionally charged personal experience, producing a rich range of language data of varying degrees of formality and complexity.

Nor was there any evidence from other sources such as written output to provide any comparison of overall language development.

It is clear that there is still much to be known about how L2 Mandarin may be expected to change over time, with no established models of stages of development in grammar or vocabulary, or reliable empirical methods to assess changes in written proficiency and oral proficiency. This paper therefore seeks to add to this new and growing field of research by reporting on an exploratory year-long study of SA effects for a group of intermediate-level students of Mandarin at a UK university, examining aspects of language development in both writing and speaking before and after SA. We assessed writing using four measures for accuracy, length, complexity and optionality. We assessed speaking using six measures for accuracy (in grammar and pronunciation), total output (number of utterances, length of utterances, lexical diversity, hesitations and speech rate.

By combining both written and spoken language changes in a SA context, we aim to provide a holistic perspective on changes in both language knowledge and language use, of benefit for future linguistic and pedagogic research in L2 Mandarin.

2. Study design

The study presented here, for reasons of space, focus on a subset of the whole research project, looking at comparisons before and after SA of assessments of written and oral proficiency. Our research questions were:

(1) How does writing performance change after Study Abroad (SA)?

(2) How does oral performance change after SA?

2.1. Participants

There were ten third-year L2 Mandarin students from a UK university (aged around 20 years old), who went to different cities in China for their study abroad. None had studied Mandarin prior to starting the course at the UK university. All the students during the SA were based in university classes, in Beijing, Shanghai, Xi'an, Chengdu, or Hainan for around eight months (two semesters, October to June); some stayed on longer after classes finished for holiday travel. Most were in mixed-language university dorms throughout their residence in China. This resulted in some lingua franca English use outside class but was comparable across the group[2]. They were therefore judged, as far as possible, to have had comparable experiences in formal exposure to Mandarin during SA.

2.2. Data collection

Given the lack of normalised measures for L2 Mandarin development, and the exploratory nature of the study, the researchers used standard university examination data pre-SA and repeated post-SA, using a battery of writing and speaking tests. These tests (detailed below) provided quantitative data on changes in writing and speaking performance for comparison in a statistically reliable pre-test/post-test design. As language learning context is known to affect language development (Kinginger 2011; Moyer 2004), individual reports of language usage in formal settings in class, and informal settings out of class, were

[2] One moved into private off-campus accommodation after one semester, but her data did not show significant differences in improvement to the others, so this is not taken to indicate that her living situation impacted on her language development significantly differently to the rest of the group

also collected throughout the year, to provide some contextual quantitative information about the amount and type of exposure the students had. The language usage reports in this study were not designed to tap exposure in qualitative detail, given the exploratory scope of the study, and the small numbers of participants, but they provide some context for the longitudinal data presented here, comparable to other research conducted using similar tools (e.g. Wright 2013). As noted above, formal exposure to Mandarin was deemed to be reasonably comparable; there may well have been individual differences in informal exposure to Mandarin, but the language usage reports did not appear to capture any significant skewing effect, and so the pre and post-SA test design is argued here to provide a valid set of reliable data for statistical comparison.

2.2.1. Written production data

At both times of testing, the participants were asked to produce two pieces of writing on similar themes in exam conditions, within half an hour – a description and a dialogue; following standard assessments, these were marked out of 15; they also completed one untimed free essay as a classroom-based assessment on expectations of life in China. These three pieces of writing yielded four measures for writing development: one overall measure of grammatical accuracy on the timed work, and three linguistically-motivated measures on the untimed piece of work: length (total characters), complexity and discourse-level optionality.

Accuracy was measured as target-like use of functional morphemes. Complexity was measured by use of two *de*-morphemes: *de*-possessive, seen as early acquired, and *de*-relative, marking relative clauses, seen as late acquired (Zhang 2005). This distinction follows standard models of SLA (and indeed child acquisition) which suggest that late acquisition, either for formal or processing reasons (Pienemann 1998; Vainikka & Young-Scholten 2005), relative clause structures are late acquired. Discourse-level optionality was marked by omission of the *shi* copula, which is required in L1 English, but can be omitted in L2 Mandarin in certain pragmatically-licensed predicate contexts, and is known to be hard to be target-like even at advanced level (Yuan 2013, p.c.).

The different measures for untimed vs. timed work were required due to lack of consistent production in the timed work (some participants did not complete both sections of the task, or had marked differences in length of writing), which made the specific linguistic analyses of length, complexity and optionality hard to compare statistically across all three pieces of written work on both times of assessment.

2.2.2. Oral production data

At both times of testing, the participants were given four tasks to complete in approximately ten minutes, tapping different aspects of preparation, planning and communicative competence. These four tasks were: a pre-prepared monologue, one of a choice of pre-prepared dialogic role plays, an unprepared, unplanned description of a photograph prompting questions about the content, and a free dialogue about expectations of life in China. The tasks were recorded in examination conditions, then transcribed for further linguistic analysis.

The tasks yielded six measures of oral development: an overall measure of average oral proficiency (aggregating standard assessment ratings out of 15 for accuracy and pronunciation); then five linguistically-motivated measures: total number of utterances, mean length of utterance or turn (MLT), lexical diversity (using a standardised measure of lexical range: D), disfluency, and speech rate.

Accuracy and pronunciation were recorded by the native-speaking examiner, using a subjective but standardised rating scheme as part of the overall university examination scoring procedure. Other measures were assessed by the researchers after transcribing the speaking tests using CHILDES conventions (MacWhinney 2000), using CLAN programs to analyse total number of utterances, mean length of utterance, and lexical diversity using D (Malvern et al. 2004). Transcripts were analysed in terms of single characters, in line with some researchers, who acknowledge that standard definitions of morphemes or words do not easily fit the Mandarin context, but that character calculations can be reliably compared to syllable-based measures of western languages (Du 2013). There is a potential problem when calculating D in Mandarin, as the distinction between single and bi-morphemes could be blurred, and making comparisons with lexical diversity in other language studies harder to manage. However, the researchers were unable to find an alternative normed measure, so D as measured by single characters was used here to aim for greatest consistency. Disfluency was calculated following Wright (2013) as the aggregated total of filled pauses (*um, er*) and repairs (repeated and reformulated expressions). The word count totals are automatically produced in CLAN analyses using standard computerised programs which measure mean length of turn (MLT) and word frequency (FREQ). The aggregated total was then divided by 100 to present a ratio (between 0 and 1) of disfluency per total length of output.

Speech rate was assessed by two native-speaker research assistants, analysing each of the four speaking tasks (each approximately 2 minutes long), which measured the number of characters produced in a 20-second segment in the middle minute of the task. We appreciate this is not the standard way of establishing speech rate (e.g. De Jong et al. 2012), which would be to use a per-minute ratio (using syllable or word). However, each task used in this study was relatively short, with no official timing constraint across the whole test or within tasks, so we found considerable variability in how each speaking test was conducted both in overall time given to each component, and in interviewer/interviewee utterances at the start and end of each task. Given this variability, we decided that using a standardised selection of the central segment of each task was a sufficiently clear length of run, not confounded by task process, which would be valid and reliable as a measure of speech rate for each participant across all four tasks at both times of assessment.

2.2.3. Data collection

We recorded participants' scores from the end of second-year university examinations, taking their results from the oral and written components; we also used a final in-class free writing assessment (Time 1). The same students took the same oral and written exam components and writing assignment at October at the start of their final year – i.e. Year 4 (Time 2) following their year in China. All scores from Time 1 and Time 2 were coded using SPSS and compared using statistical analysis. Due to student absences from written task data collection at Time 2, the written analyses refer to nine participants for the timed data and eight participants for the untimed data, while the oral analyses include all ten participants. Given the small participant set, all data are analysed using non-parametric tests.

3. Results and discussion

3.1. Research Question 1 – writing

Our first research question looked at changes in writing performance after Study Abroad. The writing performances were marked with one score for accuracy out of 15. Mean results (with SD, minimum and maximum) comparing Time 1 and Time 2 are shown in Table 1 below; percentages are also shown for ease of comparison. These (and all subsequent)

scores were tested for significant differences using non-parametric Wilcoxon signed rank analysis, due to the small sample size (significance assessed as $p < .05$ or below).

Table 1: Writing Scores

	N	Minimum	Maximum	Mean	SD
Average timed writing score Time1	8	8.75 (58.33%)	13.00 (86.67%)	10.75 (71.67%)	1.269 (8.457%)
Average timed writing score Time2	8	8.50 (56.67%)	12.25 (81.67%)	10.39 (69.26%)	1.377 (9.171%)

The writing scores reveal that even at Time 1, participants were rated at just above 70% accuracy; there was wide individual range – the highest-scoring participant achieved 87%, compared to the lowest-scoring participant with 58%. By Time 2, there was a slight overall decrease to just below 70% mean accuracy, with slightly larger individual range – the highest score decreased but still was above 80%, while the lowest scored 57%. Investigating further, it was evident that Task 1, the dialogue, showed a decrease from a mean of 11.17 at Time 1 to a mean of 10.33 at Time 2, compared to Task 2, the letter, which showed almost no change from a mean of 10.33 at Time 1 to 10.61 at Time 2.

Measures from the third piece of writing, the untimed class-based assessment, allowed us to examine development in terms of total length (number of characters) and specified target morphemes, shown in Table 2 below. Length of writing showing a marked increase by Time 2, and reduced SD (although the change was not significant, $p > .05$)

Table 2: Total length of writing assignment (total characters)

	N	Minimum	Maximum	Mean	SD
Total length Time 1	8	544	825	661.50	105.93
Total length Time 2	8	615	904	715.50	98.91

We also analysed changes in three specified morphemes used here to indicate grammatical development. The *de*-possessive marker, and the *de*-relative clause marker, were taken to indicate earlier and later stages of development (following Zhang, 2005). The third indicator of grammatical development was to check for accurate use of the *shi* copula which is optional in Mandarin – whether it is needed or not is pragmatically determined by the context, unlike English copula *be*, which is always required. The total number of characters and the total number of accurately produced target morphemes were noted and compared between Time 1 and Time 2, shown in Table 3 below.

Table 3: Total production of target morphemes in writing

	N	Minimum	Maximum	Mean	SD
de possessive morpheme Time 1	8	14	28	20.38	5.24
de possessive morpheme Time 2	8	6	19	10.00*	4.47
de relative morpheme Time 1	8	1	4	2.50	1.07
de relative morpheme Time 2	8	13	24	18.00*	3.59
shi copula morpheme Time 1	8	1	7	4.63	2.56
shi copula morpheme Time 2	8	2	8	3.25	1.98

* *Significant difference between Time 1 and Time 2, p < .05)*

There was a significant decrease in *de*-possessive, from 20 tokens to 10 ($p < .05$), and a significant increase in *de*-relative, from 2 to over 20 ($p < .05$). Although the overall number of tokens is relatively few, the wide use of the *de*-possessive at Time 1 suggests this marker is easily acquirable. The few occurrences of de-possessive at Time 1, and the significant improvement in *de*-relative by Time 2 provides some evidence for Zhang's (2005) argument that *de*-relative is acquired after *de*-possessive. It is likely that some increase in *de*-relative would have been seen in any learning situation, and further study is required to test the claim more reliably that SA fosters acquisition of specific elements of complex grammar. The decrease in *de*-possessive by Time 2 is noteworthy, but is not, of course, to do with acquisition per se. Rather, we argue that the change in distribution may reflect a shift away from over-reliance on early-acquired grammatical forms, as a wider range of complex morphemes are acquired such as aspect markers, passive constructions and other more complex grammatical forms. But future research into this data for greater lexical analysis of morpheme distribution is needed to support this hypothesis.

The cohort's grasp of the appropriate optionality of the *shi*-morpheme was markedly lower than the other target morphemes at Time 1, with no marked improvement by Time 2. This fitted expectations, since this optional morpheme was targeted to show pragmatic or discourse-level understanding of where grammaticality depends on context, and was predicted to be late acquired. However, the very small number of morphemes produced here are not sufficient evidence to validate our hypothesis; more specific tests are needed to see how far optionality remains a problem during the four years' formal study of Mandarin and whether SA could make a real difference, as it has been shown to do on other studies of contextually determined optionality, such as Regan's (1995) study of L2 dropping of *ne* in French negation).

The results here are illustrative rather than generalisable, but indicated that, in general, writing, especially timed, remained problematic for participants, which needs further investigation as to why the mode of assessment is challenging: it could be down to L2 specific problems (e.g. lexical knowledge/character familiarity), or generic problems in foreign language writing at schematic/discourse level (as noted in the L2 English writing literature, e.g. Hamp-Lyons 1991, Ferris & Hedgcock 2005). Focusing on discourse-level writing skills as such was beyond the scope of this study. However, we argue that classroom work could benefit from including specific linguistically-motivated measures to test development. We found improvements in the untimed piece of work in terms of marked increase in length, and evidence of grammatical development with significantly higher uses of the more complex *de*-relative morpheme.

This kind of linguistic analysis provides deeper insights into the specific nature of changing language knowledge in line with other SLA research, but could also be helpful for teachers and language departments to understand better how to assess evidence of progress in using increasingly complex morphemes. The improvement in quantity and complexity would not have been identified by simply comparing exam scores in the timed written tasks, which may commonly be the only formal summative evidence put on record. As this is based on only one institution's practice, we do not draw any wider conclusions about the nature of summative assessment in L2 Mandarin classrooms in general, but we suggest it would be useful to validate best practice across institutional ways of assessing L2 Mandarin, to ensure sufficient breadth of linguistic change is captured.

3.2. Research Question 2 - speaking

Speaking performances were assessed for accuracy and pronunciation, aggregated together to give an average oral score (out of 10). Again, mean results (with SD, min and max) comparing Time 1 and Time 2 are shown in Table 4 below; percentages are also shown for ease of comparison.

Table 4: Oral Scores from examiner ratings (aggregating pronunciation and accuracy)

	N	Minimum	Maximum	Mean	SD
Average oral score Time 1	10	5.00	8.20	6.54	1.103
		(50.0%)	(82.0%)	(65.4%)	(11.04%)
Average oral score Time 2	10	5.60	9.00	7.31	1.172
		(56.0%)	(90.0%)	(73.1%)	(11.7%)

The aggregated speaking scores from the examiner show clear though not statistically significant changes over time. Mean scores at Time 1 was around 65%, although the highest score was above 80%. By Time 2 the mean had increased to 73%. This also reflected a wider individual range than at Time 1 - the highest scoring participant achieved 90%, while the lowest scored less than 60%. It is interesting to note that this shifted the comparison between writing and speaking as measured purely in examination scores – participants were rated overall lower in speaking than writing at Time 1, but higher in speaking than writing at Time 2.

These differences seem to support claims that speaking is the language skill most assisted by study abroad (e.g. Freed et al. 2004). However, closer analysis of the two oral sub-measures combined here – accuracy and pronunciation – showed differences between the two sub-measures. Pronunciation improved from a mean of 7.15 to 7.5, while accuracy remained similar with a mean 6.4 at Time 1, and 6.3 at Time 2.

We then transcribed and analysed the oral tests to look further at oral proficiency, using specific linguistic measures of fluency and lexical development. The transcripts were analysed using CLAN (MacWhinney 2000) to measure total number of utterances, mean length of utterance, lexical diversity (D), mean speech rate and disfluency. These measures were calculated using characters for consistency. Speech rate scores for Time 1 and Time 2 are shown in Table 5 below.

Table 5: **Oral scores by sub-measure**

	N	Minimum	Maximum	Mean	SD
Total utterances Time 1	10	44	112	67.80	21.75
Total utterances Time 2	10	71	216	113.20**	39.61
Mean length of utterance Time 1	10	8.01	12.46	9.91	1.45
Mean length of utterance Time 2	10	6.61	13.67	9.30	2.03
D score Time 1	10	22.91	47.03	34.57	7.40
D score Time 2	10	32.51	60.02	48.76**	9.46
Speech rate Time 1	10	25.5	41.75	32.55	5.31
Speech rate Time 2	10	37.75	55.25	46.83**	6.24
Disfluency Time 1	10	.19	.46	.31	.097
Disfluency Time 2	10	.10	.56	.26	.142

** *significant difference between Time 1 and Time 2, $p < .01$)*

There were highly significant improvements in total utterances, lexical diversity (D) and speech rate ($p<.01$); but small and non-significant reductions in mean length of utterance and disfluency. In other words, they produced more overall, with wider vocabulary choice and at a faster rate. This supports the snapshot view presented earlier that oral scores evidently improved, but present interesting differentiations as to how that oral improvement is actually assessed. Particularly of note is the discrepancy between the evidence of more total spoken output, produced at a faster rate, but with no significant reduction in disfluency or increase in length of individual utterance. Taken in conjunction with the finding earlier of no significant improvement in grammatical accuracy (see Table 1 above),this suggests that improvement in oral fluency could be ascribed to easier lexical retrieval due to wider vocabulary, and faster articulation whether accurate or not rather than a more developed grammatical repertoire. However, more in-depth analyses of the lexical and grammatical range, as well as more fine-grained temporal analysis of the speech data, which are beyond the scope of this paper, are needed to substantiate this claim.

In this cohort, overall improvements were found more evidently in speaking than in writing. In order to see if these differences were due to individual variation in quantity of language exposure, we cross-checked reports of individual language usage during the SA period with the pre and post test results. The quantitative language reports (following Wright 2013) were simple forms, so as to be very quick to complete, and designed to capture hours of language use across the four language skills over each day of a typical week. Dividing the eventual total by 7 provided a daily average, presented here as overall average and average for speaking. However, we were limited in how we could use these reports due to low numbers of returns. Seven participants provided data at Time 1 (within six weeks of arrival), summarised in Table 6 below.

Table 6: Average daily hours of language use, and speaking at Time 1

	N	Minimum	Maximum	Mean	SD
Average usage	7	.32	2.81	1.45	.84
Average speaking	7	.75	4.86	2.24	1.46

Only three participants provided further reports for a mid-point and end-point of SA, so the reports cannot be used for statistical association with the linguistic measures. However, in view of the increase in oral scores discussed above, it is interesting to note the participant with the highest average hours of speaking during the whole period of study (over 4.5 hours both at the beginning and end of SA) overall did improve markedly in her oral scores (7.0 out of 10 at Time 1, 8.8 at Time 2). The participant who had the lowest average hours of speaking (below 3.5 across the SA period) did not markedly improve on her oral scores (5.0 out of 10 at Time 1, 5.6 at Time 2). The third participant who had the most marked increase in average hours of speaking across the year (2.71 at the beginning to 3.7 at the end) also consistently scored amongst the highest on all measures at Time 1 in both oral and written exam scores. These patterns suggest there may be a complex linguistic-affective "threshold" effect (Wright & Schartner 2013), where those who are already feeling proficient and/or confident before SA can then engage more fully with, and benefit more, from the target language environment, while those who do not feel proficient may not feel able to engage as fully, and may not make noticeable gains over time. Further qualitative investigations using a more substantial report method (such as Regan et al. 2009) would allow this claim to be substantiated.

4. Discussion and implications for future study

This study focused on identifying changes in written and oral production comparing pre and post Study Abroad, with mixed results as seen above. We identify three key strands emerging in evaluating these results in terms of linguistic outcomes, methodological limitations, and pedagogic issues, all of which have implications for future research in L2 Mandarin learning and teaching.

Analysing the linguistic data, we saw that oral scores improved generally, and often significantly, in line with expectations, and confirms the assumption in SA literature that oral proficiency is most clearly boosted by exposure to the target language in the target country. This exploratory study was not designed in experimental terms, so we cannot speculate here how different the findings would have been with an intensive stay at home group, as examined by Freed et al. (2004). However, we found marked individual variation between aspects of the oral measures, so more detailed analysis is needed on the specific changes in speech rate and other fluency measures found here to understand the scope and effects of these individual variations. Further analysis is also required of the improved lexical range noted here, to test whether the change represents greater use of existing taught formulaic chunks, or represents novel lexical development, although this analysis needs greater development of reliable methodological tools for investigating L2 Mandarin speech.

Evidence of development from the written data is less clear; however, within the data discussed here, there is some evidence of expected increases in proficiency in terms of overall length of writing produced, and in grammatical complexity, indicating improvements had occurred despite lack of change in the summative accuracy ratings as used for examined writing.

Some of the tasks showed wide ranges and evident individual variation in responses, which may reflect differential effects of SA on individuals, and we argue there could be a linguistic-affective 'threshold' where those with less confidence or language proficiency found it harder to engage, which then affected the rate of development. Quantitative data designed to tap individual experiences of interaction were too few to use for statistical association. Further qualitative research into the nature of language interaction and experiences in the target country would also provide a rich source of insights into individual differences, to identify some of the complex interaction of factors that would be expected to impact on language learning and individuals' rate of development (Kinginger 2011).

Methodologically, in analysing the written data, clear markers of grammatical development were not as evident as hoped, as the tasks did not yield sufficient numbers of target morphemes, and not all tasks were completed by the whole participant pool. A wider range of target morphemes, using specific elicitation tasks, designed to tap developing complexity, together with ensuring more consistent task completion, would minimise these problems in future research.

Analysing the oral data, we found that due to the nature of the spoken examination conditions, there was some lack of consistency in task performance between Time 1 and Time 2, and evidence of individual variation in how the tester conducted the test. For example two of the participants ran out of time in completing all four sub-tasks.

Further difficulties arose when looking for ways to transcribe and analyse the data reliably for fluency and accuracy. There seems to be a serious gap in existing methodological practice for analysing L2 Mandarin oral proficiency which needs addressing urgently, including establishing reliable norms for transcribing characters as single or bi-morphemes, to make comparisons with other L2 research using syllables, morphemes or lexemes, to overcome the problem of what constitutes a "word" in cross-linguistic analysis. We also could not find any existing guidance on how to transcribe L1-influenced Mandarin or other

Chinese interlanguage forms in tone and pronunciation of certain difficult phonemes. However, by establishing our own norms (co-referenced between 3 linguistically-trained speakers of Mandarin), we were able to construct and analyse the data as shown above. But it was not possible in the scope of this study to establish, for example, clear evidence of how to assess lexical frequency of content or functional elements, or evidence of formulaic chunks, which are commonly examined in other L2 speech research (Wray 2000). We call for future research into how to normalise analysis of L2 Mandarin speech to investigate such areas.

Another issue in this study is the reliability and validity of our pre/post-test design. Data collected under exam conditions (Time 1) and in non-exam conditions (Time 2) may have affected participants' attitudes to completing the tasks; failure of several participants to complete the timed written data at Time 2 suggests this may have been the case[3]. We also noted a degree of individual variation in how the interviewer and interviewee conducted the oral tasks, leading to some imbalance between the lengths of focused time on each task across the sample.

Nevertheless, we argue that the ecological validity of using the same examination format to collect data pre and post-SA gives us a useful insight into the language learning process from a pedagogic point of view; in post-hoc discussions, the teachers and students themselves expressed a clear preference to finding out how their authentic learning was progressing using the examination format, and suggested that they would have been less keen to participate if additional data elicitation methods had been required. So, in future research design, there needs to be a judicious balance between recruiting participants and constructing adequately reliable data collection instruments.

Pedagogical implications of how study abroad can measurably aid language change, as highlighted in this study, is also an area to pursue further. While we acknowledge the limited scope of this study, based on a single institution's cohort of Mandarin learners, we argue that the deeper insights into language change found in the linguistic analyses of the oral and written data compared to the examination outcomes means that it is vital to build up discussions between teachers and linguistic researchers to understand better whether and how to include linguistically-principled measures within assessments of proficiency.

This also gives impetus to much needed research to benchmark examinations in L2 Mandarin, given its rapid expansion throughout the UK and beyond, to ensure how teachers and researchers can work together to see how linguistically-informed elements can be included, in a graded, sequential way, within summative measures of L2 Mandarin proficiency, as is well established for assessments of L2 English such as the Oxford Online Placement Test (Oxford English Testing 2013, Purpura 2004).

We therefore believe it is very timely to establish a broad research agenda into teaching methods, and classroom input/output in L2 Mandarin classes. It has been known for many years that instructed input may not always lead to effective learning of language knowledge (intake, or linguistic competence) or capacity to produce (output, or communicative competence) in L2 English or European languages (Ellis et al. 2009; VanPatten 2003). We also know for many European languages there are clear predicted stages of development (Pienemann 1998; Vainikka & Young-Scholten 2005). It is assumed that instruction may speed up the rate of development, but is unlikely to alter the route of development, whether in more formal or more communicative classrooms (Norris & Ortega 2000). However, these claims have not yet, to the authors' knowledge, been validated for L2 Mandarin.

[3] As with many YA studies, comparable pre-post data collection issues are compounded by participant drop-out: out of 22 participants originally recruited at Time 1 in this study, complete data sets were only available for eight participants on the written data, and ten on the oral data.

Tracking such stages may be difficult, in addition, depending on teaching practices. Given an assumed traditional value in Chinese pedagogy placed on drilling and recitation, e.g. in L2 English (Jin & Cortazzi 2006), it could be hypothesised that current expectations of L2 Mandarin development may to some extent consist of building up greater skills in producing memorised chunks of language, tied to the content of a standard syllabus, particularly as delivered by Chinese-trained native Mandarin speakers. But we need to test this hypothesis empirically, to see how far L2 Mandarin language development reflects instructed input, or rather follows a more linguistically-driven developmental trajectory.

It is worth also exploring the range of current language pedagogies being used: it would be useful to see how far learners and teachers have similar or different attitudes to traditional teacher-fronted versus more communicative learner-centred practices, and to see if there are differences in teaching practice in different settings or between learners of Mandarin from different educational contexts. So would the Chinese learners of English referred to by Jin and Cortazzi (2006) be similar or different to the English learners of Mandarin in this current study? And how would they compare to other Asian or African learners of Mandarin, in China or in local L1 contexts? The rapid rise in Mandarin classes at universities and schools in all these different settings could provide a good opportunity to bring theory and practice together to clearly assess what constitutes most effective instruction across different global contexts when aiming to build both language knowledge (linguistic competence) and language use (communicative competence) for this new area of language learning.

5. Conclusion

This study tracked the progress of ten UK university students of L2 Mandarin in written and speaking performance before and after eight months' Study Abroad (SA) in China. Despite some methodological challenges, clear results were seen in both writing and speaking. Some development in written proficiency was partially found, especially in writing length and improvement in one measure of grammatical complexity (use of the *de*-relative morpheme), though improvement was not seen in broad-brush examination scores of accuracy. Expected significant improvement in oral proficiency were found using broad-brush examination scores for accuracy, and in linguistic analyses of speech rate, number of utterances and lexical diversity, supporting findings from SA research in other L2s. However, not all measures of oral proficiency showed consistent significant improvement; this suggests that SA effect on oral proficiency was not as robust as predicted, and that more fine-grained analysis is needed to illustrate the complex nature of oral proficiency.

These differences between proficiency as measured by examination scores, or by analyses of linguistic development, raised questions of how SLA research questions and language pedagogy intersect – how to assess what is learned through input in and out of the classroom and how it should be tested, and how to benchmark best practice in L2 Mandarin pedagogy and assessment.

Collecting and analysing the data revealed many gaps in existing L2 models and methodological conventions of linguistic knowledge as applied to L2 Mandarin; one crucial gap lies in transcribing and analysing oral data, such as how to annotate interlanguage forms, and how to equate characters to the more usual terms of syllable, morpheme or word.

Indeed, we have found very few research studies published in English relating to longitudinal research in L2 Mandarin, especially combining a linguistic and pedagogic perspective as this study sought to do. Therefore, despite its limitations, we believe the innovative nature of our study retains validity for providing a starting point in analysing the development of L2 Mandarin, and stimulating some suggestions for urgently needed lines of research.

References

Brecht, R., Davidson, D. & Ginsberg, R. (1995). Predictors of foreign language gain during study abroad. In Freed, B. (ed.), *Second language acquisition in a study abroad context*, 9-37. Amsterdam: John Benjamins.

China Scholarship Council. (2013). China Update: Chinese Government to support 50,000 International Students in 2015. http://www.csc.edu.cn/laihua/newsdetailen.aspx?cid=208&id=2339. [Accessed 1 Dec 2013].

Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition* 26(2), 227-248.

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

CUCAS. (2013). 5 Reasons to Study in China. http://www.cucas.edu.cn/HomePage/content/content_126.shtml. [Accessed 19 May 2013].

Davidson, D. (2010). Study abroad: When, how long, and with what results? New data from the Russian front. *Foreign Language Annals* 43(1), 6-26.

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

Du, Hang. (2013). The development of Chinese fluency during study abroad in China. *The Modern Language Journal* 97(1), 131-143.

Ellis, R., Loewen, S., Elder, C., Philp, J., Reinders, H.& Erlam, R. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching.* Bristol: Multilingual Matters.

Ferris, D. & Hedgcock, J. (2005). *Teaching ESL composition: purpose, process, and practice*. Mahwah, NJ: Lawrence Erlbaum.

Freed, B. Segalowitz, N. & Dewey, D. (2004). Context of learning and second language fluency in French. *Studies in Second Language Acquisition* 26(2), 275-301.

Hamp-Lyons, L. (ed.) (1991). *Assessing second language writing in academic contexts.* Norwood NJ: Ablex.

Hymes, D. (1972). Models of the interaction of language and social life. In Gumperz, J. & Hymes, D. (eds.), *Directions in sociolinguistics: The ethnography of communication*, 35-71. New York: Holt, Rhinehart & Winston.

Isabelli-García, C. (2010). Acquisition of Spanish gender agreement in two learning contexts: Study abroad and at home. *Foreign Language Annals* 43(2), 289-303.

Jin, Li. (2012). When in China, do as the Chinese do? Learning compliment responding in a study abroad program. *Chinese as a Second Language Research* 1(2), 153-316.

Jin, L. & Cortazzi, M. (2006). Changing practices in Chinese cultures of learning. *Language, Culture & Curriculum* 19(1), 5-20.

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

Liu, J.J. (2010). Assessing students' language proficiency: Anew model of Study Abroad program in China. *Journal of Studies in International Education* 14(5), 528-544.

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

Malvern, D., Richards, B., Chipere, N. & Duran, P. (2004). *Lexical diversity and language development*. Basingstoke: Palgrave Macmillan.

Marqués-Pascual, L. (2011). Study abroad, previous language experience, and Spanish L2 development. *Foreign Language Annals* 44(3), 565-582.

Moyer, A. (2004). *Age, accent and experience in second language acquisition*. Clevedon: Multilingual Matters.

Moyer, A. (2011). An investigation of experience in L2 phonology: Does quality matter more than quantity? *Canadian Modern Language Review/ La revue canadienne des langues vivantes* 67(2), 191-216.

Norris, J. & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning* 50(3), 417-528.

Oxford English Testing. (2013). The Oxford Online Placement Test: What does it measure and how? http://www.oxfordenglishtesting.com/uploadedfiles/6_New_Look_and_Feel/Content/oopt_measure.pdf. [Accessed 6 Dec 2013].

Pallotti, G. (2009). Complexity, Accuracy, Fluency: Defining, refining and differentiating constructs. *Applied Linguistics* 30, 590-601.

Pienemann, M. (1998). *Language processing and second-language development: Processability Theory*. Amsterdam: John Benjamins.

Purpura, J. (2004). *Assessing grammar*. Cambridge: Cambridge University Press.

Regan, V. (1995). The acquisition of sociolinguistic native speech norms. In Freed, B. (ed.), *Second language acquisition in a study abroad context*, 245-267. Amsterdam: John Benjamins.

Regan, V., Howard, M. & Lemée, I. (2009). *The acquisition of sociolinguistic competence in a study abroad context*. London: Multilingual Matters.

Sasaki, M. (2011). Effects of varying lengths of study abroad experiences on Japanese EFL students' L2 writing ability and motivation: A longitudinal study. *TESOL Quarterly* 45(1), 81-105.

SAT College Board (2014). http://sat.collegeboard.org/practice/sat-subject-test-preparation/chinese. [Accessed 23 June 2014].

Serrano, R., Tragant, E. & Llanes, A. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian Modern Language Review/La revue canadienne des langues vivantes* 68(2), 138-163.

Shi, F. & Baoying, W. (2009). Hanyu zuowei di er yuyan xide de yanjiu yu sikao [overview of the second language acquisition of Chinese]. *Journal of Chinese Linguistics* 37(1), 130-144.

Silver, R. & Lwin, S. (eds.) (2014). Language in education: social implications. London: Bloomsbury.

Vainikka, A. & Young-Scholten, M. (2005). The roots of syntax and how they grow: Organic Grammar, the Basic Variety and Processability Theory. In Unsworth, S., Parodi, T., Sorace, A. & Young-Scholten, M. (eds.), *Paths of development in L1 and L2 acquisition*, 77-106. Amsterdam: John Benjamins.

VandeBerg, M., Connor-Linton, J. & Paige, R. (2009). The Georgetown Consortium Project: Interventions for Student Learning Abroad. *Frontiers: The Interdisciplinary Journal of Study Abroad (special issue*), 2-75. http://www.frontiersjournal.com/documents/FrontiersXVIII-Fall09-VandeBerg-ConnorLinton-Paige_000.pdf. [Accessed 9 May 2013].

VanPatten, B. (2003). *From input to output: A teacher's guide to Second Language Acquisition*. New York: McGraw-Hill.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics* 21(4), 463-489.

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. & Schartner, A. (2013). "I can't…I won't?" International students at the threshold of social adaptation. *Journal of Research in International Education* 12(2), 113-128.

Yu, B. (2010). Learning Chinese abroad: The role of language attitudes and motivation in the adaptation of international students in China. *Journal of Multilingual and Multicultural Development* 31(1), 301-321.

Yuan, B. (2013). Personal communication with lead author (March 21, 2013).

Zhang, Y. (2005). Processing and formal instruction in the L2 acquisition of five Chinese grammatical morphemes. In Pienemann, M. (ed.), *Cross-linguistic aspects of Processability Theory*, 155-178. Amsterdam: Benjamins.

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

(Corresponding author)
Clare Wright
Department of English Language and Applied Linguistics
University of Reading
Whiteknights
PO Box 218
Reading, RG6 6AA
United Kingdom

c.e.m.wright@reading.ac.uk