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

Copyright:

This is the pre-peer reviewed version of the following article: Wright C. An investigation of working memory effects on oral grammatical accuracy and fluency in producing questions in English. *TESOL Quarterly* 2013, 47(2), 352-374, which has been published in final form at: http://dx.doi.org/10.1002/tesq.68

Always use the definitive version when citing.

Further information on publisher website: http://onlinelibrary.wiley.com

Date deposited: 30th July 2014

Version of file: Author submitted

This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License.
What can help adult learners improve their spoken English?

Abstract

This paper addresses the question of how far psycholinguistic constraints may affect learners' L2 speech improvement during a period of immersion. Data are presented testing the hypothesis that individual differences in WM capacity are associated with individual variation in rates of progress in oral grammatical accuracy and fluency in producing English questions over a year's immersion. Thirty-two Chinese adult speakers of English were tested, before and after a year's immersion, to measure accuracy and fluency using an oral question elicitation task and for WM using a battery of L1 and L2 measures. Story Recall in L1 (Mandarin) was found to be significantly associated with individual improvements in oral grammatical measures, but not in improvements in fluency, supporting the hypothesis that WM may aid individuals’ L2 development. However, on the oral measures, there was no significant mean improvement in grammatical accuracy, although there was for fluency. The findings suggest that a year's immersion helps improve oral performance in adult learners, independent of WM capacity, but is no guarantee of significant grammatical change.
Background

Teachers have long observed that their second language learners, even at reasonably advanced levels of proficiency, can “stick” at a plateau of morphosyntactic variability in L2 speech. This level may be sufficient for effective receptive language use, and general communicative competence (Alptekin 2002; Canale and Swain 1980; Hymes 1972). However, prolonged variability in grammatical accuracy and fluency in speech production can also lead to great frustration (Richards 2008). This frustration is often strongly expressed by L2 learners from traditionally-oriented learning environments such as China, where opportunities for speaking may be limited (Gu 2003; Tsui 2007).

For these learners, the goal of years of hard work may be to achieve an academic qualification in an English-speaking country. They may be lured by the prospect that studying for a Masters degree in the UK, for example, provides an opportunity to “perfect” their ability to speak English - a claim found on Graduate Prospects, a UK government-supported website, marketing international study opportunities in the UK (www.prospects.ac.uk). Their English prior to matriculation has to be good enough to pass internationally recognized exams such as IELTS or TOEFL at the required level (typically IELTS 6.5, TOEFL 575 or iBT 89 in the UK). But even with this magic pass tucked in their back pocket, the reality, for many students, is that their linguistic proficiency, particularly their L2 speech, may not be sufficient to make the most of their experience of studying abroad, either on arrival, or, more frustratingly, on departure a year later.

It is well established that students, even with advanced and effective receptive levels of reading and listening in English, can find difficulties with spoken English in academic contexts, where high levels of accuracy are expected both in writing and speech (Hinkel, 2003; Zhang & Mi, 2010). In the worst case scenario this can lead to failing modules or even
whole courses, particularly at Masters’ level in the UK, where the emphasis is almost entirely on extensive written and oral assessments.

After decades of research and teacher observations, it is far from clear how to fully account for this disparity between comprehension skills and production skills even after a period of immersion. In general terms, many different paradigms can be invoked, including learner traits such as motivation (see, e.g. Dörnyei & Ushioda, 2009), socio-cultural factors, especially for Chinese students (see e.g. Cook, 1999; Edwards & Ran, 2006; Tsui, 2007), or differences in learning environment (Hinkel, 2005; Mitchell & Myles, 2004; Swain, 2000). Research specifically looking at effects of immersion on aspects of language remains inconclusive. Some researchers have found that immersion can help grammatical accuracy to improve (Flege & Liu, 2001; Isabelli, 2004), while others have only found effects on vocabulary or fluency or listening skills (Collentine, 2004; duFon & Churchill, 2006; Freed, 1995; Freed et al., 2004, Sunderman & Kroll, 2009).

Research focusing on learner-internal cognitive factors relating to speech has suggested that specific psycholinguistic constraints can affect L2 speech production, such as level of automaticity and working memory capacity (Kormos 2006). However, it has not always been clear how these constraints interact with learner-external factors including differences in learning environments. Another puzzle is the extent of individual variation in developing L2 oral proficiency. Teachers may feel in the dark about why one learner may show greater progress in developing oral proficiency than another, and more specifically, what they as teachers can do to help learners who are feeling “stuck” (Richards, 2008). In addition, it is not always evident how these constraints affect learners’ progress when they leave the classroom to engage in typical immersion during studies overseas, and why some learners seem to progress much faster in immersion settings than others.
There is thus a growing interest in how psycholinguistic constraints may affect improvements in speaking for learners in an immersion environment (see, e.g., O’Brien et al. 2006; Sunderman and Kroll 2009). Given the growing globalization of higher education, and the rising take-up of international postgraduate qualifications, especially by Chinese students, it is particularly important to understand the potential effects of psycholinguistic constraints on L2 oral proficiency for adult L2 language users immersed in a non-native academic setting.

One linguistic structure in L2 English that causes potential difficulty for students in immersion settings is question formation. There is clear empirical evidence that questions are late acquired in L2 English (Pienemann, 1998; Vainikka & Young-Scholten, 2005), and are prone to fossilization (Han, 2004). Specific morphological problems are also evident in variable or omitted subject-auxiliary inversion and tense marking (Pienemann, 1998; author, 2010). Furthermore, the grammatical knowledge needed for English question formation, especially indirect questions (verbal inflection and auxiliary movement, do-support, embedding) have also been identified to be an element of the wider difficulties students have with the kind of complex grammar required in academic contexts (Flowerdew, 2002; Hewings, 2001; Hinkel, 2003; Johns, 1997; Shaw & Liu, 1998).

Yet question forms, especially simple subject and object questions, are a frequent and explicit focus of taught input from the earliest levels, certainly in Chinese and Taiwanese textbooks (e.g. Nani, 2006), and both simple and embedded questions in all tense forms are tested in intermediate examinations of English proficiency such as Cambridge First Certificate (Acklam, 1996). So the questions of how much exposure to input is required to learn question forms and why there is such individual variation remain open.

One possible explanation for variation in accuracy and production is based on a dual-processing model of L2 knowledge and retrieval, comparing item-based knowledge and rule-
based knowledge (Pinker, 1999; Skehan, 1998). The combination of the prevalence of instructed input, combined with the traditional grammar-drill memorization techniques common in Chinese and Taiwanese instruction (Gu, 2003) could be argued to favor the development of item-based “chunked” phrases (Myles, 2004), where simple question forms may be stored as a “holistic” memorized item, such as “Can you tell me…” or “What do you like?”. However, in terms of accurate production, the underlying grammar may not established enough to provide the speaker with the capacity to generate targetlike question forms using subject-auxiliary inversion, or different person or time reference, such as “Can she see him?”, “What did she like?” or “Who liked her?”. Until sufficient underlying grammatical competence develops, speakers may “see-saw” between a variety of compensatory strategies, often resulting in hesitant speech, as they seek to retrieve explicitly learned metalinguistic rules to express what they mean (Herschensohn, 1999; Kormos, 2006).

In terms of fluency, dependence on learned chunks and lack of automatic access to grammatical knowledge also tend to have the effect of slowing up L2 speech. Non-fluent L2 speech may indeed be accurate but very slow, or it may be effortfully produced, with hesitations, slips and restarts as a result of monitoring form as well as meaning (de Bot, 1992; Dechert, 1980; Kormos, 2006; Tavakoli & Foster, 2008; Temple, 1997; Towell et al., 1996). Such conscious speech monitoring is very demanding of cognitive processes, especially working memory which is limited in capacity (Baddeley, 2007; Kormos, 2006; see also TESOL Quarterly special issue 42 (3), 2008). The role of WM in understanding constraints on developing L2 accuracy and fluency in speech however is not yet fully understood.

Working memory research has commonly been based on Baddeley’s influential multicomponent model of WM (1986, 2007), which separates WM into short term storage and processing capacity. In Baddeley’s model, there are separate temporary storage systems for visual and verbal material; verbal information, such as a word or phrase, is stored in a
“phonological loop” that lasts about 1-2 seconds. Processing such material in a demanding complex task, such as L2 non-automatized speech, uses a limited-capacity central executive, which interacts with the temporary stores in completing the tasks, controlling any switches in attention which may be needed and inhibiting unwanted or distracting information. A recent addition to the model, the episodic buffer, is argued to provide a means of combining temporarily stored material with information from long term memory. The processing and storage functions of working memory are seen to trade off each other. More material to be stored means less capacity for processing (and vice-versa).

Tests of WM capacity or span, focusing on the trade-off between storage and processing, have been used to assess WM impact on different aspects of language proficiency including L2 speech. One of the most widely-used tasks in L1 and L2 WM research has been some version of Daneman and Carpenter’s (1980) seminal Reading Span Test, measuring capacity to recall sentence-final words while reading or listening to increasing numbers of sentences at a time. Greater capacity on this span task has been found to correlate with higher scores in general measures of L2 proficiency (Harrington & Sawyer, 1992; Kormos & Safar, 2008; Service, 1992). Greater WM capacity has also been found in some studies of L2 speech to facilitate word retrieval and speech formulation, revealed in faster, more accurate speech (Fortkamp, 1999; Kormos, 2006).

However, other studies provide contradictory findings, where no such association was found between WM and L2 proficiency, especially for grammatical accuracy and oral fluency (Mizera, 2006; Sagarra, 2000). A recent study investigating these conflicting findings is Gilabert and Munoz (2010) who investigated whether differences in WM capacity explained differences in oral performance (e.g. fluency, complexity, and accuracy) for 59 high intermediate/advanced Spanish adult learners of L2 English. Their findings showed weak but positive significant correlations between WM and both fluency ($r = .231, p<.05$) and lexical
complexity ($r=.236, p<.05$). However, they noted there was no correlation with grammatical accuracy, and that the correlation with fluency disappeared when the group was split between higher proficiency and lower proficiency levels.

Such studies reveal that there are still many problems in using WM in L2 research, not least that WM may be skill-dependent: therefore a reading span task measures reading proficiency more than WM per se (Sagarra, 2000; Yoshimura, 2001). Another concern is how far WM is language-independent - some studies (such as Fortkamp, 1999) only found correlations between language proficiency and L2 span tests, which may confound language proficiency with WM. It therefore remains unclear how applicable studies based on the standard RST may be for understanding the role of WM in L2 development.

This study pulls together all these threads by addressing the question of how far psycholinguistic constraints, specifically WM, may affect Chinese students’ L2 spoken production of questions during a period of postgraduate study in the UK.

**Method**

The research reported here is one element of a larger study following a cohort of Chinese students throughout their one-year Master’s programs at UK universities. The students’ progress in spoken English was tracked during the period of immersion to try and understand more clearly why learners from the same educational and language background may show such individual variability. The hypothesis for the study was that individual differences in WM capacity are associated with individual variation in rates of progress in oral accuracy and fluency in producing English questions over a year’s immersion. WM capacity was tested using a range of L1 and L2 tasks testing the trade-off between storage and processing, in order to shed light on the methodological issues of testing WM outlined above.
Participants

Forty students from Mainland China and Taiwan, with Mandarin as L1, were recruited from Masters programs at two British universities. This sample was targeted to provide adequate statistical reliability and validity, and also homogeneity of external factors arising from prior language learning background. Eight participants dropped out during the period of testing, so the data presented here refer to the final pool of thirty-two participants. All the participants had been in the UK less than two months at the time of first testing; the second time of testing was after eleven month’s immersion. Participation was voluntary, and confidentiality was ensured; the longitudinal nature of the study was explained, although the exact linguistic focus of the study was not specified, to avoid test effects. All the linguistic and WM tests were conducted at a single session to avoid confounding effects of different conditions between WM measures and linguistic measures. All the participants were interviewed individually.

Biodata gathered at the first time of testing were checked for factors known to affect variation in exposure: sex, country, age of starting learning (AOL), length of learning (LOL); these are are summarized in the table below.

Table 1: about here

All participants were classroom learners; they had achieved IELTS scores of 5.5 or 6 (TOEFL 525, 550) within the last year, and were therefore considered to be at advanced level. ANOVA tests showed no significant effects for sex or country on IELTS score, age or length of learning (p>.1). Thus in terms of effect of external factors arising from education history or amount of exposure prior to arrival, the participants were assumed to be as homogenous as possible.
**Data Collection**

**Linguistic task**

Oral proficiency data for accuracy and fluency in asking questions were elicited using a question and answer 2-way gap fill task, adapted from Mackey (1999), in which participants asked questions to the researcher in order to “spot the differences” between two sets of pictures. The same test was used within 4 weeks of arrival in the country, and again after 11 months’ immersion, providing semi-longitudinal data of the students’ changes in oral grammatical accuracy and fluency.

Accuracy was measured as the number of target-like question forms spoken over five minutes, using verbal inflection and inversion. In line with Pienemann’s (1998) hierarchy of acquisition (shown in the table below), questions produced by the participants were divided into two groups: Stages 1-3 (formulaic chunks, intonation only, question word fronting without head movement; double marking of verbs), and stages 4-6 (copula fronting and inversion after wh-questions, head movement and *do*-support, cancelling inversion in embedded clauses).

*Table 2: about here*

The key global measure, question total, was the total number of target-like questions produced over five minutes, indicating that learners were at stage 4 and above. Target-likeness was defined as how many questions showed accurate verbal morpho-syntax and correct fronting of the wh-word. Non-target-likeness was defined as errors with verbal morphological inflections of tense or person. The overall question total was then subdivided to compare copula questions, use of modals and auxiliaries, *do*-support for lexical questions and complex or embedded questions. These different question types act as markers of specific stages of development from stage 4 to stage 6. Differences found in the total number of
accurately formed questions at each given stage, comparing both times of testing, would show progress from lower to higher stages over the year’s immersion.

A second global measure, question ratio, was also calculated to check for the degree of task avoidance (Schachter, 1974). Question ratio was calculated by dividing the question total described above by the total number of utterances, to see how far participants’ output reflected their capacity to address the task without reliance on statements or other circumlocution. The total number of utterances was counted (not including non-propositional back channeling such as ok, sorry, or filled pauses such as mmm), and the proportion of question to utterance was thus calculated as a ratio (measured between 0 and 1).

Fluency was assessed through lexical diversity as measured by type-token ratio (Malvern et al., 2004), and evidence of hesitancy phenomena (filled pauses and repairs, following Dechert, 1980; Towell et al., 1996). The oral data were transcribed according to standard conventions of oral analysis software based on the CHILDES project (CLAN), and the fluency measures were calculated using two of the standard calculation programs (frequency or FREQ, and Mean Length of Turn or MLT).

Type-token ratio was automatically calculated using the FREQ program in CLAN, which yields a figure between 0 and 1. All words were included including content and function words, and semantically empty words used to fill pauses, specified in accordance with CLAN guidelines, to include ah, aha, uh, um, mm, mmhm, oh. A higher ratio at Time 2 would be taken to indicate increased fluency. However, type-token ratio is on its own not seen as a strongly reliable indicator of fluency (Malvern et al., 2004). Hence the second measure for hesitancy was also used. The use of a calculation for hesitancy phenomena, rather than using other measures such as Mean Length of Turn, or articulation rate, was seen as providing specific insight into the sort of conscious, monitored, non-automatized speech output typical of non-proficient L2 speech. The assumption here is that greater use of monitored speech
would be evident from higher numbers of repairs, while greater proficiency or automaticity would be revealed in fewer repairs (DeKeyser, 2001; Herschensohn, 1999; O’Brien et al., 2006; Segalowitz, 2003; Tavakoli & Foster, 2008).

Hesitancy phenomena were calculated from the number of repairs and filled pauses (adapted from Towell et al., 1996), as shown in the sample given below. Each pair of brackets () denotes a fragment. Repeated material is shown in angled brackets <>; slashes in square brackets [/] or [//] indicate the degree of repetition or rephrasing. The FREQ program calculates the total number of full words including fragments, and shows the distribution of filled pauses. The MLT program calculates the total number of words minus the fragments (i.e. a lower number). Subtracting the lower MLT word total from the higher FREQ total provides the number of fragments. This number was added to the total number of filled pauses counted in the FREQ program and yielded the total for hesitancy used here.

This example of CLAN-marked output shows typical repairs and pauses (taken from a single participant’s output):

1. **Uh <whe(re)> [/] <where is> [/] <where is her> [//] sorry **uh <how> [/] how much friend **uh of her will come?**

2. **<Why I think> [/] uh is that boy look very angry why?**

3. **And what’s the wor(d) <did she> [/] uh was she prepared to say?**

4. **<Oh> [/] <oh> [/] oh a student.**

5. **But <if the> [//] <if> [/] if he <mi(ss)> [/] uh miss the party <did> [/] <di(d)> [/] did they would separate?**

Total hesitancy phenomena (in bold in this example): 10 filled pauses and 3 fragments = 13.
Working Memory tasks

Task 1: Listening Span

This test, conducted in L2, was adapted from the concept of Listening Span (Daneman & Carpenter, 1980; Harrington & Sawyer, 1992). The test involved hearing and repeating directions, which is known to be demanding due to cognitive load (Robinson, 2001), but was designed to be simple in linguistic terms. The task was therefore identified as a naturalistic but effective test of working memory, without confounding task proficiency with linguistic proficiency.

The task consisted of sets of English sentences, all between seven and nine syllables long, using words within the 1000 most common words in the British National Corpus to avoid any comprehension difficulties. Each sentence included one of four direction words - left, right, up or down. Participants listened to each sentence and immediately repeated the direction word used (e.g. left), then at the end of each set they were prompted to recall the final words of both sentences in the correct order. Groups contained three sets of sentences, starting at two sentences per set, and increased up to a possible maximum of five sentences per set. An example set is given below:

(6) a. Turn left after the train station          LEFT (repeated immediately)
    b. Go right at the supermarket               RIGHT (ditto)

Then station and supermarket were expected to be repeated when prompted.

Responses were scored following Conway et al. (2005)’s partial credit scoring scheme, where the number of correctly recalled final words was divided by the maximum number of strings tested, giving a ratio or decimal score between 0 and 1.
Task 2: Story Recall in L1 and L2

The second task was based on a Story Recall Test which forms part of a standard clinical Adult Memory and Information Processing Battery for native speakers with aphasia and other linguistic impairments (Coughlan & Hollows, 1985). The test involves listening to a short story with no supporting visual material, and then repeating as much of the story as possible, which is scored for recall of narrative sections in the correct order (gist) and of specific words and phrases (accuracy). Although not commonly seen as a standard WM test, it is argued to provide a similar type of Span measure as the Listening Span, and also to tap the combination of long-term and short-term memory suggested for the episodic buffer (author, 2010). Using both L1 and L2 versions allows for possible effects of language dependence and task specificity. This test had been found to correlate with use of complex syntax such as subordinate clauses and adverbial phrases in L1 speakers (Fry, 2002) and bilinguals (Fehringer & Fry, 2007), and was deemed appropriate for this study (Fry, p.c.)

Two WM Story Recall tests were devised in which participants listened to two stories, one in Mandarin (54 seconds long) and a different one in English (33 seconds). The first story, closely adapted from the original Coughlan and Hollows test, was translated into Mandarin (L1) and recorded as a digital sound file by the bilingual colleague who had helped with the Digits Back task. The second story in English (L2) had a similar schematic structure, but was shortened after piloting to avoid “floor effects” due to task difficulty (Harrington & Sawyer, 1992: 28). The English version was read aloud by the researcher from a printout.

The procedure for these tests was that participants were asked to listen carefully to a short story, which they were to repeat using the exact words and phrases as far as possible. The scoring for both tests was adapted from the AMIPB scoring system, and consisted of ten points for recalling the ten narrative sections in order, and between two and three points for the syntactic and semantic elements in each section (such as connectives, verb forms and
collocations). Two bilingual colleagues were recruited to analyze the data from the Mandarin version (one of whom was the translator and “voice” for the task). The raters were trained in the scoring system used in the English task, and worked with the researcher in scoring the Mandarin version to ensure cross-linguistic scoring reliability (inter-rater reliability was scored at .873 using Cronbach’s alpha analysis within SPSS). Participants could score up to a maximum of 50 points in total, which was then converted to an overall ratio between 0 and 1.

**Results**

Linguistic and WM data were coded and analyzed using SPSS; despite the size of the group, not all the linguistic and WM measures showed non-normal distribution, so non-parametric tests were used (Spearman rho correlations and Wilcoxon signed rank analysis).

**Linguistic data**

Two key measures were taken from the oral data: total number of target-like questions (from Pienemann’s (1998) higher stages of development) and ratio (proportion of questions to total utterances). Both were measured at each time of testing, and then Time 1 scores were subtracted from Time 2 scores to show the difference over time. Scores from Time 1 and Time 2 are shown in the table below.

*Table 3: about here*

Counter to the expected improvements in question total, the data in Table 4 show that the group mean scores actually changed very little between Time 1 and Time 2. Question total showed a slight mean decrease (-0.53, SD 5.12). Question ratio showed a slight mean increase (0.04, SD 0.15). The participants thus showed a slightly better ability by Time 2 to focus on producing questions in proportion to other utterances in their overall discourse.
strategy. However, according to Wilcoxon signed rank analysis, these changes were non-significant (p>.1).

Question total was sub-divided into different question types (following Pienemann, 1998): copula questions, lexical questions, complex questions). Non-target forms were also analysed (i.e. verb omission or errors in number and tense), and mean scores are shown in the table below.

Table 4: about here

Copula questions were significantly more common than lexical questions at both Time 1 and Time 2 (p=.000). Copula questions showed a very slight but non-significant mean decrease from Time 1 (0.72), although range slightly increased as shown in the higher SD at Time 2. Lexical questions showed no mean improvement from Time 1 to Time 2, although SD also slightly increased, due to a larger range at Time 2.

The number of complex questions (as an indicator of later stages of development) showed a marked decrease from Time 1 (mean 2.5) to Time 2 (mean 1.7), and this change was statistically significant (p<.05).

Given the lack of material change in numbers of any of the target question forms, the non-target forms were also analysed. The non-target forms showed a very slight mean decrease from Time 1 (mean 5.7) to Time 2 (mean 5.6), but this was not significant (p>.1). Looking at the non-target patterns in more detail, the principal non-target patterns related to lack of target-like verbal inflection, conforming to patterns shown in Pienemann’s (1998) hierarchy of development: omitting tense marking (Stage 1), leaving verb tensed but in-situ (Stage 2), oversupplying tense marking, e.g. using do-support as well as tense marking on verb (Stage
3). Other problems were switched use of *do* and *be*, and non-agreement of number (plural verb marking on singular subjects). These are shown in the table and figure below.

*Table 5: about here*

*Figure 1: about here*

Two findings are seen as interesting here – the increase in omission of tense marking (or of any verb-related word) by Time 2, and the use of *be* instead of *do* at both Time 1 and Time 2. The overgeneralisation of *be* as a default tensed verb in early L2 interlanguage is well attested in the literature (Hawkins, 2001; Hawkins & Casillas, 2008; Ionin & Wexler, 2002). The data presented here may be taken to show a default reliance on *be* both for auxiliary and copula purposes, perhaps due to ease of processing, as *be* is identified as frequent and salient in the input (author, 2010). The increase in verb omission by Time 2 is also notable, counter to the assumption that immersion would lead to grammatical improvement (Isabelli, 2004).

Investigation of individual scores showed that thirteen participants showed positive change but nineteen showed negative change. However, these changes were mostly small: nearly half the group (thirteen participants) showed positive or negative change of less than half 1SD beyond mean (-2 to +2), while only three out of 32 participants showed a positive increase of more than 1 SD. The results suggest that the group mean scores did not obscure significant individual differences in grammatical improvement.

In sum, the pattern of production for this task showed little change over time, and a preference for simpler copula forms over any other type of question at both Time 1 and Time 2. Wide range of scores at both times showed the variability between participants in their responses, but the expected grammatical improvement across the whole group after immersion was not found.
The fluency measures, by contrast to the grammatical data, showed significant improvement across the group over time: the type-token ratio increased (p<.001), and the number of hesitancy phenomena decreased (p<.05). These data are illustrated below.

Table 6: about here

The data clearly show that speakers were using greater lexical diversity, less hesitation and fewer repairs. This may be taken as a hallmark of increasing linguistic proficiency or automaticity (Dechert, 1980; Herschensohn, 1999; O’Brien et al., 2006; Segalowitz, 2003; Towell et al., 1996). It is possible however that the increase in omission of inflected verbs, noted above, may be connected with the decrease of hesitancy – speakers may be prioritizing completing the utterance with greater interest in pragmatic or communicative intent (Schauer, 2009), than monitoring and repairing for inaccuracy of verbal inflections.

To sum up, the results from the oral task suggest that eleven months’ immersion did not materially affect all participants’ underlying grammatical knowledge, but did affect how most participants were able to use what they knew. Thirteen out of 32 participants improved in their scores for grammatical accuracy measured by question total, but there was no consistent pattern of improvement from lower to higher stages of question forms, following Pienemann’s (1998) stages of development. The findings suggest that the kind of exposure gained through immersion itself was not a key factor in triggering development. I turn now to the WM results to test the central hypothesis of this study that WM would be associated with improvements in oral grammatical accuracy and fluency.

WM correlations
The hypothesis of this study was that WM scores at Time 1 would be associated in linguistic development during immersion (following Sagarra, 2000). As discussed above, three WM tasks were used, a Listening Span task in L2, and Story Recall tasks in L1 and L2, scored as ratios between 0 and 1 (following Conway et al., 2005). The mean scores show that overall Listening Span results were the highest, and that there was a clear language effect for Story Recall, as scores were lower for Story Recall in L2 than L1. All task results showed a considerable range, but Story Recall in L2 showed the greatest range. However, these between-test differences are not simply a reflection of language proficiency. Both Story Recall scores correlated significantly with each other ($r = .402$, $p < .05$); in other words those who were good at the task in their L2 were also good at it in the L1. By contrast, there was no significant correlation between Listening Span and Story Recall in L2 ($p > .1$), suggesting that better performance on L2 tasks was not simply a product of higher L2 proficiency.

Table 7 about here

The correlations for WM at Time 1 with differences in oral scores, unsurprisingly, showed a mixed pattern over grammatical and fluency measures. Counter to expectations, Listening Span did not show any consistent or significant trends for either grammatical or fluency measures. Story Recall in L1 showed the clearest pattern of association for the grammatical measures. Correlation with mean difference in question total between Time 1 and Time 2 was moderately significant ($r = .389$, $p < .05$), and difference in question ratio was at significance ($r = .367$, $p = .05$). In addition, analysis of the non-target data revealed a strong significant negative correlation between Story Recall in L1 and verb omission at Time 1 ($r = -0.646$, $p < .01$), and at Time 2 ($r = -0.560$, $p < .01$).
So, although not many participants improved on grammatical scores, those who did improve, either through increased numbers of questions, or through fewer errors, also scored higher in Story Recall in L1 at Time 1.

However, the pattern of association with WM and grammatical measures was not found for the changes in oral fluency. No WM measures showed any consistent or significant trends with type-token ratio or hesitancy phenomena at Time 2 or with differences between Time 1 and Time 2 (p>.1). This is consistent with Mizera’s (2006) study in which he failed to find significant correlations between WM (measured through span tasks) and oral fluency in L2 Spanish learners.

Discussion

This study’s aim was to investigate a hypothesised relation between WM and improvements in oral proficiency over a year’s immersion. Significant positive correlations were found between one WM task (Story Recall in L1) and oral accuracy in questions produced at Time 2 (r=.389, p<.05), and also with lower levels of omission of verbs at Time 2 (r=-0.560, p<.01). Given that this WM task was carried out in L1 (Mandarin), it provides some support for the argument that WM is a language-independent cognitive resource associated with second language development (Miyake & Friedman, 1998).

There are two key issues arising from these data which raise further discussion. Firstly, there were few consistent and robust correlations between linguistic scores and the full range of WM tests. Therefore, we support the call for standardising measures of WM for L2 research to clarify how far WM tasks tap working memory specifically, rather than a broader spectrum of linguistic and cognitive processing demands (Gilabert & Muñoz, 2010).

Secondly, over eleven months’ immersion, there was little material grammatical change, in contrast to significant improvements in fluency. This finding is consistent with other
research that immersion tends not to favour grammatical change (Rothman & Iverson, 2007) but leads to improvements in greater automaticity (Segalowitz, 2003) or in oral fluency (Freed et al., 2004; O’Brien et al., 2006). Methodological problems with the oral task design used in the study presented here may have skewed the results, which focused on oral question production (for reasons outlined above). In this task participants relied heavily on copula forms at both times of testing, and this may also be related to the picture-based design. The majority of copula questions seemed to fit a consistent pattern for checking or describing the picture, using only 3rd person singular “is..”, such as “Why is..”, “Where is …”, “What is…?” The use of copula forms with limited variation on tense and person observed here is argued to reflect a reliance on pre-fabricated chunks (Myles, 2004). Chunks are understood to be retrieved non-consciously, without much conscious monitoring for non-target forms, since monitoring is a costly strategy (Jackendoff, 2002; Paradis, 2004; Pienemann, 1998).

However, there is also substantial evidence in these data of monitoring the output, through repairs and slow hesitant speech, so any use of automatized chunks was only one production strategy adopted by participants. It is possible that participants were adopting a practical communication strategy in responding to the task – prioritising their attention on getting their meaning across. This may have made conscious control of target-like forms too difficult for participants to manage, particularly as they had been instructed to “ask as many questions as you can in five minutes”. This conclusion is supported by comments from some participants who remarked they had asked all the questions they needed to ask well before the five minutes were completed. This could have led to a task effect of producing questions “for questions’ sake”. This task effect is reflected in the wide range of actual output produced in the time allowed (e.g. at Time 2, this ranged from 23 to 64 utterances in total, and, out of these, from four to 26 accurate questions). The oral task used here may not therefore have been a reliable measure to tap linguistic knowledge, and further research establishing
standardised fluency and accuracy measures would help to diminish the effects of this methodological problem (Pallotti, 2009).

With regard to lack of improvement, it could be speculated that participants’ exposure to input while studying at a UK university is insufficient or may be overly “passive”, through the emphasis on academic listening, reading and writing. Oral output and interaction has long been argued to be crucial in language development (Gass, 1997; Long, 1990, 1996; Mackey, 1999; Mackey et al., 2002; Swain, 1985). Masters’ programs in the UK may typically be large, with a majority of non-native students (60 or more for a business or applied linguistics program), affording little opportunity for the kind of active communicative interaction that has been argued to trigger linguistic development by Swain, Long and others as cited above. However, it has been suggested that written input does not necessarily promote slower or less accurate linguistic knowledge than oral input (Cook, 2008). In addition, although formal academic exposure to input may be somewhat passive, language in lectures, seminars and informal student contact should have provided plenty of examples of question forms as targeted in this study. Further investigation into the role of different input contexts, and the importance of individual differences in responding to input, is clearly needed to clarify these issues further.

Conclusion

The hypothesis of this study was that WM would be associated with individual variation in L2 oral proficiency for advanced Chinese speakers of English during a year’s immersion during studies in the UK. Oral proficiency was measured using an oral question production task to test accuracy and fluency. Three WM tasks were devised for this study: a Listening Span task in L2, and Story Recall tasks in L1 and L2. The results show some support for the fundamental assumption tested in this study, that WM has some association with L2 linguistic
development, in view of a significant positive correlation between a language-independent WM measure (Story Recall in L1) and individual improvements in asking target-like questions in English (r=.389, p<.05), and in lower error levels in omitting verbs. However, this conclusion remains tentative since the results were not consistent across all WM or linguistic measures. In particular, there was no association between WM and improvement in fluency (in line with other studies such as Gilabert & Muñoz, 2010; Mizera, 2006). A secondary finding was that immersion seemed to facilitate general improvements in fluency but not general improvements in accuracy, as has been found in other research into study abroad contexts for language learning.

Thus further research is required to better understand what type of learning context is most effective for triggering change in grammatical knowledge. This would be appreciated particularly by more advanced learners who may feel great frustration at their perceived lack of progress, summed up by a sorrowful participant at the end of the project reported here, “Why I still can’t questions?”

(word count 8227)
References


Author (2010).


www.ncll.org.uk/50_research/10_research_projects/chinesestudents_html (last accessed 6.6.11)


Amsterdam/Philadelphia: John Benjamins


Graduate Prospects [http://www.prospects.ac.uk/coming_to_UK.htm](http://www.prospects.ac.uk/coming_to_UK.htm), Last accessed 29 June 2011


Sorace & M. Young-Scholten (Eds.) *Paths of Development in L1 and L2 acquisition* (pp. 77-106). Amsterdam: John Benjamins.


Table 1: Summary of biodata

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>All participants</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taiwan</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mainland China</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AOL (years)</td>
<td>11.41</td>
<td>1.58</td>
<td>7</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>LOL (years)</td>
<td>11.77</td>
<td>2.91</td>
<td>5</td>
<td>18</td>
<td></td>
</tr>
</tbody>
</table>
**Table 2: Stages of development in L2 English questions (adapted from Pienemann 1998)**

<table>
<thead>
<tr>
<th>Stage</th>
<th>Formation</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rising intonation on words/formulae</td>
<td>Four children?</td>
</tr>
<tr>
<td>2</td>
<td>Rising intonation on clauses</td>
<td>The boys throw the shoes?</td>
</tr>
<tr>
<td>3</td>
<td>Placing question word at front of clause without head movement; double verb marking</td>
<td>Is the picture has two planets on top?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Where the little children are?</td>
</tr>
<tr>
<td>4</td>
<td>Copula fronting and inversion after wh-questions</td>
<td>Is there fish in the water?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Where is the sun?</td>
</tr>
<tr>
<td>5</td>
<td>Head movement of auxiliaries, modals, do-support</td>
<td>Can you tell me?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is the boy eating?</td>
</tr>
<tr>
<td>6</td>
<td>Non-movement in embedded questions</td>
<td>Can you tell me what the date is today?</td>
</tr>
</tbody>
</table>
Table 3: Analysis of oral question task: total and ratio

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th></th>
<th>Time 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>Range</td>
<td>Mean (SD)</td>
<td>Range</td>
</tr>
<tr>
<td>Question total</td>
<td>11.47 (4.62)</td>
<td>4-25</td>
<td>10.75 (4.96)</td>
<td>4-26</td>
</tr>
<tr>
<td>Question ratio</td>
<td>.27 (.13)</td>
<td>.13-.72</td>
<td>.30 (.14)</td>
<td>.11-.67</td>
</tr>
</tbody>
</table>
Table 4: Oral question task: copula, lexical complex and non-target

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Time 1 (Mean (SD))</th>
<th>Range</th>
<th>Time 2 (Mean (SD))</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td></td>
<td>Mean (SD)</td>
<td></td>
</tr>
<tr>
<td>Copula questions</td>
<td>8.35 (3.33)</td>
<td>2-17</td>
<td>7.63 (3.71)</td>
<td>1-18</td>
</tr>
<tr>
<td>Lexical questions</td>
<td>2.88 (2.03)</td>
<td>0-8</td>
<td>2.88 (2.47)</td>
<td>0-11</td>
</tr>
<tr>
<td>Complex questions</td>
<td>2.48 (1.92)</td>
<td>0-7</td>
<td>1.66 (1.57)</td>
<td>0-5</td>
</tr>
<tr>
<td>Non-target forms</td>
<td>5.68 (3.08)</td>
<td>0-12</td>
<td>5.53 (3.59)</td>
<td>0-15</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------</td>
<td>-------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>Omission Time 1</td>
<td>2.29</td>
<td>1.40</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Omission Time 2</td>
<td>2.65</td>
<td>2.19</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>In-situ Time 1</td>
<td>2.14</td>
<td>2.11</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>In-situ Time 2</td>
<td>2.24</td>
<td>1.30</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Oversuppliance Time 1</td>
<td>1.55</td>
<td>0.89</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Oversuppliance Time 2</td>
<td>1.20</td>
<td>0.56</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Be instead of do Time 1</td>
<td>1.96</td>
<td>1.22</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Be instead of do Time 2</td>
<td>1.89</td>
<td>0.68</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Do instead of be Time 1</td>
<td>1.75</td>
<td>0.89</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Do instead of be Time 2</td>
<td>1.18</td>
<td>0.40</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Number Time 1</td>
<td>2.50</td>
<td>1.64</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Number Time 2</td>
<td>1.38</td>
<td>0.52</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 6: Analysis of fluency measures

<table>
<thead>
<tr>
<th></th>
<th>Time 1 Mean (SD)</th>
<th>Range</th>
<th>Time 2 Mean (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type-token ratio</td>
<td>.41 (.05)</td>
<td>.33-.51</td>
<td>.435 (.05)</td>
<td>.33-.56</td>
</tr>
<tr>
<td>Hesitancy</td>
<td>54.09 (24.38)</td>
<td>8-129</td>
<td>42.38 (15.31)</td>
<td>10-78</td>
</tr>
</tbody>
</table>
Table 7: WM scores by task (mean and range)

<table>
<thead>
<tr>
<th>Task</th>
<th>Mean (SD)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening Span</td>
<td>0.77 (0.16)</td>
<td>0.25-1</td>
</tr>
<tr>
<td>Story Recall L1</td>
<td>0.63 (0.15)</td>
<td>0.32-0.86</td>
</tr>
<tr>
<td>Story Recall L2</td>
<td>0.41 (0.16)</td>
<td>0.04-0.74</td>
</tr>
</tbody>
</table>
Figures

Figure 1: Comparison of nontarget patterns (Time 1 and Time 2)