Factors Affecting Chinese-L2 Learners' Use of Classifiers

Jiahuan Zhang
University of Cambridge

Abstract. This study explores a number of factors influencing Chinese L2 learners' classifier acquisition: L1, task type, and classifier type. I developed a picture-based description test, including composition, free cloze, and multiple-choice cloze questions to elicit the use of classifiers. Participants were 50 Chinese L2 learners from Arabic, English, and Japanese L1 backgrounds. An analysis of potential predictors suggested that, although L1 is not a significant predictor of test performance, Japanese L1 participants performed numerically better. Certain tasks (composition) are conducive to use of test-taking strategies and reveal a higher classifier accuracy to more constrained tasks (multiple-choice). Type of task was found to interact with L1. Results also showed a universal path of acquisition of classifier types, with shape classifiers receiving higher scores.

Keywords: Chinese classifiers, second language acquisition, L1 transfer, task effect, hierarchical development

---

1 Introduction

A numeral classifier is an obligatory grammatical unit within a numeral noun phrase in classifier languages (Allan, 1977). Although ill-formed grammar may not hinder speech comprehension, semantics becomes vague if the compulsory classifier is eliminated. For example, "one person" is *yi* (one) *ge* (CL) *ren* (people) in Chinese, where classifier *ge* is a compulsory element. The expression of *yi* (one) *ren* (people) is not clear given an absence of the classifier *ge* even with the numeral "one" present. This is because, conventionally, the word "person" can be matched to multiple classifiers/quantifiers such as *yi kou ren* "one person" (only used in census); *yi qun ren* "a bunch of people"; *yi dui ren* "many people"; *yi che ren* "a car of people". Notably, 40% of frequently used nouns can be matched to the generic classifier *ge*, and because of this, young Chinese first language (L1) acquirers and adult Chinese as a second language (L2) learners often overuse *ge* in natural contexts (Erbaugh, 1986; Liang, 2009; Polio, 1994; Zhang, 2007).

The past two decades has witnessed a growing interest in the investigation of Chinese classifier acquisition in the L2, with an increasing amount of research being focused on L1 effects. The research on crosslinguistic influence aims to determine how previous language learning affects the acquisition of an additional L2, which is a useful approach for predicting errors in L2 production (e.g., Gass & Mackey, 2013; Odlin, 1989). The L1 is argued to impact on all linguistic levels, including phonology, lexis, syntax, semantics, and discourse (e.g., Chan & Leung, 2020; Cheong et al., 2019; De Vincenzi & Lombardo, 2000; Jarvis & Odlin, 2000; Jarvis & Pavlenko, 2008; Odlin, 1989; Pienemann, 2005; Quesada & Lozano, 2020). An alternative consideration is that difficulties in L2 learning can be predicted by a hierarchical development of learning whereby L1 effects are minimised (e.g., Krashen, 1983; Pienemann, 1998; 2005). From a longitudinal perspective, learners are predicted to acquire linguistic knowledge involving morphemes and syntactic structures following a universal route (e.g., Dyson, 2009).

A relative lack of research studying L1 effects on L2 acquisition of Chinese classifiers means that the impact of L1 on L2 classifier acquisition remains unclear. Methodologically, existing studies mostly focus on comparisons of a single group of participants whose L1 is a non-classifier language to that of classifier languages, which limits the possibility of comparison between speakers of different non-classifier languages (cf. Liang, 2009; Polio, 1994; Zhang & Lu, 2013). Existing studies indicate quite mixed results, which could be explained by the types of tasks employed; with most using a single task, the effect of task is poorly understood (cf. Polio, 1994; Zhang & Lu, 2013). Therefore, it remains unclear whether speakers of different non-classifier languages would vary in their acquisition of classifiers, and whether using different task types would affect the interpretation of classifier proficiency. Additionally, the potential order of L2 classifier acquisition remains underexplored (cf. Zhang & Lu, 2013).

Therefore, the aim of the present study is to contribute to and expand on existing research on the L1 effects, task, and hierarchical development of acquisition associated with the L2 use of Chinese classifiers. I use a specially designed test with three written tasks completed by three L1 groups to perform a fine-grained analysis of classifier production. As a consequence, I obtain a fuller description and understanding of the use of classifiers than that achieved in previous studies.

2 Literature review

2.1 Classifiers

Classifiers are absent or marginal in the majority of European languages, and these languages are representative non-classifier languages (Allan, 1977). In non-classifier languages such as English, only mass nouns require a unit of quantification, and there are no classifiers as a distinct part of speech (Saalbach & Imai, 2012; Shi, 2014). Instead, one of the articles *as, or the* is utilized ahead of the noun when an individual object is counted. Yet, there is "…an open class of words that are functionally similar to classifiers" (Lehrer, 1986, p. 109). If needed, a unique format "a+N1+of+N2" is used to express collective meanings when an uncountable object is counted (e.g., 'a cup of water').

On the other hand, many Asian languages are typical classifier languages, including Japanese and Chinese (see detailed description of classifier systems in Downing, 2002; Kuo & Sera, 2009; Zhang, 2007). Classifiers often categorize nouns based on their characteristics, with categorization being different across languages. There are over 900 classifiers in Chinese; and a Chinese numeral classifier is structurally obligatory within a noun phrase when the head noun is quantified, as demonstrated in (1) (Zhang, 2007). The sortal Chinese classifier *tiao* (which refers to long, thin, flexible objects) must be inserted to convey the meaning of "two ropes".

(1) liang *tiao* shengzi
two CL rope
'Two ropes'

The Chinese classifier system contains a fine-grained categorization of having shape-, animate-, and inanimate-based types of classifiers (cf. Gao & Malt, 2009). Aside from these concrete uses of classifiers, concept-typed classifiers are equally important in Chinese, which refer to the extended use of classifiers and their metaphorical function in discourse (Aikhenvald, 2003; Littlemore, 2009).

2.2 L1 transfer

Positive transfer occurs when features of the L1 and the L2 match, such that L2 acquisition is facilitated. Negative transfer (also called 'L1 interference'), on the other hand, hinders the L2 acquisition because the L1 and L2 differ in particular respects. Accordingly, classifier language speakers are anticipated to outperform non-classifier language speakers in learning Chinese classifiers. This is because non-classifier language speakers are unfamiliar with the use of classifiers owing to the lack of a corresponding category in their L1, while classifier language speakers are already familiar with the use of classifiers before learning Chinese. Evidence can be found in Liang's (2009) study, which reported that Korean L1 speakers outperformed English L1 speakers in using Chinese classifiers. The prior knowledge of using classifiers facilitates the Korean speakers' classifier learning. Likewise, in my pilot study (Zhang, 2019), the Thai L1 participant showcased a superior mastery of Chinese classifiers compared to the English L1 counterpart in a dynamic assessment writing task, although these two participants were both intermediate learners of Chinese.

On the other hand, L1-L2 similarities can impede L2 learning on the premise that similarities in languages create confusion (Andersen, 1984). Beginning L2 learners often rely on "one meaning-one-form mapping" between L1 and L2, which would easily result in L2 errors (Andersen, 1984). Tang (2005) examined L2 Chinese learners' classifier production by comparing multiple written tasks (exams, homework, and papers) across L1. It was noticed that Korean learners made some common mistakes in using classifiers, such as overuse of *ge* and incorrectly using classifiers with the same pronunciation or similar meaning to Korean classifiers. One potential explanation for this is that there are similarities and differences between the two classifier systems and Korean learners' reliance on their L1 knowledge may result in L2 mistakes where there are differences. In other words, learners are accustomed to mapping the functional or semantic similarities of L1 items onto the L2, particularly at the very early stages, which would often incur unwanted ramifications (cf. Jarvis & Odlin, 2000; Ringbom & Jarvis, 2009).

2.3 Hierarchical development of morphosyntax

Although L1 effects related accounts illustrate the reasons for some common types of errors in L2, they may depict an incomplete picture of how learners acquire an L2 from a longitudinal perspective. In this regard, one robust explanation is that the universal developmental sequence plays an essential role in both L1 and L2 learning, hence L1 transfer interacts with this developmental sequence but does not override it (Ortega, 2014). The natural order hypothesis was one of the first accounts to theorize language developmental patterns. It hypothesizes that L1 is acquired in a predictable order that is not determined by formal simplicity or influenced by instruction (Krashen, 1983). The observation of the development of individual morphemes such as pronouns and syntactic structures such as negation presents strong support for the existence of the developmental sequence (Ellis, 1994). Erbaugh's (1986) is one of the most representative L1 acquisition studies in this strand. It is a longitudinal study that recorded three young Chinese children's acquisition of classifiers in a naturalistic conversational context. Findings indicated that children might follow a trajectory of learning classifiers, especially those classifiers related to shape were earliest mastered and frequently produced from an early age; and the generic classifier *ge* can be a substitute for virtually any other classifier without causing misunderstanding in daily conversations (see also Jiang, 2017). The overgeneralization of *ge* has since been widely observed in the L2 empirical study of classifier

learning (e.g., Hu, 1994; Zhang & Lu, 2013). Ellis (1994) believes that the L1 acquisition sequence in general offers a reference point for predicting L2 acquisition sequences, while an important issue of whether L1 and L2 acquisition are exactly the same or whether there are differences remains.

The universal developmental sequence can also account for certain errors occurred in the process of L2 acquisition (Ellis, 1994). 'Processability theory', proposed and refined by Pienemann (1998), explicates the hierarchical development of morphosyntax. It posits that learners pass through six distinct developmental stages when acquiring English L1 morphosyntax, and that the same is true for both L1 and L2 learning of other languages (Pienemann, 2005). A representative illustration is Hyltenstam's (1977) investigation of L2 acquisition of Swedish negation. Results showed that the acquisition of grammatical structure is a dynamic and productive process, and learners with different language backgrounds follow the same path of acquisition as well as error production. Hence, it is argued that L2 learners potentially conform to the universal route of acquisition, in which L1 background is overshaded.

With regard to classifier acquisition, the 'numeral classifiers accessibility hierarchy' (NCAH) hypothesizes an ordered learning for different types of classifiers (Craig, 1986). That is, classifiers of 'animate human', 'animate non-human', 'shape', and 'function' type are expected to be acquired in succession. The least marked distinction (i.e., animate human) is expected to appear earliest and be retained longest in learners' acquisition, whereas the most marked distinction (i.e., function) is expected to appear last and be easiest to lose after the onset of attrition. Even though the NCAH has been formulated on the basis of a comparison of limited classifier types, a body of empirical studies substantiated that animacy classifiers are acquired first, followed by classifiers denoting inanimacy, and then concrete objects with salient features such as shape in L1 acquisition of Kilivila, Hokkien, and Cantonese (e.g., Hu, 1994; Luke & Harrison, 1986; Senft, 1996). Furthermore, Hansen and Chen (2001, p. 84) noticed that English L1 speakers' acquisition of Japanese and Mandarin Chinese, in particular, conformed to the order predicted by the NCAH. While slightly different to the NCAH order, Aikhenvald (2003) states that shape-based classifiers are acquired relatively early by speakers of Mandarin Chinese, and that classifiers referring to non-extended objects are acquired earlier than classifiers that refer to extended objects (including classifiers referring to concepts). Altogether, these contrasting findings suggest a sequential acquisition of Chinese classifiers.

2.4 Empirical studies on Chinese L2 classifier acquisition

As discussed above, it remains unclear which factor (i.e., L1 or acquisition order) plays a more important role in L2 acquisition. The investigation of Chinese classifiers in L2 may help to clarify this, and it has been of particular interest for language acquisition in general during the past decades. Many empirical studies have examined L2 classifier acquisition and have offered mixed findings as to the effect of L1 (e.g., Liang, 2009; Paul & Gruter, 2016; Polio, 1994; Zhang & Lu, 2013). Polio (1994), for example, found no significant L1 effects. Polio's (1994) study investigated L2 use of Chinese classifiers from 21 English-L1 speakers and 21 Japanese.

L1 speakers. Both groups of participants were intermediate Chinese-L2 learners, required to narrate a story for Chinese native speakers after watching a film. Findings suggested that: 1) no clear L1 effects were found in either group in terms of classifier acquisition; 2) L2 speakers did not avoid using classifiers in most cases; 3) both groups tended to overuse the generic classifier *ge*. This study offered valuable insights into classifier acquisition at a specific stage and underlying difficulties of learning classifiers from a crosslinguistic perspective. Yet, the data come from one single implicit task, and the task effects remain

unclear (cf. Quesada & Lozano, 2020). Participants could avoid difficult classifiers and just use those they were very confident with in their story-telling task (cf. Ellis, 1994).

Much contrary to Polio (1994), Liang (2009) found an L1 effect with Korean participants outperforming English participants in their use of Chinese classifiers overall. Methodologically, Liang (2009) improved on Polio's (1994) study by elaborating on learner proficiency and making a distinction between novice, intermediate and advanced levels, which helped to elucidate the effect of proficiency. It was not surprising that participants with a higher proficiency level outperformed the lower proficiency group within each L1 cohort. Equally importantly, Liang (2009) compared the results from the comprehension and production tasks, which demonstrated the potential task effects and may be the reason for the differing L1 effect finding. To be specific, intermediate English participants outperformed their Korean counterparts in the comprehension task, but not in the production task.

The above-mentioned studies offered a snapshot of classifier acquisition. Zhang and Lu (2013) administered a research study centring on adult Chinese-L2 learners' developmental acquisition of classifiers. Data were collected from a corpus of 657 essays written by L2 speakers (the majority of which were English-L1 speakers) in the same Chinese language classes throughout two consecutive academic semesters. One essential finding is the change in diversity of classifiers. Specifically, token (the number of classifiers that were used) frequency declined while type frequency increased. This supports the probability of a sequential mastery of different classifier types, although the authors had different research priorities and did not elaborate the detailed development of classifier types in their research. Compared to most other empirical studies, the data collection method (i.e., a composition test) had the merit of being situated in a naturalistic setting, as the participants were not told that the research focused on classifier acquisition (cf. Polio, 1994; Quesada & Lozano, 2020). Additionally, the study focused on written production instead of spoken communication. However, purely relying on classifier analysis from a composition task may problematize the interpretation of proficiency. Similarly, to the story narrative conducted by Polio (1994), participants could choose to use classifiers or not as per their individual preference when no instructions on the use of classifiers were provided.

Taken together, existing empirical studies find quite mixed effects of L1 transfer in Chinese L2 classifier use. On top of that, none of these studies examined speakers of more than one non-classifier language in a single study. It remains unclear whether different groups of non-classifier language speakers would perform differently in applying classifiers under the same conditions. At the methodological level, the majority of previous studies employed a single task (i.e., written composition, story narrative, or daily record) for data collection, which may incur inaccurate and incomplete interpretations (Ellis, 1994; Ellis & Barkhuizen, 2005; Johnstone, 2000). Additionally, limited studies have ascertained the existence of a developmental sequence for L2 classifier learning. Thus, this study aims to address the following research questions:

(1) Does L1 affect classifier acquisition?
(2) Does task type affect test performance in relation to classifiers?
(3) Is there a predictable hierarchical development of different types of classifiers?

3 Method

3.1 Participants

Participants (*n* = 50, female = 34) came from two language backgrounds in relation to classifiers: classifier L1 (Japanese) and non-classifier L1 (Arabic and English). All participants were taking Chinese language courses at their home universities in Australia, Egypt, and Japan: 17 Egyptian Arabic-L1 speakers (age 18-26, mean = 21.18, SD = 3.09), 15 Australian English L1 speakers (age 20-30, mean = 23.67; SD = 4.69), and 18 Japanese-L1 speakers (age 20-22; mean = 20.11, SD = 0.58). To ensure their comparability in terms of Chinese proficiency, I only recruited students enrolled in intermediate Chinese language classes, using Chinese textbooks pertaining to intermediate level, equivalent to HSK3 ('Haiyu shuiping kaoshi', Chinese Proficiency Test 3); and all of them rated themselves as intermediate learners in the demographic questionnaire (cf. Chan & Leung, 2020; Suzuki & Sunada, 2020).

3.2 Materials

There were 3 written tasks based on a description of the same picture: a composition, a free cloze test, and a multiple-choice cloze test, all intended to elicit classifiers (see Appendix). To provide a prompt for the three writing tasks, a picture (Figure 1) with many objects was designed by the first author in her pilot study. Before data collection with L2 speakers, I ran a norming study with four native Chinese speakers (aged 22-24, university students). They were invited to write a composition based on the same picture; these examples were used to develop the cloze tests. Table 1 lists all of the classifiers along with their corresponding head nouns that appear in the intermediate-level textbook 'New Practical Chinese Reader' (Liu, 2011) and that set as prompts in the test. These classifiers represent different types, including shape, animate, inanimate, and concept. Some classifiers not found in the textbooks could be elicited by the picture, such as the classifier *shan* for windows. Such classifiers were not included in the prompts to tasks 2 and 3, but any correct use of classifiers in task 1 could get a score.

Table 1: *Classifiers and types.*

Question	Classifier-object	Type
1	ge-people	animacy
2	dui-couple	concept
3	zhang-sofa	shape
4	zhi-cat	animacy
5	bei-wine	shape
6	tiao-dog	animacy
7	ge-cupboard	inanimacy
8	ke-tree	inanimacy
9	fu-calligraphy	inanimacy
10	zhang-table	shape
11	ge-vase	inanimacy
12	ge-goldfish bowl	inanimacy
13	tiao-goldfish	animacy
14	ge-home	concept
15	ge-moment	concept

Figure 1: *Picture used in the test.*

The three tasks were constructed following 'funnel' principle to narrow down the test scope. Task 1 was a short descriptive composition, instructed a compulsory word-length of 150-200 characters. It was aimed at capturing learners' naturalistic use of classifiers. Task 2 was a free cloze task intended to examine noticing of the compulsory application of classifiers in specific context. Note that task 2 accepted both formal and informal use of classifiers, which may result in a 'one-to-more mapping' in each question. The multiple-choice cloze task, task 3, aimed to ascertain accuracy in selecting classifiers, where only one formal use of the classifier was considered as correct in each question. On the whole, the research design allows data triangulation from multiple sources and thus can offer a more accurate interpretation of participant accuracy (cf. Revesz et al., 2019 for triangulation of data analysis).

3.3 Data collection

All the participants took part in one individual session in a language teaching classroom at their home university. The test was a traditional paper-and-pen test. The term 'classifier' was not used in any of the task instructions, so that participants remained unaware of the language feature that was being examined. This was to ensure a natural production of classifiers and to avoid confirmation bias. The test was not timed, but the majority finished it in around 45 minutes. Participants were prohibited to refer back to previous tasks or to consult a dictionary at any time. They could not progress to the next task until finishing the previous one either. The procedure of data collection was reviewed and approved by an (anonymised) Human Research Ethics Committee (protocol 2019/167).

3.4 Data coding and statistical analysis

Accuracy in the use of classifiers was scored as 1 or 0 for each test item. Multiple answers could be marked as correct for each head noun in tasks 1 and 2, depending on context; on the other hand, only one correct answer was accepted for each question in task 3. Considering the possible unbalanced use of classifiers (some participants wrote longer compositions), each participant was given an overall percentage score of correct use of classifiers for each task, instead of a raw correct score.

To determine the effects of participant L1 and task on the accuracy of classifier use, a linear mixed-effects regression model was fitted to the data. An interaction between task and participant L1 was used as a predictor, and participant was included as a random effect; the formula is: lmer(score ~ task*L1+(1|participant). To further examine whether there is a developmental acquisition order, a logistic mixed-effects model with L1 and classifier type as fixed effects was used to predict the scores; formula: glmer(score ~ L1*type+(1|participant) +(1|classifer). Here the random effects included individual participants and classifiers. For this model, I only consider the score from task 3 as the response variable, because there is only one correct answer assigned for each test item. The major difference between linear and logistic mixed-effects models lies in the dependent variable, with the former using score accuracy percentage and the latter employing binary accuracy response. Both models were computed using the lmerTest package in R (Kuznetsova et al., 2017; R Core Team, 2020). As the sample size was relatively small, I performed bootstrapping to validate both of the models (see Levshina, 2015 for bootstrapping in linguistic studies). The bootstrapping method suggested a stable 95% percentile type of confidence interval (0.112 - 0.351) for R^2 in the linear model and an optimism slope of 0.020 for the logistic model, both of which validated the statistical modelling.

4 Results

4.1 Classifier accuracy across L1s and tasks

Table 2 presents ranges, means and SDs of test accuracy rates in all nine (3*3) conditions. Japanese L1 participants scored highest overall with a mean of 71.94 (SD =12.47, range 42.33 - 88.67), followed by English L1 participants (mean 66.51, SD =14.65, range 33.33-89.00), and Arabic L1 participants (mean 59.35, SD = 20.05, range 30.00-93.33). In addition, each L1 group exhibited a decreasing trend from task 1 to task 3, which means that each cohort of speakers scored highest in task 1 and lowest in task 3. Descriptive analysis suggests that there are both L1 and task effects. Statistical analysis can confirm whether the observed numerical differences are in fact significant.

Table 2: *Descriptive statistics for classifier accuracy.*

L1	Task	Range	Mean	SD
Arabic	1	0.00-100.00	70.47	28.28
(N = 17)	2	20.00-87.00	54.76	21.31
	3	20.00-93.00	53.82	21.52
English	1	33.00-100.00	72.87	16.71
(N = 15)	2	27.00-93.00	69.27	17.91
	3	27.00-80.00	57.40	13.73
Japanese	1	50.00-100.00	78.56	14.91
(N = 18)	2	27.00-93.00	71.39	17.26
	3	33.00-93.00	65.89	16.09

The linear model is summarized in Table 3. To elaborate, the ‘estimate’ and the ‘standard error’ columns show the predicted score and standard error for a level, respectively. For the base level, i.e., intercept (Arabic L1 speakers performing task 1), the predicted score is 70.471. To calculate the predicted score for a different level, the respective value in the ‘estimate’ column is added or subtracted. For instance, compared to task 1, the Arabic L1 participants received a score of 16.706 lower in task 2 and 16.647 lower in task 3. These differences were significantly different ($p < 0.001$ and $p < 0.001$) as indicated in the ‘Significance’ column. This means that the scores of the Arabic participants were significantly different between tasks.

Table 3: *Summary for the linear model (Arabic L1 and task 1 as intercept).*

	Estimate	Standard error	df	t-value	Pr(>|t|)	Significance
(Intercept)	70.471	4.647	87.778	15.166	0.000	***
task2	-16.706	4.405	94.000	-3.792	0.000	***
task3	-16.647	4.405	94.000	-3.779	0.000	***
L1English	2.396	6.787	87.778	0.353	0.725	
L1Japanese	8.085	6.480	87.778	1.248	0.215	
task2: L1 English	13.106	6.434	94.000	2.037	0.045	*
task3: L1 English	1.180	6.434	94.000	0.183	0.855	
task2: L1 Japanese	9.539	6.143	94.000	1.553	0.124	
task3: L1 Japanese	3.980	6.143	94.000	0.648	0.519	

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

At the same time, no significant differences were found between the Arabic, English, and Japanese participants in task 1, which means that all L1 groups performed comparably in task 1. However, there was a significant interaction between task and L1 ($p < 0.05$), such that the drop from task 1 to task 2 was less by 13.106 for English participants in comparison to Arabic ones. The interaction between task and L1 is plotted in Figure 2.

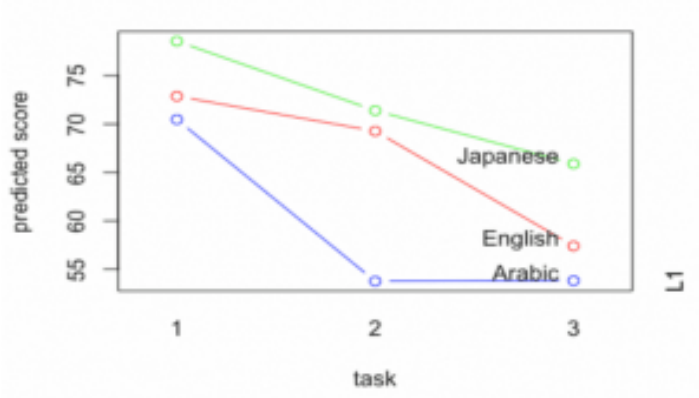

Figure 2: *Model prediction for scores of the three L1 groups in the three tasks.*

In an attempt to explain the difference between the two non-classifier language groups, I calculated the individual number of *ge* tokens in task 2, with results demonstrating that the Arabic, English, and Japanese L1 groups yielded an individual token number of 4.53, 6.13 and 4.88 in task 2 respectively, suggesting a higher usage of *ge* by the English participants. To ascertain whether the difference is significant, a Mann Whitney U Test was run. Results showed a significant difference ($p < 0.05$) between the Arabic and English participants in using *ge*.

4.2 Hierarchical development of classifier use

Table 5 shows the descriptive statistics for classifier types by L1 in task 3. Shape-type classifiers attract highest accuracy rate, followed by animacy, inanimacy, and concept. For all classifier types but shape, Japanese participants had highest accuracy score, followed by English and Arabic participants. For the shape type, however, Arabic participants were the most accurate. I run a logistic model to ascertain the significance of these effects.

Table 4: *Descriptive statistics for classifier types in task 3.*

Type	L1	Accuracy	Total
Animacy	Arabic	61.76	69.50
	English	71.67	
	Japanese	75.00	
Concept	Arabic	33.33	45.33
	English	44.44	
	Japanese	57.41	
Inanimacy	Arabic	43.53	51.60
	English	52.00	
	Japanese	58.89	
Shape	Arabic	80.39	72.67
	English	60.00	
	Japanese	75.93	

Our model suggests that neither the interaction nor L1 was a significant predictor, so I pruned the model to having type as the only fixed effect. Table 6 represents the final model. The type of concept was chosen as the reference level (the intercept). The 'estimate' column in the table represents the log odds of the dependent variable being one factor rather than the other. Positive values in the column mean a higher chance of scoring under a particular condition, while negative values mean a lower chance of scoring. The estimate for the shape type is positive at 1.617 and is significantly different from the estimate for the concept type ($p < 0.05$), as indicated in the 'significance' column. This means that classifiers of the shape type were significantly more likely to get correct scores in comparison to the concept type. There was a trend for animacy type to receive a higher score as well, but this difference did not reach significance at the traditional α level ($p = 0.063$).

Table 5: *Summary for the logistic model (concept type as intercept).*

	Estimate	Standard error	z-value	Pr(>|z|)	Significance
(Intercept)	-0.184	0.529	-0.347	0.729	
type.shape	1.617	0.747	2.164	0.030	*
type.animacy	1.278	0.686	1.862	0.063	.
type.inanimacy	0.635	0.687	0.924	0.356	

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

An effect plot (Figure 3) was made to visualise the predicted probabilities for type, with 95% error bars presented to indicate the uncertainty of the estimates. The differences in accuracy across classifier types suggest a potential acquisition order from a developmental perspective, according to which shape type classifiers are more likely to be acquired first, followed by animacy, inanimacy, and concept type classifiers.

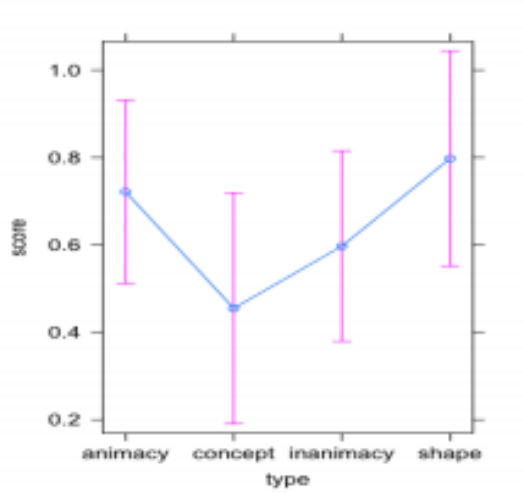

Figure 3: *Model prediction for classifier types.*

5 Discussion

5.1 L1 transfer

The descriptive statistics of the L1 effect suggest that similarity could facilitate learners' classifier acquisition. The Japanese participants scored highest numerically within each task and overall, in comparison to the English and Arabic participants. As Japanese is a typical classifier language, it exhibits frequent use of classifiers (Hansen & Chen, 2001). In contrast, Arabic and English are two non-classifier languages, so theoretically the Arabic and English participants are more likely to encounter difficulty in learning classifiers. This is in line with Liang's (2009) study of Chinese classifier acquisition, where Korean participants were found to outperform their English counterparts in producing classifiers. It also aligns well with me pilot study where the Thai participant scored much higher in using classifiers than the New Zealand participant in a short composition task.

However, statistical analysis of the data suggests that these differences did not always reach significance. While Arabic participants scored lower than Japanese in task 2, the difference did not reach significance in tasks 1 or 3 or between English and Japanese in any task. Similar conclusions can be found in another SLA research. For instance, Polio (1994) reported no clear L1 effect in the oral production of classifiers by comparing the data from the English and Japanese speakers. Liang (2009) observed a minimal difference between English and Japanese speakers' use of Chinese classifiers. That is, both groups showed a higher and correct use of shape-typed classifiers in the production task, while this varied in the other tasks

employed in his study. Hence, the current results are inconclusive in relation to L1 transfer effect and suggest that L1 effects may be modulated by task effects.

Taken together with previous research, my results suggest that L1 is not a significant predictor of classifier accuracy by itself, and classifier-L1 speakers do not outperform non-classifier-L1 speakers, at least at the intermediate level of proficiency. It may be that the L1 advantage is realised differently, perhaps, through the effort required to learn the feature or at beginner levels of proficiency.

5.2 Task effects

Task was found to be a significant predictor of test performance. Numerically, task 1 yielded the highest scores, followed by task 2 and task 3. Differences in the scores between tasks are reflections not of the different types of questions asked, but of the inherent difficulty of the tasks. From the perspective of language testing, it was not surprising that participants performed best in task 1, where there was no obligation for the use of classifiers and the score was calculated by accuracy percentage (cf. Douglas, 2014; Zhang & Lu, 2013). Participants could use the classifiers they were confident with and avoid using those they were less confident with, regardless of L1, resulting in no significant difference between the L1 groups. On the other hand, the use of classifiers became obligatory in task 2, so less space was left for the use of test-taking strategies (Downing, 2002). This task revealed a significant difference for the Arabic group. However, it becomes another story for task 3, which was intentionally designed to ascertain the correct and formal use of classifiers and there was only one fixed correct answer for each question (cf. Badger & Yan, 2012). This potentially required a superior command of classifiers in comparison to task 2 since there was no chance to choose *ge* as a placeholder without the option of *ge* being provided in the corresponding questions. The reason that participants still scored higher in task 2 than in task 3 could be the overuse of *ge* in task 2 (cf. Polio, 1994; Zhang & Lu, 2013).

In addition to the effect of task type *per se*, certain L1 groups may perform better in a specific task type than others, and test-taking strategies could affect the test performance in specific contexts (Douglas, 2014). Statistically, the linear model indicates a significant interaction between task and L1 with Arabic participants showing a substantial drop in accuracy from task 1 to task 2 and English participants from task 2 to 3. These findings could be explained in view of the combination of task nature, that the 'one-to-many mapping' allows the leeway for scoring; and test-taking strategies, namely, the overproduction of the generic classifier *ge* (Douglas, 2014). As the results demonstrated, the English participants were observed to use *ge* more frequently as a placeholder in the questions in task 2. Future research can further explore the effect of task-taking strategies.

5.3 Hierarchical development of classifiers

Results of the logistic model indicate no interaction between L1 and type, which implies a universally sequential mastery of different types, which aligns with no significant effect of L1 on accuracy. The descriptive statistics suggest a successive mastery of shape, animacy, inanimacy, and concept classifiers; and the model confirmed significant differences between shape and concept types. Accordingly, participants were more likely to get a correct score for the different types of classifiers as per the sequence predicted, and to score worse in the case of classifiers acquired later, which further implies that longitudinally the sequence could be a representation of a developmental path of L2 classifier learning. The results are in line with a longitudinal observation of L2 Chinese classifier acquisition by Zhang and Lu

(2013), who observed that the number of classifier types acquired increased over time. That is, L2 learners generally acquire different types of classifiers in succession. Although this study did not specify the sequence of classifier types, it can still, through the lens of increasing token and type frequency, support the assumption that the mastery of classifier types is developmental in nature. Moreover, Aikhenvald's (2003) study elucidates that Chinese-L1 speakers acquired shape-based classifiers earlier than the non-extended use of classifiers, which is congruent with the findings captured.

However, the findings of the current study are incongruent with the L1 acquisition of other classifier languages such as Kilivila, Hokkien or Cantonese, in which animacy classifiers are acquired first, followed by inanimacy classifiers, and then shape classifiers (e.g., Hu, 1994; 1994; Luke & Harrison, 1986; Senft, 1996). The findings also partially contradict the NCAH, which hypothesizes that classifiers of the animate type are to be acquired earlier than the shape typed ones (see also Craig, 1986). Comparison of all the classifier acquisition sequences above shows areas of agreement as well as disagreement. Debate mainly exists with respect to whether shape or animacy classifiers are acquired earlier. A consistent conclusion is that animacy classifiers are acquired relatively early, whereas concept classifiers (i.e., those that refer to abstract notions or extended meaning) are the last to be acquired.

5.4 Limitations and future research

Apart from discussing the results of the study, it is also necessary to recognise the limitations of the research. The most important one concerns the inhomogeneous distribution of the three participant groups. They were studying in three different countries, where language exposure, teaching style, or even cultural factors may affect their learning rate. In this study, this was necessary in order to recruit enough participants from different L1 groups. Future researchers are encouraged to collect data from one location to minimise the effect of such extraneous factors. Next, a gender bias clearly exists in the participant pool as the number of female participants (n = 34) was more than double that of the male participants (n = 16). Even though some studies have demonstrated that gender does not technically affect language learning outcomes (e.g., Gafni et al., 2017), future studies are recommended to recruit participants more evenly in terms of gender and other demographic variables. Moreover, the study did not take individual differences such as memory capacity and motivation to learn classifiers into account. Much research has indicated that memory capacity and motivation can lead to important influence on the rate of learning given the same learning situation and L1 (e.g., Marini, Eliseeva & Fabbro, 2016; Schuetze, 2015). However, due to the practicality and the limited scope of the current study, such data was not collected. Still, the three factors could be further explored in later research.

5.5 Conclusions

The current study focused on Chinese-L2 learners' production of classifiers by a specially designed language test and shed new light on how the performance was constrained by different factors (L1, task, and classifier type) that were not considered together in previous research. This study takes the position that L1 does not significantly affect the use of classifiers. Different types of tasks influence the classifier production because of the inherent difficulties of task setting and participants' test-taking strategies applied. Finally, the hierarchical development of classifiers is supported, with no L1 effects observed. The methodological implications of the study suggest that future research should carefully consider the pros and cons of using various types of tasks as an elicitation method and, perhaps, use several for triangulation

purposes. In the pedagogical practice, Chinese L2 language teachers may choose to introduce different types of classifiers in the order suggested here, in lieu of paying much attention to learners' L1, at least at the intermediate level.

6 References

Aikhenvald, A. (2003). Classifiers: A typology of noun categorization devices. Oxford: Oxford University Press.

Allan, K. (1977). Classifiers. Language, 53(2), 285-311.

Andersen, R. (Ed.). (1984). Second languages a cross-linguistic perspective. Rowley/London/Tokyo: Newbury House Publishers.

Badger, R., & Yan, X. (2012). The use of tactics and strategies by Chinese students in the Listening component of IELTS. Cambridge: Cambridge University Press.

Barkhuizen, G. (2005). *Analysing learner language.* Oxford: Oxford University Press.

Chan, R., & Leung, J. (2020). Why are lexical tones difficult to learn? Studies in Second Language Acquisition, 42(1), 1-27.

Cheong, C. M., Zhu, X., Li, G. Y., & Wen, H. (2019). Effects of intertextual processing on L2 integrated writing. Journal of Second Language Writing, 44, 63-75.

Craig, C. (Ed.). (1986). Noun classes and categorization. Amsterdam/Philadelphia: John Benjamins Publishing Company.

De Vincenzi, M., & Lombardo, V. (2000). Cross-linguistic perspectives on language processing. Boston: Kluwer Academic.

Douglas, D. (2014). *Understanding language testing*. New York: Routledge.

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, *7*(3), 235-241.

Dyson, B. (2009). Processability theory and the role of morphology in English as a second language development: A longitudinal study. *Second Language Research*, *25*(3), 355-376.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Ellis, R., & Gao, M. Y., & Malt, B. C. (2009). Mental representation and cognitive consequences of Chinese individual classifiers. Language and Cognitive Processes, 24(7-8), 1124-1179.

Erbaugh, M. S. (1986). Taking stock: The development of Chinese noun classifiers historically and in young children. In C. Craig (Ed.), Noun classes and categorization (pp. 399-436). Amsterdam/Philadelphia: John Benjamins Publishing.

Gafni, R., Achituv, D. B., & Rahmani, G. (2017). Learning foreign languages using mobile applications. *Journal of Information Technology Education Research*, 16, 301-317.

Gass, S. M., & Mackey, A. (Eds.). (2013). The Routledge handbook of second language acquisition. London: Routledge.

Hansen, L., & Chen, Y. L. (2001). What counts in the acquisition and attrition of numeral classifiers. *Jalt Journal, 23*(1), 83-100.

Hu, Q. (1994). Overextension of animacy in Chinese classifier acquisition. In E. V. Clark (Ed.), *Proceedings of the twenty-fifth annual child language research forum* (pp. 127-136). Stanford, USA: Stanford University Press.

Hyltenstam, K. (1977). Implicational patterns in interlanguage syntax variation. *Language Learning*, *27*(2), 383- 410.

Jarvis, S., & Odlin, T. (2000). Morphological type, spatial reference, and language transfer. *Studies in Second Language Acquisition*, *22*(4), 535-556.

Jarvis, S., & Pavlenko, A. (2008). Crosslinguistic influence in language and cognition. London: Routledge. Jiang, S. (2017). The Semantics of Chinese Classifiers and Linguistic Relativity. Abingdon: Routledge.

Johnstone, B. (2000). Qualitative methods in sociolinguistics. Oxford: Oxford University Press. Krashen, S. (1983). Principles and practice in second language acquisition. TESOL Quarterly, 17(2), 300-305.

Kuo, J. Y. C. & Sera, M. D. (2009). Classifier effects on human categorization: The role of shape classifiers in Mandarin Chinese. Journal of East Asian Linguistics, 18(1), 1-19.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1-26.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. Amsterdam: John Benjamins Publishing.

Lehrer, A. (1986). English classifier constructions. Lingua, 68(2), 109-148. doi:10.1016/0024-3841(86)90001-X.

Liang, S. Y. (2009). The Acquisition of Chinese Nominal Classifier Systems by L2 Adult Learners. (Unpublished Doctoral dissertation). University of Texas Arlington, USA.

Liu, X. (2011). *New practical Chinese reader.* Beijing: Beijing Language and Cultural University Press.

Littlemore, J. (2009). *Applying cognitive linguistics to second language learning and teaching*. London: Palgrave Macmillan.

Luke, K. K., & Harrison, G. (1986). 'Young Children's Use of Chinese (Cantonese and Mandarin) Sortal Classifiers'. In H. S. Kao & R. Hoosain (Eds.), *Linguistics, Psychology and the Chinese Language* (pp. 125-147). Hong Kong: University of Hong Kong.

Marini, A., Eliseeva, N., & Fabbro, F. (2016). Impact of early second-language acquisition on the development of first language and verbal short-term and working memory. *International Journal of Bilingual Education and Bilingualism, 22*(2), 1-12. doi:10.1080/13670050.2016.1238865

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge: Cambridge University Press.

Ortega, L. (2014). *Understanding second language acquisition*. London: Routledge.

Paul, Jing Z, & Grüter, Theres. (2016). Blocking Effects in the Learning of Chinese Classifiers. *Language Learning*, *66*(4), 972-999.

Pienemann, M. (1998). *Language processing and second language development: Processability theory (Vol. 15).* Amsterdam: John Benjamins Publishing.

Pienemann, M. (Ed.). (2005). *Cross-linguistic aspects of Processability Theory (Vol. 30).* Amsterdam: John Benjamins Publishing.

Polio, C. (1994). Non-native speakers' use of nominal classifiers in Mandarin Chinese. *Journal of the Chinese Language Teachers Association, 29*(3), 51-66.

Quesada, T., & Lozano, C. (2020). Which factors determine the choice of referential expressions in L2 discourse? Studies in Second Language Acquisition, 42(3), 1-28. doi:10.1017/S0272263120000224.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Australia. Retrieved from https://www.r-project.org

Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviors: A mixed methods study. Studies in Second Language Acquisition, 41(3), 605-631.

Ringbom, H., & Jarvis, S. (2009). The importance of crosslinguistic similarity in foreign language learning. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 106-118). Hoboken: Blackwell Publishing.

Saalbach, H., & Imai, M. (2012). The relation between linguistic categories and cognition: The case of numeral classifiers. *Language and Cognitive Processes*, *27*(3), 381-428.

Schuetze, U. (2015). Spacing techniques in second language vocabulary acquisition: Short-term gains vs. long term memory. Language Teaching Research, 19(1), 28-42. doi:10.1177/1362168814541726

Senft, G. (1996). Classificatory Particles in Kilivila. New York: Oxford University Press.

Shi, Y. (2014). Comparison of individual classifiers and collective classifiers between Chinese and English. Theory and Practice in Language Studies, 4(9), 1961-1965. doi:10.4304/tpls.4.9.

Suzuki, Y., & Sunada, M. (2020). Dynamic interplay between practice type and practice schedule in a second language: The potential and limits of skill transfer and practice schedule. *Studies in Second Language Acquisition, 42*(1), 169-197. doi:10.1017/S0272263119000470.

Tang, C. J. (2005). Nouns or Classifiers: A Non-Movement Analysis of Classifiers in Chinese. *Language and Linguistics (Taipei)*, *6*(3), 431.

Zhang, J., & Lu, X. (2013). Variability in Chinese as a foreign language learners' development of the Chinese numeral classifier system. *The Modern Language Journal*, *97*(S1), 46-60.

Zhang, J. (2019). Crosslinguistic influence on Chinese-L2 learners' acquisition of classifiers. *ANU Undergraduate Research Journal*, *9*, 156-171. Retrieved from http://studentjournals.anu.edu.au/index.php/aurj/article/view/138

Zhang, H. (2007). Numeral classifiers in Mandarin Chinese. *Journal of East Asian Linguistics, 16*(1), 43-59. doi:10.1007/s10831-006-9006-9.

7 Appendix

Materials: Test of Classifiers

Task 1: short composition
Please describe what you see in the picture with as many details as possible (150-200 characters).

Task 2: gap-filling
Please fill the blanks based on the picture (Note: only one character is acceptable for each blank).

照片里有两 1._人，他**们应该**是一 2._夫妻。他**们**坐在一 3._**黄**色的沙**发**上，女生抱着 一 4._橘**黄**色的小猫，男生手里**举**着一 5._**红**酒。他**们**看起来很幸福。旁**边**有一 6._白 色的狗，也笑得很开心。窗外的**风**景很好，看得**见蓝**天，白云和几**栋**大楼。窗台下面

有一 7._柜子，上面摆了一些绿色的小盆栽。柜子的旁边还有一 8._树。 后面的墙壁 上有一 9._大大的倒过来的“福” 字，红红的，很喜庆，渲染着过年的气氛。靠着墙壁 还放了一 10._桌子，上面有一 11._金鱼缸和一 12._花瓶。花瓶里插着几枝梅花，十 分优雅。金鱼缸里面有两 13._可爱的小金鱼，它们在水里快乐地游来游去。这应该是
一 14._很温暖的家，这家人正在享受着一 15._幸福的时刻。

Task 3: multiple-choice
Please choose the correct answer for each blank based on the picture.

照片里有两 1._人，他们应该是一 2._夫妻。他们坐在一 3._黄色的沙发上，女生抱着 一 4._橘黄色的小猫，男生手里举着一 5._红酒。他们看起来很幸福。旁边有一 6._白 色的大狗，也笑得很开心。窗外的风景很好，看得见蓝天，白云和几栋大楼。窗台下 面有一 7._柜子，上面摆了一些绿色的小盆栽。柜子的旁边还有一 8. 树。 后面的墙
壁上有一 9._大大的倒过来的“福” 字，红红的，很喜庆，渲染着过年的气氛。靠着墙 壁还放了一 10._桌子，上面有一 11._金鱼缸和一 12._花瓶。花瓶里插着几枝梅花， 十分优雅。金鱼缸里面有两 13._可爱的小金鱼，它们在水里快乐地游来游去。这应该 是一 14._很温暖的家，这家人正在享受着一 15._幸福的时刻。

1. A. 组 B. 只 C. 个 D. 位
2. A. 个 B. 位 C. 对 D. 双
3. A. 条 B. 张 C. 段 D. 面
4. A. 匹 B. 个 C. 头 D. 只
5. A. 壶 B. 杯 C. 个 D. 瓶
6. A. 个 B. 条 C. 匹 D. 头
7. A. 台 B. 面 C. 条 D. 个
8. A. 根 B. 棵 C. 行 D. 个
9. A. 页 B. 幅 C. 面 D. 本
10. A. 台 B. 面 C. 张 D. 条
11. A. 只 B. 个 C. 盆 D. 杯
12. A. 个 B. 盆 C. 支 D. 瓶
13. A. 个 B. 匹 C. 头 D. 条
14. A. 所 B. 间 C. 个 D. 座

15. A. 分 B. 点 C. 个 D. 门