

The Influence of a Semester Abroad on Reading Proficiency: A Descriptive Study

Jeffrey J. Hayden
University of Hawaii at Manoa

Introduction

Little has been written documenting the actual effects of study abroad in China on a student's Chinese language abilities, much less on a specific skill like reading comprehension. Although there are a few articles that look at the influence of study abroad for other languages (e.g., Brecht and Davidson 1992; Brecht, Davidson, and Ginsberg 1993; Carlson et al 1990), they are mainly concerned with the foremost modes of communication, i.e., speaking and listening (in particular, see DeKeyser 1991, Freed 1990, Ginsberg 1992). The articles that have been published in the area of Chinese language studies merely hypothesize possible benefits of study abroad (most recently, for example, see Kubler 1997).

A period of immersion abroad is evidently a natural catalyst for increases in overall language proficiency. Yet there is still little quantitative evidence regarding acquisition of reading skills. This study documents the semester-long progress of twenty-one foreign students studying Chinese at Beijing University. The first evaluation was taken in September 1997, and a second evaluation was made in January 1998.

Method

Participants

The participants for this study were volunteers (N = 21) from among the students studying at the Center for Teaching Chinese as a Foreign Language at Beijing University (Beida). I posted a flyer asking for volunteers. But this garnered few participants, and I had to go "door-to-door" to ask for participants and secure the final sample population. The majority (n = 12, 57%) were American students, and there were nine participants representing six other countries (seven participants from Eu-

European countries, one from Canada, and one from Mexico¹). The class levels of these participants ranged from Level 1 to Level 17.² There were fifteen males and six females ranging in age from 20 to 34 years, with a mean of 24 and a mode of 25. Ten of the participants were undergraduates, four were graduate students, and seven were currently not in any degree program but had applied to and intended to start graduate school the following fall.

This sample was taken from a foreign student population, exclusive of Japanese and Korean nationals, of approximately 200. As I wanted to observe the effects of overseas study of the Chinese language on people who were not familiar with the Chinese writing system (i.e., Chinese characters), Japanese and Korean students were not included in this study. Additionally, students who had formally studied either of these languages were intentionally kept to a minimum. Thus, only three of the twenty-one participants had studied Japanese: one participant for one semester, one participant for one year and another participant for about two years.

Finally, a caveat. Any statistics book should explain that, generally speaking and depending on the type of study being conducted, no fewer than thirty students should be obtained to achieve data significant enough to make reliable claims (see, especially, Brown 1988, or Hatch & Lazaraton 1991). Because of this, I merely offer this set of data as suggestive rather than conclusive. Any claims made in this paper are, at best, tentative.

¹ I was initially worried about these participants' English proficiency and their ability to perform the comprehension tasks I had planned, since they involved proficiency in English as well. The participants were informed that if they encountered anything they did not understand or were unclear about, they were free to ask for clarification. None of the participants asked any questions during either of the evaluations.

² The system at Beida divides classes into the following levels: Levels 1-8 (roughly equivalent to Beginning Chinese classes in the United States), Levels 9-14 (roughly equivalent to Intermediate Chinese), and Levels 15-20 (roughly equivalent to Advanced Chinese). Students at all levels are required to take 20 hours of class. Levels 1-8 attend three required sets of classes, namely, General Chinese (Hanyu, i.e., grammar and usage), Spoken Chinese (Kouyu), and Listening Comprehension (Tingli). Levels 9-14 attend two required sets of classes, namely Hanyu and Spoken Chinese, as well as an elective course, the topics of which vary from semester to semester and number from six to ten choices. Levels 15-20 attend Hanyu classes, as well as two elective courses which typically cover a wide range of academic specialties such as economics and political science, etc. Each semester at Beida is 16 weeks long. The learning period of this study was approximately 12 weeks long (base evaluation three to four weeks after the start of the semester in September, second evaluation one week before final examinations in January).

Materials

Computer Adaptive Test for Reading Chinese (CATRC)

The main instrument for evaluating the participants' reading ability was a computer program, the Computer Adaptive Test for Reading Chinese (CATRC). This test was designed and is maintained by Professor Tao-chung (Ted) Yao, now at the University of Hawaii.³ Because Beida's policy in placing students is to find a balance between a student's varying abilities,⁴ the particular levels and types of classes a student happened to be placed in at Beida only provided a general idea of a student's abilities. Therefore, the CATRC provided a more accurate picture of a student's general reading proficiency level for use in this study.

The CATRC is described in detail in Yao (1995:81), so I will only give a brief explanation of its operation here. Worth noting is that item response theory (IRT) was applied for item analysis and equating, thus reliably calibrating the test items with the abilities of examinees. The main advantage a computer-adaptive test (CAT) has over "paper-and-pencil" tests cited by Yao is that "because the computer [program] only asks the student to answer selected questions within the student's ability range, the CAT can be shorter than a conventional "paper-and-pencil" test." (p. 76) Thus, a CAT reduces the risk of undue fatigue on the examinee, concentrating on material within the examinee's ability range.

Yao's CATRC utilizes the HyperCard application on the Macintosh computer platform. There is a database that currently consists of about 450 test items. These items cover the four main proficiency levels (Novice, Intermediate, Advanced, and Superior), including all sublevels (i.e., Novice Low, Novice Mid, Novice High, etc.), as described in the ACTFL (American Council on the Teaching of Foreign Languages) Chinese Proficiency Guidelines (1987). Each level's set of items are stored in a HyperCard file called a "stack." The program randomly selects items from the different stacks to present to each examinee. When an examinee answers a question correctly, s/he is given an item of similar or greater difficulty; when an examinee answers a question incorrectly, the program then randomly selects an item of similar or lesser difficulty. In this way it closes in on and identifies an examinee's reading level.

³ I would like to thank Prof. Yao for providing me with the latest copy of the CATRC and allowing me to take it to China to use in this study.

⁴ As a simple example, let us say a student has a speaking level at Beida's Level 10 and a reading level at their Level 12. Beida might initially place that student in a Level 11 series of classes.

An example card is presented in Figure 1 below.

北京裕龍大酒店
(This is the name of a business establishment in Beijing)

What is the establishment?

- a. Department Store
- b. Hotel
- c. Wineshop
- d. Drugstore

Figure 1. A sample card from Prof. Tao-chung (Ted) Yao's Computer Adaptive Test for Reading Chinese.

The examinee uses the mouse to select the answer and clicks on it. This selection is recorded, and information such as the specific test item, the correct answer for the item and the examinee's answers are placed in a summary card for that particular examinee. There is no time limit to the evaluation, and an examinee may take as long as s/he needs to complete it. However, once the evaluation has begun, an examinee must continue until the program notifies her/him that the test is over. In other words, an examinee cannot start the test in the morning, quit the program, and then come back and finish it in the afternoon. Generally speaking, based on my experience in this study, examinees should be told that an hour to an hour and a half of time is necessary to take the test. Beginning level students may take less time, and advanced level students may take slightly longer.

There are many advantages to using such an evaluation tool. First and foremost is that it provides an immediate, if rough, assessment of a student's reading level. As soon as the student finishes and the program completes the summary processing, the examiner has most of the important information at her/his disposal. By comparison, the Chinese Proficiency Test (CPT) requires one to wait weeks before (to order a

copy of the test and set up a test site) and after its administration (to process the results), in addition to charging examinees a fee. It is also more restrictive with regard to its portability, especially in its usage outside the United States, where it is currently the test most relied on to officially assess reading and listening comprehension.

Second, the amount of storage space required to install the CATRC stacks onto any current model Macintosh computer is negligible. Provided the site of the study has Macintosh computers, installation of the stacks and scheduling of the students can be done in a matter of hours. One problem with doing such a study in mainland China utilizing the CATRC, however, is that educational institutions predominantly use IBM-PC computers. Because Beida is no exception, there are no Macintosh computer laboratories readily available. Additionally, there is no IBM-PC version of the CATRC available yet. For the purposes of this study, therefore, I brought and used my own Macintosh PowerBook 520c laptop computer.

One advantage of the CATRC not mentioned in Yao (1995) is that it can be repeatedly used with the same sample population. Unlike the CPT, the CATRC is programmed to be totally random. This means that for a longitudinal study of a group of participants, the test can be used at the beginning and end of an academic term or at the beginning, middle, and end of any academic year. And, as is shown below, the incidence of duplicate test items proved insignificant. This is important in that it does away with the "practice effect" concern mentioned in Brown (1988:35).

One possible disadvantage the CATRC has compared to the CPT, however, when considering doing statistical studies, is that each examinee of the CATRC takes an individualized version of the test. This is due to the random nature of item selection programmed into the CATRC. The selection and quantity of items and the amount of time to take the test will be different for each examinee. The implication of this is that the results of this evaluation cannot really be used for detailed comparison of one examinee with another examinee or within a group of examinees based on specific test items.

Informal Interviews

In addition to the CATRC, I interviewed the participants regarding their education in the Chinese language. This added qualitative information to the quantitative data results obtained by the CATRC. Questions asked included information regarding the classes at the participants' respective home institutions and their Chinese lan-

guage experiences at Beida. Since the students were placed in different class levels in the Beida program, it was important to elicit details regarding the materials and curriculum employed at the various levels. I also made several visits to classes in session to observe first-hand how and what students were taught.

Procedures

For the most part, participants were scheduled during times when they would be less distracted and more relaxed. Generally, this was in the afternoon or evening, or on days when they had no classes. I also tried to avoid scheduling close to their in-class test days. Evaluations were conducted in a well-lit private room free of interruptions and distractions.

The mean time used to work through the first administration of the CATRC was fifty-five minutes, with a minimum of eleven minutes and a maximum of two hours and forty-four minutes. For the second administration, the mean time was slightly longer at one hour and eight minutes, with the shortest time being sixteen minutes and the longest time being two hours and nineteen minutes.

Analyses

Descriptive statistics were generated to summarize the gathered data. A frequency distribution table was created to map the observed changes in evaluations. The ACTFL Guideline levels (both general and sublevel⁵) were converted to ordinal scales, and then frequency and means analyses (paired *t*-test) were run using the converted results of the first evaluation and the second evaluation.

Results

The initial, base evaluation resulted in one Novice, twelve Intermediate, six Advanced and two Superior readers. After the second evaluation at the end of the semester, the evaluation produced one Novice, seven Intermediate, five Advanced, and eight Superior readers. Table 1 below shows the distribution of these participants. For the main levels, there was a mean of 2.40 on the first evaluation and a mean of 3.00 on the second evaluation, with *s.d.s* of .75 and .97 respectively. At the sublevel,

⁵ General: Novice = 1, Intermediate = 2, Advanced = 3, Superior = 4. Sublevel: Novice Low = 1, Novice Mid = 2, Novice High = 3, Intermediate Low = 4, Intermediate Mid = 5, Intermediate High = 6, Advanced Plus = 8, Superior = 9.

however, there was a mean of 5.70 (Intermediate Mid) on the first evaluation and 7.05 (Advanced) on the second evaluation, with *s.d.s* of 1.89 and 2.09, respectively. The non-parametric matched *t*-test using the ordinalized scores was run where $t = 2.861$, $df = 19$, $p < .05$.

	(Second evaluation)								
	NM	NH	IL	IM	IH	A	AP	S	Totals
NM	1	0	0	0	0	0	0	0	1
NH		0	0	0	0	0	0	0	0
IL			0	4	0	0	1	0	5
IM				2	0	2	0	1	5
IH					0	0	1	1	2
A						0	0	3	3
AP				1*	0	1	0	1	3
S								2	2
Totals	1	0	0	7	0	3	2	8	21

Table 1. Observed frequencies of change between the first evaluation and the second evaluation. (*See note #7 below.)

As can be seen in Table 1 above, there were fourteen participants who showed improvement from the first to the second evaluation, representing 66.7% of the total participants. Five participants were ranked at the same level on both evaluations, representing 23.8%, and two participants indicated some degree of decline, representing 9.5% of the total. There was an average overall observed gain in reading proficiency of 1.35 sublevels (e.g., from Intermediate Low to Intermediate Mid, etc.)

Of the students who showed reading proficiency gains, the two greatest increases involved jumps from an Intermediate Low rating to an Advanced Plus rating and from an Intermediate Mid rating to a Superior rating. The qualitative data gathered on these participants is insufficient to form any conjecture regarding why the increases were so marked. (However, the reader is referred to the Discussion section of this paper to consider the possibility of language recuperation.) Of the participants who showed loss in reading proficiency, the greatest decrease involved a slide from an Advanced Plus rating down to an Intermediate Mid rating. The cause of this drop was evidently an emergency return home by the student due to illness resulting in a month of Chinese language inactivity, as well as physical and emotional hardship.

As noted above, four participants showed no gain in reading proficiency. For

two of them this comes as no surprise, since they rated at Superior and could not move up. An analysis of the range of questions the CATRC asked of the two Intermediate Mid participants who maintained the same level indicated that more questions were asked in the higher levels before settling back down to the former level. This indicates to me that although these participants apparently experienced an increase in vocabulary and character recognition, it was not to such a degree as to enable them to attain a new level. In fact, this is supported to some degree by many of the participants relating afterwards that they felt they could recognize more characters during the second evaluation.

There was a mean of 34.19 total test items given on the first evaluation, with a minimum of 19 and a maximum of 66. For the second evaluation, a mean of 34.62 total test items were given, with a minimum of 16 and a maximum of 60. Duplicate test items comprised a mean of 4.29, or 12.94% of the mean of total test items from the second evaluation, with a minimum of 0 and a maximum of 12. As mentioned above, the average amount of duplicate test items each participant encountered was determined to be unimportant. The determination that this played no important part in this participant's assessment rests on several factors. The main factor is that the duplicate items were at a much lower level of proficiency (i.e., Advanced) than the participant finally placed at (i.e., Superior). In addition, two of the duplicate items answered correctly during the first evaluation were answered incorrectly during the second evaluation. This leads me to conclude that there was no significant influence from repeated exposure to those test items.

Also revealing to me were the data obtained from the informal interviews. For the most part, participants reported that they did not do much reading apart from preparing the classroom materials. Many (n = 13) felt particularly "bored" in class (the most commonly used term used to describe their feelings) and generally dissatisfied with the classroom materials. A few (n = 6) actually went out of their way to read material outside of what was presented in their textbooks. Of these, only three participants subscribed to newspapers or magazines or bought books that might have given them an opportunity to read material closer to their personal interests. Some participants (n = 4) felt overwhelmed by the amount of time required in attending class and some even indicated that class time in China was "a waste of time". A small number of the participants (n = 3) even went so far as to stop attending some of their formal class hours by the tenth week of the semester.

Discussion

A few important things need to be said at the outset of this discussion. While an increase in ability was detected, several factors may be indirectly involved. The main consideration is, as noted in Kubler (1997:18), the fact that most students had not had formal Chinese training of any kind for about three months prior to their arrival to Beida, which may have offset the results. In essence, their first evaluation may be artificially low. Consequently, any increase observed at the end of this fall semester needs to be handled with care. In order to lessen the potential effect of the time delay in “getting back on their feet”, the participants were initially evaluated three to four weeks into the semester. This allowed them to get settled in to their new surroundings and classes, and also to regain some degree of the Chinese language they may have had. By this time, they had also been allowed by Beida to change class levels if they themselves felt they were misplaced based on their Beida placement test scores. The methodology I employed should therefore increase the reliability of the participants’ initial ratings.

Ideally, of course, an examination of participants should be done at the end of each term of formal study and the beginning of the next term of study to determine more precisely the general trends of temporary language loss. This can perhaps be the purpose of another study. However, due to the volunteer nature of the selection of participants for this study, such an examination would have been close to impossible to arrange.

Another factor may be the amount of time spent studying for class. It was difficult for participants to specify the time they spent preparing for classes. In the informal interviews, however, many of the participants admitted that they felt that they did not prepare much for their classes. Specifically, of the twenty-one participants surveyed, three said they spent no time preparing for classes, seven said they spent about two hours per week, four said they spent between three to six hours per week, and six said they spent between seven and ten hours per week. They also reported that, depending on whether tests were approaching, or other possible external activities, their “averages” could vary radically. Of course, students are probably the best judges as to how much preparation they need to make for varying class loads. This aspect of the research could be studied in more detail.

A possible factor in the overall gain in reading proficiency documented in this study may be the amount of reading for pleasure students do on their own. As previ-

ously mentioned, only three participants said that they had bought materials such as newspapers, magazines or novels to read outside of class. The other eighteen participants did not think this was worthwhile, considering their reading proficiency too low to attempt broaching such resources. However, three of those eighteen participants admitted to buying other textbook-type materials to work through on their own.

I should make one final note regarding the participants' reported motivations for coming to China to study. A majority of them stated that their main goal during this period of study abroad was oral proficiency development. Of the twenty-one participants surveyed, fourteen said that they wanted to work on their speaking abilities. Of the remaining seven participants, four said they wanted a total immersion experience, two said they wanted to specifically develop reading skills, and one listed developing writing skills as most important.

Conclusion

This study documented a certain amount of gain in reading ability resulting from a study abroad experience in Beijing among a volunteer sample of participants. As surmised by Brecht, Davidson, and Ginsberg (1993), compared to a student's oral proficiency, gains in reading comprehension appeared to be relatively small, averaging 1.35 sublevels. The reader should keep in mind that this sample is only from one school in one mainland Chinese city. Results from another school in the same city or other schools in the country may produce different results. I would urge that other researchers pursue opportunities to replicate this study. Better organization of time and arrangement of cooperation with the teachers of the target institution(s) might provide a larger sample size and more comprehensive data.

Despite the apparent weaknesses in the present study, however, I believe that it is important for several reasons. First, it encourages other researchers to go into the field and gather information. Much can be done in the comfort of one's home institution, but that only sheds light on some of the questions we have regarding language acquisition. Second, the combination of quantitative and qualitative data gives us a fuller representation of the lives of foreign students studying abroad. One can administer tests that show improvement in one modality or another, but describing the environment surrounding the students and bringing out their own thoughts about living in a foreign country helps us answer the question of why the numbers say what they say. Finally, this study is longitudinal. Many studies concentrate on taking a "snapshot" of a particular student or group of students at a specific point in their language

development, generally at the end of a defined learning period. This study monitored the progress, as well as the changing attitudes, of students involved in a study abroad experience and specifically related the findings to Chinese language acquisition.

References

- Brecht, R., & Dan Davidson. 1992. Language acquisition gains in study abroad: assessment and feedback. In E. Shohamy and A. R. Walton, eds., *Language Assessment for Feedback: Testing and Other Strategies*. Washington, D. C.: National Foreign Language Center. 87-101.
- Brecht, Richard D., Dan Davidson, & Ralph B. Ginsberg. 1993. *Predictors of Foreign Language Gain during Study Abroad*. NFLC Occasional Papers. Washington D. C.: National Foreign Language Center.
- Brown, James Dean. 1988. *Understanding Research in Second Language Learning: A teacher's guide to statistics and research design*. New York: Cambridge University Press.
- Carlson, J. S., et al. 1990. *Study Abroad: The experience of American undergraduates*. New York: Greenwood.
- DeKeyser, Robert M. 1991. Foreign language development during a semester abroad. In Barbara F. Freed (Ed.) *Foreign Language Acquisition Research and the Classroom*. Lexington, MA: D. C. Heath and Company. 104-119.
- Freed, Barbara. 1990. Language learning in a study abroad context: The effects of interactive and non-interactive out-of-class contact on grammatical achievement and oral proficiency. In J. E. Alatis, ed., *Linguistics, Language Teaching, and Language Acquisition: The interdependence of theory, practice, and research*. Georgetown University Round Table on Languages and Linguistics 1990. Washington, D. C.: Georgetown University Press. 459-477.
- Ginsberg, Ralph B. 1992. *Language Gains during Study Abroad*. NFLC Working Papers. Washington, D. C.: National Foreign Language Center.
- Hatch, Evelyn, & Anne Lazaraton. 1991. *The Research Manual: Design and Statistics for Applied Linguistics*. Boston: Heinle & Heinle.
- Kubler, Cornelius C. 1997. Study abroad as an integral part of the Chinese language curriculum. *Journal of the Chinese Language Teachers Association* 32.3: 15-30.
- Yao, Tao-chung. 1995. A computer-adaptive test for reading Chinese (CATRC): A preliminary report. *Journal of the Chinese Language Teachers Association* 30.1: 75-85.

Zeon PDF Driver Trial
www.zeon.com.tw