中 華 心 理 學 刊
民96，49卷，4期，315-334

Chinese Journal of Psychology
2007, Vol. 49, No. 4, 315-334

Chinese Readers' Knowledge of How Chinese Orthography Represents Phonology

Ming Lo, Chih-Wei Hue, Fang-Zhi Tsai

Department of Psychology, National Taiwan University

MS No.: 06030; Received: July 26, 2006; 1st revision: April 27, 2007; 2nd revision: August 16, 2007; Accepted: August 20, 2007
Correspondence Author: Chih-Wei Hue, Department of Psychology, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 106, Taiwan.（E-mail: Hue@ntu.edu.tw）

Phonetic awareness refers to a Chinese reader's knowledge of the principles governing the orthography-phonology relationships in Chinese characters (Shu, Anderson, & Wu, 2000). Shu et al. suggested that a phonetically aware reader can infer the pronunciation of an unknown character from its constituent components. In particular, they observed that a Chinese reader uses the right component of a left-and-right arranged two-component character to infer the character's pronunciation. However, analysis of seven character groups used by elementary school students and adults showed that in a left-and-right arranged two-component character, both components may provide cues to its pronunciation. The analysis also found that certain simple characters have higher validity in representing phonology than others. According to the statistical model of language learning proposed by Saffran, Aslin and Newport (1996), a learner can acquire an understanding of the statistical nature of linguistic input through repetitive use of the language. For Chinese characters, the mapping of orthography to phonology is imperfect and probabilistic. Thus, according to the statistical learning model, a Chinese reader needs to understand the statistical nature of the mapping between the pronunciation of a character and its components. We tested this assumption in two experiments. The participants in the first experiment were second-, fourth-, and sixth- graders. Their vocabulary sizes were assessed, and their guesses about the pronunciation of a group of pseudo-characters composed of two left-and-right arranged components were recorded. The experiment results indicate the following. (1) The so-called ''position strategy'' is an oversimplification of how a Chinese reader uses a character's components to infer the character's pronunciation. (2) A Chinese reader knows which simple characters are more likely to represent the phonology. (3) A Chinese

This research was supported by the National Science Council of Taiwan Grant No: NSC-94-2413-H-002-022. Some of the data used in the study is taken from Fang-Zhi Tsai's Master thesis, directed by Chih-Wei Hue.

reader knows the probabilities that the various components of a character represent the phonology. That is, the reader knows that the right component is more likely to provide cues to the character's pronunciation than the left component. (4) As a Chinese reader's vocabulary increases, the way he/she uses a character's components to infer the character's pronunciation may best be described as the so called ''position strategy''. In the second experiment, the concept of the ''position strategy'' was tested further and verified, using college students as the participants.

Keywords: ***Character pronunciation, Chinese Character, Meta-linguistic awareness, Phonetic awareness, Statistical model of language learning***

Introduction

Ever since Zhurova and Elkonin observed that there is a relationship between school children's phoneme segmentation abilities and their success in reading, a great deal of research has been devoted to study of the causal relationship between the two factors (see, Calfee & Norman, 1998). However, despite over 40 years of research, the debate about whether phonological awareness is a consequence or a determinant of literacy has not been resolved (Ehri et al., 2001; Lundberg, Frost, & Petersen, 1988). The work of Read, Zhang, Nie, and Ding (1986) is especially interesting. These researchers showed that, because Chinese orthography is not designed to represent phonology, a Chinese reader can not develop phonemic awareness without proper training in Chinese phonemes (see also, Cheung & Chen, 2004). The results of the above studies provide support for a specific training hypothesis about phonological awareness development. Furthermore, the results strongly suggest that the way phonological information is represented in the orthography of a writing system affects how a reader's meta-linguistic ability develops.

Without proper training, a Chinese reader may not be able to acquire the kind of phonemic awareness that a reader of an alphabetic writing system develops. Even so, Shu, Anderson, and Wu (2000) showed that readers of simplified Chinese characters can develop another kind of meta-linguistic ability, called ''phonetic awareness'' as they learn to read more characters. Similar findings have been reported for readers of traditional characters (Hue, 2003). Like a person with phonemic awareness, Shu et al. argued that a child who is phonetically aware has '' ... insight into the principles that govern orthography-phonology relationships in Chinese ...'', and is able to form hypotheses to guide ''...perceptual processing, strategies for learning and retrieving the pronunciations of characters, and ... to forecast the pronunciations of unfamiliar characters....(p. 57)'' They noted that the majority of frequently used characters are phonograms, which are composed of two components. The phonetic of a phonogram is usually the right-hand component, and the radical is usually the left-hand component (the upper panel of Fig. 1). As a result, a phonetically aware reader can develop a position strategy to guess the pronunciation of an unknown character composed of two left-and-right arranged components (referred to as ''LR character'' hereafter). The reader uses an unknown character's right component to guess the character's pronunciation if that component is a pronounceable character. Otherwise, the reader will infer the pronunciation of the unknown character from its neighbors that also contain the ''phonetic''.

A number of studies have shown that a learner can acquire an understanding of the statistical nature of linguistic input through repetitive use of the language (Aslin, Saffran, & Newport, 1998; Saffran, Aslin , & Newport, 1996). A similar idea was postulated by Wonnacott and Newport (2005), who also argued that language learners are constrained by two biases: 1) they expect a language to be governed by rules; and 2) they are very sensitive to the irregularities of the linguistic inputs. According to the statistical model of language

learning, a Chinese reader's phonetic awareness should include the position strategy, as well as a knowledge of the statistical distribution of a character's components; that is, how the various parts of a multiple-component character represent the phonology and which simple characters are more likely to provide phonetic cues when they are used to construct multiple-component characters. Hsiao and Shillcock (2005) observed that the position of a phonetic cue in a multiple-component character is not fixed. For example, as shown in the lower panel of Fig. 1, either component of a LR character can provide hints to the pronunciation of the character. Thus, according to the statistical learning model, the so called "position strategy" does not correctly describe how Chinese orthography represents phonology. Indeed, it may be an overly simplified ''theory'' of what a phonetically aware Chinese reader knows about how Chinese orthography represents the language's phonology.

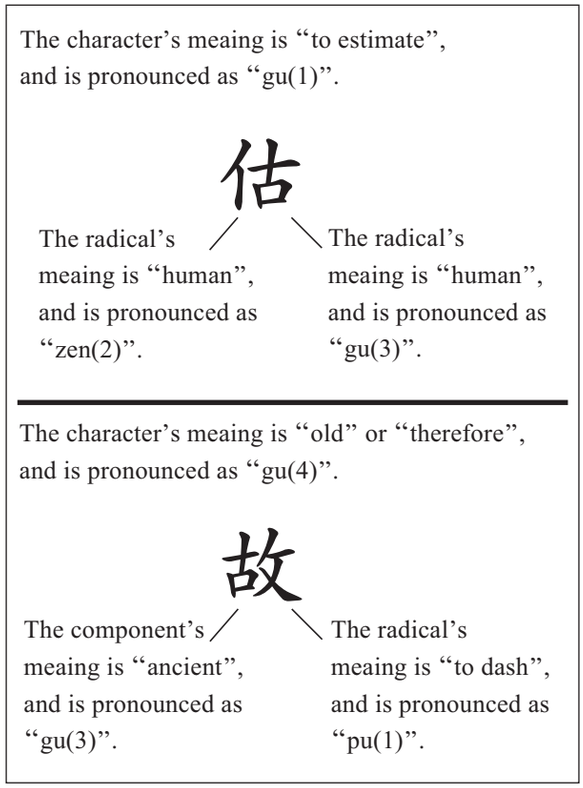

Figure 1. Two two-component characters: Top - the right component is the phonetic; Bottom - the left component is the phonetic.

Although a number of studies have analyzed the principles behind the construction of Chinese characters, none have provided data to show the probability distribution that the components in a multiple-component character represent the phonology. For example, DeFrancis (1984) summarized the results of some relevant studies and concluded that, in about 20% of the traditional characters used by literate adults, a character's pronunciation is the same (excluding the tone) as its right component; and in about another 20% of the characters, a character shares a vowel or a consonant with its right component. Similar findings were reported by Shu, Chen, Anderson, Wu, and Xuan (2003) who analyzed the simplified characters used in elementary school text books in China. According to their study, among the characters analyzed, 72% are composed of two components; and among the two-component characters, there are 72% LR characters. Furthermore, in 42% of the LR characters, the right component provides "useful" cues about the character's pronunciation. The pronunciation of these characters is either the same or similar (i.e., they have the same vowel) to the pronunciation of their right components. Interestingly, neither of the above works analyzed the probability that the left component of a LR character could provide "usefu" phonetic cues.

To verify Hsiao and Shillcock's (2005) argument and provide a basis for the present research, the characters in the frequency norms prepared by the Chinese Knowledge Information Processing Group (the CKIPG character frequency norms ,1993) and the characters in the 12 volumes of Elementary Chinese used in Taiwanese elementary schools were analyzed following the methods of Shu et al. (2003). As shown in Table 1, among the characters used by first to the sixth graders and the general public, there were 79%, 85%, 87%, 89%, 90%, 91%, and 93% two-component characters, respectively. In the LR characters of the seven-character populations, respectively, 38%, 39%, 42%, 44%, 46%, 48%, and 53% of the characters have the same vowel as the right component. Moreover, we found that in each character popula-

Table 1
Analysis of how phonology is represented in the Chinese characters used by elementary school students and educated adults

Character Population	Number of characters	Number of characters composed of two components	LR characters				Up and down arranged characters			
			Number of LR characters	Number of characters in which the right component provides useful phonetic cues [a]	Number of characters in which the left component provides useful phonetic cues	Number of characters in which both components provide useful phonetic cues	Number of characters in which the components are arranged in a vertical manner	Number of characters with a phonetic located in the upper half of a character	Number of characters with a phonetic located in the lower half of a character	Number of characters in which both components are phonetic cues
Adult	5656	5285	3436	1814	233	63	1218	209	347	21
Sixth-grader	2687	2438	1477	704	100	23	657	107	154	10
Fifth-grader	2306	2082	1251	575	85	19	567	91	130	7
Fourth-grader	1814	1619	958	422	67	14	446	70	98	6
Third-grader	1322	1156	677	287	47	8	323	46	71	3
Second-grader	896	760	448	174	32	2	210	22	42	0
First- grader	399	314	175	66	15	2	91	14	7	0

[a] The pronunciation of the component is either the same as (excluding the tone) or similar to (i.e., it has the same vowel or consonant) the character containing it.

tion, about 7% of the LR characters have the same vowel as their left components. Our analysis also showed that only some simple characters are high in "phonetic cue validity". The "phonetic cue validity" of a simple character is defined as the ratio of the number of characters that have the same vowel as the simple character to the number of characters that contain the simple character. For example, the phonetic validity value of the simple character "章" is 0.89, because it appears in 11 characters (彰, 障, 璋, 樟, 蟑, 漳, 瘴, 贛, 嶂, 幛, 獐), and the pronunciation of 10 of them ("彰", "障", "璋", "樟", "蟑", "漳", "瘴", "嶂", "幛", "獐") is the same as or similar to "章". In another example, the phonetic validity value of the simple character "每" is 0.30, because it appears in 10 characters (海, 梅, 敏, 悔, 侮, 毓, 莓, 霉, 晦, 誨), but the pronunciation of only three of them (梅, 莓, 霉) is similar to that of "每". In summary, our analysis shows that some simple characters are more likely to provide useful phonetic cues than others.

Obviously, if the statistical model of language learning is correct, the results reported by Shu et al. (2000) and Hue (2003) do not reveal the full scope of a Chinese reader's phonetic awareness. For example, these studies did not show whether their participants used the left components of the stimulus characters to infer the characters' pronunciation when making responses. There are two possible explanations for this. The first is that the statistical distribution of information about how phonology is represented in characters is different for the characters used by children and those used by adults. It is possible that the majority of the characters used by children are LR characters in which only the right components provide phonetic cues. As a result, for reading, a child is likely to develop a position strategy, which will be used as he/she grows up. However, this explanation has to be discounted because, as shown in Table 1, less than 20% of the characters used by children in the first and second years of school have a right component that provides phonetic cues. For sixth graders and adults, the ratio is about 30%.

The second possible explanation is that the participants in the above studies used a position strategy to infer an unfamiliar or a pseudo-character's pronunciation because the stimulus characters had the same orthographical structure, i.e., they were LR characters with a radical on the left and a phonetic on the right. Although all the radicals are pronounceable, in practice, very few people know how to pronounce them. Thus, in these studies, the only phonological information a stimulus character provided for a participant was the right component's pronunciation.

We tested the second explanation in our experiments. By and large, we followed the methods of Shu et al. (2000) and Hue (2003) when collecting data about how participants (elementary school students in Experiment 1 and college students in Experiment 2) pronounce pseudo-characters and real LR characters. However, to avoid possible response biases induced by the stimuli used in those works, both the right and the left components of the pseudo-characters used in our research are pronounceable. If the probability that the various components of a Chinese character represent phonology is acquired during the character-learning process, then in this research, such knowledge should be reflected in a participant's responses to the pseudo-characters. Specifically, we assume that the participants used both the left and right components of a stimulus character to infer the character's pronunciation; however, they probably used the right component more than the left one.

Experiment 1

Experiment 1 was designed to replicate the results of Shu et al.'s (2000) study. Similar to their study, the experiment used a group of elementary school students as participants. However, unlike their study, the stimuli in Experiment 1 were LR pseudo-characters, each composed of two pronounceable components. We also included a "vocabulary size"[1] test to estimate the number of characters each participant knew. By combining a

participant's pronunciation of the pseudo-characters and his/her estimated vocabulary size, we were able to infer the relationship between the person's vocabulary size and his/her phonetic awareness.

Method

Participants. The participants were from three elementary schools in Taipei. In total, there were 288 students, made up 93 sixth-graders, 63 fourth-graders and 132 second-graders. The second-year students were divided into two groups, one of which was tested at the end of the first semester (48 children), and the other was tested at the end of the second semester (84 children). All the participants were native Chinese speakers. They participated in the experiment at the request of their teachers, and they were given a small gift after the experiment.

Materials and Procedure. Apart from taking the vocabulary size test mentioned above, all the participants answered a questionnaire about the pronunciation of some real and pseudo-characters. The vocabulary test for each group consisted of 50 test characters, which were selected according to the stratified sampling method used by Hue (2003). The test characters for each group, shown in Table 2, were selected from the seven character populations listed in Table 1. For example, of the 50 test items used for the fourth-graders, five were selected from the 399 characters used in first-year Chinese text books, and five were selected from the 497 characters learnt in the second year of school (referred to as the second item set). There are 896 characters in the second-year Chinese text books, of which 399 are learnt in the first year of school. In addition, 8 characters were selected from the third item set (i.e., the 426 characters learnt in the third year of school), 10 from the fourth set (492 characters), 7 from the fifth set (492 characters), and 5 from the sixth set (381 characters). The rest of the test items (10 characters) were selected from 2,969 characters (the adult item set) that are not taught in schools; however, they are included in the CKIPG character frequency norm (1993).

In the vocabulary size test, each item was printed on a separate line. The item was printed on the left-hand side of the line, and the participant's response (the item's pronunciation and meaning) was written on the right-hand side. The participants were asked to spell the pronunciation of an item using Mandarin phonetic symbols, and to explain the item's meaning by using it to construct a meaningful multiple-character term or phrase.

A pronunciation questionnaire was designed for each of the four groups. Although the items in the questionnaires were different, the following principles were used to select the items. (1) There were 12 real- and 24 pseudo-characters, all of which were "LR characters". (2) The real characters were selected from the characters the participants were taught in school. Thus, the participants should have known how to pronounce them. (3) There were two types of pseudo-characters (12 of each). In one type, the phonetic validity of the right component was high, while that of the left component was low (called a "low-high character" hereafter). In the second type, the phonetic validity of the left component was high, while that of the right component was low (called a "high-low character" hereafter). The phonetic validity value of a component was computed separately for the four groups. For example, for the fourth-year students, the LR characters used in the first eight volumes of Chinese text books were selected, and the components used to construct these characters were identified. The following statistics were computed for

[1] In previous studies, sometimes "vocabulary size" was used to represent the amount of words that a participant could speak (Devescovi et al., 2005; Gershkoff-Stowe & Smith, 1997), sometimes the term was used to infer the number of written words that a participant has learnt (Hazenberg & Hulstun, 1996; Laufer & Nation, 1995). In this article, "vocabulary size" refers to the number of characters a Chinese reader knows.

Table 2
The principles used to select the test characters for the four questionnaires

	Groups			
Item groups	Second grader-First semester [c]	Second grader-Second semester	Fourth graders	Sixth graders
First group [a]	8	8	5	5
Second group [b]	11	10	5	5
Third group	6	7	8	5
Fourth group	5	5	10	5
Fifth group	5	5	7	8
Sixth group	5	5	5	10
Adult group	10	10	10	12
Sum	50	50	50	50

[a] These characters are included in Chinese text books used in the first year of elementary school.

[b] These characters are included in the second year Chinese text books, but not in the first year's text books.

[c] This group of participants were tested at the end of the first semester of their second year at school.

each identified component: (1) the number of characters containing the components (N); and (2) the number of characters that the component provides useful phonetic cues (n). The phonetic validity value of a component was defined as the ratio of n to N. In this experiment, the phonetic validity of a component was defined as high, if the ratio was greater than 0.6, and low, if it was lower than 0.3. The components used to construct pseudo-characters were chosen from components that were included in at least 4 characters.

In each pronunciation questionnaire, the real- and pseudo-characters were arranged randomly and printed on separate lines. A stimulus item was printed on the left-hand side of a line, and the participant wrote his/her response (i.e., the item's pronunciation) on the right-hand side.

The pronunciation questionnaire and vocabulary size test were conducted in two sessions separated by at least a week. The questionnaire was completed first. The second-year students were tested individually, and their responses to the questionnaire items were written down by a researcher. The fourth- and sixth-graders were tested in groups, and they wrote their own responses to the test items. The participants were asked to respond to the items in the vocabulary size test as best they could. For the pronunciation questionnaire, they were asked to guess if they did not know an item's pronunciation.

Results and Discussion

Data from six participants (one second-year, two fourth-year, and three sixth-year students) was excluded from the analysis because they responded

Table 3
The number of participants in the four groups based on vocabulary size

	Group of Mean Vocabulary size			
	Low	Median -low	Median -high	High
Second-graders (tested at the end of the first semester)	18	26	3	0
Second-graders (tested at the end of the second semester)	17	58	9	0
Fourth-graders	4	23	29	5
Sixth-graders	3	8	28	51
Sum	42	115	69	56
Estimated Vocabulary size	769	1691	2591	3640

to less than half the items in the pronunciation questionnaire or the vocabulary size test.

Vocabulary Size. Each participant's vocabulary size was estimated using the method proposed by Hue (2003). First, a participant's correct responses to the test items were identified. A response to a test character was deemed correct when the answers to both the character's pronunciation and meaning were correct. Next, the percentage of correctly answered items in each test item set was calculated, and the percentages were multiplied by the size of their respective item sets. Then, the results for all the item sets were summed, and used to estimate the participant's vocabulary size. The averaged vocabulary size and the standard deviations (in parentheses) for the four groups were 1,338 (547), 1,559 (531), 2,209 (670), and 3,119 (811), respectively.

To determine the relationship between a student's vocabulary size and phonetic awareness, the participants were reorganized into the following four groups according to their vocabulary size: high, medium-high, medium-low, and low. The vocabulary size of participants in the high group was one standard deviation (977) above the mean vocabulary size (2,165) of all the participants. In the median-high group the vocabulary size was above the mean, but lower than that of the high group. The vocabulary size of the median-low group was lower than the mean, but higher than one standard deviation below the mean. The rest of the participants were categorized into the low group. Table 3 shows the mean vocabulary size for each group, as well as the number of students in each group.

We also analyzed the participants' response errors. Overall, the four groups of students made 5,531 pronunciation errors and 6,042 semantic errors. The errors were classified into categories similar to that used by Hue (2003). Tables 4 and 5 show the analysis results of the pronunciation and semantic errors, respectively. Because of the large number of errors, trivial results were likely to be included if inferential statistical techniques were used to analyze the data. The following discussion of the errors is based on their descriptive statistics.

As shown in Table 4, nearly 60% of the pronunciation errors of the low group consisted of "no response". However, the participants were less likely to make such errors as their vocabulary size increased. Error types 1, 2, 3, and 4 were component-related errors, which indicated that the participants who made the errors relied on the components of a test item to infer the item's pronunciation. These errors comprised 36% of all the pronunciation errors. Analyzing the four groups of participants separately, the percentages of component related errors comprised 17%, 36%, 39%, and 53% of the pronunciation errors for the low, median-low, median high, and high groups,

Table 4
Proportion of pronunciation errors as a function of the error type and the participant group

	Participant Group			
Error type	Low	Median-low	Median-high	High
Type 1	0.06	0.12	0.10	0.12
Type 2	0.08	0.20	0.27	0.36
Type 3	0.01	0.01	0.00	0.00
Type 4	0.02	0.03	0.02	0.05
Type 5	0.05	0.05	0.07	0.08
Type 6	0.09	0.10	0.13	0.12
Type 7	0.01	0.02	0.03	0.01
Type 8	0.14	0.08	0.05	0.06
Type 9	0.56	0.40	0.33	0.20
Total number of errors	1444	2474	1068	543

Error type definitions:

Type 1: Responded with the pronunciation of the right component of the test character

Type 2: Responded with the pronunciation of a neighbor of the test character (the neighbor contained the right component of the test item)

Type 3: Responded with the pronunciation of the left component of the test character

Type 4: Responded with the pronunciation of a neighbor of the test character (the neighbor contained the left component of the test item)

Type 5: Responded with correct pronunciation , but wrong tone

Type 6: Responded with a pronunciation that had the same consonant or vowel as the test character

Type 7: Responded with the pronunciation of a frequently associated character of the test character

Type 8: Errors that can not be classified into other categories

Type 9: No response

respectively. A substantial difference in the error patterns of the low group and the median-low group was observed. The abrupt increase in the number of component-related errors in the median-low group indicates that, compared to the low group, the participants were more likely to rely on a character's components to infer the character's pronunciation. In sum, the pattern of the results indicates that a Chinese reader is likely to infer the pronunciation of an unknown character from its components. This is especially true for a person with a large vocabulary.

As shown in Table 5, in the meaning errors, the ''no responses'' were higher than that of the pronunciation errors; the proportions were 66%, 55%, 58%, and 51% for the low, median-low, median-high, and high groups, respectively. These results indicated that the participants were often unwilling to guess an unknown character's meaning. There were 27.52% component-related errors (i.e., the first three types of errors), and in these errors, radical related errors (type 1 errors) were only 7.27%. In the type 2 and 3 errors, the ones relating to the left component of a character were very few (i.e., the sum of type 2-2 & 3-2 errors), most of them were right-component related (i.e., the sum of type 2-1 & 3-1 errors). Furthermore, in the component-related errors, type 3 errors took 71.03%, which indicated that the participants were more likely to infer an unknown character's meaning from its neighbors (type 3 errors) than its components (type 2 errors). Analysis of the results separately for the four groups of participants, the component-related errors were 16%, 32%, 28%, and 33%, from the low group to the high group, respectively. Like the pattern of pronunciation errors, there was a substantial difference in the number of component-related meaning errors between the low and the median-low groups. In sum, the pattern of the results indicated that the participants of this study were often unwilling to guess an unknown character's meaning. However, if they did guess, they were more likely to infer the character's meaning from its neighbor than its radical. This finding is contrary to the common belief that a Chinese reader would guess the meaning of an unknown character from its radical, however, it is justifiable. An experienced Chinese reader knows that a radical hints to the semantic category of the character containing it. However, such knowledge should not help the person to answer the exact meaning of an unknown character, because the information is too vague (Ho, Ng, & Ng, 2003). A participant of this study was required to write down the meanings of the test items in the vocabulary size test or to use them to create multiple-character words, and thus, when an unknown character was encountered, the participant was likely not to answer it or to guess its meaning from a neighbor of the character, which carried a clear meaning.

Pronunciation task. Recall that the responses of six participants were excluded from the analysis. The remaining participants' responses to the two types of pseudo-characters in the questionnaire were divided into eight categories, after which the response ratio was computed for each pseudo-character type and each response category. Table 6 shows the means of the ratios as a function of the participants' vocabulary size.

As shown in the table, to infer an unknown character's pronunciation, a participant might have relied on one of the character's components (response types 1 and 3) or its neighbors (response types 2 and 4). Either way, the responses were ''component-related''. Thus, for the purposes of this study, the four types of responses were combined, and the results were treated in three sets of analyzes.

In the first set, the participants' tendency of using a character's components to infer the character's pronunciation was analyzed. The response ratios that were categorized as types 1, 2, 3, and 4 were combined. The data was analyzed using a mixed model of the analysis of variance (ANOVA), with Group (low, median-low, median-high vs. high vocabulary-size group) as the between-participant variable and Pseudo-character Type (low-high vs. high-low characters) as the within-participant variable. The analysis showed

Table 5
Proportion of semantic errors as a function of error type and participant group

	Participant Group			
Error type	Low	Median-low	Median-high	High
Type 1	0.02	0.02	0.02	0.02
Type 2-1	0.03	0.05	0.05	0.05
Type 2-2	0.00	0.00	0.00	0.00
Type 2-3	0.01	0.02	0.01	0.01
Type 3-1	0.05	0.10	0.08	0.10
Type 3-2	0.01	0.04	0.02	0.03
Type 3-3	0.04	0.09	0.10	0.12
Type 4	0.01	0.01	0.01	0.01
Type 5	0.02	0.03	0.03	0.03
Type 6	0.16	0.10	0.08	0.07
Type 7	0.66	0.55	0.58	0.51
Total number of errors	1423	2673	1255	691

Error type definitions:

Type 1: Responded with the meaning of the radical contained in the test character

Type 2: Responded with the meaning of a non-radical component of the test character

2-1: the component locating at the test character's right half

2-2: the component locating at the test character's left half

2-3: the component locating at neither the test character's left nor right half

Type 3: Responded with the meaning of a neighbor of the test character

3-1: the neighbor contained the right component of the test character

3-2: the neighbor contained the left component of the test character

3-3: the neighbor contained neither the left or the right component of the test character

Type 4: Responded with the meaning of a character which is frequently associated with the test character

Type 5: Responded with the meaning of a homophone of the test character

Type 6: Errors that can not be classified into the other categories

Type 7: No response

Table 6
Means of response ratios as a function of the response category, vocabulary size, and pseudo-character type

	Participant Group							
	Low		Median-low		Median-high		High	
Response types	Right [a]	Left [b]	Right	Left	Right	Left	Right	Left
Type 1	0.42	0.15	0.52	0.18	0.66	0.16	0.74	0.13
Type 2	0.11	0.10	0.14	0.21	0.18	0.29	0.20	0.30
Type 3	0.06	0.23	0.08	0.32	0.02	0.32	0.00	0.37
Type 4	0.05	0.06	0.05	0.07	0.03	0.07	0.02	0.07
Type 5	0.03	0.02	0.06	0.03	0.05	0.04	0.03	0.05
Type 6	0.07	0.11	0.03	0.04	0.01	0.01	0.00	0.00
Type 7	0.27	0.33	0.12	0.15	0.04	0.10	0.03	0.09

Response type definitions:

Type 1: Used the right component of a test item to infer the item's pronunciation

Type 2: Used a neighbor of a test item to infer the item's pronunciation (the neighbor contained the right component of the test item)

Type 3: Used the left component of a test item to infer a test item's pronunciation

Type 4: Used a neighbor of a test item to infer the item's pronunciation (the neighbor contained the left component of the test item)

Type 5: Used a character whose visual shape is similar to that of the test item to infer the pronunciation

Type 6: No response

Type 7: Responses that can not be categorized as any other type

[a] The phonetic validity values of the left and the right components of this type of pseudo-character are low and high, respectively.

[b] The phonetic validity values of the left and the right components of this type of pseudo-character are high and low, respectively.

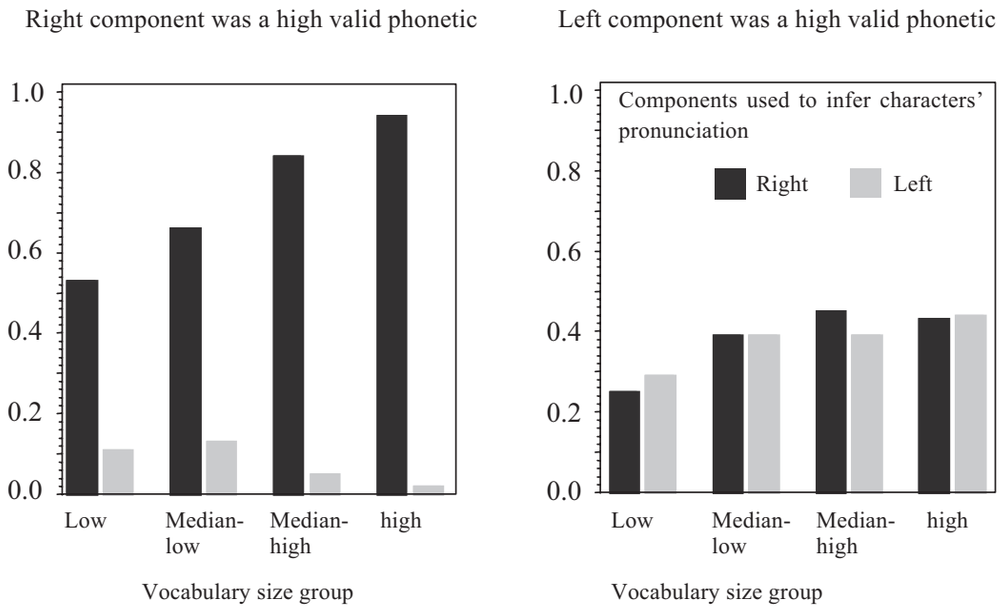

Figure 2. Means of response ratios as a function of response category, vocabulary size, and pseudo-character type.

the significant effects of Group, $F(3, 278) = 37.22$, $p < .0001$, Pseudo-character Type, $F(1, 278) = 30.08$, $p < .0001$, and the interaction between Group and Pseudo-character Type, $F(3, 278) = 3.37$, $p < .05$. For low-high characters, the results indicated that the participants had an increasing tendency to rely on a component to infer a character's pronunciation as their vocabulary size increased, $F(3, 278) = 34.29$, $p < 0.0001$. Both the high and the median-high groups had a stronger tendency to use a character's components when making a response than the median-low group, which in turn had a stronger tendency than the low group (HSD = 0.092, $p = .01$). For the high-low characters, there was also an increasing tendency for participants to base their responses on a character's components as their vocabulary size increased, $F(3, 278) = 27.36$, $p < 0.0001$. Post hoc analysis showed that the high, median-high, and median-low groups tended to use components when making responses, but the tendency was significantly weaker in the low group (HSD = 0.106, $p = .01$). The pattern of the results indicates that the participants often used a pseudo-character's components to infer the character's pronunciation, and that tendency grew stronger as the participants' vocabulary size increased.

To determine if and how the participants used the left or right component of a character to infer the character's pronunciation, the number of times they used the right component of a pseudo-character to infer pronunciation (response types 1 & 2) was analyzed separately from the number of times they used the left component (response types 3 & 4). Thus, in the second set of analyses, we combined response types 1 and 2 because they related to the use of the right component of a test item when making responses. Response types 3 and 4 were similarly combined because they related to the use of the left component. Figure 2 shows how the participants of the different groups used the left and right components to infer the pronunciation of the two types of pseudo-characters.

We used the same mixed model of ANOVA to analyze the two sets of data. For responses based on the right components of the characters, the results of ANOVA revealed the significant effects of Group, $F(3, 278) = 46.55$, $p < .0001$, and

Pseudo-character Type, $F(1, 278) = 813.77$, $p < .0001$. The interaction of the two variables was also significant, $F(3, 278) = 19.83$, $p < .0001$. Simple effect contrasts were performed separately for the two types of pseudo-characters. The results indicate that, for the low-high characters, there was an increasing tendency to use the right component of an item to make a response as the participants' vocabulary size increased, $F(3, 278) = 57.94$, $p < .0001$; the high and the median-high groups had the same ratio, which was higher than that of the median-low group. As expected, the low group had the smallest ratio (HSD = 0.097, $p = .01$). For the high-low characters, the analysis revealed a significant difference among the four groups, $F(3, 278) = 13.23$, $p < .0001$. The response ratio of the low group was lower than the ratios of the other three groups, which were all the same (HSD = 0.090, $p = .01$).

The analysis of responses based on the left components of the characters also revealed the significant effects of Pseudo-character Type, $F(1, 278) = 643.92$, $p < .0001$, and Group, $F(3, 278) = 3.99$, $p < .01$, and the interaction between them, $F(3, 278) = 15.89$, $p < .0001$. For low-high characters, the analysis showed there was a decreasing tendency to use the left component of a character as the participants' vocabulary size increased, $F(3, 278) = 17.67$, $p < .0001$. Analysis of the simple effect showed that the low and median-low groups had higher response ratios than the median-high and the high groups (HSD = 0.057, $p = .01$). However, for the high-low characters, the trend was the opposite of the low-high characters, and it was significant, $F(3, 278) = 6.23$, $p < .001$; the low group had a lower ratio than the other three groups (HSD = 0.096, $p = .01$).

The second set of analyses showed the following. (1) The idea that a Chinese reader only uses the right component of a character to infer the character's pronunciation is not supported by the experiment results, which show that both components of a pseudo-character are used. (2) The finding that the participants responded to the two types of pseudo-characters differently, especially those in the median-high and high groups, supports the assumption that a phonetically aware Chinese reader knows which simple characters have high phonetic validity. In particular, the results show that, for low-high characters, the participants based their responses predominantly on the right component. However, for high-low characters, the probability that the left component would be used to infer a character's pronunciation was about the same as that for the right component. The experiment results show that when the left component of a pseudo-character was a simple character with high phonetic validity, the participant was likely to recognize it and use it to make a response.

In the third set of analyses, the difference between the sum of the ratios of response types 1 and 2 and the sum of the ratios of response types 3 and 4 was computed for each participant, and the resulting data was analyzed following the procedure used for the second set of analyses. The analysis revealed the significant effects of Group, $F(3, 278) = 25.05$, $p < .0001$, and Pseudo-character Type, $F(1, 278) = 882.09$, $p < .0001$, and the interaction between them, $F(3, 278) = 21.05$, $p < .0001$. Simple effect analysis of the two types of pseudo-characters showed that, for low-high characters, there was an increasing tendency to rely on the right component of an item to make a response as the participants' vocabulary size increased, $F(3, 278) = 47.45$, $p < 0.001$. The median-high and high groups were more likely to select the right component of an item than the low and median-low groups (HSD = 0.129, $p = .01$). However, for the high-low characters, the trend was not significant, $F(3, 278) = 0.82$, $p = 0.485$.

The results of the third set of analyses provide further information about a Chinese reader's knowledge of the probability that the various components of a character represent phonology. For the low-high characters, the results of the third set of analyses indicate that a participant with median-high or high vocabulary size had a stronger tendency to rely on the right-component to make a response than a participant with low or median-low vocabulary size. On the other hand, for the high-low characters, in all

four groups, the probability that a participant would rely on the left component to make response was the same as that for the right component. This result indicates that the participants knew a character's right component was more likely to represent phonology than the left component. Thus, although the phonetic validity of the right component of a high-low character was low, the probability that a participant would use it to make response was as high as that of using the more phonetically valid left component.

Finally, in Experiment 1, we found that when a person's vocabulary size reaches the median-high level, he/she understands that the right component of a Chinese character is more likely to provide the character's phonology; hence the knowledge is used to infer the character's pronunciation. The above argument is supported by the following findings. (1) The second set of analyses showed that, for low-high characters, the probability that the participants would use the right component to infer a character's pronunciation increased as their vocabulary size increased, until it reached the median-high level. Conversely, the responses based on the left component decreased as the participants' vocabulary size increased. Similarly, the trend became stable when the participants' vocabulary size reached the median-high level. (2) The third set of analyses assessed whether the participants used the right or the left component to respond to the low-high characters. We found there was no significant difference between the median-high and the high groups in their use of the right component to make responses. Furthermore, the two groups were more likely to use the right component than the median-low and the low groups.

Experiment 2

The results of Experiment 1 suggest that a Chinese reader knows which simple characters have high phonetic validity and which ones do not. Moreover, such knowledge operates independently of a person's awareness that the right component of a character is more likely to represent phonology. Experiment 2 was designed to investigate these assumptions further. In addition to the two types of pseudo-characters used in Experiment 1, two more types of pseudo-characters were used in Experiment 2. In one of the new types, both components of the stimulus character were high in phonetic validity (called a "high-high character" hereafter). In the other type, both components were low in phonetic validity (called a "low-low character" hereafter). Thus, except for the high-low characters, a mature Chinese reader, who knows that a character's right component is more likely to represent phonology than the left component, should show a strong tendency to rely on the right component to infer the pronunciation of the other three types of pseudo-characters.

Method

Participants. Twenty-two National Taiwan University students taking an introductory course in psychology participated in this experiment. All were native Chinese speakers with at least 12 years of education in Taiwan. By participating in the experiment, they received course credits.

Materials and Procedure. A questionnaire comprised of 60 pseudo-characters and 24 real characters was designed. All the stimuli were LR characters. The 60 pseudo-characters were divided equally into four types based on orthogonal manipulation of the phonetic validity of the two components of a character. The definition of high and low phonetic validity values was the same as that in Experiment 1. The real characters were randomly chosen from the 3,000 most frequently used characters in the CKIPG character frequency norm (1993). The arrangement of the stimulus characters in the questionnaire and the administration of the questionnaire were the same as in Experiment 1.

Results and Discussion

The procedure used to analyze the partici-

pants' responses was the same as that in Experiment 1. The means of the response ratios for each response category are detailed in Table 7.

Three sets of analyses were conducted. In the first set, to determine if the right component of a stimulus character was used in making responses, the type 1 and 2 responses of each participant for each pseudo-character type were combined. The ANOVA results revealed the significant effect of Pseudo-character type, $F(3, 63) = 62.94$, $p < .0001$. Post hoc comparisons revealed a significant difference between the high-high and the high-low conditions (HSD = 0.101, $p = .01$), which indicated that the participants were more likely to use the right component of a pseudo-character to infer a character's pronunciation in the high-high condition than in the high-low condition.

In the second set of analyses, to determine if the left component of a stimulus character was used in making responses, the type 3 and type 4 responses of each participant for each response type were combined. The effect of the pseudo-character type, $F(3, 63) = 67.73$, $p < .0001$, was significant. Post hoc comparison showed a significant difference between the high-low and the low-low conditions (HSD = 0.108, $p = .01$), which indicated that the participants were more likely to use the left component of a pseudo-character in the high-low condition to infer a character's pronunciation than in the low-low condition.

In the third set of analyses, to determine whether the right component was used more often than the left component when making responses, the type 1 and 2 responses of each participant for each pseudo-character were combined; the type 3 and type 4 responses were similarly combined. Then, the difference between the two summed ratios was computed. The resulting data was analyzed using a repeated-measure ANOVA. The analysis showed that the effect of the Pseudo-character type was significant, $F(3, 63) = 67.99$, $p < .0001$. The means were 0.92, 0.56, 0.68, and 0.03 for the low-high, high-high, low-low, and high-low pseudo-character type conditions respectively. Post hoc comparison indicated that the tendency to use the right component to infer a pseudo-character's pronunciation was higher for the low-high condition than for the high-high and the low-low conditions. There was no difference between the high-high and the low-low conditions. Finally, the tendency to use the right component to infer a character's pronunciation was the lowest in the high-low condition (HSD = 0.205, $p = .01$). The results show that the participants tended to use the right component of a pseudo-character to infer the character's pronunciation in both the high-high and low-low conditions.

To summarize, the two main findings of Experiment 2 were as follows. First, to infer a character's pronunciation, the participants were more likely to use the left component of a pseudo-character in the high-low condition than in the low-low condition. This finding indicates that a mature Chinese reader can differentiate simple characters with high phonetic validity from characters with low phonetic validity. Moreover, it indicates that a mature Chinese reader can use the phonological information in a character to infer the character's pronunciation, no matter whether the information is in the right or the left component. Second, in both the high-high and low-low pseudo-character conditions, the participants were more likely to use the right component of a pseudo-character to infer the character's pronunciation. This finding indicates that a mature Chinese reader probably uses the position strategy to search for the phonological cue in a character.

General Discussion

In this paper, we have investigated whether Chinese readers know how Chinese orthography represents phonology. In particular, we investigated whether the statistical nature of the mapping between the phonology of a character's components and that of the character itself is acquired by Chinese readers.

The following findings are significant. First, the participants in the study used either component (the right or the left) of a pseudo-character to infer

Table 7

Means of the ratios as a function of response category and phonetic validity

Phonetic validity of the right component	High		Low	
Phonetic validity of the left component	Low	High	Low	High
Response type				
Type 1	0.77	0.66	0.38	0.21
Type 2	0.18	0.12	0.44	0.29
Type 3	0.01	0.17	0.07	0.36
Type 4	0.03	0.05	0.07	0.11
Type 5	0.00	0.00	0.00	0.00
Type 6	0.00	0.00	0.00	0.00
Type 7	0.01	0.01	0.03	0.03

Response type definitions:

Type 1: Used the right component of a test item to infer the test item's pronunciation

Type 2: Used a neighbor of a test item to infer the item's pronunciation (the neighbor contained the right component of the test item)

Type 3: Used the left component of a test item to infer the test item's pronunciation

Type 4: Used a neighbor of a test item to infer the item's pronunciation (the neighbor contained the left component of the test item)

Type 5: Used a character whose visual shape is similar to that of the test item to infer the pronunciation

Type 6: No response

Type 7: Responses that can not be categorized as any other type

Pseudo-character type

Low-High: The phonetic validity values of the left and right components of a pseudo-character of this type are low and high, respectively.

High-High: The phonetic validity values of both components of a pseudo-character of this type are high.

Low-Low: The phonetic validity values of both components of a pseudo-character of this type are low.

High-Low: The phonetic validity values of the left and the right component of a pseudo-character of this type are high and low, respectively.

a character's pronunciation. This finding indicates that the so called ''position strategy'' is an oversimplified assessment of how a Chinese reader uses a character's components to infer the character's pronunciation. Second, we found that, for the high-low condition, the participants' probability of using the left component of a stimulus character to make a response was higher than that of the low-high condition. This suggests that a Chinese reader knows which simple characters are high in phonetic validity. Moreover, when such a character is used to construct another character, the reader is likely to use it to infer the pronunciation of the new character, irrespective of its position in that character. Third, in the high-high and low-low conditions in Experiment 2, the participants were much more likely to use the right component of the stimulus pseudo-characters to make responses. This finding indicates that a Chinese reader knows that a character's right component is more likely to provide cues to the character's pronunciation than the left component. Fourth, Experiment 1 found that the participants in the median-high and high groups were more likely to use the right component than those in the median-low and the low groups to infer the pronunciation of a low-high character. In addition, the probability of them using the left component to make a response was lower than that of the median low and the low groups. These findings indicate that when a Chinese reader's vocabulary size reaches the median-high level, the way he/she uses a character's components to infer the character's pronunciation can best be described as the so called ''position strategy''.

The results of the present study discredit the idea that a Chinese reader relies on the right component of a multiple-component character to infer the character's pronunciation, as mentioned earlier. However, Experiment 2 found that if both components of a character provide, or do not provide, useful phonological cues to the character's pronunciation, a mature reader is likely to infer the character's pronunciation from the pronunciation of the character's right component. Thus, we do not exclude the possibility that a Chinese reader will use the position strategy when reading. This argument is consistent with Wonnacott and Newport's (2005) idea that a language learner tends to regularize and form rules to guide the way they speak when the structure of the linguistic input is inconsistent. Once a rule has been created, it tends to be used in all situations. In the case of Chinese, when a reader knows that the right component of a character is more likely to represent the phonology than the left component, he/she may construct a ''position strategy'' and use it when reading. The tendency for Chinese readers to use the ''position strategy'' when reading has been demonstrated in a number of research works using on-line tasks. For example, Chen and Allport (1995) demonstrated that a Chinese reader tends to focus on the right component of a character when performing a phonological task. Similarly, Wang and Ching (2005) showed that, in an explicit recognition task, a Chinese reader only showed a bias toward a ''phonetic cue'' if it was the character's right component. The critical difference between the above studies and the present research is the ''time pressure'' involved. In a time-limited task, a participant has to use a strategy that can derive the correct responses rapidly. In contrast to these studies, our study allowed the participants in the pronunciation task to make responses without a time limit so that all the information in a character was likely to be considered and used to form a response. Thus, the pattern of our research results indicates that when reading Chinese, the position strategy is a natural choice, whether or not there is a time limit.

In sum, the present study shows that learning how Chinese orthography represents phonology is the same as learning the other characteristics of a language. In other words, learning Chinese characters also reflects the process described by the statistical model of language learning proposed by Aslin et al. (1998).

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional statistics by 8-month-

old infants. *Psychological Science*, *9*, 321-324.

Calfee, R. C., & Norman, K. A. (1998). Psychological perspectives on early reading wars: The case of phonological awareness. *Teacher's College Record*, *100*, 242-275.

Chen, Y. P., & Allport, A. (1995). Attention and lexical decomposition in Chinese word recognition: Conjunctions of form and position guide selective attention. *Visual Cognition*, *2*, 235-268.

Chinese Knowledge Information Processing Group. (1993). *Corpus-based frequency count of characters in journal Chinese: Corpus-based research series no.1*. Taipei, Taiwan: Academia Sinica Institute of Information Science.

Cheung, H., & Chen, H. C. (2004). Early orthographic experience modifies both phonological awareness and on-line speech processing. *Language and Cognitive Processes*, *19*, 1-28.

DeFrancis, J. (1984). *The Chinese language: Fact and fantasy*. Honolulu, HI: University of Hawaii Press.

Devescovi, A., Caselli, M. C., Marchione, D., Pasqualetti, P, Reilly, J., & Bates, E. (2005). A crosslinguistic study of the relationship between grammer and lexical development. *Journal of Child Development*, *32*, 759-786.

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the national reading panel's meta-analysis. *Reading Research Quarterly*, *36*, 250-287.

Gershkoff-Stowe, L., & Smith, L. B. (1997). A curvilinear trend in naming errors as a function of early vocabulary growth. *Cognitive Psychology*, *34*, 37-71.

Hazenberg, S., & Hulstun, J. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics*, *17*, 145-163.

Ho, D. S. H., Ng, T. T., & Ng, W. K. (2003). A ''radical'' approach to reading development in Chinese: The role of semantic radicals and phonetic radicals. *Journal of Literacy Research*, *35*, 849-878.

Hue, C. W. (2003). Number of characters a college student knows. *Journal of Chinese Linguistics*, *31*, 300-339.

Hsiao, J. H., & Shillcock, R. (2005). Differences of split and non-split architectures emerged from modeling Chinese character pronunciation. *Proceedings of the Twenty Seventh Annual Conference of the Cognitive Science Society* (pp. 989-994). Mahwah, NJ: Lawrence Erlbaum Associates.

Laufer, B., & Nation, P (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307-322.

Lundberg, I., Frost, J., & Petersen, O. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly*, *23*, 263-283.

Read, C, Zhang, Y., Nie, H., & Ding, B. (1986). The ability to manipulate speech sounds depends on knowing alphabetic spelling. *Cognition*, *24*, 31-44.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, *274*, 1926-1928.

Shu, H., Anderson, R. C., & Wu, N. (2000). Phonetic awareness: Knowledge of orthography-phonology relationships in the character acquisition of Chinese children. *Journal of Educational Psychology*, *92*, 56-62.

Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y. (2003). Properties of school Chinese: Implications for learning to read. *Child Development*, *74*, 27-47.

Wang, M. Y., & Ching, C. L. (2005). *Implicit versus explicit word recognition: They differ in patterns of attentional bias for word components*. Manuscript submitted for publication.

Wonnacott, E., & Newport, E. L. (2005). Novelty and regularization: The effect of novel instances on rule formation. In A. Brugos, M. R. Clark-Cotton & S. Ha (Eds.), *BUCLD 29: Proceedings of the 29th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

中文讀者對於中文字形表徵字音的構形知識：一個探討表音覺識的研究

羅明 胡志偉 蔡方之

國立台灣大學心理學系

根據Shu等人（Shu, Anderson, & Wu, 2000）的研究，當一位中文讀者的識字量增多時，他能夠發展出一種被稱做表音覺識（phonetic awareness）的後設語言知識。亦即，該讀者能夠瞭解中文字構形表徵語音的原則，並根據此原則猜測生字的讀音。本研究將延續Shu等人的研究，並根據「語言學習的統計分佈理論」（statistical model of language learning；下文簡稱「統計學習理論」）進一步探討表音覺識的發展和內容。根據統計學習理論的觀點，一位語言使用者會根據其所接收到的語言訊息，建構訊息與訊息之關係的機率分配表徵。而這些機率分配不但會影響該使用者的語言行爲，更代表了他對語言結構的內隱知識。

本研究分析了包含於國小一到六年級的國語課本，以及一個成人字庫中的中文字。分析結果顯現，不論哪一種字庫中的左右排列雙部件字，右部件提供文字讀音訊息的機率均比左部件高。但分析也發現，左部件在約7%的左右排列雙部件字中提供字音訊息。另外，本研究發現，有一組文字不但可以成爲其它中文字的部件，而且具有高表音性；亦即，當作爲文字部件時，它們經常爲文字讀音提供有效線索。由此觀之，中文是一種深層文字，其構形和讀音之間具有一種不完全對應的機率關係。如果統計學習理論是對的，則中文讀者會在文字學習的過程中，習得中文字之形音對應的機率關係。

本研究以兩個實驗來檢驗上述的想法。實驗一以國小學童爲受試者評估其識字量，並收集他們對兩類假字的標音反應。實驗結果顯示，當受試者的識字量增多時，他們會發展出下面的知識：（一）有些中文字做爲部件時，具有高表音性，（二）左右排列雙部件字的右部件和左部件都可能提供字音的訊息，但右部件提供字音訊息的機率較高。在上述的知識下，本研究發現中文讀者會發展一種以右部件猜測文字讀音的便捷策略；實驗二以大學生爲實驗對象，驗證了這種想法。

關鍵詞：語言學習的統計分佈理論、後設語言知識、語音覺識、文字習得、部件表音性