Research Forum

Japanese EFL Learners' Test-Type Related Interlanguage Variability

Akihiro Ito
Hiroshima University

The purpose of this article is two-fold: 1) to investigate the effects of differences in test-types on the accuracy rates in interlanguage performance of Japanese EFL learners, and 2) to examine the reliability and validity of a grammaticality judgment test. Three grammar tests of relative clauses with three different test-types were assigned to 41 Japanese high school students. The tests were constructed on the basis of amount of attention to linguistic form from reviewing recent SLA works on task variation theory. From the results of the investigation, it is argued that 1) the participants showed different accuracy rates in the three test-types according to the expected order, and 2) the grammaticality judgment test showed relatively high reliability (rxx'=0.792) and moderate validity. The article also discusses the pedagogical implications of the findings and future direction of research.

この論文の目的は二つある。一つは日本人英語学習者の中間言語パフォーマンスへのテスト・タイプの効果を調査することであり、もう一つは文法性判断テストの信頼性と妥当性を検証することである。3種類の関係詞節に関するテストを41人の日本人高校生に実施した。テストは最近の第二言語習得研究におけるタスク要因理論を参考にして、言語形式への注意の度合を変数として作成された。調査の結果、3種類のテストにおいて仮説通りの順番で正確さの度合が異なり、文法性判断テストは比較的高い信頼性（rxx'=0.792)と中程度の妥当性を示した。本稿はさらにこの研究の教育への示唆と将来の研究の方向についても論じている。

When we try to estimate language learners' competence by measuring performance on a certain grammatical item, we often find it quite difficult to decide which test-type should be used. Recent research on second language acquisition theory has suggested that the accuracy rates of interlanguage performance systematically vary according to the kind of test-type[1] (Kameyama, 1987; Ohba, 1994a, 1994b). This phenomenon has been mainly explored in

an area of second language acquisition research called task variation theory (Ohba, 1987).

In the early to mid-1990s, researchers in the field of language testing claimed that research on language achievement tests had been neglected (Weir, 1993; Negishi, 1995) because most language testing professionals were more interested in measuring learners' general or overall proficiency.[2] Researchers tried to construct tests with potential for explaining learners' proficiency. As a consequence, little attention was being paid to test-type.[3] From the late 1970s through the 1980s, findings from SLA research implied that it was dangerous to measure learners' achievement level on a grammatical item or feature through use of a test containing only one test-type[4] (c.f. Kameyama, 1987; Nunan, 1992; Ohba, 1994a, 1994b). This was because of the reported systematic performance variability shown in foreign or second language learners' interlanguage according to test-type (Tarone, 1979, 1982, 1983, 1985, 1988; Sajjadi & Tahririan, 1992). Hence, the accuracy rate of the learners' interlanguage performance measurements can be affected by the test-type even if intended to measure a similar trait (Bachman, 1990; Ohba, 1994a, 1994b). This has served as the motivation for the present research. In this study relative clauses are the subject for three different test-types constructed to examine the effect of difference in test-type on Japanese EFL learners' interlanguage performance.

Test-type Classification

Recent research on test-types in SLA began with the dichotomous catergorization of test types into Natural Communication Tasks and Linguistic Manipulation Tasks (Dulay, Burt, & Krashen, 1982). A Natural Communication Task required learners to pay attention to content in order to use language for communication. A Linguistic Manipulation Task asked learners to pay attention to linguistic manipulation of form. This categorization was based on the *Monitor Model* (Krashen, 1977a, 1977b, 1978a, 1978b, 1981, 1982; Krashen & Terrell, 1983). The *Monitor Model* "predicts that the nature of second language performance errors will depend on whether monitoring is in operation" (Krashen, 1982, p. 152). Therefore, the Linguistic Manipulation Task, which encourages use of the monitor, measures to what degree learners have mastered grammatical rules and permits them to show higher accuracy. On the other hand, the Natural Communication Task, which does not permit monitor use, measures subconsciously-learned grammatical rules, which results in lower accuracy. In other words, Dulay, Burt, & Krashen (1982) suggest that test-types be classified rather like an on-off switch based on whether the test-type allows the monitor to be in operation or not.

Criticisms of test-type classification based on this dichotomy were quick to appear. Tarone (1983) first modified the dichotomy of test-types from the Labovian sociolinguistic perspective, arguing that there was no clear on-off point. She then proposed that the accuracy of interlanguage performance ranges on a continuum from vernacular to careful style based on the *Interlanguage Continuum Model.* Following this model, Hyltenstam (1983) proposed a more detailed categorization of experimental data, suggesting the following eight test-types:

a) Elicited production, often with pictorial stimuli, such as the *Berko Test*, the *Bilingual Syntax Measure*, guided composition;
b) Manipulation of given linguistic material such as sentence combining and sentence completion;
c) Intuition or grammaticality judgment test;
d) Introspection;
e) The cloze procedure;
f) Imitation;
g) Dictation or partial dictation; and
h) Translation. (p. 58)

Hyltenstam (1983), examining research on the relationship between interlanguage performance and selection of test-types, argued that it was impossible to divide test-types into two groups and reasonable to categorize them according to how much each required attention to linguistic form and content. According to this classification, as learners move toward translation, and pay greater attention to linguistic form, the resulting accuracy rate in the test is higher.[5]

Research on Interlanguage Variability

Most research on the effect of test-types on learners' interlanguage performance has not stemmed from research on testing but from other perspectives of SLA (Ohba, 1994a, 1994b). Ohba (1987, 1994a, 1994b) makes a rough division of what kind of grammatical items are chosen for the purpose of second language acquisition research, categorizing them into three types. The first, concerned with phonology, investigates the relationship between accuracy rate of a particular phoneme and the type of tasks provided to the participants (Dickerson, 1975; Beebe, 1980; Sato, 1985; Schmidt, 1987; Shizuka, 1993). The second examines the acquisition of morphology (Larsen-Freeman, 1975; Kameyama, 1987; Tomita, 1988; Inoi, 1991; Takamiyagi, 1991). The third discusses the

acquisition of syntax (Bailey, Eisenstein & Madden, 1976; Schmidt, 1980; Hueber, 1985). The results from these studies have supported the idea that the accuracy rate of language learners' performance changes according to the *Interlanguage Continuum Model,* though some morphological features such as articles have been shown not to follow it.

Recently, Ohba (1994b) investigated the effect of different test types on the interlanguage performance of Japanese EFL learners. The noteworthy points in this study were:

1. the use of a larger number of participants (N=370) in order to generalize the findings to a statistical population;
2. the division of participants by proficiency level (higher, average, lower) from scores on the *STEP* placement test (Obunsha, 1987, 1989) to determine whether higher-level learners show differences in accuracy rates;
3. the use of three test-types a) Grammaticality Judgment, b) Sentence Combining, and c) Picture Description;
4. the selection of a complex syntactic structure, relative clauses, to examine the effect of test-types because accuracy rates for morphology such as articles are easily influenced by discourse (Long & Sato, 1984; Tarone, 1985; Tarone & Parrish, 1988; Ohba, 1994a, 1994b), making it difficult to determine if the difference affects interlanguage performance.

Though Ohba (1994a, 1994b) made a significant contribution to this field, there were, nevertheless, a few drawbacks to his research. First, since different sentences were used for each test, it cannot be concluded that the scores were affected only by test-type. Second, in the Picture Description test, learners were asked to produce subject-type relativised sentences for sentences containing relative pronouns because these are considered easiest for L1 and L2 learners of English (Schachter, 1974; Keenan & Comrie, 1977; Gass, 1980) and avoided other kinds of relative clauses. The other test-types, Grammaticality Judgment test and Sentence Combining test, consisted of a mixture of sentences with four locations of the head noun phrase to be relativised: subject, direct object, object of preposition, and possessive. Therefore, though the Picture Description test was classified as not requiring attention to linguistic form, the characteristics of relative clauses may have resulted in higher than expected accuracy rates. Third, though Ohba said the low time pressure might have activated the learners'

monitor in the Picture Description test to explain the accuracy rate, Krashen (1985) said that time pressure may be unrelated to monitor activation. Therefore, it is difficult to view lack of time pressure as the reason for the increased accuracy rate.

In response to these concerns, the effects of test-types alone on learners' interlanguage performance need to be examined.

The Study

Purpose and Research Questions

The purpose of the present study was threefold: first, to replicate and expand Ohba's (1994a, 1994b) findings under more strictly controlled conditions; second, to examine the reliability coefficients of the three tests, and third, to determine what the grammatical judgment test employed by Ohba examines.

It was expected that the participants' interlanguage performance would vary according to Hyltenstam's (1983) theoretical framework.

Grammaticality judgment tests tend to be criticized (Ohba, 1994b), since some researchers (Ellis, 1991, 1994; Chaudron, 1983) are skeptical of their reliability. Moreover, some researchers question what grammaticality judgment tests measure in comparison to tests with different test-types. In spite of such criticism, grammaticality judgment tests are widely used because they are easy to construct and are believed to be a useful tool for eliciting learners' linguistic knowledge (Ohba, 1994b). To address these concerns, a grammaticality judgment test constructed on the basis of findings concerning relative clauses (Gass, 1980; Kawauchi, 1988) was used to examine the reliability. In order to determine what the grammaticality judgment test really examines, the correlations between the grammaticality judgment test and other tests, which target the same trait, but with different test-types, were calculated.

The research questions are:

1. Does the accuracy rate of the three tests follow the pattern of: Cloze > Grammaticality Judgment > Sentence Combining?
2. What does the Grammaticality Judgment measure? Is this test's reliability low?

The small sample ($N=41$) necessitated a conservative treatment of statistical analyses. Therefore, the alpha level for all statistical decisions was set at $a<0.01$.

Method

Subjects: The subjects in this study (N=41) were second year high school students (10th grade) enrolled in an English reading class at a high school attached to the Aichi University of Education. All were native speakers of Japanese. The average age was 16. All had completed at lease four years of formal English courses. The sample was thus homogeneous with regard to nationality, language background, educational level, and age. The group consisted of 21 males and 20 females. One male student was absent throughout this study. Though he later took the tests, his scores were not included.

In general, Japanese students are required to learn the usage of relative clauses: subject (SU), direct object (DO), indirect object (IO), and genitives (GEN), in junior high school, and relative clauses of preposition (OPrep) in high school. Therefore, it was concluded that the Ss had basic ability regarding use and comprehension of the English relative clause.

Materials: The following three 24-item relative clause tests were administered: 1) Cloze, 2) Grammaticality Judgment, and 3) Sentence Combining. These tests are a modification of the tests in Ohba (1994b) (see Appendix). To prevent use of only the easiest type of relative clause, subject type (SU), the same sentences were used for each test and the Picture Description test was replaced by the Cloze.[6]

In the process of constructing the three relative clause tests, careful attention was paid to the Noun Phrase Accessibility Hierarchy: SU (subject) > DO (direct object) > IO (indirect object) > OPrep (object of preposition) > GEN (possessive) > OComp (object of comparative particle) (Keenan & Comrie, 1977). The location of the head noun phrase is considered an influential component in the degree of difficulty associated with relative clauses in both L1 and L2 English acquisition (c.f. Schachter, 1974; Keenan & Comrie, 1977; Gass, 1980; Eckman, Bell, & Nelson, 1988; Akagawa, 1992; Sadighi, 1994; Aarts & Schils, 1995).

The tests: The 24 questions in the Sentence Combining test contained six pairs of sentences to be combined into sentences containing a relativised SU, six into sentences containing a relativised DO, six into sentences with a relativised OPrep, and six into sentences containing genitive cases. In the Cloze test, an appropriate relative pronoun, who, which, whose, whom, was required. In the Grammaticality Judgment test, a determination of each sentence's grammaticality, using either "O" or "X" as markers for correctness and incorrectness respectively, was required. For sentences judged incorrect, Ss were asked to make

necessary corrections.Typical errors in relative clauses are universal. They are categorized as: 1) relative clause marker omission, 2) pronoun retention, 3) wrong selection of relative clause marker, and 4) adjacency (Gass, 1980; Kawauchi, 1988). There were 12 correct and 12 incorrect sentences.

Test administration: In response to the shortcomings of recent studies, the ordering of the three test papers was taken into consideration. Since in the three tests the same sentences appear, the possibility of Ss memorizing the orthography to gain a higher score in later tests was considered. In order to reduce the potential order effect, the 24 items in each test were divided into three groups. Eight items from each of the tests were combined to make a 24-item test. In the first session, Cloze items 1-8, Grammaticality Judgment items 9-16, and Sentence Combining items 17-24 appeared; in the second session, Grammaticality Judgment items 1-8, Sentence Combining items 9-16, and Cloze items 17-24; and in the third session, Sentence Combining items 1-8, Cloze items 9-16, and Grammaticality Judgment items 17-24. Ss were allowed 20 minutes to complete each test. There was a one week interval between testing sessions, assumed to be long enough for Ss to forget some of the orthography and decrease the negative order effect. Ss were not told the dates of the sessions. Each session was conducted at the beginning of English reading classes by the instructor and his assistant. Though Ss were not informed of the purpose of the tests, they were encouraged to answer as many questions as possible. It is noteworthy that the Ss showed a great deal of interest on all the tests.

Scoring procedure: The tests were scored by the author. Scoring was based on whether the Ss had used the appropriate pronouns following an established criterion (Celce-Murcia & Larsen-Freeman, 1983; Quirk, Greenbaum, Leech, & Svartvik, 1985; Ohba, 1994a, 1994b). Therefore, local errors such as spelling mistakes were ignored as long as the meaning was clear.

Reliability estimation based on internal consistency: The Spearman-Brown split-half method was used to estimate the tests' reliability coefficients. The split-half method can be used when each test item is regarded as independent and also can contribute to the total score independently (local independence). In all three tests, each item is clearly independent. Thus, the split-half method is a permissible estimating procedure. The author scored the odd- and even-numbered items separately and first examined the Pearson's product-moment correlation (r) (Ito, 1996; Brown, personal communication, February 22, 1996). Each value was

then corrected for the reduction to half-test length using the Spearman-Brown prophecy formula ($rxx' = 2r/1+r$).

Results and Discussion

In this section, descriptive statistics of the tests are shown and the research questions addressed. Before discussing the results, however, two other aspects must be considered:

1. the effects of sample homogeneity on reliability and estimated correlations among the three tests, and
2. the small size of the sample ($N=41$).

Table 1: Reliability Coefficient, Mean, Maximum Score & Standard Deviation of Tests ($N=41$)

Tests	Reliability *rxx'*	Mean (*M*)	Max. Score	*SD*
CL	0.596	14.976	24	4.891
GJ	0.797	12.976	24	3.309
SC	0.693	10.732	24	6.490

CL = Cloze test; GJ = Grammaticality Judgment test; SC = Sentence Combining test.

Table 2: Analysis of Variance Summary Table ($N=41$)

Source	*SS*	*df*	*MS*	*F*-ratio	*p*
Subjects	2333.659	40	58.341		
Test-Type	498.260	2	249.124	24.215	<0.000
Error	823.073	80	10.288		
Total	3654.073	122			

Table 3: Multiple comparison test (Ryan's method) summary table ($N=41$)

Pair	*r*	Nominal Level	*t*	*p* level	Significance
CP-SC	3	0.003	6.920	0.000	s.
CP-GJ	2	0.007	2.823	0.005	s.
GJ-SC	2	0.007	4.097	0.000	s.

CL = Cloze test; GJ = Grammaticality Judgment test; SC = Sentence Combining test.

1. Does the accuracy rate of the three tests follow the pattern of: Cloze > Grammaticality Judgment > Sentence Combining?

Table 1 shows that in the relative clause tests the Cloze test had the highest mean ($M = 14.976$), the Grammaticality Judgment test a lower mean ($M = 12.976$), and the Sentence Combining test the lowest ($M = 10.732$). The accuracy rate seems to have changed according to the expected order. In order to determine statistical significance, one-way analysis of variance was performed. Table 2 shows an overall significant difference in the three test scores ($F(2,80) = 24.833$, $p < 0.000$). A multiple comparison test using Ryan's method was performed to determine where the difference lay. Table 3 shows it existed in each pair of the three tests at $p < 0.01$. Therefore, hypothesis 1 was confirmed.

2. What does the Grammaticality Judgment measure? Is this test's reliability low?

Table 1 shows the reliability coefficients of the three relative clause tests. Unexpectedly, the Grammaticality Judgment test showed the highest reliability among the three tests ($r = 0.798$), with reliability high enough for it to be regarded as a reliable testing device. The Cloze test ($r = 0.597$) and the Sentence Combining test ($r = 0.694$) showed moderate reliability. Ranked from highest to lowest in reliability coefficients, the reliability was: Grammaticality Judgment > Sentence Combining > Cloze.

Table 4: Correlations between each pair of three tests ($N = 41$)

Tests	r (exploratory rate %)	p
CP & GJ	0.626 (39.189)	<0.01
GJ & SC	0.696 (48.442)	<0.01
SC & CP	0.698 (48.720)	<0.01

CL = Cloze test; GJ = Grammaticality Judgment test; SC = Sentence Combining test.

Results indicate that a Grammaticality Judgment test can be relatively reliable. However, validity also needs to be investigated. Table 4 displays the correlation for each pair of three tests. The correlations measured show almost the same magnitudes: $0.60 < r < 0.70$. Results reveal that each pair of tests shared the same amount of trait or ability needed for producing or comprehending relative clauses (exploratory rate or coefficients of determination: $r^2 = 39.189$ to 48.720%). [The square of

value of r (r^2) *100 indicates what percentage of similar traits or abilities each pair of tests share]. The Grammaticality Judgment test shows a relatively high reliability coefficient and moderate correlation with the other two test-types.

Though Grammaticality Judgment tests are still controversial with regard to reliability and validity, this research indicates they have far-reaching potential as reliable and valid elicitation tools. However, test designers must be aware of the universal error types or typical errors from learners' L1 transfers in order to construct appropriate tests. In addition, Table 1 reveals that the Grammaticality Judgment test shows the lowest standard deviation here, implying it has limited discriminative ability.

Conclusions and Pedagogical Implications

The results of this study indicate that, for Japanese EFL high school students: 1) the accuracy rate follows the expected pattern (Cloze > Grammaticality Judgment > Sentence Combining); and 2) unexpectedly, the Grammaticality Judgment showed fairly high reliability, with moderate correlations between the two other test-types. However, since its discriminative ability seems limited, it should be used with extreme care.

In a pedagogical sense, the results indicate that the manifestation of learners' interlanguage competence, their performance, varies according to test-type. As a consequence, teachers may well underestimate or overestimate learners' knowledge or ability to use a grammatical item if they rely on only one test-type. Moreover, in order to characterize the learners' actual abilities in the target language, it is necessary to employ a variety of test-types.

Limitation and Suggestions for Further Research

It should be acknowledged that one of the limitations of the present investigation is that it focused only on the Ss' performance. Thus, the results can be generalized only for Japanese students. However, in many studies conducted in the past, various language backgrounds, ages, and educational backgrounds were mixed. As a result, the findings have often been hard to interpret because they can only be generalized to the single situation in which the data was collected. In addition, the results of this study may be influenced by two internal and external factors:

1. the nature of reliability in measures in general, and
2. restrictions in the range of ability that was sampled in the investigation.

Generally, tests are not simply reliable and valid but they can be reliable and valid for specific types of students and specific ranges of ability (Brown, personal communication, February 22, 1996). In this regard this research should be replicated with a larger sample of participants from a much wider population.

The following three general research questions are posed in the hope that other researchers will pursue further investigations.

1. Does the proficiency level affect the magnitude of inter-language variability with regard to accuracy rates? If so, does the degree of the variability in the target language decrease as proficiency level increases (i.e. higher level learners < average level learners < lower level learners?).[7]
2. What really causes the difference in accuracy rates? The amount of attention to linguistic form by monitoring? Or the difference in cognitive processes and demands required of subjects?
3. How high is the construct validity of each test? That is, does each of the tests measure what it is constructed to measure?

Acknowledgments

The author sincerely thanks Norihisa Matsumoto, Shogo Miura, Yoshito Nakamura, Tadashi Nishida, Nobukazu Matsu-ura, Hiroshima University, and J.D. Brown, University of Hawaii, for their valuable critical comments on earlier drafts of this paper. Thanks also go to Hiromasa Ohba, Health Sciences University of Hokkaido; Kenji Ohtomo, University of Tsukuba; and Susan Gass, Michigan State University for their communications on this topic, as well as Toshifumi Takizawa, Takao Imai, Yuya Koga, Shizuya Tara, Hiroyuki Araki, Hidetoshi Inoue, and Fumiyo Yoshikawa, for their encouragement and cooperation, and two anonymous reviewers of the JALT Journal *for their valuable comments. I alone, however, am responsible for the analysis herein.*

Akihiro Ito, M.A., Hiroshima University, is currently a Ph.D. candidate in the Department of English Language Education, Hiroshima University. His research interests include language testing, evaluation, and second language acquisition research methodology.

Notes

1. Throughout this paper, "test-type" refers to any task type. Tasks for data elicitation are widely used in language testing conditions. Terms such as test format, test method, or method facet can refer to the concept of test-type used here since there has been a great amount of variability in classification.
2. Following Bachman's (1990) definition, "performance" is used as an indicator of a long-standing ability or competence, a person's knowledge of the language, which can be estimated indirectly by a test score and its valid interpretation (p. 33). On the basis of Bachman's definition, I take the position that competence is basically homogeneous and unitary, and that interlanguage variability manifested in different test-types is essentially a phenomenon of performance. The variability of performance can be observed by examination of change in accuracy rate or score on tests.
3. Some language testing and/or curriculum experts (Brindley, 1989) claim that the distinction between proficiency and achievement is not clear-cut. However, in my view, an achievement test is given to language learners at the end of a program to check if they have mastered the targeted items or skills. In this sense, an achievement test is, as Alderson, Clapham, and Wall (1995, pp. 11-12) argue, similar to a progress test. Proficiency tests are not based on language programs or classes. The main purpose of proficiency tests is to examine overall language ability.
4. For overviews of research on task variation theory, see Ohba (1987, 1994b).
5. Tarone's (1983) categorization of elicitation tools is comprised of tasks which are more differentiated for Hyltenstam (1983). Other test-types such as cloze tests or translation test-types in Hyltenstam's classification of experimental data can be located on a continuum with regard to accuracy rate, though some must be placed in the careful-style area.
6. Some might question if the cloze test-type in Hyltenstam (1983) allows higher accuracy than grammaticality judgment tests or sentence combining tests. DeKeyser (1990), however, argues that the fill-in-the-blank test-type, which is similar to the cloze, works as a more valid measure of monitored knowledge. He discusses the effectiveness of the fill-in-the-blank test-type for examining monitored knowledge compared to the other test-types. In this regard, my interpretation of Hyltenstam's theoretical framework is a combination of Tarone's (1983) interlangauge continuum model and DeKeyser's proposals.
7. Reexamination of the data collected to determine the effects of proficiency level on the magnitude of variability of accuracy rates in participants' interlanguage performance is currently underway.

References

Aarts, F., & Schils, E. (1995). Relative clauses, the accessibility hierarchy and the contrastive analysis hypothesis. *International Review of Applied Linguistics, 33*, 47-63.

Akagawa, Y. (1992). Avoidance of relative clause by Japanese high school students. *JACET Bulletin, 23*, 1-18.

Alderson, J.C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Bachman, L.F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bailey, N., Eisenstein, M., & Madden, C. (1976). The development of wh-questions in Adult second language learners. In Faneslow, J., & Crymes, R. (Eds.), *On TESOL '76* (pp. 1-9). Washington, DC: TESOL.

Beebe, L.M. (1980). Sociolinguistic variation and style shifting in second language acquisition. *Language Learning, 30,* 178-194.

Brindley, G. (1989). *Assessing achievement in the learner-centered curriculum.* Sydney: National Centre for English Language Teaching and Research.

Celce-Murcia, M., & Larsen-Freeman, D.E. (1983). *The grammar book: An ESL/EFL teacher's course.* Rowley, MA: Newbury House.

Chaudron, C. (1983). Research on metalinguistic judgments: A review of theory, methods, and results. *Language Learning, 33,* 343-377.

DeKeyser, R. (1990). Towards a valid measurement of monitored knowledge. *Language Testing, 7,* 147-157.

Dickerson, L.J. (1975). The learner's interlanguage as a system of variable rules. *TESOL Quarterly, 9,* 401-407.

Dulay, H., Burt, M., & Krashen, S.D. (1982). *Language two.* Oxford: Oxford University Press.

Eckman, F.R., Bell, L., & Nelson, D. (1988). On the generalization of relative clause instruction in the acquisition of English as a second language. *Applied Linguistics, 9,* 1-20.

Ellis, R. (1991). Grammar judgments and second language acquisition. *Studies in Second Language Acquisition, 13,* 161-186.

Ellis, R. (1994). Data, theory, and applications in second language acquisition research. In Ellis, R. (Ed.), *The study of second language acquisition* (pp. 669-691). Oxford: Oxford University Press.

Gass, S. (1980). An investigation of systematic transfer in adult second language learners. In Scarcella, R.C., & Krashen, S.D. (Eds.), *Research in second language acquisition* (pp. 132-141). Rowley, MA: Newbury House.

Hueber, T. (1985). System and variability in interlanguage syntax. *Language Learning, 35,* 141-163.

Hyltenstam, K. (1983). Data type and second language variability. In Rongbon, H. (Ed.), *Psycholinguistics and foreign language learning* (pp. 57-74). Abo (Turku), Finland: Abo Akademi.

Inoi, S. (1991). Variation in interlanguage with special reference to articles and pronouns. *Annual Review of English Language Education in Japan, 2,* 1-10.

Ito, A. (1996). Testing English tests: A language proficiency perspective. *JALT Journal, 18,* 183-197.

Ito, A. [in press]. A study on the variability of test performance of Japanese EFL learners: A combination of two theoretical frameworks. *CELES Bulletin, 26.*

Kameyama, T. (1987). *Bunpo tesuto niokeru tesuto keishiki niyoru keitaiso seitoritsu no henka* (The change of accuracy rate in grammar tests by the

difference of test-types). *CELES Bulletin, 17*, 248-252.
Kawauchi, C. (1988). Universal processing of relative clauses by adult learners of English. *JACET Bulletin, 19*, 19-36.
Keenan, E.L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry, 8*, 63-99.
Krashen, S.D. (1977a). The monitor model for adult second language performance. In Burt, M., Dulay, H. & Finocchiaro, M. (Eds.), *Viewpoints on English as a second language* (pp. 152-161). New York: Regents.
Krashen, S.D. (1977b). Some issues relating to the monitor model. In Brown, D.H., Yorio, C.A., & Crymes, R.H. (Eds.), *On TESOL '77* (pp. 144-158). Washington, DC: TESOL.
Krashen, S.D. (1978a). Individual variation in the use of the monitor. In Ritchie, W.C. (Ed.), *Second language acquisition research: Issues and implications* (pp. 175-183). New York: Academic Press.
Krashen, S.D. (1978b). The monitor model for second language acquisition. In Gingras, R.C. (Ed.). *Second language acquisition and foreign language teaching* (pp. 1-26). New York: Center for Applied Linguistics.
Krashen, S.D. (1981). *Second language acquisition and second language learning*. Oxford: Pergamon Press.
Krashen, S.D. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon Press.
Krashen, S.D. (1985). *The input hypothesis: Issues and implications*. London: Longman.
Krashen, S.D. & Terrell, T. (1983). *The natural approach: Language acquisition in the classroom*. Oxford: Pergamon Press.
Larsen-Freeman, D.E. (1975). The acquisition of grammatical morphemes by adult ESL students. *TESOL Quarterly, 9*, 409-419.
Long, M.H., & Sato, C.J. (1984). Methodological issues in interlanguage studies: An interactionist perspective. In Davies, A., Criper, C., & Howatt, A.P.R. (Eds.), *Interlanguage* (pp. 253-279). Edinburgh: Edinburgh University Press.
Negishi, M. (1995). *Ri-dingu no tesuto to hyoka* (Reading test and evaluation). *Eigokyoiku (The English Teachers' Magazine), 1*, 29-31.
Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.
Ni kyu yosou mondaishu (Collections of Second Grade STEP Tests). (1987). Tokyo: Obunsha.
Ni kyu yosou mondaishu (Collections of Second Grade STEP Tests). (1989). Tokyo: Obunsha
Ohba, H. (1987). *Tasuku barieshun riron ni tsuite* (On the task variation theory). *CELES Bulletin, 17*, 43-49.
Ohba, H. (1994a). *Nihonjin eigo gakushusha no chukangengo kahensei: Tasuku keishiki no kanten kara* (Japanese EFL learners' interlanguage variability: With reference to task-type). *CELES Bulletin, 24*, 187-192.
Ohba, H. (1994b). *Tesuto keishiki no chigai niyoru eigo gakushusha no pafomansu no kahensei* (Task-related variability in interlanguage by Japanese EFL learn-

ers). *STEP Bulletin, 6*, 34-48.
Powers, D.E. (1982). Selecting samples for testing the hypothesis of divisible versus unitary competence in language proficiency. *Language Learning, 32*, 331-335.
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language.* London: Longman.
Sadighi, F. (1994). The acquisition of English relative clauses by Chinese, Japanese, and Korean adult native speakers. *International Review of Applied Linguistics, 32*, 141-153.
Sajjadi, S. & Tahririan, M.H. (1992). Task variability and interlanguage use. *International Review of Applied Linguistics, 30*, 35-44.
Sato, C.J. (1985). Task variation in interlanguage phonology. In Gass, S., & Madden, C.G. (Eds.), *Input on second language acquisition* (pp. 181-196). Rowley, MA: Newbury House.
Schachter, J. (1974). An error in error analysis. *Language Learning, 24*, 205-214.
Schmidt, M. (1980). Coordinate structures and language universals in interlanguage. *Language Learning, 30*, 397-416.
Schmidt, R.W. (1987). Sociolinguistic variation and language transfer in phonology. In Ioup, G., & Weinberger, S.H. (Eds.). *Interlanguage phonology* (pp. 365-377). Rowley, MA: Newbury House.
Shizuka, T. (1993). Task variation and accuracy predictors in interlanguage phonology production. *Bulletin of Kanto-Koshin-etsu English Language Education Society, 7*, 63-77.
Takamiyagi, T. (1991). A study of task related variation in interlanguage by Japanese university students in an instruction only environment. *Joetsu University of Education: Kenkyu Ronshu (Bulletin of Language Studies), 6*, 47-64.
Tarone, E. (1979). Interlanguage as chameleon. *Language Learning, 29*, 181-191.
Tarone, E. (1982). Systematicity and attention in interlanguage. *Language Learning, 32*, 69-84.
Tarone, E. (1983). On the variability of interlanguage systems. *Applied Linguistics, 4*, 142-164.
Tarone, E. (1985). Variability in interlanguage use: A study of style-shifting in morphology and syntax. *Language Learning, 35*, 373-403.
Tarone, E. (1988). *Variation in interlanguage.* London: Edward Arnold Publishers.
Tarone, E. & Parrish, B. (1988). Task-related variation in interlanguage: The case of articles. *Language Learning, 38*, 21-44.
Tomita, Y. (1988). *Nihonjin kokosei no chukan gengo nitsuite no ichikousatsu: Keitaiso shutoku junjo kenkyu* (A study on the high school students' interlanguage: With reference to the acquisition order of morphologies). *CELES Bulletin, 18*, 208-213.
Weir, C.J. (1993). *Understanding and developing language tests.* London: Prentice-Hall International.

(Received Jan. 3, 1996; revised Oct. 7, 1996)

Appendix:Test sentences

Cloze

1. The policeman has caught the girl ( ) stole the car.
2. The author ( ) books I haven't read yet is well known.
3. The magazine from ( ) I got the information is Newsweek.
4. He still had the pen ( ) I had given to him.
5. I came across some students ( ) names I couldn't remember.
6. She paid the man from ( ) she had borrowed the money.
7. The wine ( ) you brought to our party was excellent.
8. You will receive the kind letter ( ) your mother wrote.
9. The book to ( ) I referred can be obtained from the library.
10. The girl ( ) is sitting at the reception desk is pretty.
11. The city from ( ) that boy came is far from here.
12. I met a friend ( ) mother was a famous designer.
13. The building ( ) stands near the lake is our hotel.
14. The man ( ) feet were very large has just bought new shoes.
15. I need someone ( ) can help me clean the house.
16. The woman ( ) John will marry next month is Japanese.
17. The man ( ) was injured in the accident is in the hospital.
18. I saw a man ( ) bag was the same as mine.
19. I have just found the key ( ) I lost yesterday.
20. The book ( ) I bought the other day is interesting.
21. The woman with ( ) he was talking was Mrs. Miller.
22. The police interviewed the lady from ( ) the diamonds had been stole.
23. The boy ( ) essay I corrected has entered Hokkaido University.
24. I know the children ( ) are playing in the yard.

Grammaticality Judgment

1. ( ) The policeman has caught the girl stole the car.
2. ( ) The author whose books I haven't read yet is well known.
3. ( ) The magazine which I got the information from it is Newsweek.
4. ( ) He still had the pen which I had given to him.
5. ( ) I came across some students names I couldn't remember.
6. ( ) She paid the man from whom she had borrowed the money.
7. ( ) The wine was excellent which you brought to our party.
8. ( ) You will receive the kind letter whose your mother wrote.
9. ( ) The book to which I referred can be obtained from the library.
10. ( ) The girl who is sitting at the reception desk is pretty.
11. ( ) The city is far from here which that boy came from.
12. ( ) I met a friend which mother was famous designer.
13. ( ) The building which stands near the lake is our hotel.
14. ( ) The man has just bought shoes whose feet were very large.

15. ( ) I need someone who can help me clean the house.
16. ( ) The woman whom John will marry next month is Japanese.
17. ( ) The man is in the hospital who was injured in the accident.
18. ( ) I saw a man whose bag was the same as mine.
19. ( ) I have just found the key which I lost yesterday.
20. ( ) The book which I bought it the other day is interesting.
21. ( ) The woman with whom he was talking was Mrs. Miller.
22. ( ) The police interviewed the lady from which the diamonds has been stolen.
23. ( ) The boy whose essay I corrected has entered Hokkaido University.
24. ( ) I know the children which are playing in the yard.

Sentence Combining

1. The policeman has caught the girl. She stole the car.
2. The author is well known, I haven't read his books yet.
3. The magazine is Newsweek. I got the information from it.
4. He still had the pen. I had given it to him.
5. I came across some students. I couldn't remember their names.
6. She paid the man. She had borrowed the money from him.
7. The wine was excellent. You brought it to our party.
8. You will receive the kind letter. Your mother wrote it.
9. The book can be obtained from the library. I referred to it.
10. The girl is pretty. She is sitting at the reception desk.
11. The city is far from here. The boy came from it.
12. I met a friend. His mother was a famous designer.
13. The building is our hotel. It stands near the lake.
14. The man has just bought new shoes. His feet were very large.
15. I need someone. Someone can help me clean the house.
16. The woman is Japanese. John will marry her next month.
17. The man is in hospital. He was injured in the accident.
18. I saw a man. His bag was the same as mine.
19. I have just found the key. I lost it yesterday.
20. The book is interesting. I bought it the other day.
21. The woman was Mrs. Miller. He was talking with her.
22. The police interviewed the lady. The diamonds had been stolen from her.
23. The boy had entered Hokkaido University. I corrected his essay.
24. I know the children. They were playing in the yard.