Examining the reliability of processability theory-based procedure for use in Japanese SLA assessment

Judith Preston Ishigami

Abstract

This study explains the method of examining the interrater reliability of a procedure for assessing the stage of acquisition of Japanese as a second language (JSLA) of adult learners of Japanese. The assessment model was based on an application of processability theory (PT) of grammar acquisition introduced by Pienemann (1998) and developed for Japanese SLA by Preston (2004, 2008). The assessment targeted the emergence of 12 morphosyntactic procedures in the spontaneous speech samples of second language (L2) learners of Japanese.

After designing a brief, task-based elicitation procedure to collect speech samples from eleven JSL learners residing in Japan, the researcher performed a detailed linguistic investigation of the grammar in the speech samples based on distributional analyses. Next, a workshop on processability theory was offered to adults enrolled in language pedagogy courses in a large city in Japan. The participants listened to audio recordings of eight L2 Japanese learners and simultaneously judged the stage of acquisition of each sample based on the 12 target Japanese grammar criteria derived from the PT study. Results indicated a strong correlation between the rater assessments and the results of the PT-based detailed linguistic research analyses (T_c=.919, $p \leq .001$).

1. Introduction

The language processing model of second language acquisition introduced by Pienemann (1998) provides a theoretical framework for predicting the order of acquisition of grammatical processing skills by L2 learners. The theory makes specific claims regarding the developmental path of L2 grammar, stating that the sequence of acquisition of the target language grammar is inherently constrained by the implicational nature of grammar processing procedures that are essential to real-time human language production.

An understanding of the rationale and nature of language processing theory is crucial to the design of an assessment procedure based on PT's implicational hierarchy. This point was emphasized to the participants in the interrater reliability study below, who in preparation for using JSLA assessment procedure completed approximately two hours of workshop training in PT.

2. Overview of the language processing approach to SLA

Second language acquisition research using a theory of language processing seeks to explain the developmental path of L2 learners based on a cognitive framework of the automaticization of procedural skills (McLaughlin, 1987; Pienemann, 1998, 2005). On this view, second language learning requires the acquisition of a complex skill (or procedure) made up of a number of sub-skills. It is the automaticization of these skills that is presumed responsible for language acquisition by the L2 learner, as well as responsible for the production of well-formed, fluent speech characteristic of adult native speakers of a language.

The theory that a model of language processing involving a hierarchy of grammar procedures could predict the order of acquisition of the target language grammar by L2 learners was expounded in Pienemann (1998). Processability theory has since been applied in research on SLA by adult language learners of typologically diverse languages such as English (Pienemann, 1998, 2005), German (Pienemann, 1998), Swedish (Hakansson, 2002), Arabic (Mansouri, 2002), Italian (DiBiase & Kawaguchi, 2002) Chinese (Zhang, 2002), and Japanese (Kawaguchi, 2002; Preston, 2008).

In the majority of the studies above, the focus has been on examining the validity of using PT's implicational hierarchy of procedural grammar to predict target grammar acquisition orders as evidenced in longitudinal and cross-sectional empirical data collected from L2 speakers. However, that processability theory may be used to inform pedagogical considerations regarding language learning and teaching is undoubtedly its most practical and therefore valuable potential. This aspect provides impetus to the study here, which developed a PT-based assessment protocol for use in measuring JSLA based on grammatical production in spontaneous speech.

2.1 Processability theory and procedural grammar

The theory of language processing used in this study relates to the incremental procedural grammar (IPG) developed by Kempen and Hoenkamp (1987), who invented a cognitive model for formulating output of spontaneous speech (Dutch) using semantic-based lexical input. IPG's processing hierarchy was adopted to explain the order of acquisition of L2 grammar by Pienemann (1998), who tested empirical data from English and German SLA studies against predictions he derived from the implicational nature of the procedural grammar hierarchy. The nature of the procedural grammar hierarchy also framed the JSLA profiling study by Preston (2008), who posited four stages of grammar processing for Japanese and investigated their implicational relationship based on the emergence of twelve grammatical criteria in speech samples collected from adult learners of Japanese.

In the implementation of the theoretical framework in SLA research, a key argument for PT rests on the assumption that the order of emergence of target language grammar forms is evidence of an underlying hierarchical sequence of grammar processing procedures. The implicational relationship of these procedures derives from the fact that the output of lower-level (or stage) procedures serves as input to higher-level ones. In this way the "value return hierarchy" of IPG (Kempen & Hoenkamp, 1987) is related to the procedural stages of SLA: Acquisition of a lower-stage grammar skill is prerequisite for acquisition of the next higher-stage one. Consequently, no stage in the grammatical hierarchy can be skipped during the L2 acquisition process. A concise summary of the consequences of adopting a grammar processing hierarchy-based approach is made by Pienemann (1998):

"The implicational nature of the hierarchy derives from the fact that the processing procedures developed at one stage are a necessary prerequisite for the following stage: A word needs to be added to the L2 lexicon before its grammatical category can be assigned. The grammatical category of a lemma is needed before a category procedure can be called. Only if the grammatical category of the head of phrase is assigned can the phrasal procedure be called. Only if the latter has been completed and its value is returned can Appointment Rules determine the function of the phrase after which it be attached to the S-node. Only after appointment Rules are refined by 'Lemma functions' can subordinate clauses be formed—with their own structural properties." Pienemann, 1998, p. 87 (emphasis removed)

In short, any implementation of the theory should be based on the definition of the abstract grammatical procedures underlying production of a variety of language-specific target grammar forms representative of each stage of L2 development. The researcher can categorize language-specific morphosyntactic features into sets of target grammar procedures representative of the abstract stages.

Although detailed discussion of the theory is beyond the scope of this paper, some important assumptions of the PT approach to SLA should be stated now to help clarify the implementation of PT in second language acquisition research.

i. The procedural grammar hierarchy, inherent in processing the target language, constrains the order of acquisition of the L2 grammar regardless of both the learner's native language and the second (or other) language of study.

ii. The language processing hierarchy described for SLA by Pienemann (1998) addresses only the developmental path of L2 *grammar* in *spontaneous speech*. The relationship between acquisition of the grammar acquisition of other processing modules (such as phonology and semantics) is relevant to the formulation of a total theory of language acquisition. However, because each acquisition module has a distinct nature, an acquisition order predicted by any additional modules (if and when they are established) should not contradict the acquisition order predicted by the grammar processing model.

iii. Factors such as learner motivation, instructional method and feedback, and frequency and mode of exposure of the L2 likely effect the rate of grammatical or other module acquisition by learners. However, such factors cannot alter the fundamental order of acquisition of grammatical processing procedures.

2.2 Language processing and JSLA

In constructing a profiling procedure using a PT model to determine the stage of acquisition of Japanese by adult L2 learners, Preston (2008) defined the language processing hierarchy for Japanese and then selected 12 Japanese morphosyntactic features to represent four stages of a grammar processing in JSLA. The procedural stages and their associated target grammar criteria are presented in Table 1.

Table 1. Hierarchy of JSL grammar processing procedures. (cf. Preston, 2008: 137).

Stage	Procedure	Target criteria
4	Sentential: Exchange of grammatical information between clausal (S) head and its sisters[1]	Emergence of SUBJ and OBL grammar functions (marked -GA and -NI, respectively) in obligatorily S contexts, evidenced by morphosyntactically headed S: 1) -(RA)RE (passive), 2) -(SA)SE (causative), 3) -TE YARI/MORAI (benefactive auxiliary)
3	Phrasal: Grammatical processing of NP and VP head; canonical processing of complex event paths; modification of within- and between-canonical event relationships	1) Emergence of 3-argument canonical structures (AGERU, MISERU) and event-modifying words (DEMO, -KEDO, -GA) and nouns (KINOU, KORE, TOKIDOKI) 2) Use of -TE form on verbs preceding verbs 3) Use of -NO on nouns preceding modified nouns 4) Canonical structures adjoined by TOKI (ATO, MAE)
2	Categorial: Category-based canonical constructions; semantic-based case on N category	1) Canonical N N V structures 2) Semantic -TO conjuncts two nouns 3) Semantic -NO to mark possessive relationships
1	Memorized formulae; Words	1) Production of repetitive lexical context in use of any of the above criteria above 2) Single words, word strings lacking morphosyntactic features necessary for determining semantic relationships between words

[1] See Bresnan (2001: 109-112) for a discussion of features of exocentric (small-clause), S.

Targeting production of the 12 grammar items in Table 1, a task-based elicitation procedure was designed and used in interviews to collect spontaneous speech samples from 42 adult learners of Japanese. The interviews were transcribed, and a distributional analysis of the emergence of each item in each sample was performed. An implicational analysis of the results of emergence was used to test the hierarchy's validity. Results showed no evidence against the order of emergence defined by the hierarchy.

In every learner sample collected, one or more of the 12 target criteria emerged. Moreover, the emergence of just one criterion belonging to the highest stage in the PT/JSLA hierarchy predicted correctly that at least one target criterion of every lower-stage procedure was also present in the sample. This provided theoretical rationale for using the target criteria to develop a real-time, rapid profiling procedure that could be used by Japanese language educators and researchers to perform JSLA-based assessments in an educational setting.

2.3 Language profiling and theoretically-based SLA assessment

Traditionally, linguistic profiling takes place in two stages: 1) the elicitation of the spontaneous speech sample and 2) the analysis of its contents for evidence of target criteria. The evolution of profiling from a strictly empirical approach, to its use in a general language processing approach, to its application in the grammar-based PT model is directly related to both the theory's rationale and the analytic model introduced in section 3.

The linguistic profiling method was originally introduced as a diagnostic by Crystal, Fletcher and Garman (1976), who investigated the production of English syntax by native (L1) speakers with mental disabilities. Crystal et al. (1976) designed a task-based interview procedure to elicit approximately 30-minute speech samples from their subjects. They then performed a detailed linguistic analysis of each sample to examine the forms and number of times of emergence of each form in each sample. Based on the implicational pattern of emergent syntactic forms in the samples, each speaker could be assigned a final score according to the highest-ranked (stage 1-7) criteria emerging in their sample.

The profiling methodology was modified for used in a processability-based SLA assessment procedure designed by Pienemann, Johnston and Brindley in 1988. For their study, Pienemann et al. (1988) collected 30 minutes of spontaneous speech samples from 16 adult L2 learners of English residing in Australia. After analyzing the learners' interview data linguistically for evidence of use of 14 target grammar criteria, they assigned each a final stage of English acquisition based on the highest-ranked criteria that had emerged in the speech sample based on the language processing strategies approach known as the "Multidimensional Model" (Meisel, Clahsen & Pienemann, 1981). This model explained sequence of target grammar forms emerging in German as a second language using a cognitive strategies form of processability theory.

Pienemann, Johnston and Brindley (1988) then trained 15 ESL education specialists to use the processing strategies-based criteria to perform an assessment of real-time speech production by ESL learners. The rater assessments were then used in an interrater reliability investigation of the procedure. An analysis of the final stage of assessment assigned by the raters (divided into two groups) showed moderate degree of correlation with the results of the assessment based on the detailed linguistic analysis (Group 1, r_S = .68 $p \leq .01$; for Group 2, r_S = .557, $p \leq .01$; Pienemann, Johnston, & Brindley: 236-237).[2] A stronger correlation was found between the raters observations (r_S = .86, $p \leq .01$). Moreover, upon inspection of the actual number of tics ("|") marked in the columns next to the target grammar criteria to indicate positive evidence (+) and negative evidence (-) observations, Pienemann et al. (1988:237) found that the judges' actual observations (tics) were more strongly associated with the linguistic analysis than their final ratings were (r_S =.74, $p \leq .01$).

Despite well-founded criticisms,[3] the Pienemann, Johnston and Brindley (1988) study was seminal because it uniquely adapted a language processing model for use in informing L2 assessment. By developing the use of linguistic profiling in simultaneous assessments of L2 speech, the researchers established a base protocol for further

[3] Criticisms are both methodological and theoretical. The methodology raises concerns because the researchers used two observation forms (hence, two group correlations reported) and applied a non-criterion based statistic (Spearman's r) in calculating interrater reliability. The theoretical shortcoming of the study is due to its underlying framework. As discussed in Pienemann (1998: 49-53), the Multidimensional Model (Clahsen, Miesel an Pienemann, 1981) assumed that UG/Minimalist theory based transformations (syntactic permutations) were related to development of word order in L2 grammar of German.

research to develop SLA profiling for use in practical educational settings. Soon after, the SLA-based assessment procedure for ESL was developed as "Rapid Profile" (Pienemann, 1992; Pienemann and Mackey 1993). And since the introduction of a grammatical processing theory (Pienemann, 1998) to replace the strategies approach, the theoretical justification for further pursuing and fine-tuning the application has gained momentum (Pienemann, 2003; Pienemann and Keßler, 2004). Important developments in language processing-based assessments have shown considerations of the following points.

1. Negative evidence observed in the learner's production of obligatory contexts provides important information regarding the steadiness of a learner's acquisition stage. Moreover, negative evidence may interact with lexical variation. The development of the "Rapid Profiling" (Pienemann, 1992; Pienemann and Mackey, 1993; Pienemann and Keßler, 2004) application is intended to aid in this interpretation by automatically tracking the observation data.

2. Attention to emergence of non-obligatory contexts: The absence of emergence of a given target grammatical criterion does not constitute negative evidence regarding the learner's use of the target criterion. However, for criteria which are assessed based on production in obligatory contexts, the learner often produces a non-obligatory context at an earlier time in the acquisition process than that in which the obligatory contexts emerges. Training judges to pay attention to this phenomenon, known as the learner's hypothesis space (Pienemann, 1998: 231) helps them focus their observations to the window of evidence regarding emergence of specific processing procedures.

3. As a practical assessment tool for use educational settings, the profiling interview should be capable of eliciting dense data regarding the learner's stage of acquisition in a reasonable amount of time. Eliminating learner-dominated small-talk and replacing unstructured interviews with communicative tasks was important to developing "Rapid Profiling" for ESL (Pienemann and Mackey, 1993; Pienemann and Keßler, 2004), as well as for the procedure developed here for JSL.

4. During the interview, learner production must be original and spontaneous. All instructions and cues should be given prior to each task and be fully comprehensible to the learner. Preston (2008) argues that a language that is not the target of the acquisition study should be used for all procedural instructions.

The empirical goal of rapid profiling is identical to that of the longer, research-based agenda: to gain accurate information regarding the learner's language processing stage at the time of the profile. The difference lies in their practical applications, the rapid version being a practical tool in real-time assessment of L2 speech in educational settings. Moreover, the rapid assessment protocol promises the potential for transforming the acquisitional experience itself, particularly through the development of teachability and learnability applications which address the language processing-based hypotheses of the learner at each stage of L2 acquisition.

On the surface, this presents at odds with traditional assessment goals which reward test-takers for their production of memorized material (formulae), their ability to interpret criterion not related to the measurement tool (such as instructions given in the L2), and paying attention to subtle cues provided in instructions or other linguistic contexts resulting from use of the L2 in the protocol. Such sources of procedural error are detrimental to the design and goal of a second language acquisition assessment procedure rooted in processability theory.

Fortunately, the recipe for successful profiling is rooted in the theory itself. The cognitive demand placed on working memory involved in production of spontaneous speech acts as a built-in control on the learner's ability to carry out speech production tasks, particularly when those tasks incorporate visual cues.

3. Design of the JSLA profiling study

3.1 Task-based elicitation

A brief, communicative task-based interview was created for use in an elicitation interview to collect speech samples from non-native speakers of Japanese. Incorporating the recommendations regarding interviewing procedures above, six tasks were designed to elicit spontaneous speech samples from L2 Japanese learners for use in the validity study of the PT hierarchy for JSLA (Preston, 2008). Four of the tasks were based on graphically designed materials.

Tasks:
1: Picture story narration, WATATASHI-NO UN-NO WARUI HI ("My bad hair day"), 17 pages
2: Picture story narration, OKAASAN-NO ICHINICHI ("A day in the life of my mother"), 17 pages
3: Picture comparison, "In the Park" (Hadfield, 1999), 2 drawings
4: Picture description, TANAKASAN-NO UCHI ("Tanaka-san's house"), 4 pictures
5: Interview question elicited in written in English, "Talk about your home" 3 cues)
6: Interview cue written in English, "Talk about your weekend"(3 cues)

A pilot study of the tasks' efficiency to elicit sufficient data about the Japanese learners' use of each target criteria was performed based on detailed linguistic analysis of 15 interview samples collected from 3 learners at each stage of JSLA and 3 Japanese monolingual adults (Preston, 2008: 239-244).

3.2 Collection of L2 samples

Following the pilot study of the task-based interview procedure, the researcher collected spontaneous Japanese speech samples from eleven adult L2 learners of Japanese enrolled in a public university in Japan. The eight recordings used in the assessment reliability study were produced by native speakers of English. Two samples of speech that had been scored a final rating for each of the four PT/JSLA stages were selected for this purpose. The other three samples (two elicited from non-native speakers of English) were used as recordings for practice assessments during the rater training session. Each learner sample varied in duration, lasting from between about 5-10 minutes. The average time length of the 11 samples was slightly more than seven minutes.

3.3 Analysis of L2 speech samples based on emergence criteria

After recording and transcribing the speech samples, the researcher performed a detailed linguistic analysis based on a distributional analysis of information regarding the target grammar criteria observed in each of the eleven samples. A final stage of acquisition was assessed based on the highest stage in the PT/JSLA hierarchy (see Table 1) for which positive evidence of emergence of at least one criterion's use by the learner could be established.

The researcher evaluated emergence of target forms based on emergence criteria stemming from measurement of two phenomena for each target grammar procedure. First, the sample was examined for evidence of emergence of target forms (see Table 1) in more than one lexical context. Next, the emergence of each target form was categorized as either obligatory or non-obligatory, and in the case of the latter, a rate of suppliance of greater than 50% was established as the benchmark for determining whether or not the evidence of the associated language processing stage was sufficient.

The application of criteria to assess the evidence regarding the emergence of each of the target grammar forms was a main emphasis of the rater workshop and training. The criteria were displayed in a handout arranging the stages, target forms, and examples of L2 utterances (positive and negative evidence) in a chart. For each criterion, one or two examples key to assessing the lexical variation condition was also included. This handout was the primary reference used for discussion during the PT/JSLA profiling workshop lecture and training.

Examples of the use of emergence criteria to assess hundreds of utterances relative to the production of the 12 JSLA target grammar forms used in processability theory research constitute the main body of empirical data in Preston (2008). The reader should refer to that work for detailed definition of emergence criteria for each target form exhibited in Table 1. Two examples will be presented here.

Instances of the emergence of the canonical structure criterion (ref. Table 1) are exemplified in (1a) and (1b), below. Moreover, because the verbs used are different (IKIMASU, -DESU), the condition of lexical variation is satisfied. Hence, a learner sample containing both (1a) and (1b) would be assessed positively (+) for emergence of the grammar criterion as well as its associated language processing stage (in this case the categorial procedure, or Stage 2). However, examples (1b) and (1c) do not differ lexically--both end in the polite form of the copula (a verb). Therefore, in spite of more than one observation of the target criterion's emergence, a learner sample containing multiple instances of this type of canonical utterance (N N -DESU) would not meet the emergence criteria for lexical variation for N N V.[4] Now, because Japanese allows relatively free word order, the target criterion, "canonical structures," is non-obligatory: The utterance in (1d) cannot be considered positive evidence of the

[4] These analytic criteria are referred in other literature as "tokens and types" (Pallotti, 2007).

canonical word order procedure, but neither is it considered negative evidence, i.e. evidence that the target form has not yet emerged in the grammar processing procedure of the learner's language.

1. Lexical variation and non-obligatory context for Stage 2 canonical utterances:
   a. WATASHI-WA GAKKOU-NI IKIMASU (I go to school) (+)
   b. KORE-WA HON DESU (This is a book) (+)
   c. KORE-WA TEREBI DESU (+, but no new lexical context)
   d. GAKKOU-NI IKIMASU, WATASHI-WA (not targeted)

Supposing that for one learner's speech sample the production of all four utterances in (1a)-(1d) were observed, a final assessment of the emergence of the "canonical structure" criteria and its procedure would be calculated based on the condition of lexical variation for non-obligatory contexts:

Total number of observances of target criteria emergence: 3
Lexical variation: Yes

For criteria defined by obligatory contexts, criterion's emergence is also contingent upon observations of both negative and positive evidence present in the learner's speech sample. For example, the production of the genitive case marker -NO hinges upon the procedure for processing the modifying N as a head of its phrase in the morphology, implicitly evidenced by the resulting NP constituent. Here, it is not sufficient that the learner process the categories of the words, because those words enter into the correct (target-like) relationship if the morphosyntax makes it so, as in (2a) and (2c), below. Omission of the case marker -NO in (2b) is negative evidence of emergence, demonstrating that the L2 speaker does not utilize the grammatical procedure necessary for producing a NP. The modifying element is processed as a bare category N (EIGO) rather than as a headed NP phrase (EIGO-NO).

2. Positive and negative evidence in obligatory context
   a. (+) NIHONJIN-NO TOMODACHI-TO IKIMASHITA (I went with my Japanese friend)
   b. (-) EIGOSENSEI DESU (I am an English teacher)
   c. (+) SHIGOTO-NO TAME-NI TOUKYOU-NI IKU (I'll go to Tokyo for work)

In an obligatory context, only if positive evidence accounts for 50% or more of the total observations is a grammatical procedure (the phrasal procedure of Stage 3, in the case of -NO) to be considered relevant to its production. The distributional analysis of the target criterion for a learner speech sample in which only (2a)-(2c) were observed yield the following:

Positive evidence of target criterion emergence (2)
Total number of obligatory contexts produced (3)
Rate of target criterion production: 66%
Lexical variation: Yes

Defining lexical variation and the method of observing emergence in obligatory and non-obligatory contexts is central to the study of PT's validity in general, and imperative to the development of a reliable tool for evaluating JSL speech samples using a real-time, rapid profiling assessment protocol. Consequently, these methodologies were elaborated upon in great detail during the training workshops and practice assessments in which the assessors of the study below participated.

3.3 Assessor recruitment

38 assessors were recruited for participation in this study. The assessors participated as part of their educational training at three locations in a large city of Japan. One group of raters was enrolled in a program for Japanese teacher training at a local non-profit organization (N=14). The other two groups of raters were graduate students who enrolled in courses related to Japanese language education and foreign language assessment (N=13, N=11).

3.4 Assessor training

In order to train the participants for real-time assessment of JSL adult speech based on PT, the researcher offered a workshop consisting of three 50-60 minute blocks. In the first block, the researcher briefly introduced the nature of processability theory to the participants, and then distributed a print-out chart listing the target procedures and grammatical criteria for JSLA (Table 1) showing emergence criteria for each target form (as described in section 3.1). The PT-based grammatical phenomena observed in JSLA research by Preston (2004, 2005) were explained. Group discussion followed.

An observation form was then drafted in large writing onto a chart written on a whiteboard in the front of the room. The chart displayed top-to-bottom, from highest to lowest, the four PT Stages (as in Table 1). To check the participants' knowledge of the criteria, the researcher elicited the target forms from the participants, jotting them down in abbreviated form next to their corresponding stage procedures. To the right, two columns were made, one with a (+) and the other with a (−) at the top. The participants were asked to identify for each target grammar procedure whether or not there was an obligatory context, and state an example of positive and negative evidence.

Next, the researcher simulated use of the target criteria in a real-time assessment. During playback of a practice speech recording, at each observation of information regarding the target criteria a tic ("|") was made in the appropriate column in the chart. Additionally, potentially relevant information regarding lexical context was written down adjacent to the tic. After the recording, the researcher vocally assessed which procedures had been sufficiently evidenced to have emerged in the sample based on the number of tics placed next to each target criterion in the (+) and (−) columns and the notes made regarding lexical variation. Finally, based on highest stage-based procedure (Stage 3, for practice sample 1) for which positive observation of emergence had been observed in the sample a "final assessment" was recorded in the top-left corner of the chart.

In the second training block, the handout on target emergence criteria was reviewed and participants were asked if they had any questions regarding the assessment procedure before they began practice assessments of their own. At this time, it was emphasized that raters should consider the theoretical implications in order to guide their observations: Once it had been established with a fair degree of certainty that a learner sample contained positive evidence of emergence of a given processing stage, attention should be directed toward information regarding production of the next higher stage of emergence, thereby establishing a window of observation tuned to the hypothesis space of the learner.

The group then began practice of real-time profiling assessments. They were given a blank observation form printed on paper, resembling the observation chart drafted on the whiteboard, and assessed alongside the researcher the same practice sample used

in the first block. They then scored one more practice sample (that of a Stage 1 learner), and discussed in small groups their assessments as marked on their observation charts. The whole group then re-listened to specific parts of the samples containing evidence pertinent to differences in observations they had claimed, and disagreements were resolved by reference to the emergence criteria. Lastly the training sample was one final time in its entirety. This procedure was repeated for one final training sample.

In the third session, raters were given a few minutes to ask any final questions they had regarding the assessment procedure and target criteria. They were then told they would assess eight samples of spontaneous speech, and that their ratings would be used by the researcher to investigate the viability of the assessment protocol based on an interrater reliability study.

Raters were then given an observation chart, similar to that used in the practice sessions. Before each sample was played, raters marked their form with an anonymous ID and indicated the sample number to be rated. They were instructed to mark their observations with a tic in the appropriate column next to the criteria they observed, and jot down any lexical contexts considered important to their window of observation, time permitting. Each sample was played once. No discussion during playback or final scoring of samples was allowed. After listening to a sample, raters assigned it a final, highest-stage of acquisition, and submitted it immediately to the researcher.

4. Data analysis and results

Because each stage of the four stages of language processing in the PT/JSLA hierarchy in Table 1 represents a qualitatively distinct grammar procedure, the final-stage assessment data might be considered nominal (categorical) in nature. However, given the hierarchical nature of grammar processing in human language, the target morphosyntactic criteria were classified into four implicational stages. Moreover, the assessment protocol emphasized that raters use their understanding of processability theory to guide their observations, in particular to obtain evidence regarding the hypothesis space of the learner. For these reasons the final ratings have strong ordinal characteristics.

The data analysis includes additional considerations. For one, because the raters were trained to use the same criteria that the researcher used in the detailed linguistic analysis in order to determine the highest stage of grammar acquisition for each of the eight learner samples, the interrater reliability study must account for both 1) the correlation among rankers (n=38) and 2) the agreement (concordance) with the linguistic analysis across all 304 rater assessments. Moreover, given that the "correct" assessments, i.e. those of the linguistic analyses, were already known, the study implies more than mere interest in a correlation coefficient, but rather an investigation of whether or not the real-time, rapid assessment protocol was as powerful as the detailed linguistic method at obtaining valid results, based on the criterion. This presents an asymptotic relationship between the datasets: the linguistic analysis is presumed 100% efficient, so its results, against which the raters' final-stage assessments are compared, assume rejection of the null hypothesis.

As demonstrated in Table 2, rater results were not in perfect agreement with the linguistic analysis. Across 304 comparisons made, 12.8% of the raters' final assessments were adjacent (off by one stage) to the linguistic analysis results.

Table 2. Count of agreements and disagreements between linguistic and assessors' criterion-based rankings of eight L2 Japanese samples. [5]

	Raters' final assessment of samples					
	Stage	1	2	3	4	Total
Linguistic analysis rankings	1	76	0	0	0	76
	2	0	67	9	0	76
	3	0	14	58	4	76
	4	0	0	12	64	76
	Total	76	81	79	68	304

In short, he data does not meet the assumptions of normal distribution (characteristic of standardized tests), the null hypothesis is not assumed, ordinal properties of the data must be addressed given the (assumption of) validity of processability theory, and one should expect a great number of ties both among raters and between raters and the linguistic assessment. Kendall's Tau-c (T_c), a chi-square measure of the coefficient of

[5] The comparison of 76 samples across each of the four stages is based on the fact that the linguistic analysis was performed on two (2) learner samples for each stage, and those were rated by all 38 judges.

rater agreement, was chosen for its appropriateness in this case.(See Siegel and Castellan, 1988: 281-284 for discussion.) The results of the analysis indicated that the strength of association expressed by the correlation coefficient is strong (T_c=.919, p<.001).

5. Discussion and concluding remarks

The study here is both novel and noteworthy for two reasons. First, as a pioneer application to JSLA, it represents an important field of further research to examine assessments in general in light of psycholinguistic phenomena. The study here is one of only a few applications of rapid profiling, and the only one used in Japanese. The ESL procedure has been developed over several decades for ESL, but yet the JSL procedure introduced here shows itself to be equally as reliable. Particularly given the fact that the JSL assessment procedure is so young, the finding that the correlation between the raters' final assessments and the linguistic analysis is strong testifies to the viability of further developing the procedure for use in PT-based assessments of learners of Japanese as a second language.

Secondly, if the research design is looked at wholly for its theoretical value, the use of Kendall's T_c statistic in a study of second language assessment is unprecedented. Recall that in this study raters were specifically instructed to use their knowledge of the theory to help fine-tune their observations. This would hopefully lead to rater observations particularly focusing on data regarding adjacent stages, the higher of which might be newly emerging in the speaker's Japanese. Using the standard procedures of analysis one obtains varying results. For example, an analysis of the same dataset using Kendall's T_b, the statistic recommended for examining interrater agreement based on use of ordinal scales (Surface & Dierdorff, 2004), yields approximately the same correlation coefficient (T_b=.918, p<.001); however, the usual non-parametric Spearman's correlation was strikingly higher (r_s =.947), p<.001).

It should be noted that no rater scores were discrepant (differing by more than one stage) from the linguistic analysis score, and all adjacent final scores were in only one

direction (either higher or lower than the linguistic analysis) for any one speech sample. However, that the correlation is not perfect shows that some raters had some problems in rating some of the samples. For example, that adjacent scores for any sample were unidirectional may be linked to under- or over-observations of specific target criteria. Or, perhaps some raters did not properly apply emergence criteria for obligatory and non-obligatory contexts, or overlooked the lexical variation condition in their observations. A detailed examination of the raters' observation forms can shed light on some of these problems, and is crucial to the improvement of the procedure overall.

In conclusion, the study attests to the benefits of adopting an interdisciplinary approach to examining interrater reliability. The results present in a convincing way the viability of developing means to improve language assessments so that they better reflect psycholinguistically plausible models of language learning and are measured for their usefulness based on appropriate analytic tools--that is models in line with theoretical assumptions. The further investigation of the method for development as a reliable, practical guide for teachers and researchers of Japanese as a second language using PT is exciting. Only by examining precisely which raters' assessments did not agree with those of the linguistic analysis can we posit potential sources of error. Additionally, further study will reveal whether or not group effects such as the class of participants, varying experience teaching Japanese, and rater's native language influenced their reliability.

References:

Clahsen, H. (1984). The acquisition of German word order: A test case for cognitive approaches to L2 development. In R. Anderson (ed.) *Second Languages.* Rowley: MA.

Crystal, D., P. Fletcher, P. M. Garman. (1976). *The grammatical analysis of language disability*. London: Arnold.

DiBiase, B. and Kawaguchi, S. (2002). "Exploring the typological plausibility of Processability Theory: language development in Italian second language and Japanese second language." *Second Language Research 18(3)*, 274-302.

Hadfield, J. (1999). *Beginners' Communication Games.* Essex, UK: Longman.

Hakansson, G. (2002). "Learning and teaching of Swedish: a processability perspective." In B. DiBiase (ed.) *Developing a Second Language. Australian Studies in Language Acquisition, 10,* 7-16.

Kawaguchi, S. (2002). "Grammatical development in learners of Japanese as a second language." In B. DiBiase (ed.) *Developing a Second Language. Australian Studies in Language Acquisition, 10,* 17-29.

Kempen, G. and Hoenkamp, E.(1987). "An incremental procedural grammar for sentence formulation." *Cognitive Science,11.* 201-258.

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge, MA : MIT Press.

McLaughlin, B. (1987). *Theories of Second-Language Learning.* Edward Arnold: New York.

Miesel, J., H. Clahsen and M. Pienemann. (1981). "On determining developmental stages in natural second language acquisition." *Studies in Second Language Acquisition, 3(2)*, 109-135.

Miesel, J.M. (1995) "Parameters in Acquisition". In P. Fletcher and B. MacWhinney (eds.), *The Handbook of Child Language.* Cambridge Mass: Blackwell, 10-35.

Mansouri, F. (2002). "Exploring the interface between syntax and morphology in second language development." In Bruno DiBiase (ed.) *Developing a Second Language.* Australian Studies in Language Acquisition, 10, 59-73.

Pallotti, G. (2007). "Operational Definition of the Emergence Criterion. *Applied Linguistics 28(3)*, 361-382.

Pienemann, M. (1992). "Assessing second language acquisition through rapid profile." *LARC Occasional Papers, No. 3.* Feb. 1992.

Pienemann, M. (1998). *Language Processing and Second Language Development: Processability Theory.* Studies in Bilingualism 15. Amsterdam: John Benjamins.

Pienemann, M. (2003) Rapid Profile: Revised labels for the phenomena; revised examples. Retrieved May 15, 2004, from http://www-fakkw.upb.de/institute/ Anglistik Amerikanistik/Personal/Pienemann/PT-Stages.pdf

Pienemann, M. & Keßler, J. (2004) "Rapid Profile." Flyer produced by the MILES team, Paderborn Univeresity. Retrieved Jan 15, 2004, from http://groups.uni-paderborn.de/rapidprofile/docs/English_Flyer.pdf Jörg-U.

Pienemann, M., DiBiase, B. and Kawaguchi, S.(2005). "Extending Processability Theory." In M. Pienemann (ed.) *Cross-Linguistic Aspects of Processability Theory.* Amsterdam:John Benjamins, 199-252.

Pienemann, M. and Johnston, M. (1987). "Factors influencing the development of language proficiency." In D. Nunan (Ed.) *Applying second language acquisition research.* Adelaide: National Curriculum Resource Center, Adult Migrant Education Program. 45-141.

Pienemann, M., M. Johnston and G. Brindley. (1988). "Constructing an Acquisition-based Procedure for Second Language Assessment." *Studies in Second Language Acquisition, 10,* 217-243.

Pienemann, M. and A. Mackey (1993). "An empirical study of children's ESL development and Rapid Profile." In P. McKay (ed.), *ESL Development. Language and literacy in Schools, Vol. 2.* Commonwealth of Australia and National Languages and literacy Institute of Australia, 115-259.

Preston, (2004). Developing a profiling procedure for Japanese second language acquisition. Paper presented at the *4th Annual Symposium on Processability, Second Language Acquisition and Bilingualism*, April 13-16, 2004, University of Sassari.

Preston, (2005). 2004-NENDO NIHONGO-KYOUIKU JISSHU HOUKOKUSHO [2004 Report on Teacher Training Program]. Retrieved from: http://www.lang.nagoya-u.ac.jp/nichigen/menu5_folder

Preston, J. (2008). PT-NI MOTODUITA PUROFAIRINGUTEJUN-NO KOUCHIKU: DAINIGENGO-TOSHITE-NO NIHONGOSHUUTOKU-NO BAAI. [Ms., 259 pp.] HAKUSHIGAKUIRONBUN. NAGOYADAIGAKUDAIGAKUIN KOKUSAIGENGOBUNKAKENKYUUKA NIHONGENGOBUNKASENKOU.

Zhang, Y. (2002). "A processability approach to the L2 acquisition of Chinese grammatical morphemes. " In Bruno DiBiase (ed.) *Developing a Second Language*: *Australian Studies in Language Acquisition, (10).* 29-44.