
事例研究論文

日本人英語学習者のための文法診断テストの妥当性の主張に向けて

Toward Validation Argument for Test Interpretation and Use Based on Scores of a Diagnostic Grammar Test for Japanese Learners of English

小泉 利恵1, 酒井 英樹2, 井戸 聖宏3, 太田 洋1, 羽山 恵5, 佐藤 正俊4, 根本 章子7
Rie Koizumi 1, Hideki Sakai 2, Takahiro Ido3, Hiroshi Ota4, Megumi Hayama5, Masatoshi Sato6, and Akiko Nemoto7

1常磐大学, 2信州大学, 3早稲田中学・高等学校, 4駒沢女子大学, 5獨協大学, 6山梨県立笛吹高等学校, 7茨城県龍ヶ崎市立愛宕中学校

1Tokiwa University, 2Shinshu University, 3Waseda Junior & Senior High School, 4Komazawa Women’s University, 5Dokkyo University, 6Fuefuki High School, Yamanashi, 7Atago Junior High School, Ibaraki

Correspondence concerning this article should be sent to: Rie Koizumi, Tokiwa University, 1・430-1 Miwa, Mito, Ibaraki, Japan 310-8585. E-mail: koizumir@tokiwa.ac.jp
The purpose of this paper is to review recent trends in validity and validation, particularly with regard to the argument-based approach to validity, and by using this approach, to construct a validity argument for test interpretation and use based on scores of the English Diagnostic Test of Grammar (EDiT Grammar) for Japanese learners of English. The EDiT Grammar focuses on the knowledge of basic English noun phrases, especially their internal structures. Using Chapelle, Enright, and Jamieson’s (2008a) validation framework, we first formulated the interpretive argument, and conducted two studies, one using verbal protocol analysis and the other using Rasch analysis. The results suggest that test-taking processes employed by test-takers are in accordance with the expectations drawn from the test specifications and that all the items and test-takers fit the Rasch model. These two positive results provided two strains of evidence to support inferences of evaluation and explanation, which are synthesized into the validity argument for the EDiT Grammar.

Keywords: validity, diagnostic test, noun phrases, verbal protocol analysis, Rasch analysis
1. Introduction

Since Messick (1989) published his seminal chapter on validity, the unified view of validity has been accepted by validity theorists and practitioners. This unitary view of validity has affected how test developers advance the validation process. Several researchers have furthered the idea of validation frameworks that provide direction for combining multiple sources of evidence (e.g., Bachman & Palmer, 2010; Messick, 1996; Mislevy, Steinberg, & Almond, 2002; Weir, 2005). Among them is Kane (e.g., 1992, 2006), who proposed an approach that emphasizes the aspect of validity as an argument. This approach enables test developers to take systematic steps to persuasively argue for the validity of test interpretation and use. His argument-based approach has gained strong support from validity theorists and practitioners, especially among language testers (e.g., Bachman & Palmer, 2010; Chapelle, Enright, & Jamieson, 2008a; Xi, 2010). In the language testing field, Kane’s framework was extended by Chapelle et al. (2008a), by elaborating upon Kane’s (2006) concept and adjusting his components to fit into the case of language test validation. The current paper reviews recent trends in validity and validation, specifically, the validity frameworks of Kane (2006) and Chapelle et al. (2008a). Further, we describe a procedure based on Chapelle et al. (2008a) for validating the English Diagnostic Test of Grammar (EDIT Grammar), a diagnostic test for Japanese learners of English. Before discussing validity theory, we briefly explain what the EDIT Grammar is.

2. Context of the current validation: EDIT Grammar

2.1. Noun phrases and the noun phrase groups

The EDIT Grammar is a diagnostic grammar test targeted primarily for Japanese junior and senior high school (especially 8th- to 12th-year) students learning English as a foreign language. The test aims to provide diagnostic feedback to teachers and students on students’ grammatical weaknesses, which can be utilized for remedial teaching (see Koizumi, Sakai, Ido, Ota, Hayama, Sato, & Nemoto, 2011, for details). The test was constructed with the support of the Association for English Language Proficiency Assessment (ELPA, n.d.).

In this test, we target basic English noun phrases (NPs), especially NP internal structures. For example, the sentence I saw a woman eating ice cream has the NP a woman eating ice cream. Precise comprehension of this sentence requires readers to understand that the phrase eating ice cream modifies another phrase a woman; otherwise, learners probably misunderstand and think that it means I saw a woman and ate ice cream. Word order is also important: a woman precedes, and eating ice cream follows and modifies a woman (i.e., postmodifies a woman). The postmodification merits attention because it does not basically exist in Japanese (Ogasawara, 1965). We selected the knowledge of NP structures as the focus of measurement because previous studies (e.g., Kimura, Kanatani, & Kobayashi, 2010) have shown that Japanese learners tend to face difficulty acquiring this grammatical structure, although it is presented in the early stage of English learning for students (by the early stage in the 9th year).

In the test, there were initially four NP groups (see Table 1), and the difficulty order was predicted as follows: Groups 2 < 3 < 4 < 5, in ascending difficulty. To put it briefly, this order was determined, on the basis of Pienemann (1998) and Kimura et al. (2010), as follows: Group 5 items are the most difficult because they contain subordinate clauses. Group 4 is next in level of difficulty because these items contain structures in which syntactic roles of a sentence (e.g., subject and object) are assigned to phrases. Since Groups 3, 4, and 5 have postmodified phrases, these groups were predicted to be more difficult than Group 2, which contains less difficult premodified phrases.
Table 1. Noun phrase groups tested

<table>
<thead>
<tr>
<th>Group</th>
<th>Structure</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 2</td>
<td>(Determiner) + premodifier + head noun</td>
<td>my friend’s bike</td>
</tr>
<tr>
<td></td>
<td></td>
<td>which beautiful dress</td>
</tr>
<tr>
<td>Group 3</td>
<td>NP + prepositional phrase</td>
<td>a surprise party for my brother</td>
</tr>
<tr>
<td>Group 4</td>
<td>(1) NP + present/past participle</td>
<td>books read by many people</td>
</tr>
<tr>
<td></td>
<td>(2) NP + to-infinitive</td>
<td>chances to meet his mother</td>
</tr>
<tr>
<td>Group 5</td>
<td>NP + relative clause</td>
<td>flowers (that) I bought today</td>
</tr>
</tbody>
</table>

Note. Group 1 (Determiner + head noun) was not tested. Group 4 (2) NP + to-infinitive was later renamed as Group 6.

Koizumi et al. (2011) found that Group 4 should be divided into “NP + present/past participle” and “NP + to-infinitive” because “NP + to-infinitive” was found to be more difficult than the Group 5 structure. Accordingly, the latter subcategory was renamed as Group 6, and the difficulty order was found to be as follows: Groups 2 < 3 = 4 = 5 < 6 (with nonsignificant results with marginal effect sizes expressed as =).

2.2. Test format and diagnosis

The EDIT Grammar is a multiple-choice test and requires test-takers to match a Japanese phrase with one of four English options. Each test item belongs to one of the four NP groups and is dichotomously evaluated to appraise learners’ understanding of the targeted NP group structure. In the latest version of the test, there are 10 or 12 items for each NP group and 46 items in total.

For the purpose of diagnosis, the test uses not only proportions correct for each NP group, but also the selection rate of distractors, which can suggest test-takers’ patterns of misunderstanding NP structures. An example follows.

Instruction: Choose the most appropriate English phrase for the underlined part of the Japanese sentence from 1 to 4.

Example item 1

Stem: 山田先生は、理科の先生です。 (NP Group 3)
(Mr. Yamada is a teacher of science)
Options: 1. a teacher of science*
2. a teacher science of [Phrase]
3. science of a teacher [Japanese]
4. of science a teacher [Postmodification]
(The instruction was originally written in Japanese. Square brackets show a distractor type of each item used for diagnosis, which is not presented to test-takers. See Appendix A, for more examples).

Each distractor belongs to one of five distractor types seen in Table 2: Premodification structure, Phrase structure, Clause structure, Postmodification, and Japanese word order. For example, Option 3 has a Japanese word order: *rika [science] + zo [of] + sensei [a teacher]. Consistently choosing several such options suggests that test-takers do not seem to know appropriate English word order.

Table 2. Distractor types

<table>
<thead>
<tr>
<th>Problem</th>
<th>Characteristics</th>
<th>Example [Answer]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-modification structure</td>
<td>Select an option with an incorrect premodification structure</td>
<td>the name girl’s [the girl’s name]</td>
</tr>
<tr>
<td>Phrase structure</td>
<td>Select an option with an incorrect phrase structure (the room in)</td>
<td>the cat the room in [the cat in the room] Item 1. Option 2</td>
</tr>
<tr>
<td>Clause structure</td>
<td>Select an option with an incorrect clause structure (who the piano play can)</td>
<td>any friends who the piano play can [any friends who can play the piano]</td>
</tr>
<tr>
<td>Post-modification structure</td>
<td>Select an option with an incorrect modification. In this structure, a modifier (in the room) does not modify their head NP (the cat)</td>
<td>in the room the cat [the cat in the room] Item 1. Option 4</td>
</tr>
<tr>
<td>Japanese word order</td>
<td>Select an option having the Japanese word order (heya [the room] + nonakano [in] + neko [the cat])</td>
<td>the room in the cat [the cat in the room] Item 1. Option 3</td>
</tr>
</tbody>
</table>
We expect that when test-takers solve questions, they parse NP internal structures, identify subcomponents of phrase structures, and think of which phrase modifies which phrase. Whether this process is seen in the actual test-taking behavior is examined in Study 1.

After the test, a diagnostic profile is given to teachers and students (see a student’s example shown in Appendix B). The diagnostic profile consists of (a) an overall result, (b) a score profile, and (c) an error pattern profile. The overall result is mainly for the use of students, and the score profile and the error pattern profile are basically targeted for teachers. If such remainder was provided to students, the explanation on the diagnostic profile would require more elaboration.

The first feedback section, overall results, shows the proportion correct with all NP groups combined. The second feedback section, the score profile, indicates the proportion correct in each NP group and the type of NP groups a student has mastered. The cut-off points for mastery are set at 60% and 80% (see Appendix B for an example of a student passing only the Group 2 structure). These percentages were based on the phi lambda (Φ (λ)), which represents the extent of agreement of decisions. The cut-off point of 80% reflected high agreement (.85 or above; Koizumi et al., 2011), whereas that of 60% indicated a lower value (.68 to .85). Despite this slightly lower agreement of decisions at the 60% cut-off point, we decided to include this measure to enhance the diagnostic value.

The third component of the feedback is the error pattern profile, with cut-off points of 30% and 50%. The low proportion indicates positive results, suggesting that test-takers do not have error patterns. For example, Appendix B shows an example of a student receiving two square marks, which indicates he had proportions between 30% and 50%. This suggests that he had a moderate tendency to select distractors whose postmodification is not correct and distractors that conform to Japanese word order; thus, he is unlikely to comprehend postmodified NPs in English word order well. Different error patterns indicate the need for teachers to guide students in different ways. By using the EDIT Grammar and examining the diagnostic feedback, various patterns of score profiles as well as error profiles can be revealed, as was demonstrated in Sakai and Koizumi (2010).

3. Validation framework

3.1. Validity and validity argument

According to Sireci (2009), validity theory has changed considerably, along with the procedures of validation (see also Chapelle, 1999; Kane, 2001; Messick, 1989; Shepard, 1993, for the history). Traditionally, validity was defined as the degree to which a test measures the trait it intends to measure (e.g., Henning, 1987). It was often categorized into three main types: criterion-related validity (including concurrent validity and predictive validity), content validity, and construct validity. Presenting evidence for one of the three types was previously considered sufficient for demonstrating test validity. Later, construct validity was regarded as the most important of all the types of validity, which led to the conceptualization that “all validity is construct validity” (Sireci, 2009, p. 25) and that validity should be considered unitary.

Messick (1989) played a major role in systematizing validity theory. With the concept of unified validity in mind, he referred to validity as “the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores” (Messick, 1989, p. 13). This definition suggests two main shifts in establishing validity. First, validity is examined using both empirical evidence and theoretical rationale, and it is expressed as a matter of degree, rather than as zero or perfect. Presenting multiple sources of empirical and theoretical evidence is required in validation. It should also be noted that reliability is considered one essential source of evidence in this framework.
Second, it is not the test but the interpretation (also called inference) and action (also called use and decision) based on test scores that are validated. In other words, we examine not the validity of a test but the validity of interpretation and use based on test scores. This is because “the same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third” (Cureton, 1951, as cited in Kane, 2009, p. 42). The inclusion of test use in the validity concept led to the recognition of social consequences, both intended and unintended effects, as vital lines of validity evidence.

This view toward validity was not always accepted. Borsboom and his associates argued for moving away from Messick’s (1989) conceptualization and returning to the early concept of validity, for example, by excluding social consequences from the validity concept and focusing only on causal relationships in the validation (e.g., Borsboom, Cramer, Kievit, Scholten, & Franić, 2009; Borsboom, Mellenbergh, & van Heerden, 2004; see Lissitz, 2009, for other views). However, Messick’s theory has garnered numerous advocates and remained in the mainstream (Lissitz, 2009). Based on his conceptualization, many researchers presented a validation framework or procedures to conduct validation (e.g., Bachman & Palmer, 2010; Chapelle, 1999; Chapelle et al., 2008a; Mislevy, Steinberg, & Almond, 2002; Weir, 2005). Messick (1996) himself proposed a framework consisting of six aspects of validity. He argued that “unified validity” could be separated “into several distinct aspects to underscore issues and nuances that might otherwise be downplayed or overlooked” (Messick, 1996, p. 248). The six aspects he proposed were content, substantive, structural, generalizability, external, and consequential aspects. His framework was explicated (e.g., Chapelle, 1999; Hirai, 2006; Murayama, 2006) and employed in previous studies (e.g., Beglar, 2010; Guerrero, 2000; Hasselgren, 2005; Koizumi, 2005; Miller & Linn, 2000).

However, Messick’s (1989, 1996) conceptualization and validation have increasingly caused frustration among practitioners because the concept is difficult to grasp and apply in real-life validation practice (Lissitz, 2009). One major reason validation is seen as a conundrum is that it is not clear “what kind of evidence to gather and how much evidence is enough” for validation (Chapelle, 2008, p. 321). Messick (1989) stated that validation involves examining and reporting many diverse sources of evidence, which should be integrated into an overall judgment of validity. Kane (2009) explained, “some statements in the literature can be interpreted as saying that adequate validation requires that every possible kind of validity evidence be developed for validation to be complete” (p. 49). In other words, validation could be regarded as an everlasting endeavor (Sireci, 2009), which requires test developers to accumulate “as many kinds of evidence as possible, or worse, . . . all possible kinds of evidence” (Kane, 2009, p. 49).

In response, researchers have attempted to revise validation procedures while maintaining the current concept. An exemplary researcher is Kane (e.g., 1992, 2006, 2009), who proposed an argument-based approach to validation, although the concept of validity as argument was originally suggested by Cronbach (1988). Kane (2009) emphasized that validation does not require test developers (or test validators) to present all possible strands of evidence: instead, he underscored that “the evidence required for validation depends on the test and the interpretation/use, and therefore, different interpretations/uses will require different kinds and different amounts of evidence for their validation” (p. 40). Kane (2009) detailed the procedure as follows:

If the interpretations and uses are simple and limited, involving a few plausible inferences, the validation can be simple and limited. If the interpretations and uses are ambitious, involving a more extensive network of inferences, the evidence needed for validation would also be more extensive, including for example, inferences to non-test and/or future outcomes. (p. 48)

An example of simple and limited interpretation and use is when the test’s only purpose is to describe
observed performance, which can be “evaluated using judgments about the representativeness of content sampling” (Kane, 2009, p. 43). An ambitious example is when a test aims to “estimate or predict some non-test performance,” which requires “evidence linking test scores to a criterion measure of performance” (p. 44), in addition to content evidence. Kane (2006) further stated that more evidence does not lead to a stronger argument since inferences and assumptions that are already coherent enough do not require more evidence, which is clearly contrasted with Messick’s (1996) view of considering more evidence to be better.

By this approach, Kane (2009) aimed “to provide a realistic and pragmatic framework for evaluating the interpretations and uses of test scores” (p. 49). Sireci (2009) also positively evaluated Kane’s approach as “a compromise between sophisticated validity theory and the reality that at some point, we must make a judgment about the defensibility and suitability of use of a test for a particular purpose” (p. 29).

3.2. Kane’s (2006) approach to validity

Kane’s (2006, 2009) argument-based approach to validity is based on Toulmin’s informal argument structure (2003), in which informal arguments are used to build a case for a particular conclusion by constructing a chain of reasoning in which the relevance and accuracy of observations and assertions must be established and the links between them need to be justified.

(Chapelle et al., 2008b, p. 6)

The argument structure used for Kane’s framework, in principle, consists of seven components: claim, grounds, inference, warrant, assumption, backing, and rebuttal (Chapelle et al., 2008a). A claim refers to “a conclusion or action” (Kane, 2009, p. 47) about test scores or test-takers to be examined. A claim is made based on grounds, that is, data or observations relevant to the claim. The claim and the grounds are associated through the inferential link (also called inferential bridges or simply, inference), which needs to be justified by a warrant. The warrant is “a law, generally held principle, rule of thumb, or established procedure” (Chapelle et al., 2008b, pp. 6-7). Underlying the warrant, one or more assumptions exist, which need to be corroborated by proper evidence, which is called backing. Thus, in order to argue for the claim, adequate backing should be collected, which provides support for the assumptions and warrant, which then vindicates the inference derived from the grounds to the claim. A rebuttal, or negative evidence for the inference, can be presented in the argument, and the rebuttal “weakens the inferential link between the grounds and the claim” (Chapelle et al., 2008b, p. 8). If a rebuttal is demonstrated, a firmer backing should be counter-demonstrated to support the inference. Additionally, stronger backing is indispensable in the validation of test interpretation and use of a high-stakes test.

Employing Toulmin’s (2003) argument structure, Kane’s (2006, 2009) argument-based framework primarily consists of two facets: an interpretive argument and a validity argument. An interpretive argument acts as a framework to holistically evaluate the test and justify test interpretation and use. With this framework, test developers describe proposed test interpretation and use and “the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances” (Kane, 2006, p. 23). Test developers begin to determine and depict how the test will be interpreted and used, which is termed proposed test interpretation and use. They next describe inferences and assumptions associated with the proposed interpretation and use, sometimes coupled with rebuttals or alternative interpretations that may refute the inferences. They should further present a plan for how they will produce evidence for these warrants and assumptions and how the overall argument can be persuasively made. The resultant interpretive argument needs to be assessed in terms of “the clarity, coherence, and completeness” overall and “the plausibility of each of the inferences and assumptions in the interpretive argument” (Kane,
The validity argument is made on the basis of the interpretive argument. Simply put, “a validity argument is an interpretive argument in which backing has been provided for the assumptions” (Chapelle, Enright, & Jamieson, 2010, p. 5). In the validity argument, test developers obtain theoretical rationales and empirical results, judge whether the backing for the inferences, corresponding warrant and assumptions, and a chain of reasoning are plausible and defensible, and make a case for validity of interpretation and use based on test scores. Positive backing supports an assumption, whereas negative backing weakens it. When their evidence reasonably supports all the assumptions and warrants, they sufficiently satisfy requirements for the proposed inference. This means that they bridge proposed inferences and that, accordingly, they successfully make the validity argument. On the other hand, if there are some assumptions not tenable and inferential bridges are difficult to make, test developers need to modify or recreate the interpretive argument and try the validation argument again (Kane, 2006).

Kane (2006, 2009) cautions against “the begging-the-question fallacy,” which occurs when “a relatively modest interpretive argument is assumed for purposes of validation, but a more ambitious interpretation is employed in practice” (Kane, 2009, p. 50). Imagine that a test is developed to describe observed performance and the validity argument is successfully completed. Although this test can only be interpreted as reflecting observed performance, the same test may more ambitiously be interpreted as predicting future performance. This gap between the originally intended interpretation and use and the actual interpretation and use is problematic. Consequently, test developers need to be vigilant not only that their proposed test interpretation and use are supported by evidence in the validity argument, but also that test interpretation and use after validation do not go beyond those proposed.

Kane (2006, 2009) brought about a paradigm shift in validation attempts from accumulating all the evidence (Messick, 1989, 1996) to the sophisticated act of pinpointing and presenting the evidence required for the inferences pertaining to the interpretive argument. Chapelle, Enright, et al. (2010) summarized four major differences between the argument-based approach to validity and Messick’s approach, the latter of which is also accepted by Standards for Educational and Psychological Testing (AERA/APA/NCME, 1999; hereafter, Standards). First, to “frame the intended score interpretation” (Chapelle, Enright, et al., 2010, p. 3), the argument-based approach uses the interpretive argument, whereas the Standards employs the construct. The interpretive argument is more manageable than the construct as a foundation for validation research since the construct can be defined in various ways and it tends to be too ambiguous to produce testable propositions. Second, the argument-based framework produces assumptions firmly associated with specific inferences that enable test developers to systematically frame their past and future investigations into validity claims. On the other hand, the Standards type provides “a list of propositions” (Chapelle, Enright, et al., 2010, p. 5) and fails to present “explicit guidance on how to formulate an intended interpretation and the propositions that are supposed to point to the types of evidence that would ultimately contribute to the . . . validity argument” (p. 6). Third, the argument structure grants “a step-by-step movement across ‘bridges’ . . . to a conclusion about test score use” (p. 9), whereas the Standards style does not show the path to validity argument after collecting multiple evidential results. Fourth, in the argument-based approach, there is a place for rebuttal (i.e., counterevidence) in the interpretive argument, whereas the framework displayed in the Standards does not provide direction to test developers for systematic incorporation of counterevidence. Chapelle, Enright, et al. (2010) concluded that the argument-based approach serves better and recommends adopting this framework for the next version of the Standards.
In the language testing field, Kane’s (2006) framework was extended by Chapelle et al. (2008a), by elaborating upon and including inferences essential for language tests (see also Bachman & Palmer, 2010; Kim, 2010; Pardo-Ballester, 2010). In their approach, there are six types of inferences in the interpretive argument: domain definition, evaluation, generalization, extrapolation, and utilization (see Figure 1). Using these inferences, test makers are required to make an interpretive argument and a validity argument. Chapelle et al.’s (2008a) framework has been employed for the Test of English as a Foreign Language Internet-based test (TOEFL iBT; Chapelle et al., 2008a) and also for a test of productive grammar ability (Chapelle, Chung, Hegelheimer, Pendar, & Xu, 2010).

Figure 1. Inferences in the interpretive argument.

According to Chapelle et al. (2008a), their framework differs from Kane’s (2006, 2009) mainly in the number of inferences in the interpretive argument: In addition to Kane’s (2009) three inferences (evaluation, generalization, and extrapolation), Chapelle et al. (2008a) added domain definition, explanation, and utilization. We argue that the framework of Chapelle et al. (2008a) is superior to that of Kane, at least in the language testing field, because the three additional inferences are crucial for language tests. First, the domain definition inference is vital in delineating the target domain, which should be properly positioned in the interpretive argument. Second, the explanation inference enables or requires test developers to include a construct, which is “the source of meaningful interpretation for test scores, guidance for test development, and the cornerstone for validation research” (Chapelle, Enright, & Jamieson, 2007, p. 1). It should be noted that the construct was the fundamental basis of validity inquiry in Messick’s (1989, 1996) theory and in the Standards (AERA/APA/NCME, 1999). In Kane’s framework, the role of the construct is less explicit, but the construct needs to be clearly situated in the validation, as seen in the framework of Chapelle et al. (2008a). Third, the utilization inference relates test use to the validation process. Test use is an essential perspective in validation because a test is conducted to use its test scores for certain purposes and validation encompasses the examination of its use as well as its interpretation. These three additions made the framework of Chapelle et al. (2008a) more comprehensive to cater to a variety of test validation scenarios. For this reason, we decided to adopt their framework in the current validation study.

3.3. Interpretative and validity argument structure

This section describes components of the interpretive argument in the case of the EDiT Grammar by using terms and expressions utilized in Chapelle et al. (2008a), along with illustrations of the available backing for the assumptions. The proposed interpretations of the EDiT Grammar are
(a) that the test scores reflect a construct of grammatical knowledge of NP internal structures in English as a foreign language, and (b) that the scores explain the quality of reading comprehension in English. The intended test use is as follows: that the test is used for low-stakes decisions regarding instruction, and that the test scores are useful for aiding diagnostic decisions, guiding English-language instruction, and especially, for giving teachers information on students who need remedial activities. The first proposed interpretation relates specifically to the domain definition inference up to the explanation inference, whereas the second one adds one more inference to be made (the extrapolation inference). The intended test use adds the utilization inference. Although all six inferences are necessary for the EDiT Grammar, the test is low-stakes, and the backing required for the validity argument may not necessarily be very strong. Table 3 summarizes the relevant inferences, assumptions, and methods to draw backing in the interpretive argument of the EDiT Grammar.

### Table 3: Summary of the interpretive argument of the EDiT Grammar

<table>
<thead>
<tr>
<th>Inference</th>
<th>Assumptions underlying the warrant</th>
<th>Method to seek backing for the assumptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Domain</td>
<td>1. The target domain that was focused on covers critical knowledge of NP structures needed for reading</td>
<td>1. &amp; 2. Domain analysis by expert judgment</td>
</tr>
<tr>
<td>definition</td>
<td>comprehension in English.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. A test item format in the EDiT Grammar requires important knowledge of NP structures relevant to</td>
<td></td>
</tr>
<tr>
<td></td>
<td>and representative of the target domain.</td>
<td></td>
</tr>
<tr>
<td>2. Evaluation</td>
<td>1. Scoring rubrics are appropriate for providing evidence of targeted knowledge.</td>
<td>1. Expert judgment</td>
</tr>
<tr>
<td></td>
<td>2. The item statistical characteristics are appropriate.</td>
<td>2. Item analysis</td>
</tr>
<tr>
<td></td>
<td>3. The scaling model used in scoring fits the data.</td>
<td>3. Rasch analysis (Study 2)</td>
</tr>
<tr>
<td>3. Generalization</td>
<td>1. A sufficient number of items are included on the test to provide stable estimates of test-takers' performance.</td>
<td>1. Generalizability study</td>
</tr>
<tr>
<td></td>
<td>2. Test specifications are well defined so that parallel items are created.</td>
<td>2. Expert judgment</td>
</tr>
<tr>
<td>4. Explanation</td>
<td>1. The test-taking processes observed accord with the expectations by test developers.</td>
<td>1. Verbal protocol analysis (Study 1)</td>
</tr>
<tr>
<td></td>
<td>2. The means of the NP groups vary as predicted.</td>
<td>2. Analysis of NP group difficulty</td>
</tr>
<tr>
<td></td>
<td>3. Scores are related positively as intended with the scores on the test assessing similar or different constructs.</td>
<td>3. Correlational analysis</td>
</tr>
<tr>
<td>5. Extrapolation</td>
<td>The test scores are related to other test scores that reflect reading comprehension in English.</td>
<td>Correlational analysis</td>
</tr>
<tr>
<td>6. Utilization</td>
<td>1. The meaning of the test score is clearly interpretable by teachers and test-takers.</td>
<td>1. Expert judgment and survey analysis</td>
</tr>
<tr>
<td></td>
<td>2. The test will have a positive influence on how English grammatical knowledge, especially of NP internal structures is taught.</td>
<td>2. Washback studies</td>
</tr>
</tbody>
</table>

Note: Terms and expressions are adopted from Chapelle et al. (2008a).
Toward Validity Argument for Test Interpretation and Use Based on Scores of a Diagnostic Grammar Test for Japanese Learners of English

(i.e., observation of test responses or performance derived from the test). The observation can be described from two perspectives: characteristics of a student’s performance and those of test items. In order to support the domain definition inference, we posit the warrant 1, that is, observations of performance on the EDiT Grammar reveal relevant knowledge in situations representative of those in the target domain of language use. For this warrant to be supported, Kane (2009) recommends focusing on “how the domains are defined, and in particular, on the relationship between test content and the domain content” (p. 51). We identified two assumptions to be satisfied, which correspond to the two perspectives of the observation mentioned above. First, the target domain on which we focused covers critical knowledge of NP structure needed for reading comprehension in English; second, the test item format in the EDiT Grammar requires test-takers to demonstrate important knowledge of NP structures relevant to and representative of the target domain. The backing for these assumptions can be drawn by analyzing the target domain using expert judgment. Koizumi et al. (2011) described the procedures for analyzing the target domain and determining the NP internal structure as an essential testing area, and we considered this assumption to be satisfied and the domain definition inference to be successful.

The second inference, the evaluation inference (scoring in Kane’s [2009] terminology) bridges the argument from the observation to an observed score (i.e., a test score including raw and scaled scores). The warrant for the evaluation inference (warrant 2) is associated with “a scoring rule” (Kane, 2009, p. 58), which can be elaborated as, “observations of test performance are evaluated in a manner that provides relevant observed scores” (Chapelle, Chung, et al., 2010, p. 451). The possible assumptions that can be laid out in general are as follows: that the scoring “rule is appropriate, that it is applied correctly, that it is free of bias, . . . that any scaling model used in scoring fits the data” (Kane, 2009, p. 57), and that “the statistical characteristics of items, measures, and test forms are appropriate” for interpretation and use (Chapelle et al., 2008b, p. 21). Among the possible assumptions, the EDiT Grammar uses three: (a) that scoring rubrics are appropriate for providing evidence of targeted knowledge, (b) that the item statistical characteristics are appropriate, and (c) that the scaling model used in scoring fits the data. The backing for the first assumption was obtained during the test development processes of trialing and modifying items and scoring keys. The second assumption was confirmed by item analysis (Koizumi et al., 2011), where the item discrimination (point-biserial correlation) was sufficient ($M = .45, SD = .13$) and all the options were properly selected. The backing for the third assumption is examined using Rasch analysis in Study 2 in the current research.

The third inference, the generalization inference chains the argument from the observed score to an expected score. The expected score is construed as “a stable score” (Chapelle, Chung, et al., 2010, p. 451) or as “expected performance over the universe of generalization” (Kane, 2009, p. 57). Related to the generalization inference, the warrant 3 is made that “a stable score . . . would be obtained by examinees’ [sic] on parallel tasks, test forms, administrations and rating conditions” (Chapelle, Chung, et al., 2010, p. 451) and “observed scores are estimates of expected scores for parallel tasks, tests, and administration and rating conditions” (Chapelle et al., 2008b, p. 22). Four possible assumptions underlying the warrant are “that (a) a sufficient number of tasks are included on the test to provide stable estimates of test takers’ performances; (b) the configuration of tasks on measures is appropriate for [sic] intended interpretation; (c) appropriate scaling and equating procedures for test scores are used; and (d) task, test, and rating specifications are well-defined, so that parallel tasks and test forms are created” (Chapelle et al., 2008b, p. 22). We used the following two assumptions for the EDiT Grammar: (a) A sufficient number of items are included on the test to provide stable estimates of
The fourth inference, the explanation inference connects the argument from the expected score to (test) construct. This inference is supported by the warrant 4 that the expected score would reflect the construct, grammatical knowledge of understanding NP structures. The assumptions can be made by using the prediction that the construct “fits logically into a network of related constructs and theoretical rationales” (Chapelle et al., 2008b, p. 16). While “a theoretical construct should serve as the basis for score interpretation for a large-scale test with high-stakes outcomes” (Chapelle et al., 2008b, p. 1), the construct of the EDiT Grammar is not well developed. Thus, to posit and test relationships with the theory of the construct, for instance, by investigating the factor structure of the test scores, is not really possible. Therefore, we posited the following three assumptions that can be empirically tested. First, the test-taking processes observed are in accordance with the expectations of the test developers. Second, the means of the NP groups vary as predicted. Third, scores are related positively, as intended, with the scores on a test assessing similar or different constructs. The first assumption can be examined by analyzing verbal protocol analysis, which is reported later in Study 1. Kane (2006) suggests that the first assumption can also be analyzed by obtaining test-takers’ reactions to the test. As for the second assumption, weak backing was derived in Koizumi et al. (2011). As mentioned in section 2.1, the difficulty order of NP groups was found to be as follows: Groups 2 < 3 = 4 = 5 < 6. Although we had expected a more distinguished difficulty between Groups 3, 4, and 5, the trend was overall consistent with the initial prediction with Group 2 being the easiest. The future revised test should examine whether the same difficulty order can be replicated so that the test structure, in terms of difficulty, is stable and whether this result can be extended to the theory. The evidence for the third assumption is to be tested in future research by comparing the test scores and criteria that measure similar or different traits and showing convergent and discriminant evidence for validity (Kane, 2006). One possible criterion for assessing similar knowledge would be a task of asking test-takers to parse English sentences with premodification and postmodification.

The fifth inference, the extrapolation inference, bridges test performance to nontest performance and relates the construct to a target score. The target score is “interpreted as indicative of performance and scores that they would receive in the target setting” (Chapelle et al., 2008b, p. 11). This inference is based on the warrant 5 that “measures of this construct are relevant to language performance of interest beyond the test setting” (Chapelle, Chung, et al., 2010, p. 451). Two types of assumptions underlying extrapolation can be made. The first type is that “test task characteristics are authentic relative to tasks in the target language domain” (Chapelle et al., 2008b, p. 12), which is examined using “a logical analysis of the correspondence between test tasks and relevant tasks beyond the test” (p. 451). The second assumption is that test scores are related to other test scores that reflect nontest performance. The future validation of the EDiT Grammar can posit
Rie Koizumi, Hideki Sakai, Takahiro Ido, Hiroshi Ota, Megumi Hayama, Masatoshi Sato and Akiko Nemoto

Toward Validity Argument for Test Interpretation and Use Based on Scores of a Diagnostic Grammar Test for Japanese Learners of English

the latter assumption and examine whether test scores on the EDiT Grammar are correlated with reading comprehension test scores.

The sixth inference, the utilization inference, bridges the target score with test use. This inference is qualitatively distinct from other inferences in “moving the interpretive argument from interpretation of score meaning to actual score use” (Chapelle et al., 2008b, p. 12). Test use includes “educational decisions made” based on the test scores (Chapelle et al., 2008b, p. 17), such as placement and admission. The warrant states that test scores obtained from the EDiT Grammar are useful for making decisions about diagnosis and about necessary remediation. Two assumptions are made. First, the meaning of the test score is clearly interpretable by teachers and test-takers. The second assumption is that the test will have a positive influence on how English grammatical knowledge, especially of NP internal structures is taught. The backing for the first type is that the EDiT Grammar diagnostic profile was made (see section 2.2) and the quality was examined through expert reviews. In order to retrieve firmer evidence, however, it is necessary to take a principled approach to creating the diagnostic profiles based on Roberts and Gierl (2010) and to examine whether the diagnostic profile format is easy to understand, for example, through conducting surveys. It is also essential to further develop materials for score interpretation and instructions. With respect to the second assumption, the impact of diagnosing the grammatical knowledge of NP structures needs to be explored by conducting washback studies and checking unintended side-effects of the diagnosis (Kunnan & Jang, 2009).

The validation of the EDiT Grammar is a work in progress. The current paper reports on the two lines of backing for the inferences of evaluation and explanation.

3.4. Current study

We investigate the validity of the interpretation based on the test scores of the pilot and the revised versions of the EDiT Grammar, using Chapelle et al.’s (2008a) argument-based framework. While Koizumi et al. (2011) discussed validity evidence for the EDiT Grammar, they did not relate such evidence to the validation framework, which the current study attempts, by documenting two other pieces of empirical evidence. The present study intends to demonstrate one case of validation using the argument-based framework that other validation attempts can utilize. We examine two research questions (RQs) in two studies, which Koizumi et al. (2011) did not examine and which can support the validity argument.

RQ1: Are the test-taking processes consistent with the expectations based on the test specifications? (which corresponds to the explanation inference; Study 1)

RQ2: Do the items and test-takers fit the Rasch model? (evaluation inference; Study 2)

Table 4. Comparisons Between Studies 1 and 2

<table>
<thead>
<tr>
<th>Test used</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research question</td>
<td>1 (test-taking</td>
<td>2 (item and</td>
</tr>
<tr>
<td></td>
<td>processes)</td>
<td>person fit to</td>
</tr>
<tr>
<td></td>
<td></td>
<td>the model)</td>
</tr>
<tr>
<td>Method</td>
<td>Protocol analysis</td>
<td>Rasch analysis</td>
</tr>
<tr>
<td>Inference</td>
<td>Explanation</td>
<td>Evaluation</td>
</tr>
</tbody>
</table>

As summarized in Table 4, RQ1 was examined using verbal protocol analysis, whereas RQ2 was investigated using Rasch analysis. Study 1 used a pilot test; Study 2 used a revised test. The entire test items can be obtained in Kanatani and EDiT Development Group (2006) and Kanatani (2008).

4. Study 1: Verbal protocol analysis using the pilot test

4.1. Study 1 method

The participants were seven students in the first or second year at a private university (three males and four females, aged from 18 to 19). Five of them majored in English, whereas the rest majored in
international studies. Their proficiency levels ranged from low to intermediate-low levels (roughly estimated from Eiken Grade 3 to Grade 2 levels), which broadly corresponds to target test-takers in the 9th to 12th year. Although the EDIT Grammar mainly targets junior and senior high school students, we recruited university students in this study because explaining the test-taking processes of students at the secondary school level was found to be a difficult task, which we learned through informal trials of eliciting protocols from them. The test-takers volunteered to take the test; they received a gift card after the experiment.

One of the authors conducted the pilot test of the EDIT Grammar (24 items) and asked each student to orally express their thinking process in their L1 (Japanese) while they solved questions. The students practiced producing protocols before taking the test. We analyzed the utterances that the test-takers spoke without any prompting questions. The test-takers responded to test items well, with the proportions correct of .67 to 1.00 (M = .91, SD = .12). The reliability of the pilot test of the EDIT Grammar was high (α = .80).

Table 5 shows the coding scheme we used, consisting of four categories: (a) predicted process, (b) partially predicted process, (c) unpredicted process, and (d) process difficult to categorize. Category (a) predicted process means the process that can be expected from the test specifications. In the example in Table 5, Female A explicitly stated that she understood the chunk to read and mentioned the order in which books comes before to read. These processes are predicted because the test intends to assess NP internal structures by having test-takers explicitly analyze NP internal structures (especially a modifying relationship between a head NP and a modifier, specifically in this item, to infinitive modifying the head noun). It should be noted that even when the selected option was wrong, the process intended in the test development was categorized here. Thus, another example belonging to this category is Female B's, in which to infinitive is mentioned and there is evidence that she analyzed the NP internal structure, although she did not explain why she selected Option 1.

Table 5. Examples of the protocols and corresponding categories

| Item | 2. 読むべき本が机の上にあります。  
(NP Group 4) | (There are books to read on the desk.) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. to read books  [Postmodification]</td>
<td>2. books to read*</td>
<td></td>
</tr>
<tr>
<td>3. read books to  [Phrase]</td>
<td>4. read to books  [Japanese]</td>
<td></td>
</tr>
</tbody>
</table>

(a) Predicted process
Female A: “Because yomu beki [to read] is to read, I put books in the beginning. Option 2.”

Female B: “Option 1 because with to-infinitive, the meaning yomu beki is expressed.”

(b) Partially predicted process
Male C: “Because of the expression yomu beki hon [books to read], probably to comes in between, and maybe to read is right. But I wonder if Option 2 or 4 is right, but when I consider yomu beki hon, books can also appear after. But with the pronunciation in mind, books to read sounds natural, so I select Option 2.”

(c) Unpredicted process
Male C (revised): “With the pronunciation in mind, books to read sounds natural, so I select Option 2.”

(d) Process difficult to categorize
Female D: “yomu beki hon has the same meaning yomu tameno hon, so Option 1”

Male E: “I wonder which. Probably Option 2.”

Category (b) partially predicted process refers to the process in which some parts of the utterances can be predicted from the test specifications but some parts cannot. In Male C’s example, the underlined part is in accordance with the test specifications, because the test-taker mentioned the use of to in English for the expression beki in
Japanese and wondered which English expression was appropriate for the Japanese presented in the stem (yomu beki hon), books to read and read to books. However, in the last utterance, he decided to select one option based on naturalness as an English expression, and determining the answer by its naturalness is not the process we intended to elicit. Overall, both predicted and non-predicted processes were seen, which led us to code this utterance as partially predicted process.

Category (c) unpredicted process is used when all the processes are not predicted, for example, when test-takers selected the option by its naturalness when pronounced or by intuition.

Category (d) process difficult to categorize is employed when a test-taker did not provide any clear explanation. In the example, Female D paraphrased yomu beki hon in Japanese and did not relate it to English phrases; therefore, it is difficult to ascertain why Option 1 was selected.

The tape-recorded utterances were transcribed by one of the authors or by a university student, whose transcripts were later double-checked by the same author. After discussing the criteria of the coding scheme, two of the authors coded 17% (29/[24*7] = 29/168) of the transcribed protocols. The interrater reliability was high, with the agreement ratio of 79% and Kappa coefficient of $\kappa = .53$, which indicates "a fair to good agreement" based on Landis and Koch (1977, as cited in Fleiss, Levin, & Paik, 2003, p. 604). The remaining transcribed protocols (83%) were coded by one author.

4.2. Study 1 results and discussion

Table 6 shows that the overall percentage of predicted agreement exceeded 60%. This percentage was considered to be acceptably high for two reasons. First, having a quarter of protocols in the process difficult to categorize seems unavoidable because test-takers do not always explain all their mental processes explicitly (Green, 1998).

Second, the second and third category (Partially predicted and unpredicted processes) had only 10% in total. There were three students who uttered protocols belonging to these categories, and each had a unique tendency for solving questions. Female F solved one item saying that it is a fixed expression. We found this problematic as an item to test NP internal structures and decided to exclude such items in the revised test. Female D answered six items by using irrelevant grammatical structures to select an answer; for example, when the answer to the item was the boy in the room, with other options being in the room the boy, the in boy the room, the room in the boy, she said, “The boy in the room is an answer mainly because there is only one boy.” The number of boys is not related to a prepositional phrase or word order; thus, this protocol was categorized as unpredicted process. Male C chose seven options using his experience of learning English, intuition, and naturalness as English expressions (see one example in Table 5). Although the processes of Female D and Male C were partially or completely unpredictable, they seem to be derived not due to problems of item quality, but rather due to the test-taker’s structures of grammatical knowledge or item selection strategy. Therefore, we would argue that their protocols are not negative evidence for the EDiT Grammar. Thus, the answer to RQ1 (Are the test-taking processes consistent with the expectations based on the test specifications?) can be positively stated: The processes were mostly consistent with the expectations.

<table>
<thead>
<tr>
<th>Category</th>
<th>Number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Predicted process</td>
<td>107 (64)</td>
</tr>
<tr>
<td>(b) Partially predicted process</td>
<td>5 (3)</td>
</tr>
<tr>
<td>(c) Unpredicted process</td>
<td>9 (5)</td>
</tr>
<tr>
<td>(d) Process difficult to categorize</td>
<td>47 (28)</td>
</tr>
<tr>
<td>Total</td>
<td>168 (100)</td>
</tr>
</tbody>
</table>

5. Study 2: Rasch analysis using the revised test

5.1. Study 2 method

The revised test of the EDiT Grammar (46 items) was conducted with 107 ninth-year public junior
high school students. They had already studied all the grammatical structures in the test. The test was administered as part of a normal English lesson by a teacher who wanted to obtain the information on students’ weaknesses. After the test, we scored test-takers’ responses dichotomously and estimated test-takers’ ability and item difficulty on the logit scale using Facets (Linacre, 2008a), a software program for conducting Rasch analysis. The reliability was high for items and test-takers (both .91).

5.2. Study 2 results and discussion

The infit mean squares for items were found to range from 0.73 to 1.33 ($M = 1.00$, $SD = 0.14$), whereas the infit mean squares for test-takers ranged from 0.66 to 1.36 ($M = 0.99$, $SD = 0.15$). When the infit mean squares of 0.50 and 1.50 were considered appropriate (Linacre, 2008b, p. 191), the analysis indicated no misfitting items and no misfitting students. We thus considered the test items and test-takers to fit the Rasch model.

When the relationship between the test-takers and items was examined, the test-takers’ ability ranged from -4.05 to 5.32 logits ($M = 0.18$, $SD = 1.40$), whereas the item difficulty varied from -1.46 to 1.78 ($M = 0.00$, $SD = 0.76$). Although the range of item difficulty was narrower than that of test-takers’ ability, the former covered most (84%, $n = 90$) of the test-takers. The result also shows that the peak of item distribution was similar to that of the test-takers’ distribution, suggesting that the distribution of items in terms of difficulty mostly correspond to that of test-takers.

6. Conclusion

In order to present the degree of validity of test interpretation and use based on test scores of the English Diagnostic Test of Grammar (EDiT Grammar), we have constructed an interpretive argument for the EDiT Grammar and reported two lines of empirical evidence that can be used in support of the validity argument. In the first research question (RQ1), we examined whether the test-taking processes employed by test-takers were in line with the expectations derived from the test specifications. We found the processes mostly consistent with the expectations. RQ2 investigated the item and person fit to the scoring model, where it was shown that all the items and test-takers fit the Rasch model successfully.

Table 7 summarizes all the backing collected so far, including the two strands of positive backing obtained from the current study. Each backing can contribute to the justification of the corresponding assumptions and warrants, which enhances the plausibility of each inference. In the case of the EDiT Grammar, we generally received favorable evidence strong enough for a low-stakes test, up to the explanation inference. Therefore, we would argue that a chain of inferences from the domain definition to the explanation was formulated and that the following test interpretation is tenable: The EDiT Grammar scores reflect a construct of the grammatical knowledge of NP internal structures in English as a foreign language. Thus, we claim that the test is sufficiently valid for interpreting the test scores as such. However, it is also clear that backing for the extrapolation and utilization inferences should be gathered in order to interpret the test scores as accounting for the quality of reading comprehension in English, and to use the scores for assisting diagnostic decisions and for guiding remedial instruction. Such test interpretation and use based on extrapolation and utilization inferences without the backing for these two inferences would lead to committing the begging-the-question fallacy (Kane, 2006, 2009).

The current study employed an argument-based approach to validity, originally proposed by Kane (1992, 2006, 2009) and further developed by Chapelle et al. (2008a). The advantages and usefulness of the argument-based validation framework, articulated by Chapelle, Enright, et al. (2010) were confirmed in our validation; namely, utilizing the interpretive argument as a basis for validation, strongly unifying assumptions with
specific inferences, having a clear direction toward the validity argument, and having the explicit role of the counterevidence. Therefore, the two problems that Chapelle (2008) mentioned about Messick’s (1989) framework—“what kind of evidence to gather and how much evidence is enough” (p. 321)—can be better addressed. In Kane’s framework, we can take a systematic approach to select types of evidence required for creating the interpretive argument, and to synthesize all the evidence collected into a validity argument. In addition to the aforementioned advantages of the argument-based validation framework, two strong points can be suggested. First, this framework can be used not only for a test consisting of discrete items intended to assess rather simple constructs like the EDiT Grammar, but also for a performance test that deals with complex constructs like the TOEFL iBT (Chapelle et al., 2008a). A second strength is that the framework excels in demonstrating weak areas that require backing for the validity argument.

<p>| Table 7. Summary of the backing for inferences in the interpretive argument |
|-------------------------------------------------|-------------------------------------------------|</p>
<table>
<thead>
<tr>
<th>Inference</th>
<th>Method to seek backing for the assumptions</th>
<th>Backing for each assumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Domain definition</td>
<td>1. &amp; 2. Domain analysis by expert judgment</td>
<td>1. Experts (English teachers at secondary schools and researchers in SLA and language testing) identified the knowledge of NP structures as required and important. 2. A relevant and representative test item format was designed by the experts.</td>
</tr>
<tr>
<td>2. Evaluation</td>
<td>1. Expert judgment</td>
<td>1. The scoring rubrics were made appropriate through trialing and revisions.</td>
</tr>
<tr>
<td></td>
<td>2. Item analysis</td>
<td>2. Item discrimination was appropriate</td>
</tr>
<tr>
<td></td>
<td>3. Rasch analysis (Study 2)</td>
<td>3. The test items and test-takers fit the Rasch model.</td>
</tr>
<tr>
<td>3. Generalization</td>
<td>1. Generalizability study</td>
<td>1. In most NP groups, a sufficient number of items were included; in the other groups, the number of items needed was identified.</td>
</tr>
<tr>
<td></td>
<td>2. Expert judgment</td>
<td>2. The test specifications were found to be well defined for creating parallel items.</td>
</tr>
<tr>
<td>4. Explanation</td>
<td>1. Verbal protocol analysis (Study 1)</td>
<td>1. The test-taking processes observed accorded with the expectations by the test developers.</td>
</tr>
<tr>
<td></td>
<td>2. Analysis of NP group difficulty</td>
<td>2. The means of the NP groups varied basically as predicted.</td>
</tr>
<tr>
<td></td>
<td>3. Correlational analysis</td>
<td>3. (To be examined in future studies)</td>
</tr>
<tr>
<td>5. Extrapolation</td>
<td>Correlational analysis</td>
<td>(To be examined in future studies)</td>
</tr>
<tr>
<td>6. Utilization</td>
<td>1. Expert judgment and survey analysis</td>
<td>1. We produced diagnostic profiles that were judged by the experts to be transparent and interpretable for teachers and test-takers. (Survey analysis is to be conducted in future studies)</td>
</tr>
<tr>
<td></td>
<td>2. Washback studies</td>
<td>2. (To be examined in future studies)</td>
</tr>
</tbody>
</table>
However, we have to admit that the argument-based validation framework is conceptually rigid and difficult to grasp; thus, it requires intensive reading of Kane (2006, 2009) and Chapelle et al. (2008a). More exemplary validation studies in various contexts would help practitioners use this approach in an informed and principled way. This concurs with a request by Chalhoub-Deville (2009) and Lissitz (2009) to include various examples in the future Standards.

Moreover, the argument-based approach still has some ambiguity. For example, it is not clear how many and what type of assumptions need to be presented in one inference in the interpretive argument. It seems that the number and types of assumptions, as well as the required strengths of the backing, vary depending on the stakes of the test and types of constructs (e.g., overall language proficiency or grammatical knowledge) and decisions (e.g., placement or diagnosis), but these were not very clear when reading the relevant literature. As one of the reviewers pointed out, the lack of clear rules in selecting assumptions underlying the warrant may tempt some test developers to choose only assumptions that produce positive evidence. Nonetheless, we believe this is where rebuttals play a part. Test developers themselves or critical observers of validation process can pose rebuttals for inferences or warrants that are related to missing assumptions. Then, test developers need to reveal logical or empirical backings to respond to the rebuttals; otherwise, their validity argument weakens, or worse, it may become difficult to sustain. As Kane (2006) admits, test developers are inevitably likely to have a confirmationist bias in the test development stage; however, test developers, together with critical outsiders, should begin by inspecting the test critically, presenting rebuttals, and reexamining the validity argument in the appraisal stage. It is hoped that the current validation study contributes to validity inquiry by showing an example of a low-stakes diagnostic grammar test. It is also hoped that the EDIT Grammar will further receive feedback, particularly with suggestions of possible rebuttals in the upcoming appraisal stage.

Acknowledgement
We highly appreciate ELPA, especially Etsuro Ikawa for supporting us for the test construction and giving us permission to publish the results. We are also thankful to Yo In'nami and three reviewers for their insightful comments to our paper.

References


Rie Koizumi, Hideki Sakai, Takahiro Ido, Hiroshi Ota, Megumi Hayama, Masatoshi Sato and Akiko Nemoto

Toward Validity Argument for Test Interpretation and Use Based on Scores of a Diagnostic Grammar Test for Japanese Learners of English

29th Annual Language Testing Research Colloquium, Barcelona, Spain.


Kanatani, K., & English Diagnostic Test (EDiT) Development Group. (2006). Eigo shindan test kaihatsu heno michi [Road to the development of the EDiT: Trajectory of the ELPA English Diagnostic Project]. Tokyo: ELPA.


Appendix A. Sample Test Items

Choose the most appropriate English phrase for the underlined part of the Japanese sentence from 1 to 4.

1. 彼女はすてきな歌を5曲歌いました。
   (NP Group 2)
   1. five songs nice [Premodification]
   2. songs nice five [Premodification]
   3. nice songs five [Japanese word order]
   4. five nice songs*

2. これは、平和について語っている人々の写真です。
   (NP Group 4)
   1. peace about talking people [Japanese word order]
   2. people talking about peace*
   3. talking people about peace [Phrase]
   4. talking about peace people [Postmodification]

3. 軽井沢で人気の高い店を教えてあげます。
   (NP Group 5)
   I will show you shops which are popular in Karuizawa.
   1. are popular in Karuizawa which shops [Japanese]
   2. shops are popular in Karuizawa which [Clause]
   3. shops which are popular in Karuizawa*
   4. which are popular in Karuizawa shops [Postmodification]
Appendix B. Example of diagnostic feedback (English version)

<table>
<thead>
<tr>
<th>Results of the EDIT Grammar</th>
<th>Name</th>
<th>Tsuto Yamada</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Result</td>
<td>14 out of 46 points (30% correct)</td>
<td></td>
</tr>
</tbody>
</table>

- **Acquisition rate for each group**
  - **Premodification**
    - 92% ○
    - Example: [Tom’s beautiful] [car]
    - Noun phrases premodified by premodifiers (e.g., adjectives)
  - **Postmodification**
    - Prepositional phrase
      - 10% ×
      - [a surprise party] [for my brother]
      - Modified by prepositional phrases
    - Present/past participle
      - 0% ×
      - [a couple] [having dinner]
      - Modified by present participles
      - [books] [written in English]
      - Modified by past participles
  - **Relative clause**
    - 17% ×
    - [things] [(that) I bought yesterday]
    - Modified by relative clauses
  - **To-infinitive**
    - 0% ×
    - [chances] [to go to parties]
    - Modified by to infinitive phrases

**Description**

It is likely that this student can accurately understand English sentences with premodified noun phrases but NOT the ones with postmodified noun phrases.

- **Characteristics of error patterns:** Lower percentages mean better results. If the student has × marks, s/he needs immediate remedial instruction for the areas.
  - ○ = Less than 30% [Pass]; △ = 30~50%; × = 50% or more

| Can s/he understand premodyifying structures? | 8 | ○ | Correct: the boy’s name
Selected: the name boy’s |
| Can s/he understand internal structures within preposition phrases and/or present/past participle? | 18 | ○ | Correct: the boy in the room
Selected: the boy the room in (S/he does not understand the internal structure within in the room) |
| Can s/he understand internal structures within relative clauses? | 17 | ○ | Correct: any friends who can play the guitar
Selected: any friends who the guitar play can (S/he does not understand the internal structure within who can play the guitar) |
| Can s/he understand structures in which one modifies another? | 38 | △ | Correct: the boy in the room
Selected: in the room the boy (S/he does not understand the structure in which in the room postmodifies the boy) |
| Can s/he understand English word orders, without being affected by Japanese ones? | □2 | △ | Correct: the boy in the room
Selected: the room in the boy (S/he selected the option with the Japanese word order) |