Title	What L2 grammar assessment teaches language teachers
Sub Title	文法能力測定と言語教育
Author	中村, 優治(Nakamura, Yuji)
Publisher	慶應義塾大学日吉紀要刊行委員会
Publication year	2009
Jtitle	慶應義塾大学日吉紀要. 言語・文化・コミュニケーション (Language, culture and communication). No.41 (2009.) ,p.31- 43
JaLC DOI	
Abstract	
Notes	
Genre	Departmental Bulletin Paper
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=AN10032394-20091218-0031

What L2 grammar assessment teaches language teachers

Yuji Nakamura

1. Theoretical aspect of grammar assessment

Grammar has played a critical role in L2 teaching, learning and assessment. This paper will show how L2 grammar assessment is related to language teachers.

Research and theory related to the teaching and learning of grammar have made significant advances over the years. In applied linguistics, our understanding of language has been vastly broadened by the work of corpus-based and communication-based approaches to language study, and this research has opened new pathways into recent pedagogical grammar theory (Purpura, 2004). In language teaching, there has been a shift of focus from structures and metalingtuistic terminology to greater emphasis on comprehensible input, interaction, and no explicit grammar instruction (Purpura, 2004). Current research in grammar instruction involves investigation into the effects of teaching grammar explicitly or implicitly, reactively or proactively, and integrated in the curriculum punctually or sequentially (Doughty, 2002).

All these developments have implications as to the optimal assessment of grammatical ability, and how it should be used. However, theoretical discussions on the nature of grammatical ability have been to this day not extensive(Purpura, 2004).

Purpura (2004) discusses how grammar assessment has evolved over the years and how it has begun to change, demonstrating some theoretical and practical issues and challenges that language educators face in assessing grammatical ability. In the last fifty years, language testers have dedicated a great deal of time to debating the nature of language proficiency and the testing of the four skills, the quality of test usefulness (e.g., reliability, authenticity), the relationship between

test-taker or task characteristics and performance, and numerous statistical procedures for examining data and providing evidence of test validity. In all of these discussions, very little has been said about the assessment of grammatical ability and, unsurprisingly, not much has changed since the 1960s. In other words, for the past fifty years, grammatical ability has been defined in many instances as morphosyntactic form and tested in either a discrete-point, selected-response format—a practice initiated by several large language-testing firms and emulated by classroom teachers—or in a discrete-point, limited-production format, typically by means of cloze or other gap-filling tasks. (Purpura, 2004).

In recent years, the assessment of grammatical ability has taken an interesting turn in certain situations. Grammatical ability has been assessed in the context of either speaking or writing sections. One example of this kind of test is the new iBT TOEFL where structural separation was abandoned in favor of a more integrated version, testing all four language skills. In this case, grammatical ability is not scored separately by means of an analytic rubric; rather, it contributes holistically to a language performance score in support of other components of language knowledge (Purpura, 2004).

The issue of grammar versus vocabulary should be further explored. Whereas vocabulary is words (specific items we find in a dictionary), grammar consists in putting these words together. Grammar seems to have a more general character involving rules that apply to whole categories of items (Swan, 2005). Rather than seeing grammar and vocabulary as totally separate and distinct, linguists tend increasingly to conceptualize the two in terms of a continuum. At one end there are facts about language that are indisputably lexical. At the other, there are purely grammatical phenomena. Between the two extremes lies the grammar of individual words; there are words that are grammar as much as they are vocabulary; there are structures that are bound to small groups of words; there are structures that, in practice at least, are more vocabulary than grammar. As with most elements in the world that seem easy to distinguish on first sight, vocabulary and grammar are undoubtedly of the different kind, but at no moment no one can assert "This is where one stops and the other starts" (Swan, 2005).

According to Purpura (2004), we need to take the five issues and challenges associated with assessing grammar into account: a) defining grammatical ability, b) scoring grammatical ability, c) assessing meanings, d) reconsidering grammar test tasks, and e) assessing the developmental grammatical ability. They can be roughly recategorized into three: 1) definition of grammatical ability, 2) scoring form and meaning of grammatical ability in new complex performance tasks and 3) assessing

acquisitional development of grammatical ability.

The first and foremost challenge revolves around how grammatical ability has been defined both theoretically and operationally in language testing. What is the construct of grammatical ability? In the 1960s and 1970s language teaching and language testing maintained a strong syntactocentric view of language rooted largely in linguistic structuralism. Grammatical knowledge was determined solely in terms of linguistic accuracy (Pupura, 2004). Since the 1980s the description of language knowledge has been significantly broadened with a push towards communicative language teaching and with theoretical models of communicative competence (cf. Canale and Swain, 1980; Bachman, 1990; Bachman and Palmer, 1996). In most of these models, grammatical knowledge was defined in terms of phonology, morphology, lexis, syntax and sometimes semantics. Nevertheless, in spite of the emphasis on meaning in communicative language teaching and learning, the role of meaning in those models of language proficiency has been difficult to locate, because in none of those models has the role of meaning or its relationship to linguistic form been defined or specified (Purpura, 2004).

Purpura (2004) proposed a model of grammatical knowledge that includes both grammatical form and meaning on the sentence and discourse levels, since in communicative language testing, assessing both grammatical form and meaning provides teachers and learners with a more complete assessment of the test-takers' grammatical ability than just providing information on form or on meaning alone. As Purpura (2004) pointed out, with some grammatical structures learners find the form more challenging to learn than the meanings, whereas with other structures, they find the meanings more difficult. It is inevitably important for language educators to make clear distinctions between the form and meaning components of grammatical knowledge and to incorporate these distinctions in construct definition.

The second challenge is scoring grammatical ability. This deals with scoring form and meaning of grammatical ability in new complex performance tasks. In other words, it includes 1) the dichotomous scoring or partial-credit scoring and 2) the scoring of grammatical ability in complex performance tasks. This challenge, furthermore, is related to the design of test tasks that are capable of both measuring grammatical ability and providing authentic and engaging measures of grammatical performance.

Since the specification of both form and meaning is likely to influence the ways in which grammar assessments are scored, language teachers might need to adapt their scoring procedures (i.e. dichotomous or partial-credit) to reflect the

two dimensions (form and meaning) of grammatical knowledge (Purpura, 2004). Also, the advantage and disadvantage of using complex performance tasks should be well taken into account when it comes to the scoring of grammatical ability. The advantage is the generalizability of the inferences the highly authentic tasks allow us to make about grammatical ability. The disadvantage is, on the other hand, the potential lack of accuracy with which we are able to infer what test-takers know about specific grammatical knowledge (Purpura, 2004).

While there is a place for discrete-point tasks in grammar assessment, language teachers have long used a wide range of simple and complex tasks in which to assess test-takers' explicit and implicit knowledge of grammar (Purpura, 2004). As Purpura (2004) pointed out, in reflecting upon the design of authentic tasks, we must acknowledge that not all grammar assessments in large-scale or classroom contexts need to have communication as their primary assessment goal. There are times, in fact, when we simply wish to know if students have understood the meaning of a form or if they have acquired explicit knowledge of a particular grammatical structure without the complexities of rich context or on-line, spontaneous performance (Purpura, 2004; cf. Nakamura, 2007). When language teachers and testers begin with a grammar assessment goal, the challenge is to identify tasks within the target language use (TLU) domain that elicit only that aspect of the grammatical ability that we wish to measure.

The third challenge revolves around the argument made by some researchers that grammatical assessments should be constructed, scored and interpreted with developmental proficiency levels in mind. Researchers argue that the acquisitional development of learners should be a major consideration in the L2 grammar testing. While it makes sense that the challenge for language testers is to design, score and interpret grammar assessments with a consideration or developmental proficiency, what basis can we use to infer progressive level of development ? It is also argued that the research based on developmental orders and sequences is vastly incomplete and at too early a stage for use as a basis for assessment (Purpura, 2004; Hudson, 1993).

Although the suggestion that grammar-test tasks be designed to give credit to learners who demonstrate knowledge of advanced interlangugae forms is well taken, and even relevant information that relates to development could be incorporated into the rating-scale descriptors, Purpura (2004) still recommends that grammar assessments based on developmental sequences be used only in research, not for decision-making. He maintains that acquisitional developmental levels need not be the only basis upon which to make inferences about grammatical

development (cf. Nakamura, 2006).

2. Empirical data analysis of grammar tests

The second half of this paper shows how L2 grammar assessment is related to language teachers. It reports an empirical study that examined 75 grammatical items in five placement tests which were administered to more than 4000 students. The data was analyzed in terms of the grammar categorization, an item analysis (item difficulty and item discrimination), and students' proficiency level setting.

Method

Subjects

More than 4000 freshman university students in total for three years in K University

Materials/ Instruments

A grammatical knowledge test (15 items) as part of a placement test for measuring students' English reading ability as well as grammar and vocabulary knowledge. As for Item Banking, 25% of the 15 items were linked to equate two tests, thus, 4 items in total were retained and 11 items were replaced in each test. Therefore, the total number of items is 75 (including linked ones), but the genuinely new items are 57 in total in the data. We need to choose items based on item locations, in other words, item difficulties.

Procedure

The grammar items were chosen by taking into consideration almost all of the grammar items that were supposed to have been mastered at the high school level. The test was a multiple-choice format rather than a response construct test, and the scoring was done using the optical mark reader in an objective way.

Data Analysis

The data was analyzed in terms of the grammar categorization, an item analysis (item difficulty and item discrimination), and students' proficiency level setting.

Table 1 The ratio of four levels (percentage) in each test

Advanced level 7%
Level 3 30%
Level 2 58%
Level 1 5%

Total approximately 850 students in each test were grouped into four levels according to their raw scores (number right scores)

Table 2 Item characteristics

Discrimination: item discrimination

Proportion: item difficulty

PC level 1-PC advanced level: proportion of students who got each item correct

Category: grammatical category that should explain the testing point of each item

PT	item number	discrimination	Proportion of correct response(PC)	PC level 1	PC level 2	PC level 3	PC advanced level	category
PT1	G01	0.5	86	13.2	85	96.9	98.4	verbs
PT1	G02	0.47	68	10.5	58.1	87.1	98.4	conjunctions
PT1	G03	0.41	72	21.1	64.7	88.6	96.7	verbs
PT1	G04	0.2	87	73.7	83.4	93.7	93.4	relative adverbs
PT1	G05	0.44	81	18.4	79.0	92.2	95.1	collocations
PT1	G06	0.44	86	23.7	84.2	96.9	93.4	modals
PT1	G07	0.32	56	23.7	46.7	70.6	86.9	verbs of perception
PT1	G08	0.3	84	31.6	84.2	89.8	88.5	verbs
PT1	G09	0.38	89	36.8	88.0	94.9	98.4	comparison
PT1	G10	0.17	55	44.7	51.3	60.0	70.5	adjectives
PT1	G11	0.3	40	18.4	31.3	52.2	80.3	collocations
PT1	G12	0.07	44	36.8	40.5	48.2	55.7	prepositions
PT1	G13	0.44	68	2.6	63.1	85.5	82.0	verbs
PT1	G14	0.39	80	34.2	75.8	91.4	98.4	inversion
PT1	G15	0.47	86	23.7	83.2	96.9	100.0	concessive clauses

PT2	G16	0.31	51	13.3	42.3	58.9	82.3	verbs
PT2	G17	0.3	85	51.1	79.6	91.6	97.5	subjunctive
PT2	G18	0.17	58	51.1	46	65.4	83.5	passive
PT2	G19	0.18	35	26.7	25.9	41.7	53.2	relative pronouns
PT2	G20	0.24	82	40	79.9	87.2	88.6	verbs of perception
PT2	G21	0.47	83	24.4	74.6	94.1	98.7	collocations
PT2	G22	0.14	69	55.6	64.9	72.6	78.5	nouns
PT2	G13	0.45	62	4.44	51.7	75.7	88.6	verbs
PT2	G23	0.3	69	31.1	59.5	79.1	83.5	relative pronouns
PT2	G02	0.5	73	15.6	62.7	87.5	93.7	conjunctions
PT2	G24	0.16	20	15.6	14.9	19.9	44.3	adjectives
PT2	G25	0.39	80	28.9	71.1	91.6	93.7	prepositions
PT2	G14	0.5	78	15.6	67.9	91.3	98.7	inversion
PT2	G11	0.33	46	8.9	33.6	56.7	70.9	collocations
PT2	G07	0.3	64	20	56.7	73.8	79.7	verbs of perception
PT3	G26	0.11	96	92.9	94.7	97.1	100	collocations
PT3	G21	0.47	83	21.4	81.5	95.5	94.5	collocations
PT3	G27	0.31	92	54.8	92.7	95.9	98.6	infinitives
PT3	G28	0.32	74	35.7	70	85.7	84.9	subjunctive
PT3	G16	0.31	56	21.4	49.9	66.5	82.2	verbs
PT3	G29	0.46	83	21.4	79.5	95.9	97.3	verbs/passive
PT3	G30	0.18	36	14.3	33.1	39.6	58.9	verbs
PT3	G31	0.44	89	33.3	89.5	97.1	98.6	conjunctions
PT3	G19	0.17	30	31	22.3	40	50.7	relative pronouns
PT3	G32	0.22	61	38.1	56	70.6	76.7	conjunctions
PT3	G33	0.16	23	11.9	19.3	29.8	39.7	comparison
PT3	G34	0.28	91	64.3	88.6	96.7	98.6	verbs
PT3	G07	0.37	64	19	57.6	77.6	93.2	verbs of perception
PT3	G11	0.34	42	19	34.1	55.5	72.6	collocations
PT3	G35	0.31	77	33.3	74.2	89.4	83.6	adverbs

PT4	G36	0.21	53	16.3	49.8	59.5	55.7	verbs
PT4	G37	0.27	47	14	41.7	55.6	70.5	subjunctive
PT4	G38	0.3	91	55.8	89.8	95	98.4	causative verbs
PT4	G39	0.22	89	65.1	86.2	94.6	98.4	relative adverbs
PT4	G40	0.28	57	23.3	51.2	65.2	85.2	verbs
PT4	G41	0.19	63	32.6	60.8	68.1	73.8	modals
PT4	G27	0.28	89	44.2	88.2	92.8	98.4	infinitives
PT4	G30	0.18	40	32.6	32.7	44.8	72.1	verbs
PT4	G42	0.19	62	30.2	58.7	68.1	72.1	comparison
PT4	G33	0.21	24	18.6	18.3	30.8	45.9	comparison
PT4	G43	0.33	71	25.6	65.9	80.3	91.8	prepositions
PT4	G44	0.4	84	34.9	78.9	93.9	98.4	conjunctions
PT4	G45	0.33	72	46.5	62	85.3	95.1	prepositions
PT4	G46	0.19	68	48.8	64.2	70.6	85.2	conjunctions
PT4	G07	0.36	61	11.6	54.5	72	90.2	verbs of perception
PT5	G47	0.26	89	67.4	87.1	94.4	95.2	verbs
PT5	G37	0.22	65	37	61.7	73.4	77.8	subjunctive
PT5	G48	0.3	67	39.1	58.8	79.4	96.8	infinitives
PT5	G49	0.26	47	19.6	41.3	56	79.4	relative pronouns
PT5	G50	0.29	75	50	68.5	86.5	93.7	collocations
PT5	G51	0.25	85	56.5	82.5	91.3	95.2	pronouns
PT5	G40	0.27	55	34.8	48.1	63.1	90.5	verbs
PT5	G52	0.3	70	30.4	66	81	82.5	inversion
PT5	G53	0.29	72	21.7	71.5	82.1	79.4	subjunctive
PT5	G54	0.19	9	8.7	5	7.5	39.7	collocations
PT5	G55	0.31	85	50	82.7	94	98.4	chance
PT5	G56	0.34	91	65.2	89.4	98	100	verbs of perception
PT5	G45	0.34	71	34.8	64.8	83.3	96.8	prepositions
PT5	G44	0.38	89	50	87.3	96.8	98.4	conjunctions
PT5	G57	0.37	77	30.4	73.8	90.1	85.7	verbs

Table 2 supplies many pieces of information of each item and each level of students as follows:

1. The item discrimination informs us how each item discriminates the upper and the lower level students as a whole group.
2. The item difficulty shows us how difficult each is (or how easy each item is). This can also mean how many students get each item correct.
3. The proportion of students who got each item correct in four different levels provides teachers with text book choice information, class target information, diagnostic information of grammatical knowledge, inter-group difference information as well as intra-group information.
4. The category information indicates that what part of grammatical knowledge has been mastered, what has not been acquired. This information is very useful when choosing the textbook and writing test items.

The following five figures also provide a number of useful pieces of information of items and four levels of students.

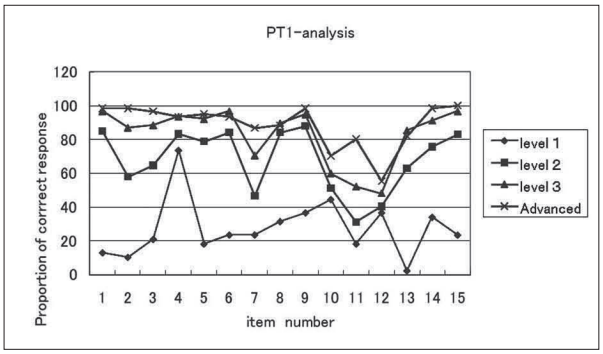

Figure 1 Grammar Test 1

Figure 1 shows the relative difference among four levels of students. The level 1 students almost constantly are placed at the lower end of the graph. The two top levels (level 3 and advanced level) move rather similarly. Although there is an item difficulty index in Table 2 as a whole group, this figure gives us detailed information how each group is behaving for each item.

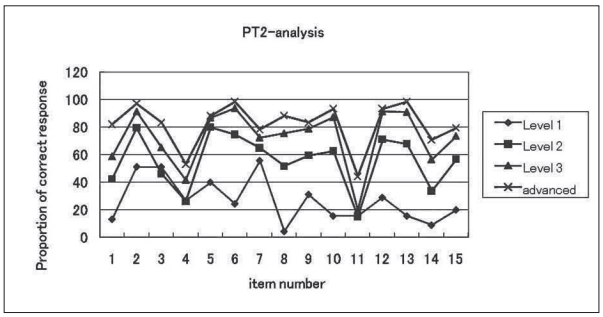

Figure 2 Grammar Test 2

Figure 2 indicates an idiosyncrasy of item 4 and item 11. In these two items, all the four levels find them rather difficult, and furthermore, the difficulty level of these two items is almost the same for the two student levels. One further noticeable thing is that in item 3, apparently level 1 students are behaving better than level two students.

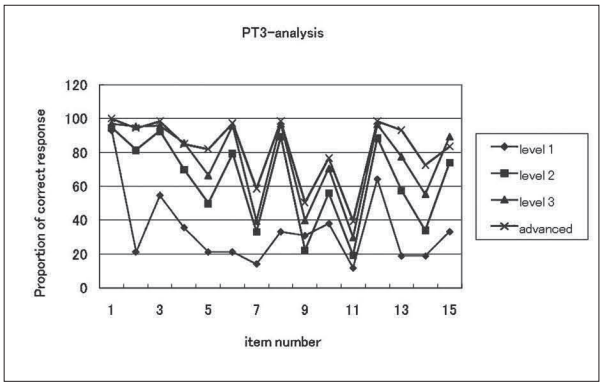

Figure 3 Grammar Test 3

Figure 3 shows that item 1 is very easy for all of the four levels of students, while item 11 is rather difficult for the same four levels. This figure also indicates that level 3 and advanced level students behave similarly except in item 15 they are behaving oppositely, in other words level 3 is better than the advanced level.

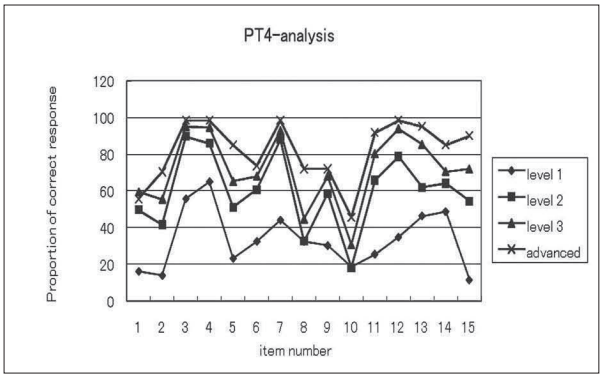

Figure 4 Grammar Test 4

Figure 4 demonstrates two typical examples of items in which four levels of students are separated in a well-balanced way. This figure also tells us that three levels (advanced, 3 and 2) are behaving in a more or less similar way, whereas level 1 is behaving similarly in a distant level.

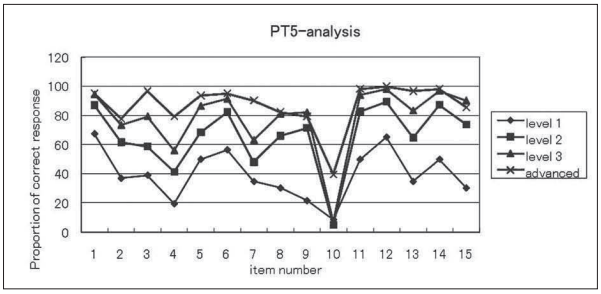

Figure 5 Grammar Test 5

Figure 5 indicates us almost the same pattern as shown in the fires above. This figure convinces us that the information of the four levels can give us different results from the difficulty and discrimination mentioned in Table 2 where the information of each item from the whole group is available.

Implications

The result in the categorization will be useful for language teachers, while the item analysis will provide the test item writer with important information. Still, the item discrimination and difficulty will be more interesting for the second language

acquisition researchers.

3. Conclusions

This paper dealt with two issues. One was the theoretical aspect of grammar assessment by taking into researchers concepts. The other was the empirical data analysis of grammar tests. On the theoretical side, three issues were discussed: 1) definition of grammatical ability, 2) scoring form and meaning of grammatical ability in new complex performance tasks and 3) assessing acquisitional development of grammatical ability. All of these topics should be well taken in the assessment of grammar knowledge. On the empirical side, not only the classical test theory-based analysis but also the item response theory-based analysis was demonstrated. Basically both have complementary roles which can help each other to better assess students grammar knowledge.

Simple discrete-point grammar assessment can benefit teachers in various ways. Firstly, the discrete point grammar test provides the teachers with diagnostic information on students' grammatical knowledge which can be used for students' feedback. Secondly, the item characteristics information (item difficulty and item discrimination examined by Classical Test analysis) still can give test users basic and useful information on each item in a traditional way.

Furthermore, the information from level setting differences (four levels of students) provides different types of useful and practical information for the teachers. Additionally, item calibration through Item Response Theory can help equate tests with anchor items linked by the IRT analysis.

Both the theoretical idea and the empirical data make a complementary role with each other and they are both beneficial not only for language teachers but also for testers and SLA researchers.

Acknowledgement

I am grateful to Prof. Kyoko Yoshida and Prof. Satoko Tokunaga for their invaluable help with the data analysis. I would also like to thank all the members of the Keio Placement Test Research Group for their cooperation in editing and administering placement tests.

Bibliography

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

Bachman, L.F. and Palmer, A.S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1–47.

Doughty, C. (2002). Focus on form: the latest word for teachers. Third Annual Apple Lecture, Teachers College, Columbia University, New York City.

Grabe,W. (2000). Reading research and its implications for reading assessment. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp.226–62). Cambridge: Cambridge University Press.

Hudson, T. (1993). Nothing does equal zero. *Studies in Second Language Acquisition, 15,* 461–93.

Nakamura, Y. (1998). Components of Reading Ability. *Educational Studies* 40, 259–281. International Christian University.

Nakamura, Y. (2006). *An Investigation of the Construct of L2 Reading Comprehension*

Ability. Proceeding of the JACET summer seminar. The Japan Association of College English Teachers.

Nakamura, Y. (2007). A Rasch-based analysis of an In-house English Placement Test. In T. Newfields (Ed.), *2007 JALT Pan-SIG Conference Proceedings.* Tokyo: The Japan Association for Language for Language Teaching (pp. 1–14).

Purpura, J. E. (2004). *Assessing Grammar.* Cambridge: Cambridge University Press.

Swan, M. (2005). *Grammar.* Oxford, Oxford University Press.