DEVELOPING THE CAN-DO SYSTEM BASED ON THE NCUEE TEST RESULTS: AN APPLICATION OF THE NEURAL TEST THEORY

Naoki Sugino
(gwisno@is.ritsumei.ac.jp)
Ritsumeikan University, Japan

Kenichi Yamakawa
(kyamakaw@yasuda-u.ac.jp)
Yasuda Women's University, Japan

Hiromasa Ohba
(hohba@juen.ac.jp)
Joetsu University of Education, Japan

Kojiro Shojima
(shojima@rd.dnc.ac.jp)
National Center for University Entrance Examinations, Japan

Yuko Shimizu
(yukos@ec.ritsumei.ac.jp)
Ritsumeikan University, Japan

Michiko Nakano
(nakanom@waseda.jp)
Waseda University, Japan

Abstract

As a preliminary study in the research project to explicate the relationship between Japanese EFL learners' grammatical proficiency and their functionality in the target language, the present study applies Neural Test Theory (Shojima, 2008a, 2008b, 2009), a test theory for grading academic achievement on an ordinal scale, to the actual data of the test takers of National Center for University Entrance Examinations (NCUEE) Test. Conventionally, test results are given on a continuous scale. It is, however, extremely difficult to discern what a learner with a score of, say, 78 out of 100 can do, and how his/her ability differs from the ability of another learner with the score of 83 in the same test. Employing this new test theory, learners' achievements are evaluated in ordinal grades, or "latent ranks." This analysis also yields item reference profiles (IRPs), which express each test item's correct answer ratios for each of the latent ranks. By sorting the items according to the correct response ratio and the IRPs, we can obtain groups of test items that learners at a latent rank can answer correctly but those at lower latent ranks cannot. Based on this information, we can describe the ability of the learners at each of the latent ranks.

1 Introduction

This study is carried out as a part of the authors' collaborative research project in which they attempt to explicate what learners of English as a foreign language can do with their attained level of grammatical proficiency. So far, the authors have developed several grammaticality judgment tests, each of which focuses on a particular grammatical feature, such as dative alternation, psych verbs, unergative/unaccusative verbs, *wh*-question construction, and relative clause construction, and have administered them to university-level Japanese EFL learners. The results obtained were then equated by using Item Response Theory (IRT) so that direct comparisons of the items become possible on one difficulty scale (Yamakawa et al., 2008).

In order to investigate relationships between learners' grammatical proficiency and how their grammatical knowledge is turned into use in language skills, it is required to clearly state what a learner can do in the target language. Thus, the aim of the present study is to apply a newly-developed test theory to a set of actual data from a large-scale test, and by so doing, to prove that this test theory is a powerful tool in constructing a system of can-do statements based on the data.

2 Brief outline of neural test theory

If we are to measure physical quantity, such as weight, a measurement tool is highly reliable: Someone weighing 61 kg can always be distinguished precisely from someone weighing 62 kg, and we can, with high accuracy and much confidence, rank them on a continuous scale. On the other hand, tests used in the fields of psychology, pedagogy or behavioral science usually intend to measure characteristics which cannot be observed directly, and they cannot guarantee the same degree of measurement accuracy as weighing scales do. For example, we cannot exactly specify the differences between two test takers with the scores of 61 and 62 in terms of their ability (See Figure 1). As the reliability studies in the Classical Test Theory have clarified, any test intending to measure psychological traits rarely achieves the reliability of 0.9. Roughly said, 10% of what is measured is measurement error. Tests do not have sufficient resolution to evaluate human ability nor to rank test takers on a continuous scale.

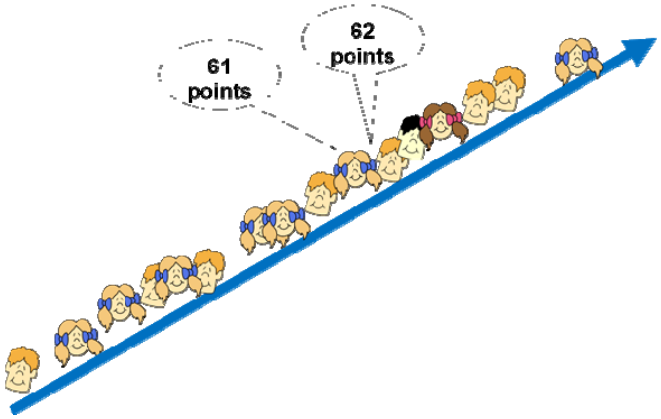

Fig. 1. Test scores placed on a continuous scale

Despite this limitation, scores of a test are mostly taken and regarded as something reliable and absolute, something similar to the weights. This is even more so if the tests are large-scale and standardized. In test standardization, IRT is employed in equating difficulty levels of the test items. However, IRT assumes theta, a latent continuous variable, and therefore, is a statistical model that evaluates examinees' abilities on a continuous ability scale. The resolution issue of a test and the difficulty in relating scores to abilities remain unsolved.

Originating from concern about this resolution issue, Kojiro Shojima and his colleagues developed Neural Test Theory (NTT, henceforth) that evaluates academic achievements in ordinal grades (Shojima, 2008a, 2008b, 2010). Simply put, NTT is a theoretical model to detect latent ranks into which test takers are classified, and simultaneously, to provide rich information on each of the test items. Ranks are "latent," because they are hidden behind the scores. NTT uses the self-organizing map or generative topographic map mechanism to detect the ranks and place each examinee on the ordinal scale. Latent ranks can be schematized as in Figure 2 below.

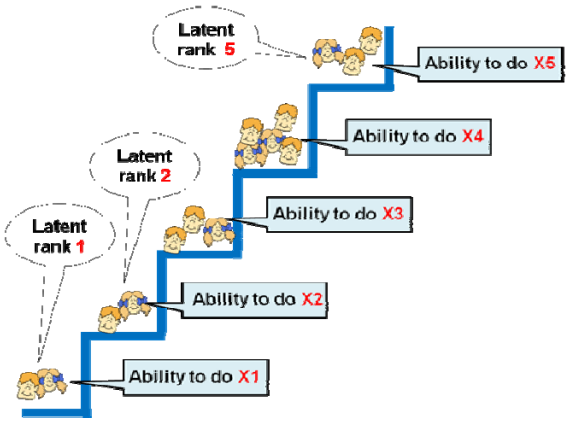

Fig. 2. Test scores placed on an ordinal scale

Of many kinds of output NTT produces, Item Reference Profile (IRP) is of crucial importance in the present study. IRP is an item-wise statistical information and is given as a set of correct response ratios from examinees at each latent rank; Figure 3 below is a graphic representation of IRPs of sample items.

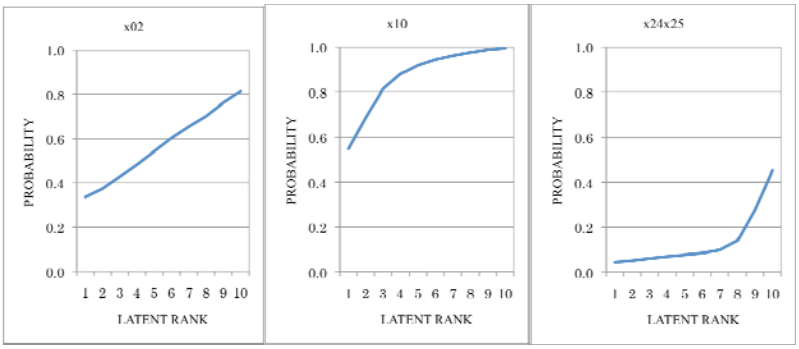

Fig. 3. Graphic representations of IRPs of Items 2, 10 and 24-25

3 The present study

3.1 The NCUEE test

The National Center for University Entrance Examination (NCUEE) has been developing a nationally-applied university entrance examination (henceforthe, the NCUEE Test) used by all public universities as well as some private universities in Japan since 1979. Since then, increasing numbers of universities and colleges, both public and private, are using the NCUEE Test, and currently more than 500,000 students take the test every year.

The NCUEE is an independent administrative institution and responsible not only for developing and administering the NCUEE Test, but researching various aspects to make the tests more valid and reliable. In accordance with the Courses of Study developed by the Japanese Ministry of Education, Culture, Sports and Technology (henceforth the MEXT), the NCUEE Test consists of six subject areas including English as one of the five foreign languages. Its purpose is to measure test takers' achievement levels of what they have learned throughout high school education. Test takers are to take the subjects specified by the university or universities to which they are to apply. Among the subjects, English is the most commonly taken; more than 97% of the test takers choose English as one of the test subjects. The NCUEE Test is held annually in mid-January over a period of two days using high schools and universities all over Japan as test sites. However, the entire process of administration is supervised by the NCUEE to ensure security as well as uniformity.

Much about the history of the NCUEE or the NCUEE Test will not be detailed in this paper (cf. Guest, 2008a, 2008b); it would suffice here to mention that the test has gone through several phases since 1979. Focusing on the English test, the year 2006 was epoch-making since a listening section (50 points) was added to the original section (200 points) for the first time in the last quarter century. Since we use the original paper-pencil section of the English test in this study, we will provide detailed information of the test format.

3.2 Item specifications of the 2004 test

The English test consists of six sub-tests (see Table 1). Sub-test 1 focuses on test takers' phonological/phonetic awareness. Sub-test 2 consists of 10 discrete-point items of grammar and idioms, three dialogue completion items and three items to rearrange scrambled words and phrases to form a sentence. Sub-test 3 mainly assesses test takers' understanding of coherence of short passages. Sub-tests 4, 5 and 6 assess reading comprehension of different types of texts. All items use a multiple-choice format for response. Table 2 below summarizes skills and knowledge being tested by each item. Although task types, weightings and the total number of items differ slightly from year to year, the basic organization is mostly maintained.

		Contents and Points	# of items	point allocation
1	A	Word accent (@ 2 pts.)	2	16
	B	Sentence stress in verbal interaction (@ 3 pts.)	4	
2	A	Idiom/grammar (@ 2 pts.)	10	38

	B	Dialogue (@ 2 pts.)	3	
	C	Scrambled sentences (paired; @ 4 pts.)	3	
3	A	Coherence of short passages (@ 2 pts.)	2	
	B	Coherence of short passages (@ 5 pts.)	2	34
	C	Coherence of short passages (@ 6 pts.)	3	
4	Comprehension of an expository text with visual prompt (@ 7 pts.)		5	35
5	Comprehension of long discourse (@ 6-7 pts.)		5	32
6	Comprehension of narrative (@ 5-6 pts.)		8	45
		Total	50	200

Table 1. Overall structure of the 2004 test and the abilities/knowledge tested

3.3 Procedure

Of the 535,944 actual test takers, 40,000 cases were randomly selected for the analysis. The use the actual test takers' data had been approved by the Information Disclosure Committee of the NCUEE, on the condition that identities of the test takers be kept anonymous throughout the research. The data were analyzed using *Exametrika* 4.2 (Shojima, 2010).

4 Results

Table 3 below shows IRPs of each item, sorted by the correct response ratio (CRR). Considering the number of items and the sample size, the number of latent ranks was set at ten.

IRPs should be understood as the probabilities learners at a particular latent rank can arrive at the correct response for that item. See, for example, the Item 10's row; Rank 1 learners' probability to answer correctly in Item 10 is 54.8%, while that of Rank 10 goes up to 99.5%. Cells are shaded if the probability is above 0.500, and the shades become darker as the probability improves. As is indicated by the thick line at items with probabilities above 0.600, learners at a higher latent rank have more items within their capabilities, as well as have more chance to be able to answer those items correctly.

Combining Tables 2 and 3 yields Table 4, based on which abilities and knowledge of the learners in each rank can be described as follows:

Rank 1 learners have insufficient knowledge in vocabulary, grammar, accent patterns, and sentence stress, and their reading comprehension skills are limited.

Rank 2 learners shows some understanding of the basic sentence structure (Item 5), and started to displays their understanding of dialogues (Item 19). Also, if they do not need to read across paragraphs, they can understand information in the text with the aid of some visual prompt (Item 33).

Rank 3 learners are characterized by their understanding of, and making appropriate responses in, dialogues (Items 17, 18 and 19), and their successful performance in placing sentence stress (Items 3, 4 and 5). However, being able to effectively use sentence stress to

emphasize a point (Item 6) needs to wait till much later (at Rank 9). With the visual aid, they can understand a longer passage (Item 39, and although slightly less successful, in Item 36).

Rank 4 learners start to display their ability to make inference based on information within a paragraph (Item 43), which is strengthened by a visual prompt when their understanding needs to be based on information across paragraphs (Item 36).

Rank 5 learners are differentiated from Rank 4 learners because of their wider knowledge of vocabulary and collocation (Items 13 and 15). They can use the linguistic markers to express logical relations within a sentence (Item 12) as well as within a paragraph (Item 27).

Rank 6 learners display their knowledge of grammar necessary to produce more complex sentences (Item 20-21). They start to show the ability to sequencing sentences within a paragraph (Item 29).

Rank 7 learners are characterized by improved discourse competence as indicated by Items 26, 30 and 31. Their repertoire of logical connectors appears to be wider, and they can form coherent passages with several paragraphs.

Rank 8 learners' knowledge of grammar becomes refined (Items 11, 16 and 22-23), and their global understanding of passages is improved as evidenced by Items 41-42 and 32.

Rank 9 learners outperform other learners in most of the items they were capable of correctly answering. Furthermore, learners can use sentence stress effectively to make a point (at least, they have the knowledge to do so) at this rank (Item 6).

Rank 10 learners demonstrate better performance in Item 48-50 and Item 40, which implies their understanding of the longer passages are more accurate and precise.

5 Concluding remarks

What emerges from above analysis and description of the learners is their dependence on top-down processing and the better performance in "communicative" language than in expository and narrative texts. The fact that the CRR for the discrete items measuring the learners' knowledge in grammar and vocabulary is lower than those for the other items and that the visual prompts help their better understanding of the texts both locally and globally implies that their understanding is rather holistic and insufficiently supported by the linguistic evidences therein, and consequently, less accurate. Furthermore, that the learners' performance was better with the dialogues than with the expository texts and narratives may be the reflection of the communicative shift prevalent in the Japanese EFL context.

The purpose of the present study was to prove that the NTT is a powerful tool in constructing a system of can-do statements. The above description of each of the latent ranks would sufficiently support this claim; by matching the item specifications and the

IRPs obtained from the actual data, it has been demonstrated that it is possible to describe with evidence what learners in a particular latent rank can and cannot do. The authors believe that the present study has profound implications both for future test development and evaluation of learners' achievements.

Acknowledgement
The present study is supported by Grant-in-Aid for Scientific Research (B) (22320114) from Japan Society for the Promotion of Science.

References

Guest, M. (2008a). A comparative analysis of the Japanese university entrance *Senta Shiken* based on a 25-year gap. *JALT Journal*, *30* (1), 85-104. Retrieved from http://www.jalt-publications.org/archive/jj/2008a/art5.pdf

Guest, M. (2008b). Japanese university entrance examinations: What teachers should know. *The Language Teacher*, *32* (2), 15-20. Retrieved from http://jalt-publications.org/archive/tlt/2008/02_2008TLT.pdf#page=16

Shojima, K. (2008a). Neural test theory: A latent rank theory for analyzing test data. *DNC Research Note*, 08-01. Retrieved from http://antlers.rd.dnc.ac.jp/~shojima/ntt/Shojima2008RN08-01.pdf

Shojima, K. (2008b). The batch-type neural test model: A latent rank model with the mechanism of generative topographic mapping. *DNC Research Note*, 08-06. Retrieved from http://antlers.rd.dnc.ac.jp/~shojima/ntt/Shojima2008RN08-06.pdf

Shojima, K. (2009). Neural test theory. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New Trends in Psychometrics* (pp.417-426). Tokyo: Universal Academy Press.

Shojima, K. (2010). Exametrika 4.2 [Computer software] Retrieved from http://antlers.rd.dnc.ac.jp/~shojima/exmk/index.htm

Yamakawa, K., Sugino, N., Ohba, H., Nakano, M., & Shimizu, Y. (2008). Acquisition of English grammatical features by adult Japanese EFL learners: The application of Item Response Theory in SLA research. *Electronic Journal of Foreign Language Teaching*, *5* (1), 13-40. Retrieved from http://e-flt.nus.edu.sg/v5n12008/yamakawa.htm

Table 2. Item specifications of the 2004 test

Sub-tests	Item #	Skills and knowledge tested
1A	1	accent patterns of *network* and *fascinating*
	2	accent patterns of *comments* and *particular*
1B	3	sentence stress in "*He didn't send me one.*"
	4	sentence stress in "*… we could study a bit.*"
	5	sentence stress in "*… it's going to be a surprise party.*"
	6	sentence stress in "*… it was a surprise party.*"
2A	7	idiom: *make sense*
	8	grammar: *devote O to Ving*
	9	vocabulary: *common* (as differentiated from *familiar, popular,* and *broad*)

	10	grammar: *have O Ven*
	11	grammar: past perfective
	12	vocabulary: *unless* (as differentiated from *if, whether,* and *since*)
	13	vocabulary: *mind Ving*
	14	grammar: subjunctive
	15	collocation: adverb before *any-/every-/no-/some- thing*
	16	grammar: tense and aspect, plus a psych verb
2B	17	discourse comprehension: response to a negative prospect
	18	discourse comprehension: response to an unexpected schedule change
	19	discourse comprehension: making a promise
2C	20-21	syntax/word order: relative clause construction
	22-23	syntax/word order: "*used to*" and "*with the light on*"
	24-25	syntax/word order: "*what it is that is making the noise*"
3A	26	logical connector: *as a matter of fact*
	27	logical connector: *therefore*
3B	28	coherent sequencing of sentences in one paragraph
	29	coherent sequencing of sentences in one paragraph
3C	30	coherent sequencing of sentences across several paragraphs
	31	coherent sequencing of sentences across several paragraphs
	32	coherent sequencing of sentences across several paragraphs
4	33	matching information in an expository text to a visual prompt; local
	34	sentence completion based on information in an expository text; local
	35	sentence completion based on information in an expository text; local
	36	sentence completion based on inference from a visual prompt; global
	37	sentence completion based on inference from an expository text; global
5	38	social norm of interaction; back channeling markers
	39	matching understanding of the interaction to visual prompt
	40	identification of pronouns' referents
	41-42	overall understanding of the interaction
6	43	inference of a character's feeling; local
	44	inference of a character's intention; local
	45	inference of a character's feeling; local
	46	inference of a character's feeling; local
	47	inference of a character's feeling; local
	48-50	fact finding; local

Table 3. Item reference profiles of the 2004 test items sorted by correct response ratio

Item #	CRR	Item Reference Profile									
		Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
10	0.907	0.548	0.686	0.814	0.881	0.919	0.944	0.962	0.976	0.988	0.995
5	0.897	0.592	0.694	0.796	0.860	0.904	0.934	0.950	0.961	0.974	0.984
33	0.889	0.561	0.683	0.793	0.849	0.887	0.916	0.939	0.958	0.976	0.989
19	0.885	0.519	0.647	0.765	0.829	0.877	0.917	0.948	0.971	0.987	0.995
39	0.861	0.374	0.545	0.717	0.812	0.873	0.914	0.941	0.961	0.980	0.991
45	0.839	0.308	0.439	0.612	0.748	0.847	0.912	0.951	0.974	0.987	0.993
17	0.796	0.440	0.559	0.669	0.731	0.775	0.812	0.841	0.871	0.914	0.950
4	0.779	0.455	0.525	0.600	0.660	0.725	0.790	0.842	0.882	0.922	0.948
7	0.770	0.344	0.446	0.566	0.658	0.734	0.797	0.845	0.884	0.928	0.960

3	0.763	0.288	0.389	0.521	0.637	0.738	0.814	0.864	0.899	0.930	0.950
14	0.762	0.416	0.509	0.610	0.679	0.731	0.773	0.807	0.842	0.894	0.937
18	0.750	0.463	0.559	0.641	0.678	0.703	0.732	0.771	0.815	0.871	0.917
36	0.740	0.338	0.415	0.496	0.564	0.646	0.737	0.818	0.886	0.945	0.979
12	0.735	0.302	0.401	0.527	0.624	0.704	0.766	0.811	0.850	0.893	0.926
43	0.727	0.294	0.408	0.552	0.656	0.723	0.765	0.791	0.816	0.861	0.903
34	0.719	0.408	0.450	0.513	0.575	0.639	0.700	0.760	0.824	0.900	0.951
15	0.696	0.426	0.482	0.547	0.596	0.644	0.694	0.738	0.775	0.822	0.864
13	0.648	0.301	0.371	0.462	0.538	0.604	0.657	0.697	0.734	0.796	0.856
27	0.643	0.248	0.322	0.418	0.488	0.546	0.612	0.687	0.766	0.857	0.922
44	0.640	0.310	0.389	0.489	0.564	0.622	0.666	0.699	0.727	0.756	0.775
29	0.636	0.398	0.464	0.517	0.542	0.563	0.592	0.633	0.691	0.777	0.846
20-21	0.635	0.167	0.256	0.368	0.466	0.557	0.634	0.699	0.769	0.860	0.928
47	0.633	0.374	0.414	0.466	0.513	0.556	0.599	0.644	0.703	0.793	0.867
38	0.619	0.229	0.293	0.374	0.444	0.518	0.598	0.677	0.751	0.834	0.894
2	0.612	0.336	0.374	0.429	0.484	0.544	0.603	0.655	0.702	0.763	0.814
35	0.609	0.360	0.403	0.439	0.454	0.476	0.523	0.594	0.691	0.819	0.911
26	0.606	0.339	0.371	0.418	0.463	0.511	0.565	0.624	0.690	0.784	0.863
30	0.578	0.193	0.229	0.269	0.309	0.374	0.479	0.618	0.764	0.896	0.966
31	0.572	0.208	0.244	0.294	0.343	0.404	0.494	0.612	0.735	0.849	0.918
11	0.564	0.376	0.413	0.443	0.452	0.462	0.488	0.536	0.608	0.722	0.819
16	0.551	0.285	0.319	0.354	0.385	0.429	0.487	0.555	0.638	0.754	0.846
46	0.540	0.265	0.278	0.298	0.326	0.372	0.439	0.528	0.647	0.805	0.916
41-42	0.530	0.139	0.181	0.244	0.316	0.399	0.488	0.578	0.668	0.782	0.871
22-23	0.528	0.164	0.213	0.267	0.314	0.371	0.444	0.538	0.655	0.803	0.908
32	0.505	0.155	0.176	0.205	0.240	0.299	0.392	0.520	0.666	0.821	0.921
37	0.497	0.282	0.287	0.303	0.331	0.372	0.428	0.497	0.580	0.692	0.784
6	0.484	0.120	0.152	0.214	0.289	0.371	0.446	0.514	0.592	0.717	0.827
28	0.451	0.210	0.239	0.278	0.316	0.360	0.411	0.466	0.525	0.613	0.696
1	0.403	0.250	0.256	0.280	0.315	0.354	0.385	0.406	0.431	0.501	0.586
48-50	0.342	0.055	0.078	0.114	0.151	0.191	0.238	0.301	0.405	0.597	0.765
9	0.320	0.325	0.328	0.321	0.308	0.298	0.297	0.304	0.317	0.336	0.356
40	0.313	0.144	0.135	0.134	0.143	0.166	0.208	0.272	0.365	0.519	0.659
8	0.236	0.239	0.215	0.187	0.169	0.158	0.155	0.167	0.205	0.317	0.450
24-25	0.160	0.045	0.052	0.062	0.070	0.077	0.085	0.098	0.138	0.276	0.450

Table 4. Item specifications of the 2004 test items sorted by correct response ratio

Item #	Skills and knowledge tested	CRR	Rk.
10	grammar: *have O Ven*	0.907	2
5	**sentence stress** in "… *it's going to be a surprise party.*"	0.897	
33	matching information in an expository text to a visual prompt; local	0.889	
19	**discourse comprehension:** making a promise	0.885	
39	matching understanding of the interaction to visual prompt; **global**	0.861	3
45	**inference** of a character's feeling; local	0.839	
17	**discourse comprehension:** response to a negative prospect	0.796	
4	**sentence stress** in "… *we could study a bit.*"	0.779	
7	idiom: *make sense*	0.770	
3	**sentence stress** in "*He didn't send me one.*"	0.763	
14	grammar: subjunctive	0.762	
18	**discourse comprehension:** response to an unexpected schedule change	0.750	
36	**sentence completion** based on **inference** from a visual prompt; **global**	0.740	4

12	vocabulary: ***unless*** (as differentiated from *if, whether,* and *since*)	0.735	
43	**inference** of a character's feeling; local	0.727	
34	**sentence completion** based on information in an expository text; local	0.719	
15	collocation: adverb before *any-/every-/no-/some- thing*	0.696	
13	vocabulary: *mind Ving*	0.648	5
27	**logical connector:** *therefore*	0.643	
44	**inference** of a character's intention; local	0.640	
29	**coherent sequencing** of sentences in one paragraph	0.636	
20-21	syntax/word order: relative clause construction	0.635	
47	**inference** of a character's feeling; local	0.633	6
38	social norm of interaction; back channeling markers	0.619	
2	accent patterns of *comments* and *particular*	0.612	
35	**sentence completion** based on information in an expository text; local	0.609	
26	**logical connector:** *as a matter of fact*	0.606	7
30	**coherent sequencing** of sentences **across several paragraphs**	0.578	
31	**coherent sequencing** of sentences **across several paragraphs**	0.572	
11	grammar: past perfective	0.564	
16	grammar: tense and aspect, plus a psych verb	0.551	
46	**inference** of a character's feeling; local	0.540	8
41-42	**overall** understanding of the interaction	0.530	
22-23	syntax/word order: "*used to*" and "*with the light on*"	0.528	
32	**coherent sequencing** of sentences **across several paragraphs**	0.505	
37	**sentence completion** based on **inference** from an expository text; **global**	0.497	
6	**sentence stress** in "*… it was a surprise party.*"	0.484	9
28	**coherent sequencing** of sentences in one paragraph	0.451	
1	accent patterns of *network* and *fascinating*	0.403	
48-50	fact finding; **global**	0.342	10
9	vocabulary: *common* (as differentiated from *familiar, popular,* and *broad*)	0.320	
40	identification of pronouns' referents	0.313	
8	grammar: *devote O to Ving*	0.236	
24-25	syntax/word order: "*what it is that is making the noise*"	0.160	