Linking Can-do Statements with Language Tests Using Neural Test Theory

Yukie Koyama* and Tetsuo Kimura**

* *Nagoya Institute of Technology*
koyama@nitech.ac.jp
***Niigata Seiryo University)*
(kimura@n-seiryo.ac.jp

Abstract
The objective of this study is to examine the appropriateness of Neural Test Theory (NTT) (Shojima, 2007) in order to analyze Can-do Statements (CDSs) for the purpose of linking them with the corresponding language tests. Firstly, NTT is explained, and then compared with Item Response Theory (IRT). Two different data sets in case studies are analyzed using IRT and NTT in this study. One case study consists of CDSs in EIKEN Can-do List and the English placement test. The other consists of CDSs for science and technology English and the final examination of that course. The outcomes of these analyses show IRT and NTT are generally consistent, for example, estimated latent abilities of examinees are highly correlated. However, NTT has a unique feature of Rank Membership Profile (RMP), which provides useful diagnostic information, in terms of ranking, to both test takers and teachers.

I. Introduction

In the history of testing, IRT is an epoch-making theory in terms of breaking an impasse on the limitation of Classical Test Theory (CTT), which had long been the primary and only scientific tool to analyze test results. The limitation is that CTT analysis is always based on the population. IRT, on the other hand, can estimate the difficulty parameter of each item independently from the population.

However, IRT is based on a continuous scale, which is not always appropriate for analyzing test data. For example, as Shojima (2007) claims, ordinary tests always include standard error of measurement to some extent and do not have the high resolution. Therefore NTT, which is based on an ordinal scale instead of a continuous scale, was introduced as a more appropriate test theory for the tests in educational settings.

Since Common European Framework of Reference for Languages (CEFR) was published in 2001, CDSs have attracted attention of many researchers in the world as an innovative method to establish a goal of each level of language learning, or to enhance autonomous learners' self-awareness. The CDSs of CEFR are considered to be valid because they are conducted together with an online language test, DIALANG, which includes measures of reading, writing, grammar, and vocabulary skills. (Alderson, 2005) This kind of external evaluation test is indispensable when confirming the validity of a list of CDSs. Although many different types of CDSs have been developed in Japan as well, the validity of CDSs has not been studied so much. Dunlea (2009), for example, conducted a survey on 20,000 EIKEN test-takers in order to make a Can-Do list for the interpretation of EIKEN levels.

In short, this study aims to examine the validity of CDSs using NTT through two case studies.

II. Case Study 1 : EIKEN Can-Do List

A. Participants

In April 2010, 295 freshmen were asked to evaluate their own English skills by answering whether they considered they could do each of 109 CDSs or not. After eliminating data of 62 participants who didn't answer to the end and those of 13 participants who answered aberrantly, data of 220 participants remained for data analysis. The majors of the 220 participants were engineering (64), nursing (54), and social welfare and psychology (102). A placement test was also administrated to them a week before in order to stream them into five groups.

B. CDS on the EIKEN Can-Do List

CDSs on the EIKEN Can-Do List, which was developed and published in Japan by the Society for Testing English Proficiency (STEP) in 2008, were used in this study. The 109 CDSs were categorized into four main skills: 26 reading CDSs, 27 listening CDSs, 30 speaking CDSs, and 26 writing CDSs. The CDSs were divided into five levels: EIKEN Test Grade 4, Grade 3, Grade pre 2, Grade 2 and Grade pre 1 (see Table1).

Table1. Number of CDSs

	R	L	S	W	Total
Grade Pre 1	6	4	4	4	18
Grade 2	5	5	7	6	23
Grade Pre 2	4	6	6	5	21
Grade 3	6	6	6	5	23
Grade 4	5	6	7	6	24
Total	26	27	30	26	109

C. Placement Test

The placement test consisted of the following four types of questions: (1) vocabulary and grammar (Vgm), (2) listening comprehension with dialogue (Dlg), (3) listening comprehension with monologue (Mlg), and (4) reading comprehension (Rdg). All the items were adopted from the EIKEN Test Grade 3, Grade pre 2, Grade 2 and Grade pre 1 in 2007 and 2008, under the permission of STEP. Four versions of the placement test were constructed so that all items could be equated with common anchor items that were calibrated in the previous study (Kimura, 2009).

D. Procedure

The participants' dichotomous responses to the CDSs for each skill were analyzed by the dichotomous model NTT that uses a self-organizing map (NTT-SOM) mechanism using *Exametrika Ver.4.4* (Shojima, 2010). The number of latent ranks was set at five because the original CDSs were divided into five different EIKEN grades. *Exametrika* can specify the target latent rank distribution as either uniform or normal distribution. In this study, however, the target latent rank distribution was not specified because the numbers of CDSs in each grade were not the same and because there was no theoretical evidence which allowed assumption of normal distribution for the target latent rank. The responses to Vgm, Dlg, and Mlg of the placement test were also analyzed by the dichotomous model NTT-SOM in the same manner. However, the responses to Rdg of the placement test were analyzed using the graded model NTT-SOM because items in Rdg were testlet style: one passage followed by two to five questions.

E. Results and Discussion

The result of item analysis by NTT is best described in item reference profile (IRP) and its plot. However, IRP indices produced by Kumagai (2007) are useful for roughly grasping the shape of each IRP and understanding the characteristic of each item without viewing the IRP plot. One of the IRP indices, β, which is the location of the latent rank when the IRP value is closest to 0.5, simply expresses item difficulty. The power of item discrimination is represented by another IRP index, a, which is the maximum difference in the IRP value among all adjoining rank pairs.

In order to see the consistency between difficulty of CDSs analyzed by NTT and the EIKEN grade that the corresponding CDSs belong to, the ordinal numbers 1, 2, 3, 4, and 5 were assigned to grade 4, grade 3, grade pre 2, grade 2, and grade pre 1, respectively. Then Spearman's rank-correlation coefficients between β and grade of CDSs were calculated and found to be very high: .93 for Reading, .94 for Listening and Speaking, and .95 for Writing.

IRP index a, which shows discrimination power of CDSs, was varied from .02 to .34. There were two types of CDSs with low discrimination power: too difficult or too easy for most of the participants (see Figure 1.1 & Figure 1.2). The CDSs with high discrimination power tend to have moderate difficulty (see Figure 1.3).

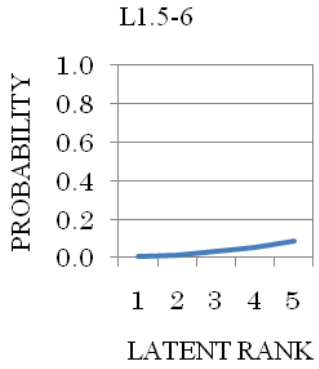

Figure 1.1. IRP with low discrimination

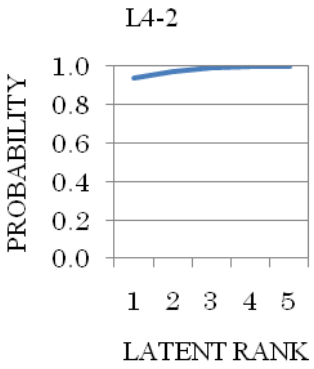

Figure 1.2. IRP with low discrimination and low difficulty

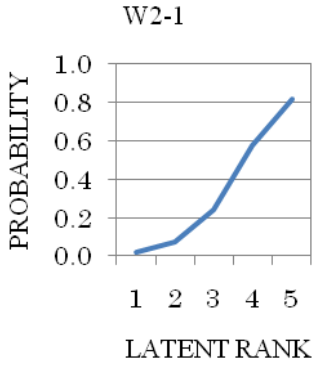

Figure 1.2. IRP with high discrimination and moderate difficulty

The estimation of latent ability analyzed by NTT is best described in rank membership profile (RMP) and its plot. RMPs show person's membership probability for each latent rank. The latent lank of the highest membership probability is summarized as the person's current estimated latent rank. Spearman's rank-correlation coefficients between self-evaluations based on CDSs and result of placement tests were calculated and found to be very low: .22 for reading and .28 for listening. This is likely to be a result of many participants' over- or under-estimation in the self-evaluation. Low correlation does not connote the low reliability of the CDSs and placement tests. Both tests show rather high Alpha reliability coefficient: the highest .92 and the lowest .81.

Figure 2.1, Figure 2.2, and Figure 2.3 show the RMPs of three persons whose current estimated latent rank for self-evaluated speaking ability are all rank four. Person 001 has still more than 40% of membership probability to the latent rank three and only 5% to the latent rank five. Person 039 has more than 70% of membership probability to the latent rank four and less than 20% to both the latent rank three and five. Comparing to Person 038, Person 001 doesn't feel confident enough to be able to achieve CDSs in the latent rank four. On the other hand, Person 092 seems to be in transition from the latent rank four to five since the person has almost the same membership probability to the latent rank four and five.

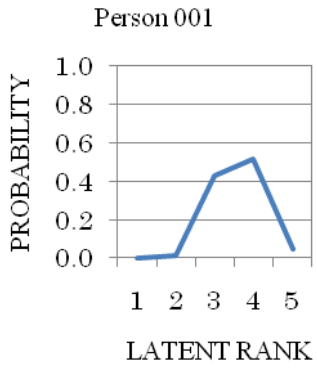

Figure 2.1. RMP of Person 001

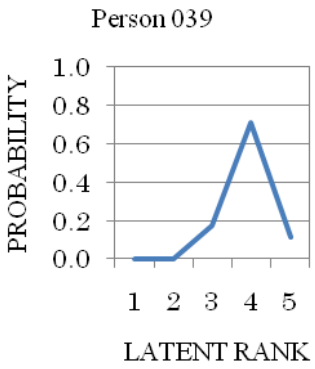

Figure 2.2. RMP of Person 039

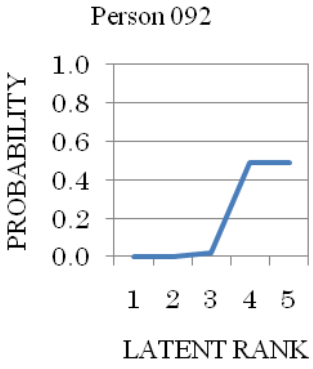

Figure 2.3. RMP of Person 092

III. Case Study 2 : General Science Can-Do List

A. Participants

The participants of CDSs were 882 first-year students at Nagoya Institute of Technology (NIT). The CDSs were given during the class time in October, 2007, and there was no limit of time. As for a Unified Final Examination (EXAM), the number of participants was 942. The allotted time for EXAM was 90 minutes, and it was given to the same students as CDSs in February, 2008.

B. CDS

The list of CDSs used in this study was developed by English teachers at NIT to give students an opportunity of self-evaluation. (Koyama, 2008) It is to evaluate a student's English ability of general science and technology. The list includes Reading, Listening, Writing, and Speaking sections, with five statements respectively. For example, the first statement of Reading is "I can read and understand a menu.", and the fifth statement is "I can read and understand a scientific article of a newspaper." These statements are answered on five-point scale; from 1 (I cannot do it at all) to 5 (I can do it quite easily). Mark sheets and mark sheets reader were used for marking and storing data.

C. Unified Final Examination (EXAM)

Together with the CDSs, EXAM was analyzed as an external evaluation test of the CDSs. EXAM was written by English teachers as a final examination of English for Science & Technology course at NIT, which is designed as a required course for the 2nd year students. EXAM is composed of 30 Listening items and 70 Reading and Vocabulary items, all of which are multiple-choice type questions, mainly 4 choices. For marking EXAM also, mark sheets and mark sheets reader were used.

D. Procedure

For the purpose of analysis of both CDS and EXAM, CTT, Rasch model, and NTT were used. The analysis tools are respectively Winsteps (Linacre, 2009) for Rasch model and *Exametrika* for NTT. Since the data size is relatively small, NTT-SOM is adopted for the NTT analysis. In addition, because the CDS data were polytomous in this case,

the data were analyzed using Andrich's Rating Scale Model for Rasch model, and Graded Neural Test (GNT) model for NTT.

E. Results and Discussion

First of all, the results of the CDS and EXAM analysis using CTT are shown in Figure 3.1 and 3.2. These figures show the difference in score distribution between the CDS and EXAM. The relatively small dispersion and the high mean score of EXAM are prominent, but it is reasonable because it is an achievement test.

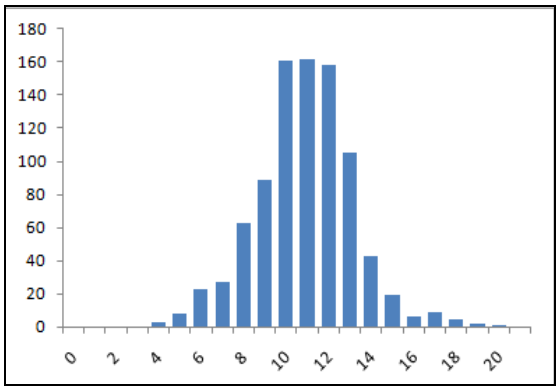

Figure 3.1 Distribution of CDS scores by CTT

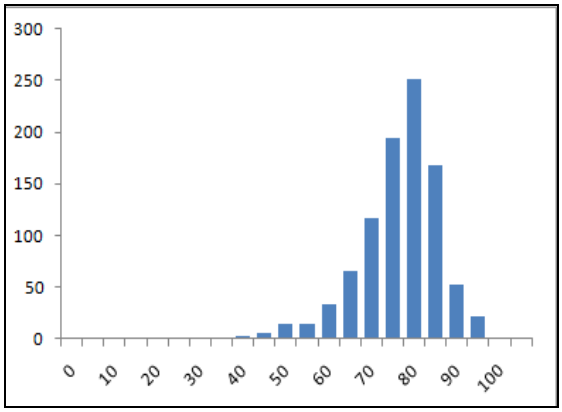

Figure 3.2 Distribution of EXAM scores by CTT

The results of the next analysis using Rasch model show that both the CDS and EXAM have the unidimensionality of all items. The infit values of all the items of the CDS and EXAM lie between 0.6-1.4, which means all the items measure the unite construct. Then the relationship between item difficulty and students' ability was examined. The Rasch person-item map show that the CDS items to discriminate lower level students are insufficient. However, the peaks of these distributions are close, which means the difficulty level of the CDS is relatively appropriate for the students. On the other hand, though the distribution of item difficulty of EXAM is large and the range of item difficulty values is from -1.66 to 1.72, the distribution of ability values is from .99 to 4.28. This results in the peaks being placed at different values, which means the items are, on the whole, too easy for the test-takers' ability. This tendency was explained also in the CTT analysis, but the relations between specific items and students became clear by the results of Rasch analysis.

When conducting NTT Analysis, the number of ranks was set as five both for the CDS and EXAM because the number of items of CDSs of twenty is not large enough to set a larger number of ranks. It is also desirable to have the same number of ranks for the CDS and EXAM when they are compared. Figure 4.1 and 4.2 are the Test Reference Profile (TRP), the expected score of test-takers in each latent Rank, of the CDS and EXAM. As can be seen in Figure 4.1, which is the result of the CDS, the expected scores of Rank 1 and 5 differs only a little more than 20 percent of the total score. However, in the case of EXAM, the lowest Rank 1 scores more than 60 percent and the difference between Rank 1 and 5 are even smaller. These result indicate EXAM is, on the whole, easier for the test-takers' levels.

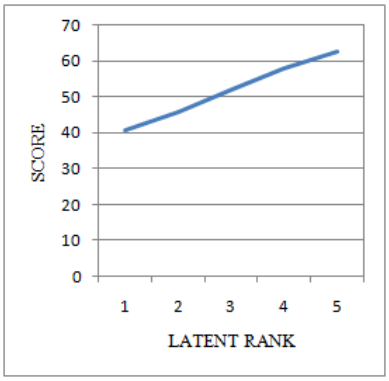

Figure 4.1 NTT-TRP of the CDS

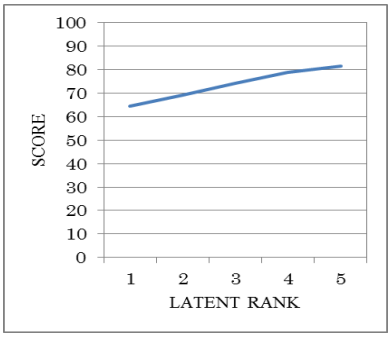

Figure 4.2 NTT-TRP of EXAM

In order to examine the relation between the NTT results of the CDSs and EXAM, further investigation was made specifically on the ranking of the test-takers. Because the small number of ranks is desirable for the purpose of comparison of two parameters, the five ranks of the original settings for the CDSs and EXAM were organized into three groups. Table 1 shows the cross tabulation of the three groups of the CDSs and EXAM ranks after regrouping was carried out twice. Through this procedure, Figure 6.2 was finally obtained.

This Figure 6.2 explains the percentage of students, who were analyzed to belong the rank(s) of the CDS, in each group (Rank 1 and 2, Rank 3 and 4, and Rank 5) of EXAM. As EXAM rank group gradually goes up, such as from the

group of Rank 1 and 2 to that of Rank 3 and 4, and then from the group of Rank 3 and 4 to that of Rank 5, the ratio of the group of Rank 1 and 2 of the CDSs gradually decreases. On the other hand, The ratio of the group of Rank 5 gradually increases. This result indicates the CDSs are linked with EXAM, an external evaluation test, through NTT analysis. In other words, this can be the proof that the validity of the CDSs are examined using NTT which focuses on ranking based on ordinal scale.

Table 1. Cross Tabulation of Three Groups of CDSs and EXAM

		CDS			
		R1&R2	R3&R4	R5	合計
EXAM	R1&R2	123	98	40	261
	R3&R4	106	141	81	328
	R5	79	123	84	286

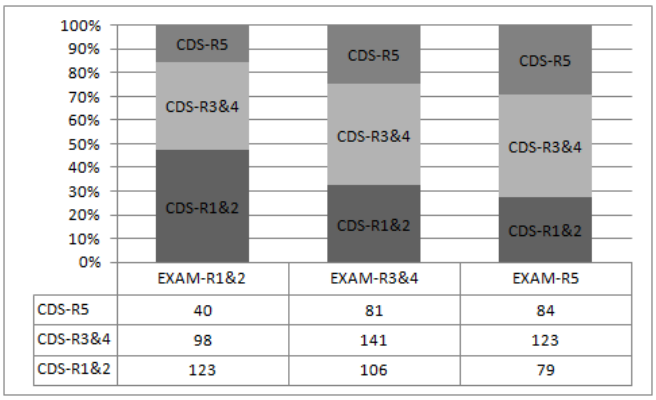

Figure. 6.2 Relation Between Ranks of CDSs and EXAM

IV. Conclusion

Comparing to the 1PLM IRT or Rasch model, NTT can provide item characteristics flexibly with IRP, which shows not only the difficulty but also the discrimination power. IRP and other outcomes of NTT can be calibrated even with rather small numbers if the number of latent ranks is limited (Shoujima, 2010). The most essential and prominent advantage of NTT is that it is much easier to divide both items and persons into several levels because it works on an ordinal scale rather than an interval scale from the beginning.

The outcome of analysis provided by NTT is generally consistent with that provided by IRT or Rasch model. For example, the order of item difficulties is mostly the same. In addition, estimated latent ability calibrated by NTT and Rasch are highly correlated (Kimura, 2009; Koizumi & Iimura, 2010). In the same way that IRT or Rasch model has its extended models for polytomous data or nominal data, NTT has its extended models such as graded model for polytomous data, which is useful for analyzing testlet items and Likert-type variables of psychological questionnaires, and nominal model for analyzing nominal-polytomous data, which can be used for evaluating the statistical feature for incorrect choices of multiple-choice items.

One of the unique profiles NTT provides is RMP, which is useful for reviewing the behavior of each examinee's membership probability for each latent rank. As we see in Figure 2.1, Figure 2.2, and Figure 2.3, RMPs can describe the status of an individual learner's progress in membership probabilities. If test results are reported not only in latent ranks but also in RMPs, test takers and teachers can read more useful diagnostic information from RMPs.

One of the limitations and disadvantages of NTT is that the methodology of item fit analysis or aberrant responses detection has not been developed sufficiently yet. In addition, cumulative of empirical data analysis based on NTT is not enough. However, as stated above, NTT has an essential advantage as a testing theory to analyze the test results to be classified into ranks such as those of CDSs.

Acknowledgements

A part of the present research has been supported by a Grant-in-Aid for Scientific Research for 2010-2012 (No. 22520590 and No. 22520561) from the Japan Society for the Promotion of Science. The Institute of Statistical Mathematics as well sponsored partly this research in 2010. We are deeply grateful to the above organizations, and our special thanks go to Kojiro Shojima for his helpful advices for data analysis.

References

Alderson, J.C. (2005). *Diagnosing foreign language proficiency: the interface between learning and assessment.* New York: Continuum

Dunlea, J. (2009). The EIKEN can-do list: improving feedback for an English proficiency test in Japan. In L. Taylor & C.J. Weir (Eds.), *Studies in language testing 31: Language testing matters*, 245-262.

Kimura, T. (2009). Neural test riron niyoru eigo placementtest no sakusei to hyouka (Construction and evaluation of an in-house English placement test from a neural test theory perspective). *KATE Bulletin*, 23, 23-34.

Kumagai, R. (2007, November). Neural test riron wo risanhennsuugata IRT tominasitatoki koumokutokusei wo simesu shihyou nituite (Item characteristic index and parameter estimation in NTT when considering discrete variable IRT). Paper presented at the workshop "Neural Test Theory."

Koizumi, R. Iimura, H. (2010). Neural test riron no tokuchou: kotenteki test riron, Rasch modeltono hikaku (Characteristics of neural test theory: comparison with classical test theory and rasch modeling). *JLTA Journal, 13*, 91-109.

Koyama, Y. (2008). Can-Do statements no datousei kenshou : ESP no kanten kara (Validity of can-do statements: an ESP perspective). *25th Anniversary Journal of JACET Chubu Branch 2008*, 177-187

Linacre, J. M. (2009). WINSTEPS [Computer software]. Retrieved Februaly 16th, 2009, from

http://www.winsteps.com/, originally developed by Wright, B.D., and Linacre, J. M. (1998). Chicago:MESA Press,

Shojima, K. (2007). Neural test theory. DNC Research Note, 07-02.

Shojima, K. (2010). Exametrika (Version. 4.4) [Computer software]. Retrieved October 22, 2010, from http://www.rd.dnc.ac.jp/~shojima/exmk/

Shojima, K. (2010). Neural test riron: gakuryoku wo dankaihyouka surutameno sennzai ranku riron (Neural test theory: latent rank theory for evaluation of academic ability). In Ueno, M. & Shojima, K. *Gakushu hyokano shin chouryu (New trend in learning evaluation)* (pp.83-111). Tokyo: Asakura Shoten.

Society for Testing English Proficiency. (2008). *The EIKEN Can-do List.* Tokyo, Japan: The Society for Testing English Proficiency.