JLTA Newsletter No. 32
日本言語テスト学会
The Japan Language Testing Association

JLTA Newsletter No. 32　　発行代表者:浪田克之介　　2012 年 (平成 24 年) 3 月 20 日発行
発行所:日本言語テスト学会 (JLTA) 事務局

〒389-0813 長野県千曲市若宮 758 TEL 026-275-1964 FAX 026-275-1970
e-mail: youichi@avis.ne.jp URL: http://jlta.ac

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

有意義な秋の日ー第15回全国研究大会を終えてー

研究会運営委員長･実行委員長
島田勝正（*桃山学院大学*)

大学の７階にある私の研究室からは関西国際空港が見えている。だから、空の便はいいはずだ。ここは大阪とは言え、大学の近辺にホテルがなく、ホテルのある堺市までは小一時間かかる。また、新幹線新大阪駅からは１時間半ほどかかる。そんな訳で地の利を得ているとは決して言えない。果たして、こんな条件の良くない会場にどれだけの人が集まってくれるのだろうか。私の心配は尽きなかった。当日の参加者は、米国、英国、韓国、中国、イラン等々と国際色豊かなものになり、８０名を超える言語テストの研究者で賑わった。私の心配は杞憂に過ぎなかった。

主会場となる一号館の英国風の学舎に囲まれた中庭には、秋の青空が広がっているはずだ。この中庭を取り巻くように研究発表会場を配置した。この配置は全国大会を本学でとのお話を頂いた時から既に私の頭の中にあった。この中庭での野外パーティも考えたが季節と天候を考えると躊躇せざるを得なかった。

前日のワークショップでは、澤木泰代氏(早稲田大学)による「探索的因子分析」の説明に聞き入った。大会当日、開会式での諸連絡のジョークが受けてすっかり気をよくした私は、いい気分で一日を過ごした。第１５回研究大会のテーマは、“University Entrance Examinations and Language Testing”と設定されている。そのテーマに関する基調講演では、Fred Davidson 氏 (University of Illinois at Urbana-Champaign) による “Test specifications in university entrance examinations”と題する興味深い話に耳を傾けた。この講演は日本英語検定協会のご支援があったからこそ実現したものである。研究発表は２１件あり、５室で行われた。そのうち英語による発表は１２件であった。事前審査で関心を抱いた発表がいくつかあったが、当日は裏方の仕事で走り回り、耳を傾ける余裕が十分にはなかったことは残念である。シンポジウムでは、笠原究氏(北海道教育大学)と静哲人氏(埼玉

大学)が Fred Davidson 氏を交えて、日本の大学の入学試験に関する問題点やその改善策を話し合うのを聞きながら、何度も頷いた。

秋の陽はつるべ落としである。総会が終わった頃には外は薄暗くなっていた。毎回思うことだが二日分に相当する盛り沢山のプログラムを一日に圧縮して、早朝から夕刻まで一気に消化するのは勿体ない気がする。

懇親会場は金剛山や和泉山脈の遠望できるペテロ館最上階で催した。参加された先生方には少しでも喜んで頂けたかと密かに自負している。特に否定的な評価がないのは、暗闇と酔いと談笑とでその景観を楽しんで頂いた方が少なかったからなのかも知れない。

ご支援、ご協力頂いた関係各位にお礼申し上げるとともに、行き届かなかった点はお詫び申し上げる。来秋は専修大学でお会いしましょう。

第 33 回日本言語テスト学会
研究例会報告

2011 年 7 月 2 日(土)

於：新潟青陵大学

共催：JALT 新潟支部

会場校の木村哲夫先生,お世話になりました。

（事務局,広報委員会）

講演

能力記述文構築のためのテスト理論とその分析ツール

荘島宏二郎
（独立行政法人 大学入試センター 研究開発部）

本講演では、荘島氏が開発した「潜在ランク理論（latent rank theory: LRT)」の背景と、Exametrika という分析ツールの説明があった（ http://www.rd.dnc.ac.jp/~shojima/exmk/index.htm よりダウンロード可)。

潜在ランク理論は、以前は「ニューラルテスト理論（neural rank theory: NNT)」と呼ばれていた、段階評価を行うためのテスト理論である。能力特性を順序尺度で捉えるのが特徴で、連続尺度で表現する「古典的テスト理論」や「項目応答理論」と異なる。潜在ランク理論には、2 値データ・名義データ（例：多肢選択式回答の分析）・段階データ（例：スピーキングの評価尺度の分析）を扱う潜在モデルがある。荘島氏は、連続尺度と順序尺度の利点・弱点を考慮し、分析の目的に合わせて理論を選ぶことの重要性を指摘した。

潜在ランク理論では、指定したランク数が適切かを調べる際に有用である「適合度指標」、項目ごとやテスト全体の特性、各受験者や全受験者でのランク確率などが示される。それらの情報を生かせば、学習者や教師に提示できる診断情報を抽出することも可能である。特に、ランク・メンバーシップ・プロファイルは、受験者がどの潜在ランクに所属するかについての確率を示し、同じランクに位置している受験者の中でも下のランクに近い受験者や、上のランクに上がれそうな受験者を見分けることができる。また、項目参照プロファイル（item reference profile）では受験者のランクと正答率の関係が項目ごとに表示される。ある潜在ランクに属する受験者が正答できる確率と、あるテスト項目で測ることを意図した能力とを突き合わせて解釈することにより、あるランクの受験者が何ができ、どのような能力を持つかを示す Can-Do Chart を作成し、ランクの解釈や診断に活用することができるそうで

ある。言語テスティング分野において、潜在ランク理論を用いた研究のさらなる発展が期待できよう。

報告者　小泉利恵
（*常磐大学*）

研究発表

ＮＴＴに基づく CAT の開発とシミュレーションによる特性評価

秋山 實
（*東北大学大学院／株式会社ｅラーニングサービス*）
木村哲夫
（*新潟青陵大学*）
荘島宏二郎
（*独立行政法人 大学入試センター*）

本発表の趣旨は、Neural Test Theory (NTT) に基づく Computerized Adaptive Test (CAT)のシステム及びシミュレーターを開発し、小規模なテストへの可能性について論じることであった。次の段階として、この CAT システムを使ったテストを実施し、結果を分析することによって、その信頼性・妥当性を考察する必要があるが、この発表ではまずシミュレーションによって回答データを生成し、その結果分析により NTT に基づく CAT の適性を評価するものである。

本発表では、まず CAT のメリットとデメリットや一般的な仕組みについて説明がなされた。CAT は、受験者の解答状況に応じて、予め設定された項目困難度に従って次の問題が選定され、またその解答状況が終了条件に達した時に能力測定値が決まるコンピュータを使用したテストである。秋山氏によって挙げられた CAT のメリットは、通常のテストの約半分の項目数で受験者の能力を測定できる、受験者によって出題内容が異なるため問題の露出が比較的少ない、等の点である。またデメリットとしては、項目の難易度を確定するための予備テストの実施、広範囲の難易度から成るアイテムバンクの必要、システム構築の技術的なハードルが高いため小規模のテストでは採用しにくい、等が挙げられた。

今回の CAT システムは、online 上で moodle の小テスト機能を使用して作成されたが、項目選定アルゴリズムには項目応答理論における Urry の方法、即ち、暫定推定能力と同じ困難度をもつアイテムを選択して出題する方法からの類推によりアイディアを得た木村（2010）の方法を採用している。また能力推定アルゴリズムには最尤推定法、ベイズ法等があるが、本研究で使われたのは荘島（2007）の最尤推定法である。終了条件は受験項目数が 20 に達するか、あるいは受験する項目がそれ以上ない、というものである。

NTT は能力を段階的に評価する理論であり、項目困難度は Item Rank Profile の値βによって示される。本発表の CAT ではこのβ値によって１ランクから１項目が選択され出題された。また、シミュレーションを行うことのメリットは、真値が分かっているため推定誤差が評価できる、低コストであり試行錯誤が可能、という点である。本研究では NTT をベースにモンテカルロ法によるシミュレーションを行い、50 アイテムで 500 名の回答者のデータを生成した。ランク数は 5、従って各ランクの項目数は 10 である。

このシミュレーションの結果を最尤推定法によって推定し、潜在ランクを求めたところ、真値との誤差の 0.49 で、Exametrika による誤差 0.39 と大きな差はなかった。しかし項目選択アルゴリズムに関しては、項目数が全体で 50 という少数であるため、終了条件の 20 項目になる前に終了するケースが多

く、アルゴリズムの改善が必要であることが今後の課題として明らかとなった。

報告者　小山由紀江
（*名古屋工業大学*）

Testing lexical fluency at the psycholinguistic level: A practical approach and insights

David COULSON
(*University of Niigata Prefecture*)

The primary purposes of Dr. Coulson's presentation were (1) to provide an overview of the rationale and design of Q_Lex, a newly developed software application for assessing lexical fluency taking a psycholinguistic approach and (2) to examine the degree to which data obtained from Q_Lex can offer useful information about vocabulary development of university-level English learners. First, as the principle behind the design of Q_Lex, Dr. Coulson presented the notion of lexical space that comprises various dimensions of word knowledge. While some dimensions in the lexical space such as breadth and depth of lexical knowledge have often been tested, another dimension of lexical access called lexical fluency has been recognized as a relatively difficult area to assess. As an attempt to address this challenge, Dr. Coulson developed Q_Lex, which allows administration of lexical fluency tasks to learners simply by using PCs.

Q_Lex measures learners' reaction time and its variability on a series of word detection tasks employing a masking technique. Along with reliability and validity data, results of three experiments conducted with Japanese university-level learners of English were presented. Key findings discussed by Dr. Coulson included the general tendency observed in the obtained data, where the variability of higher-level learner groups' reaction time on the word detection tasks was found to be significantly lower than that of lower-level learner groups'. This result, which suggests that the higher-level groups' reaction time was stable across various word detection tasks, was consistent with what one might expect from previous research findings. This result further indicated that the higher-level groups had reached a threshold marked by stability of lexical access efficiency. Based on the results, Dr. Coulson concluded that Q_Lex may offer a practical solution for testing English learners' lexical fluency as well as useful data for group diagnosis of vocabulary development for university-level English language learners.

Reported by Yasuyo SAWAKI
(*Waseda University*)

A report from the Language Teaching Research Colloquium 2011 held in the University of Michigan.

David COULSON
(*University of Niigata Prefecture*)

This presentation reported some of the main presentations from the LTRC event held Ann Arbor in June 2011. The keynote speaker was Dr. John Michael　Linacre. He has been one of the most important figures in the history of Rasch analysis in language testing. Latterly, he created the FACETS software which has become very significant in the analysis of testing situations such as rater reliability and test characteristics.

This presentation reported secondly on a workshop by the prominent Japanese researcher Toshihiko Shiotsu under the title "Test-taker and

task characteristics as sources of variability in reading comprehension and speed."

Reported by David COULSON
(*University of Niigata Prefecture*)
Adapted by JLTA NL editors

日本言語テスト学会
Workshop
於：桃山学院大学
2011 年 10 月 28 日（金）

Exploratory factor analysis

Yasuyo SAWAKI
(*Waseda University*)

This year, Professor Yasuyo Sawaki of Waseda University conducted a three-hour intensive workshop on the application of exploratory factor analysis (EFA) to language testing data. The workshop commenced with a discussion of the historical background of factor analysis, followed by a discussion of the previous applications of EFA to applied linguistics (e.g., Bachman, Davidson, Ryhan, & Choi, 1995; Vandergrift, Goh, Mareschal, & Tafagodtari, 2006); a discussion of the logic behind EFA (identification of an optimal number of factors that describes the pattern of relationships among a set of variables); the key steps involved in the application of EFA (study design and data type, number of factors to extract, extracting and rotating factors, and interpreting results); and a comparison of EFA with confirmatory factor analysis. Subsequently, participants were introduced to the interpretation of language-testing data, with reference to Bachman and Kunnan (2005). In response to questions from the audience, Professor Sawaki made the following observations:

Multivariate normality (kurtosis), as represented by standardized Mardia's coefficients available in EQS (in the "normalized estimate" in the output), needs to be below five to seven, approximately. One cannot detect multivariate normality with SPSS. However, Mahalanobis distance, which can be calculated using SPSS, could indicate multivariate normality. When analyzing (ordered) categorical data, one needs to calculate a tetrachoric correlation matrix for dichotomous data or a polychoric correlation matrix for polytomous data before one can submit these data to SPSS for factor analysis. In other words, one cannot calculate these matrices by using SPSS. Principal component analysis is designed to create a composite variable, whereas factor analysis is used to examine the underlying factor structure. Oblique rotation should preferably be used even when the correlations among factors are low. Interpreting a scree test is not easy and involves subjective judgment about the point where the discontinuity in eigenvalues occurs. Researchers are therefore advised to use several extraction methods (e.g., Kaiser's criterion, scree test, and parallel analysis), let alone interpretability of factors.

Overall, the workshop was very well received, and the question-and-answer session was truly insightful.

References:

Bachman, L. F., Davidson, F., Ryan, K. and Choi, I.-C. (1995). *An investigation into the comparability of two tests of English as a foreign language*. Cambridge, UK: Cambridge University Press.

Bachman, L. F., & Kunnan, A. J. (2005). *Statistical analyses for language assessment: Workbook and CD-ROM*. Cambridge, UK: Cambridge University Press.

Vandergrift, L., Goh, C. C. M., Mareschal, C. J., & Tafagodtari, M. H. (2006). The metacognitive awareness listening questionnaire: Development and validation. *Language. Learning, 56,* 431-462.

Reported by Yo IN'NAMI
(*Toyohashi University of Technology*)

第 15 回日本言語テスト学会
全国研究大会 研究発表
於：桃山学院大学
2011 年 10 月 29 日（土）

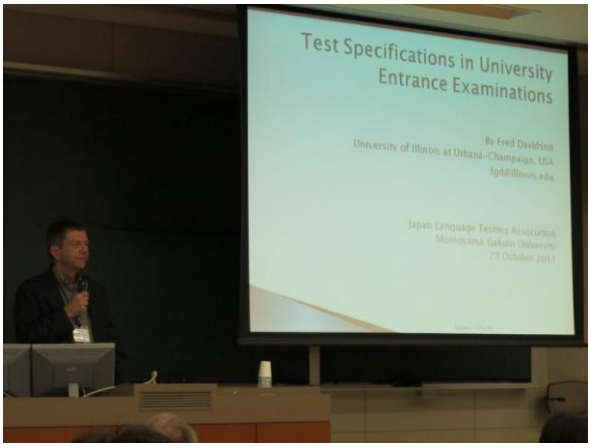
Keynote Speech

Test Specifications in University Entrance Examinations

Fred DAVIDSON
(*University of Illinois at Urbana-Champaign*)

The plenary lecture by Fred Davidson was informative, entertaining and offered many future research possibilities. Specifications (blueprints), designed to control equivalency of items and tasks, evolve through many stages of feedbacks, and enhances validity. (When test specifications are not available, they can (or rather, should) be created from sample examples of the test by reverse engineering. Together with validity, test specifications increase transparency but this creates the axiomatic tension: proprietary knowledge and language test security *versus* the public's desire to know about tests. We may term the issue about this tension as 'releasability'. Releasability involves aspects like scope and focus; audience; comparability, outcome; and tradition, each of which offers good research questions for the future. For further information, you may refer to Fulcher, G. and F. Davidson (2007) *Language Testing and Assessment: An Advanced Resource Book* (London: Routledge) or Davidson, F. and B.K. Lynch. (2002) *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications* (New Haven: Yale U.P).

Reported by Yosuke YANASE
(*Hiroshima University*)



研究発表

Korean Secondary School English Teachers' Perceptions about the Speaking/Writing Tests of the New National English Ability Test

Oryang KWON
(*Seoul National University*)

Professor Kwon spoke on English Teachers' Concerns About the Speaking and Writing Tests of the National English Ability Test (NEAT) that is to be administered to 1,200,000 Korean high school students during the 2012 academic year. See my summary of the presentation by

Professor Wonkey Lee for more details of the test itself.

After presenting a brief outline of the test and mentioning that there is concern about it, Professor Kwon presented his questionnaire research on the opinions of middle school and high school teachers (N = 169). He found that more experienced teachers claimed to have greater awareness of the speaking and writing section of NEAT, that a clear majority of the teachers surveyed were in favor of introducing the test, and that younger, less experienced teachers, were more in favor of the new test than older, more experienced teachers. Professor Kwon attributed this last finding to the higher competence in English that younger teachers have to demonstrate in order to pass the new employment tests. The positive correlation between a favorable attitude toward the introduction of NEAT and English proficiency was also demonstrated in data from the 63 teachers for whom TOEIC and iBT TOEFL results were available. The survey also showed that teachers had concerns about the introduction of the new test (what Professor Kwon called "some psychological burden" and that opinion was divided on willingness to participate in the rating of the students' test performance. Older teachers expressed less willingness than younger ones.

Professor Kwon concluded by pointing out that it was natural for teachers to feel anxious about the introduction of the NEAT since it would be used to determine university entrance and that it is necessary for teachers to be given appropriate in-service training concerning NEAT.

Reported by Randy THRASHER
(*Okinawa Christian University*)

The Speaking Test of Korea's NEATs 2 & 3: Its Usability for University Admission Qualifications

WonKey LEE
(*Seoul National University of Education*)

Professor Lee's presentation dealt with the speaking test of the Korean National English Ability Test (NEAT); the need for it, its nature, and what is hoped to be accomplished with it. He begins by pointing out that the use of an internet-based speaking test for university admission is unprecedented and therefore controversial. However, in this information age the ability to communicate is crucial. Thus the Korean government is attempting to reform education in that country by introducing a test that focuses on communication—particularly the ability to speak and write in English. He claims that the conventional way of teaching and testing of English in Korea has not produced competent English speakers, so NEAT is being introduced to attempt to reform education through washback. He states that the "purpose of the NEAT is to activate the teaching of speaking and writing of English at schools". Using teachers to rate the speaking and writing samples is expected to further this purpose. He also pointed out that the test has to be large-scale because there are 600,000 seeking college admission in Korea each year and each candidate will be given 2 chances to take the NEAT. However, testing 1,200,000 candidates each year means that the speaking tasks cannot be interactive which he claims reduces the validity of the test.

The NEAT speaking test was developed through a series of 6 pilot test conducted over 3 years. The results will be reported on an A, B,

C, and F (failure) scale. Professor Lee believes that this reduction in discrimination power (compared to the present university entrance exam) will force universities to employ a more discriminating device for final selection of candidates.

He concluded his presentation with examples of the various tasks to be used in the NEAT.

Reported by Randy THRASHER
(*Okinawa Christian University*)

Employing multiple test-centered standard-setting methods in relating exams to the CEFR

Jamie DUNLEA
(*Society for Testing English Proficiency*)

The study reports the process and partial results of a series of ongoing research projects that have been implemented by the Society of Test for English Proficiency (STEP). The main purpose of this project was to provide a common frame that links the EIKEN tests to the CEFR, which can fulfill the role of a good communication tool to help test users to better understand the meaning of certification at each EIKEN grade. The CEFR can also serve test developers as a starting point for the test assembly process, but as the presenter emphasized, many authors have pointed out that the CEFR cannot be used in its present form for this purpose as it lacks enough detail to be used as a constructor-oriented scale. The method STEP employed was standard-setting, using two standard-setting methods, the Basket Method and a modified Angoff procedure. The panelists went through stages of standard-setting sessions by examining the difficulty of test items in reference to the levels of other tests and standards. They have drawn a tentative conclusion that Grade 2 of EIKEN is relevant to CEFR B1, while EIKEN Grades 3 and 4 are relevant to CEFR A1. The study is significant in at least three respects. First, this type of linking facilitates appropriate interpretations and use of the test scores for high-stakes decisions. Second, it also helps the test takers to choose and consider the suitability of the tests they are going to take. Third, the use of standard-setting methods for linking several existing tests and standards opens up new possibilities in the application of standard-setting procedures. The presentation drew a large audience, indicating a wide range of interest in STEP's efforts.

Reported by Hidetoshi SAITO
(*Ibaraki University*)

Investigating Spelling Performance with Production Tests and Recognition Tests

Sachiyo TAKANAMI
(*Saitama Prefectural University*)

The paper was an informative discussion of a topic not often discussed in EFL research forums. The presenter suggested that knowledge of spelling rules needs to be measured with formats designed for that purpose. The study used six formats, four from previous research on L1 spelling, and two added by the researcher to address spelling knowledge from an EFL perspective. Twelve words were administered in the same order to 93 university students. The order of the formats followed a hypothesized order of difficulty ranging from receptive formats (easier) to productive formats (more difficult).

The formats were moderately to highly correlated, with Pearson correlations ranging from .67 to .92. An analysis of variance showed a significant effect for format, with post-hoc tests showing that the story format, in which learners are required to fill in appropriate words in a passage, and the timed dictation format were the most difficult. The matching task was the easiest. *Implicational scaling* was then used to investigate the scalability of the words in terms of their acquisition order. The *coefficient of scalability* derived from this analysis should be above 0.6 for words to be considered scalable. Of the 12 words, 8 met this criterion, showing an order of acquisition consistent with the hypothesis that receptive spelling knowledge would be acquired before productive spelling knowledge of the same word.

The paper prompted comments and questions from the audience. One member welcomed the focus on spelling, noting that it fails to receive much attention in teaching materials. Several comments focused on steps that could be taken to address potential issues in the research methodology. It was suggested that a fixed order of administration for the different formats could result in an order affect, and some possible ways of avoiding this problem in future data collection were also discussed.

Reported by Jamie DUNLEA
(*Society for Testing English Proficiency*)

Assessing the effectiveness of a DCT pragmatics test

Fred S. TSUTAGAWA
(*Seikei University*)

This presentation was about an interesting attempt to validate a DCT (open-ended discourse completion test): a modified version of the one created by Hudson, et. al (1995). Eight situations (items) representing 3 variables (degree of imposition, social distance, and emotion) were included with 6 pragmatic abilities being assessed: ability to use the correct speech acts, ability to use typical expressions, amount of speech and information given, levels of formality, levels of directness, and levels of politeness.

A quite heterogeneous population of 144 subjects was used in contrast to the previous studies that mostly used homogenous populations and 2 raters rating all the responses. Then, the correlations among all the situations and 6 evaluative components were shown together with inter-rater reliability statistics. The inter-rater reliabilities were rather low, except for the ability to use typical expressions and use the correct speech acts. This is probably because these two expression-related components are closely-related, and at the same time encompass some abilities overlapping with other pragmatic evaluative components. The researcher attributes the generally low inter-rater reliability to insufficient rater training, lack of clarity in some evaluative descriptors as well as the inherently interrelated nature of these components in realistic communication. Also, one situation (item) was found to be a bit skewed in its representation of variables, which may indicate the necessity for validity check.

It seems that inter-reliability will increase if the researcher adds more raters, but this line of study is very important for establishing the construct of pragmatic ability, which is included in any kind of communicative test.

Reported by Kahoko MATSUMOTO
(*Tokai University*)

A Critical Review of the Status Quo of ELT in Japan: A Meta-Analysis

Mehrdad Amiri
(*Graduate School, Islamic Azad Univ., Iran*),
Parviz Maftoon
(*Islamic Azad University, Iran*)

本研究は、政府支援による多大な財政とエネルギーをかけているにもかかわらず、日本の英語教育が未だ成功していないこと指摘し、他国から見てこの不可解な現象に課題を投げかけた発表となった。世界的で有名な「日本人の英語べた」という汚名挽回のため、「『英語が使える日本人』の育成のための戦略構想」のもと、2002 年度から文部科学省によって「スーパー・イングリッシュ・ランゲージ・ハイスクール」や「高等学校等における先進的な英語教育の実践研究や研修」が実施された。

本報告では、「中学・高校の全英語教員６万人に対し、集中的に研修を実施」、「JET プログラムによる ALT の有効活用」など様々な試みを検証し、結論として、「成功していない」と報告された。

政府主導型では成功しないということが示唆され、他国のアジア諸国でも同じことが言えるのではないかとも指摘された。報告者の出身国イランの英語教育制度は、日本よりはるかによい結果を出している。社会的資源や教育的資源が日本のそれより乏しいにもかかわらず、民間レベルの語学教育が政府主導の教育を凌駕している。「ネィティブスピーカーに頼らない英語学習」、「市民レベルの教育実践」は日本が学ぶべき点である。グローバル市場では、日本は、韓国、中国と比べ、確実に劣勢に位置する。韓国、中国では政府主導型教育制度がよい結果を出しており、「高い英語力と交渉力」でグローバル市場での「勝ち組」であることは間違いない。

報告者　李洙任
（*龍谷大学*）

Text and auditory processing characteristics affecting item difficulty in EFL listening comprehension—Analyzing TOEIC® short conversations and short talks

Ken NORIZUKI
(*Shizuoka Sangyo University*)
Akihiro ITO
(*Seinan Gakuin University*)
Hiroshi SHIMATANI
(*Kumamoto University*)
Satoshi MONZAWA
(*Hiryu High School, Mishima*)

This is a very original, multi-faceted study about the complex relationships between textual features and auditory processing characteristics in EFL listening comprehension. The researchers replicated their first inquiry with 86 Japanese university students taking TOEIC® Parts 3 and 4, using more students (225 subjects) and seeking correlations between item difficulty, textual features, and auditory processing characteristics.

The first study using 86 subjects showed a general tendency of the factors that make listening comprehension of TOEIC® Parts 3 and

4 easier or more difficult. Naturally, more factors were involved with Part 4 where longer, more complicated texts are auditorily processed. In the next phase, many more insights were obtained by the follow-up study with 225 students, in which they were divided into 3 levels of upper, middle, and lower groups. In this second study, it was found that different text and auditory processing characteristics affect the students' listening comprehension either positively or negatively, meaning that students with different listening abilities would benefit from different feedback and assistance to overcome their weaknesses. What they felt difficult or stumbled upon in auditory processing was also demonstrated by representative retrospective protocols.

The impressive point of this study is that the researchers tried to find the reasons for different types of difficulty faced by different-level students in both textual features and subjects' auditory processing characteristics, using a myriad of indices. This study certainly shed light on the intricate relationships between aural input and its cognitive processing, and waits for more replication and validation with different subjects and assessment tools.

Reported by Kahoko MATSUMOTO
(Tokai University)

Washback effects of the National Center Listening Test on Japanese students' listening ability and their attitudes toward studying English listening

Akiyo HIRAI
(*University of Tsukuba*)
Ryoko FUJITA
(*Graduate School, University of Tsukuba*)
Hideaki MATSUZAKI
(*Graduate School, University of Tsukuba*)

The objective of this research is to clarify the washback effects of the National Center Listening Test, which was introduced in 2006, on Japanese High School students' English study through discussing the results of study 1, 2 and 3. The research questions raised by the authors are:

- RQ1. Did the students' listening ability improve after the introduction of the Center listening test? - Related to Study 1&2
- RQ2. Did the Center listening test affect the students' attitudes and motivation toward studying listening? - Related to Study 3
- RQ3. Did the Center listening test influence the curricula of high schools? - Related to Study 1&3

The subjects of Study 1 are total of 1123 high school students from 2005 to 2011, and the changes in CASEC scores are examined. However, the results do not identify the influence clearly.

Study 2 examines the changes in listening ability of 1600 university students from 2003 to 2011 using the scores of the placement test of a university. The results show that the Center listening test seems to have influence on the improvement of students' listening ability, and

the effect lasted at least for three months, though the influence varies depending on the major.

Study 3 includes a questionnaire of 13 questions, as follows, given to 95 undergraduates; Question 1: Ever taken the Center listening test? Question 2: Difficulty level of the test, Question 3-9; Preparation & effectiveness of the preparation, Question 10-13; Pros and cons of the listening test. The results show that students are motivated to study listening, and they study listening more in school curricula; they also recognize the importance of listening skills.

In sum, the answer to RQ1 is that there is gradual improvement, though they vary depending on the major. The answer to RQ2 is that positive washback is observed on students' motivation, and the answer to RQ3 is that more time is spent on listening. However, the causal link between the change of the Course of Study and students' listening ability is not clear, and further studies of the content of English education at both high school and university are necessary to examine the washback effects more clearly.

Reported by Yukie KOYAMA
(*Nagoya Institute of Technology*)

Error recognition tests as a predictor of learners' writing ability

Adel Dastgoshadeh
(*Islamic Azad University, Iran*)
Kaveh Jalilzadeh
(*Islamic Azad University, Iran*)

This presentation was canceled.

The Relationship between the Scores on the TOEIC Bridge and TOEIC tests

Hiroko YOSHIDA
(*Osaka University of Economics*)

The aim of the presentation is to examine the relationship between the scores of the TOEIC Bridge and TOEIC tests and to present a formula that can predict the TOEIC scores using the TOEIC Bridge scores. The TOEIC Bridge, developed by Educational Testing Service (ETS), is said to be suitable for those who score less than 450 points in the TOEIC test. However, only a small number of studies have investigated the actual relationship between the TOEIC Bridge and the TOEIC. In Professor Yoshida's research, 292 non-English major university students took both the TOEIC Bridge and TOEIC tests in 2009. Her results indicate that the scores of both tests were moderately correlated, but show a different distribution of the scores from those calculated by the ETS research in Korea and Japan in 2009. The differences were reported among the high-scorers of the university students. In a Q&A session, Professor Yoshida clarified some points regarding her research design, such as the research budget, the test takers, and the interval between the two tests. Taking the TOEIC Bridge was part of a university-funded program, so all freshmen took the TOEIC Bridge at the beginning of the term. Taking the TOEIC was also part of a university-funded program, but students were not required to take the TOEIC test. The interval between the two tests was two months. Since the topic was intriguing, questions from the floor continued until the time was up. It seems that further studies on the

TOEIC Bridge and TOEIC tests will definitely be needed.

Hiroshi SHIMATANI
(*Kumamoto University*)

The Effect of the Interview Test to Improve Students' Speaking Ability and How to Reflect it to Their Grade

Midori NISHIKO
(*Shizuoka Toyoda Junior High School*)

Ms. Nishiko presented a case study of the essential interaction among choices that EFL instructors are compelled to make about reconciling methods for both teaching and assessing students that can produce positive learning outcomes. The focus was to develop the speaking ability of 2nd yr. Jr. High School students from 5 different schools in Shizuoka. The method of assessment was to administer structured 2-minute interview tests to each student in November, and again in January. In the interim between the tests, classroom activities were conducted daily to develop the communicative abilities that were to be demonstrated during each assessment.

The interviews had two components. The first was to present an oral monologue on a non-academic topic related to personal experience or interest. The pool of topics had been revealed a few days earlier, but use of written notes, etc. was not allowed during the interviews. The second part was to interact with an interlocutor in a short Q&A session about the topic. The rating of performance was analytic based on criteria that included fluency, accuracy and degree of engagement. Daily classroom activities simulated aspects of the interviews and exercises in functionally useful grammar and vocabulary related to the expository and interactive modes of speech tasks in the interviews.

The results showed a noticeable gain in performance between the first and second interviews. This was attributed to coordination between the content of daily lessons as well as familiarizing the test takers with tasks required in the interviews and how they would be graded.

Reported by Jeff HUBBELL
(*Hosei University*)

Stakeholder input and test design: Investigating the effect of group member familiarity on test scores in a group oral discussion test

Dennis KOYAMA
(*Kanda University of International Studies*)
Eric SETOGUCHI
(*Kanda University of International Studies*)

Test takers are stakeholders and affected by test results. Therefore, they need to have a voice in testing practices. Nevertheless, careful investigation is required to determine whether incorporating test-takers' input into test development affects the validity of inferences drawn from test scores and to what extent. This case study examines this effect, focusing on how interlocutor familiarity affects scores of a group oral discussion test in a university EFL program. Students were randomly assigned to a familiar (classmate) or an unfamiliar (non-classmate) group. Their performance was rated in terms of pronunciation, fluency, lexical and grammatical correctness and communication skills. No statistical difference in scores was found across categories, and similar reliability estimates were obtained. These results suggest the

appropriateness of including test-taker input into the test development process.

In response to questions from the audience, Professors Koyama and Setoguchi stated the following: The assignment of students—to either group—was at random. At Kanda, students are in the same class throughout the year, with an average class size of 25, and thus get to know each other very well. This however, does not necessarily imply close bonding. Raters are systematically assigned to a group of students unknown to them. Contract learning is not practiced at Kanda. However, students can seek personal learning assistance from learning advisers. This presentation reports and analyzes the data from the test administered at the end of the freshman year. Interlocutor familiarity would probably be an issue if the data were analyzed according to ability groups; students with low proficiency would be most susceptible to group familiarity.

Overall, the presentation was very well received, and the question-and-answer session was truly insightful.

Reported by Yo IN'NAMI
(*Toyohashi University of Technology*)

Issues in Rating Junior High School Students' Speaking Performance in a Discussion Contest: A Case of the Ibaraki Interactive English Forum

Hidetoshi SAITO
(*Ibaraki University*)

The present study dealt with the issue of rating speaking performance in the context of the Ibaraki Interactive English Forum, a discussion contest, which has been held annually since 1999. Saito was concerned about the lack of sufficient amount of rater training and conducted the present research, which had two major purposes. The first purpose was to examine the rater effects of discussion performances, and the second was to propose a plan for improving the rating process and its criteria based on the finding. The data consisted of two sets, including prefectural finals and regional finals containing the total of 162 second and third year junior high school students with a total of 24 raters. During the contests each participant goes through three stages, while all raters rate all the participants at least once. Multi-faceted Rasch analyses were carried out to examine the effects of raters on rating. The results showed high reliabilities of raters with only one clear misfitting rater overall, though further analyses indicated the presence of problems in the contest's initial rater bias and rater-participant interactions. Amongst a number of interesting findings, most important were as follows. First, out of two ratings raters were more unstable in the first rating than the second and third ratings. Second, there were individual differences in the preference among raters as to the performance of particular individuals. Third, the three items (i.e. expression, content and cooperativeness) used for rating exhibited virtually identical item difficulties. And fourth, almost half of the twenty rating categories used for each of the three items were never used and there was disordering of the scale categories.

The presentation was followed by comments and questions from both professional test researchers (including Professor Fred Davidson) and in-service English teachers.

Reported by Yoshinori WATANABE
(*Sophia University*)

CEFR に基づく英語スピーキングテストの開発

跡部 智
(*慶應義塾大学外国語教育研究センター*)
島﨑 のぞみ
(*慶應義塾大学外国語教育研究センター*)
小林 夏子
(*教育測定研究所*)
池田 直樹
(*教育測定研究所*)

慶応義塾大学外国語教育センターは、文部科学省学術フロンティア推進事業「行動中心複言語学習（AOP）プロジェクト」の取り組みの一環として、CEFR（ヨーロッパ共通参照枠）の項目に基づいた日本人英語学習者向けスピーキングテストの開発を行っており、その進捗状況に関して以下のように報告された。スピーキングテストの対象となる受験者は中学生、高校生、大学生で、到達段階などを把握する目的で作成された。研究の中では、①スピーキングテストの課題、②教員用手引き、③生徒への準備教材、の 3 つが開発された。テストの課題は、A1 から B2 までの 4 段階のレベルに合わせて各段階で 1～3 つ作成され、試験官により与えられたテーマに（1）受験者が制限時間内に 1 人でスピーチを完結させる形式、（2）受験者が試験者とロールプレイを行う形式、の 2 種類がある。指示言語は母語（日本語）で統一されている。どちらの形式においても、課題終了後に試験官によって follow-up question が与えられ、受験者の発言を引き出すよう構成されていた。また、テスト実施後に、教員が生徒に対してすぐに評価を与えることができるよう、テストには評価表の雛形が添付されており、フィードバックの与え方にも工夫が施されている。評価は、Language, Content, Strategy の 3 つの構成概念について、それぞれ到達度を 5 段階で評価する。点数評価に加えて具体的な改善点を生徒に示すことで、生徒の学習意欲および能力の促進に繋がることが期待されている。準備教材として、生徒用には「生徒練習用教材」、教員用には「教師トレーニングマニュアル」が作成されている。本研究で開発されたスピーキングテストを現場で運用するための情報提供が行われている。

以上のような発表内容に対し、今後のテスト実施に向けての課題点などについて、フロアから多数の質問が上がり、活発な意見交換が行われた。

報告者　髙波　幸代
（*埼玉県立大学*）

自動採点ポートフォリオ評価と CEFR 基準評価を融合したライティング学習プログラムの構築

工藤 洋路
(*東京外国語大学*)
長沼 君主
(*東京外国語大学*)
高野 正恵
(*東京外国語大学英語学習支援センター*)
増田 斐那子
(*東京外国語大学英語学習支援センター*)

本発表はライティング学習におけるコンピュータ使用評価と採点者による評価との関連性を示し、その融合可能性を論じた。東京外国語大学英語学習支援センターのライティング学習プログラムの一環として使用している Criterion 自動採点と CEFR に基づくライティングタスクを用いた採点者による評価を比較し、量的、質的の両面から分析を行った。量的な分析では、Criterion と採点者の相関は高かったが、書く量によって

Criterionの得点が変動することがあるため、CEFR との整合性が困難となり、質的な調査が必要となる。質的調査では discourse marker の使用状況によって Criterion の得点が変動するなど、つなぎ語使用と文章展開の関係や、使用文体と英文の質の関係という点で、Criterion の得点と CEFR のレベルが必ずしも一致しないことも指摘された。発表は分析結果に留まらず、実際のリサーチで使用したエッセイを提示し、自動採点と採点者の感覚との差異が生じる点を説明した。例えば、順序、追加、結果といったつなぎ語の接続副詞の頻繁な使用により、 Criterion では構成面から得点に導かれているものが、実際は内容や一貫性という点から見て高いとは言い難い例が提示され、また、少ないつなぎ語使用や無生物主語使用のために Criterion で評価されていないものも、CEFR では B2 レベルに相当するといった具体例も挙げて論じた。近年、教育機器の発展により、評価のデジタル化やコンピュータ化が進む一方で、指導者による評価の重要性認識も同時に促し、両評価を相補的に活用しながら統合させようとする、大変意義のある発表であった。両評価を融合させたライティング学習プログラムの構築が今後大いに期待される。

報告者　宮崎啓
（*慶應義塾高等学校*）

日本語学習者の文完成問題の自動採点へ向けての検討

酒井 たか子
（*筑波大学*）

本発表は，キャンセルされました。

潜在ランク理論による診断的テスト結果の提示

木村 哲夫
（*新潟青陵大学*）

木村氏は、潜在ランク理論(latent rank theory, LRT)のひとつであるニューラルテスト理論(neural test theory, NTT) により分析したテスト結果をどのように受験者や教育者に提示すべきかについて論じた。

NTT による分析では、その受験者がどのランクに所属するかという推定潜在ランクだけでなく、各潜在ランクに所属する事後確率を示すランク・メンバーシップ・プロファイル(rank membership profile, RMP)も得られる。同じ分野のテストを同一の受験者が複数回受験した場合、複数の RMP を提示することで、その受験者の能力の変化をより精細に示すことができる。

以上のような前提に基づき、発表では、大学1年生が4月と8月に受けたテストの一部の結果を NTT により分析し、一人ずつの受験者に2つの RMP を示すことで、どのように能力が変化したかを示す試みについて報告がなされた。RMP で示されたのは 26 問の語彙文法問題への解答であり、両テストは事前の分析により項目特性を調べて固定されたアンカー項目を6項目ずつ含み、等化が図られていた。この２つのテスト以外にも同様に等化が図られたテストが4つあり、いずれも受験者数は 250～300 人ぐらいで、4 月と 8 月のテストを両方とも受験した 70 名に対してのテスト結果が報告された。

結論として、2回の推定潜在ランクが同じであっても、RMP の違いによって異なる診断情報を伝えられること、推定潜在ランクに変化があったとしても、どの程度の確からしさ能力が変化したのかも伝えられることが分かった、との報告がなされた。

質疑応答では段階の区別の明確化やプレイスメントテストへの応用について活発に議論がなされた。

報告者　中村優治
（*慶應義塾大学*）

シミュレーションによるアダプティブテストの評価

秋山 實
（*東北大学大学院生*）

アダプティブテストは、受験者全員が同じテスト項目を受験するリニアテストと比べ、受験項目数は半分以下になり、テスト項目の露出度もそれに応じて低くなるため、様々なテストに適用できる。しかしその構成要素は多く、これまでに、最適な構成要素の組み合わせについて十分な議論がなされていなかった。本研究では、アイテムバンクの識別力や項目難易度が正規分布している条件の下、項目応答理論（２値パラメータロジスティックモデル）に基づく CAT の代表的なアルゴリズムを、小規模（290 アイテム）、中規模（853 アイテム）、大規模（2586 アイテム）アイテムバンク別に、モンテカルロシミュレーションを使って網羅的な評価を試みた。評価ツールとしては、CAT の構成要素を包括的に指定できる、オープンソースの統計ソフトウェア catR（Magis & Raiche, 2010）を使用した。分析の結果、大規模アイテムバンク使用時には今まで優れているとされてきたアルゴリズムの優位性が確認されたが、小規模アイテムバンクにおいては様相が異なり、最適なアルゴリズムの組み合わせは、実際に使用するアイテムバンクや評価される受験者の能力分布を考慮して決定することが望ましいことが示唆された。今後の課題としては、繰り返し回数（本研究では時間的制約のため 20 回）を増やして再評価する、項目選択アルゴリズムに　‘K-L Information Criterion’ を追加する、潜在テスト理論に基づく CAT との比較を行う、実際のテストデータで評価する、catR を高速化することが指摘された。発表後は、実際のプレースメントテストにおける運用等に関して、活発な質疑応答が行われた。

報告者　法月健
（*静岡産業大学*）

問題項目作成者が想定する困難度と IRT 分析による困難度とのズレ
－聴解問題項目の特徴－

申 貞恩
（*筑波大学*）
今井 新悟
（*筑波大学*）

日本語聴解テストにおいて、項目作成者が想定した困難度と IRT により受験結果を分析して得られた困難度のズレが生じるケースを抜き出し、どのような特徴をもつ項目がこのズレを生じさせる原因になっているかを考察し、日本語聴解テストの困難度を左右させる要因を探った。分析対象は、日本語学習者を対象とする J-CAT の聴解テストの一部で、旧日本語能力試験に準じた設定級ごとの平均値を基準とし、ある級の問題として作成された項目が、その級と隣接する級の平均を上回るか下回る場合に、項目作成者が想定した困難度と IRT 分析による困難度にズレがあると判定した。ズレがあると判定された 33 項目の特徴を分析すると、以下のことが分かった。

困難度が想定よりも高くなる項目は、１）問題指示（PreQ）が抽象的なもの、２）PreQ で一部の情報だけが提示され残りの情報は音声の後半で提示されるもの、3）音声の言

語情報から推測を要するもの、4）複数の情報を組み合わせて選択肢と比較するもの、であった。困難度が想定より低くなる項目は、1）概ね PreQ が明確で推測が不要なもの、2）正解の手がかりが二度繰り返されるもの、であった。

以上のことから、PreQ の構造と推測を伴う設問か否かは、聴解テストの困難度に影響を与える要因であると考えられる。また、問題項目作成者に対しては、これらの要因を意図的に調整することで、困難度を調整できることを示唆している。

報告者　木村 哲夫
（*新潟青陵大学*）

映像を用いたリスニング指導における 2 種の先行オーガナイザーの効果

荒金 房子（*植草学園大学*）

大学での英語教育用の英語教材を用いる指導の中で、学習者が映像を見る前に導入資料としての先行オーガナイザーを提示した場合の効果を検証する研究の報告であった。

74 名の学習者を 3 つの等質グループに分け、2つの実験群は、導入資料の先行オーガナイザーとして①日本語によるサマリーを提示するグループ（線上的提示）と②グラフィック・オーガナイザー（I 以下、GO）を提示するグループ（空間的提示）に分け、また残りのグループは③導入資料を提示しない統制群とした。5 分程度の映像教材を 2 回視聴後、3 種の下位テスト（3 肢択一法 5 問、真偽法 4 問、ショート・アンサー形式 3 問）からなる内容理解度の確認テストを実施し、その結果を比較したところ、真偽法では異なる先行オーガナイザー間での有意な差が見られ、GO の使用の有効性が観察された。また、等質検定で使用したテスト結果をもとに、上位群と下位群を比較したところ、下位群では、真偽法の結果にのみ GO の有意性が検証された。上位群では、真偽法で GO グループが有意に高い数値を示したが、ショート・アンサー形式のテストでは、サマリーグループに有意性が出た。このことから、より表層の設問レベルでは GO が効果的な働きをし、より詳細を理解するにはサマリーの活用が補助となることが示され、特に上位群への指導でのサマリー利用の効果が示された。なお、事後のアンケート調査では、GO の使用がリスニングへの集中を促すとの意見が学習者から出ていたとの報告があった。

視聴覚教材を用いた聴解指導における先行オーガナイザーの研究は稀有であり、今後、より信頼性の高い測定方法を用い、広いレベルの学習者を対象に研究が継続されることで、聴解指導の方法などへの示唆を多く与えていくと期待する。

報告者　清水裕子（*立命館大学*）

中国人日本語学習者の「話す」能力における自己評価に関する一考察
ーJF Can-do-statements を利用してー

張 毅（*九州大学大学院生*）

本発表では、Nunan の学習者中心の言語教育という教育理念に基づき、中国の大学の日本語学習者を対象に、言語の熟達度をどう認識するかという評価段階に焦点を当てて、「みんなの『Can-do』サイト」で提供している JF Can-do の項目を利用して、学習者主導型の自己評価法の可能性を探った。

具体的には、中国の日本語教育現場における新たな自己評価法の可能性について、以下の 3 点を明らかにすることを目的とした。1）日本語「Can-do」アンケート調査を行い、学習者の日本語「話す」技能に関する自己評

価の実態がどうであるか、２）学習者の自己評価にはなんらかの特徴が見られるか、３）Cdsによる自己評価方法にはどの程度の信頼性があるか。

その結果、信頼性の高いJF Can-doの自己評価により、１、2、3 年生の「話す」能力に関する自己評価はA1、A2、B1、また 4 年生の調査にJF-cdsを利用するのは不適当であることなど中国人日本語学習者の日本語「話す」技能の大まかなレベルを把握できた。また、1級合格者は日本語以外の社会科学知識への要求が非常に強いなどのいくつかの特徴も見られた。そして、JF Can-doの利用により、学習者中心の言語教育の可能性がある点、及び妥当性に関してはさらに検証を重ねていく必要がある点が示唆された。

発表後フロアーからは、１）学年を追うに連れて自己評価が高くなるのではないか、2）質問紙には訳がついていたのかどうか、ついているとしたら日本語の能力も反映されてしまうのではないか、3）調査対象者の数が少ないのではないか、4）ポートフォリオはCdsの妥当性を検証できるのかといった質問やコメントが寄せられた。

報告者 荒金房子
（*植草学園大学*）

日本語Can-do statementsに含まれる要素と妥当性との関係
―韓国人JFL学習者の「聞く」技能に焦点を当てた項目記述の分析から―

入江 友理（*名古屋大学大学院生*）

本発表ではCan-do statementsを用いた自己能力評価の妥当性に関して、その構成する要素と聴解能力との関係について、韓国人日本語学習者を対象として行なった分析結果が報告された。

先行研究から自己評価にばらつきをもたらす要因のうち質問項目要因としては、「できること」を問うよりも、「難しいこと」を問うことや、詳細な具体例をつけることで有用性があがるとの指摘もあるが、Can-do statements の記述としては適切ではなかったり、経済性を損なうおそれがある。

そこで本研究では、詳細な具体例をつけることなく、ばらつきの少なくなる質問項目を作成するにあたり、Can-do statementsの構造に着目し、どの程度詳細な記述をすることで学習者の能力をより反映するのか、また、どのような要素がより能力を反映するのかの検討を行なった。

調査対象は韓国の大学で日本語を学ぶ韓国人学習者185名であり、質問紙調査に加え、日本語能力試験の過去問題の1級から3級の問題からなる聴解問題を作成し、実施した。

Can-do statements としては、ニュース、ドラマ、映画、会話、指示・説明、討論・発表のカテゴリーの基本記述に加えて、程度、具体例、身近さ、補助の有無のそれぞれの要素のうち、1 つまたは 2 つを加えた計 66 項目の質問項目が作成された。また、ダミー項目も44項目作成され、合わせて実施された。

要素数と聴解能力の相関分析の結果、要素なし、要素1つ、要素2つと相関の向上が見られたが、統計的に有意な結果ではなかった。また、各要素間の相関に関しては、各要素と基本記述との相関の統計的有意差は見られなかったが、身近さより、程度や補助の有無で相関が高い結果であった。

フロアからは本研究のサンプルの能力層による天井効果やフロア効果の影響についての指摘や、質問項目の経験の有無による回答の信頼性への指摘があった。今後より幅の広い能力層での検証が必要となるだろう。

報告者　長沼 君主
（*東京外国語大学*）

CASEC Can-Do リストの開発

野上 康子
(*教育測定研究所*)
林 規夫
(*教育測定研究所*)

英語コミュニケーション能力判定テスト CASEC のスコアを解釈するために受験者に提供される Can-Do リストが開発された。はじめに CASEC および 2011 年 10 月より新たなサービスとして提供されることになった Can-Do リストの概要が紹介された後，Can-Do リストの開発過程が報告された。

CASEC は 4 つのセクション（語彙の知識，表現の知識，リスニングにおける大意把握，具体情報の聞き取り能力）で構成されており，Can-Do リストの記述文として(1)セクション別記述文（計 130 文）と(2)CASEC の問題で多く使われる 5 つの場面における言語活動に焦点を当てた場面別記述文（計 65 文）の 2 種類が作成された。記述文の作成には塩澤・石司・島田(2010)のガイドラインが参考にされた。

Can-Do 記述文と CASEC スコアを対応付けるために，CASEC の受験者を対象として Can-Do 記述文に対する自己評価アンケートが実施された。受験者は各記述文に対して「1. できない，2. あまりできない，3. ある程度できる，4. できる」などの 4 件法で回答したが，回答が 1 または 2 の場合は 0，3 または 4 の場合は 1 として 2 値型に変換し，ロジスティック回帰分析を用いて CASEC スコアと Can-Do 記述文に「できる」と回答する確率との対応づけが行われた。これらの分析結果をもとに，各記述文について「できる」と回答する確率が一定程度以上になる CASEC スコア が求められ，Can-Do リストが完成された。

報告者　村上 京子
（*名古屋大学*）

Symposium

University entrance examinations and language testing in Japan

Coordinator & Panelist
Kiwamu KASAHARA
(*Hokkaido University of Education*)
Panelist
Tetsuhito SHIZUKA
(*Saitama University*)
Discussant
Fred DAVIDSON
(*University of Illinois at Urbana-Champaign, USA*)

笠原究氏

大学入試では Cognitive Academic Language Proficiency (CALP) を求めているが，中学では Basic Interpersonal Communication Skills (BICS)，高校側では BICS と少しの CALP を教育しており隔たりがある。大学入試では英文和訳をはじめ日本語産出能力が問われていることが多く，大学入試で改善すべき点として，日本語の CALP を測定するのではなく，英語の CALP をもっと測定すべきである。

また，大学入試センター試験で最近リスニング・テストが導入されたが，ライティ

ング・テストとスピーキング・テストがまだ導入されていない。韓国の NEAT を見習うべきで，ライティング・テストとスピーキング・テストの導入が望ましい。といった提案が笠原氏からなされた。

靜哲人氏

靜氏は大会要綱で(a)テスト作成時の組織内手順の改善，（b）英文和訳問題出題の廃止，(c) テスト項目分析の恒常的実施，(d) テスト問題実施後の公開制度の廃止，（e）専門授業（他教科）の英語による実施を増やす，の 5 つの提案をされていたが，本発表では時間の関係で特に（a）と（c）の 2 つの点に絞ってお話をされた。

（a）に関しては，一人のテスト作成者がテスト問題の元となる題材集めからテスト項目作成まで一貫してすべておこなうと，そのテストが作成者の創作物という側面が出てきて，作成者の面子を潰さないようになどの配慮が必要になり，テスト項目について純粋に学術的に批判やフィードバックがしにくいという現状がある。よって，たとえば題材を集める人とテスト項目作成者を分けて役割分担制にするなど，一人の創作物とならないような方法が求められる。

（b）に関しては，大学入試テスト実施後に，基本的なテスト項目分析すらされていないケースが大部分である。大学入試テスト実施のたびに，テスト項目分析がなされることが望まれる。といった提案が靜氏からなされた。

Fred DAVIDSON 氏

日本の大学入試に関する笠原氏・靜氏からの提案を受けて，Davidson 氏は主に 3 つのことに関して述べられた。（1）大学入試問題作成過程はチームで行うものである。よって，個人のエゴは脇に置いておこう。（2）人々はテストが公平に実施されていることを知る権利がある。よって，（Davidson 氏の午前の講演におけるキーワードであった）releasability が大切である。（3）テスト作成・分析・研究等においては，小さなことから始めよう。テストに関して，死ぬまでに小さなことを 3 つ改善できたらもう幸せである。そして最後の締めくくりとして，本学会に対して「JLTA 15 歳おめでとう！」とエールが送られた。

報告者　片桐一彦

（*専修大学*）

日本言語テスト学会
第2回最優秀論文賞
結果と経過

JLTA 最優秀論文賞の授与式が 2011 年 10 月 29 日開催の日本言語テスト学会 15 回大会（於桃山学院大学）で行われた。受賞論文は“Exploring item-examinee response characteristics in search of diagnostic functions of TOEIC ® tests for university students in Japan”、著者は法月健（静岡産業大学）、伊藤彰浩（西南学院大学）、島谷浩（熊本大学）の３名の研究者である。本賞は昨年 2010 年度より設けられ、今回は第２回であった。以下に選定経過を報告する。

小職が事務局より正式の依頼を受けたのは８月下旬であった。選定にあたって、事務局からは、考査の対象となった４篇の論文（各々について第一稿、査読結果、査読委員のコメント、最終稿の計４点）が送付されてきた。５名の委員で評価するに先だって評価の方法について決めなければならないことが２点ほどあった。第一に、初稿から最終稿にいたるまでのプロセスをどのように考慮するか。最も簡単なのは、査読委員に訂正する必要なし（accept as is）と判定された論文を受賞論文とすることである。しかし、実際には多少改訂して受理（accept with minor revisions）と判定された論文でも、かなりの量の改訂を求められている場合などもある。事務局と相談の上、結局のところ最終稿のみを考察の対象とすることとした。

以上の経緯を委員に説明した上で、規定に従い３つの規準（独自性 originality、研究上の貢献 contribution to academic field、教育上の貢献 contribution to education）それぞれについて、５段階で評価するよう依頼した。ただし、委員のうち１名は選定対象論文の著者でもあったため、残り３篇の論文の評価のみを行った。対象となったのは固よりどれも優れた論文である。受賞論文決定においては委員全員の得点を単純に平均したものを採用したが、加えて３つの規準についてバランスがとれていることも重視した。

総会開会直後に行われた授与式では、浪田克之介会長より３名の受賞者に副賞の盾と賞金が贈られた。それに引き続き、受賞者を代表して法月健氏から受賞の喜びの言葉が述べられた。懇親会では伊藤彰浩氏、島谷浩氏

渡部委員長による JLTA 最優秀論文賞の発表

受賞者記念撮影
（左から，島谷氏，法月氏，伊藤氏）

からもご挨拶をいただいた。３氏は学内外の教育・研究その他で多忙の中、ひとつのテーマについて数年以上にわたり地道に研究を続け、成果を公にしてこられた。当初は、木下正義先生が声をかけられて始まったプロジェクトであるとも伺っている。

判定委員という立場を離れ個人の研究者として、今回受賞の方々から継続と蓄積の力、そして協力の重要さを学んだ。研究の成果はすぐに目に見える形となっては現れないかもしれない。しかし地道に続けていればその成果はいずれ明らかにされる。チームワークで研究を進めるのは容易ではない。しかし１人では決してできない。単独で発表する研究であっても多くの方々の協力を仰いでいるという意味で共同の成果なのだということを心に留めておきたい。その意味で法月氏が受賞の言葉でテスト受験者の学生たちも含めて感謝の言葉を述べていらしたのは印象深かった。

蓄積、継続、協調が必要であることを学んだと述べたが、これは私たちの学会にも正しく当てはまることだと思う。今後とも継続し受賞者の方々の成果がすなわち本学会の蓄積となり、過去から未来への橋渡しとなるよう祈願する次第である。

報告者
2011 年度第２回 JLTA 最優秀論文賞委員会
委員長　渡部良典
（*上智大学*）

島田勝正研究会運営委員長・実行委員長による挨拶

＜編集後記＞

今号では，研究会運営委員長･全国大会実行委員長による巻頭言，JLTA研究例会報告，全国研究大会報告の執筆をお願いしました。ご執筆くださった先生方，ご快諾くださり誠にありがとうございました。

日本言語テスト学会事務局
〒389-0813　長野県千曲市若宮　758
TEL 026-275-1964　　FAX 026-275-1970
e-mail: youichi@avis.ne.jp
URL: http://www.avis.ne.jp/~youichi/JLTA

編集：　広報委員会
委員長　片桐一彦（専修大学），副委員長　齋藤英敏（茨城大学）
委員　　秋山實（東北大学大学院／株式会社ｅラーニングサービス），
佐藤臨太郎（奈良教育大学），長沼君主（東京外国語大学）