日 本 言 語 テ ス ト 学 会 (JLTA)

第 8 回 全 国 研 究 大 会（2004 年度）プ ロ グ ラ ム

The Eighth Annual Conference

of

The Japan Language Testing Association

大会テーマ：「言語テストとその妥当性：妥当性へのアプローチ」

Approaches to Validity of Language Testing

日時： 2004 年 9 月 19 日（日）8:45 ～ 17:45

会場： 麗澤大学1号館
(〒277-8686 千葉県柏市光ケ丘2-1-1)
TEL: 04-7173-3601（代） FAX: 04-7173-1100

日本言語テスト学会 (JLTA)
The Japan Language Testing Association (JLTA)

事務局
〒389-0813 長野県埴科郡戸倉町芝原 758
TEL: 026-275-1964 FAX: 026-275-1970
E-mail: youichi@avis.ne.jp
URL: http://www.avis.ne.jp/~youichi/JLTA.html

全 国 研 究 大 会 本 部 委 員

Randolph Thrasher（沖縄キリスト教短期大学・国際基督教大学名誉教授）
中村　優治　　(東京経済大学)
中村　洋一　　(常磐大学)
島谷　浩　　(熊本大学)
法月　健　　(静岡産業大学)
大坪　一夫　　(麗澤大学)
片桐　一彦　　(麗澤大学)

全 国 研 究 大 会 運 営 委 員

島谷　浩　　(熊本大学)
法月　健　　(静岡産業大学)
中村　優治　　(東京経済大学)
大坪　一夫　　(麗澤大学)
小山　由紀江　　(名古屋工業大学)
塩川　春彦　　(北海学園大学)
伊藤　彰浩　　(愛知学院大学)
藤田　智子　　(東海大学)
Soo-im Lee　　(龍谷大学)
林　孝憲　　(東京経済大学非常勤)
峯石　緑　　(広島国際大学)

全 国 研 究 大 会 実 行 委 員

大坪　一夫　　(麗澤大学)
片桐　一彦　　(麗澤大学)

研 究 発 表 審 査 委 員

中村　優治　　(東京経済大学)
島谷　浩　　(熊本大学)
法月　健　　(静岡産業大学)

第 8 回 大 会 プ ロ グ ラ ム

9 月 18 日（土）

16:30 ～ 18:00　理事会　（れいたくキャンパスプラザ　ティーラウンジ）

9 月 19 日（日）

8:45 ～　　**受 付**（1 号棟 5 階中央廊下）
(PC 利用発表者：発表教室で機器接続確認)

9:15 ～9:30　　**開会行事**（1503 室）

総合司会　**島谷　浩**（大会運営委員長・熊本大学）

挨拶　**Randolph Thrasher**（JLTA 会長・沖縄キリスト教短期大学・国際基督教大学名誉教授）

9:35～11:00　　**研究発表**（発表 30 分, 質疑 10 分）　　発表 I　9:35 ～10:15
発表 II　10:20～11:00

第 1 室（1501 室）　　司会　**Soo-im Lee**（龍谷大学）

[1]　発表 I　Approaches to the Validation of Language Tests for College Admissions
S. J. Ross　(Kwansei Gakuin University)

[2]　発表 II　Applying Rasch Measurement to Entrance Exams
Eton Churchill　(Kanagawa University)
David Aline　(Kanagawa University)

第 2 室（1502 室）　　司会　**渡部　良典**（秋田大学）

[3]　発表 I　Reasons for “Saying No” to Peer Assessment of EFL Presentations
Hidetoshi Saito　(Ibaraki University)

[4]　発表 II　Aspiration Factors Affecting TOEIC Score Gains: Influences of the Peer Group
Yoko Kozaki　(Part-time Lecturer, Mukogawa Women's Univ.)
S. J.　Ross　(Kwansei Gakuin Univ.)
Sumie Matsuno　(Part-time Lecturer, Nagoya Univ.)

第 3 室（1601 室）　　司会　**大津　敦史**　（福岡大学）

[5]　発表 I　　Can-do-statement における評定尺度の等間隔性について
脇田　貴文　(名古屋大学大学院 [院生])
野口　裕之　(名古屋大学大学院)

[6]　発表 II　　日本語 Can-do-statements に対する IRT 多値型モデルの適用
野口　裕之　(名古屋大学大学院)
熊谷　龍一　(名古屋大学大学院 [院生]・
学校法人 河合塾嘱託研究員)
脇田　貴文　(名古屋大学大学院 [院生])
和田　晃子　(早稲田大学大学院 [院生])

第 4 室（1602 室）　　司会　**塩川　春彦**　（北海学園大学）

[7]　発表 I　　対称性を持たせて共通項目法により段階反応モデルの項目母数を等化する方法
服部　環　(筑波大学)

[8]　発表 II　　スピーキングテストの並存的妥当性(Concurrent Validity)の検証―直接テスト SST と半直接テスト T-SST における検証―
金子　恵美子　(株式会社アルク)

第 5 室（1603 室）　　司会　**卯城　祐司**　(筑波大学)

[9]　発表 I　　L1 と L2 のメンタルレキシコン：英単語仕分け課題の結果にみられる semantic clustering の特徴
折田　充　(熊本大学)

[10]　発表 II　　ＥＡＰ語彙の広さと深さの関係
島田　勝正　(桃山学院大学)

第 6 室（1503 室）　　司会　**法月　健**　（静岡産業大学）

[11]　発表 I　　速読力の測定におけるテキスト構造と項目タイプとの関連性の検証
長沼　君主　(清泉女子大学)
和田　朋子　(東京外国語大学大学院 [院生])

[12]　発表 II　　A Comparison of Summarization and Free Recall as L2 Reading Test Tasks
Yasuyo Sawaki　(University of California, Los Angeles・現 ETS)

11:00～11:10　　**休　憩**

11:10～12:40　**基 調 講 演**（1503 室）

司会　**中村　優治**（JLTA 副会長・東京経済大学）

紹介　**Randolph Thrasher**（JLTA 会長）

演題：　**コミュニケーション能力論の諸問題**

講師：　**柳瀬　陽介**　（広島大学）

12:40～13:50　**昼　食**　（役員会： 1504 室、休憩室：1505 室）

13:50～15:15　**研究発表**　（発表 30 分, 質疑 10 分）　発表 III　13:50～14:30
発表 IV　14:35～15:15

第 1 室（1501 室）　司会　**S. J.　Ross**　(関西学院大学)

[13]　発表 III　Group Oral Testing: Amount of Floor Time and Score Variance
Miyoko Kobayashi　(Kanda University of International Studies)
Alistair Van Moere　(Lancaster University)

[14]　発表 IV　Application of Toulmin Argument Structure to the Case of TOEIC Validity
Mark Chapman　(Hokkaido University)

第 2 室（1502 室）　司会　**澤木　泰代**　(ETS)

[15]　発表 III　A FACETS Analysis of a Performance Test Measuring EFL Pronunciation
Hiroko Yoshida　(Osaka Jogakuin College)

[16]　発表 IV　The Effects of Task Types on Listening Test Performance: A Retrospective Study
Yo In'nami　(Graduate Student, University of Tsukuba)

第 3 室（1601 室）　司会　**小山　由紀江**（名古屋工業大学）

[17]　発表 III　多肢選択式聴解試験の項目様式効果： TOEIC リスニングテスト Part3 と Part4
柳川　浩三　(神奈川県立伊勢原高校・早稲田大学大学院[院生])

[18] 発表 IV　TOEIC IP テスト受験者の業務経験と自己評定の関連性の研究
伊東　田恵　(豊田工業大学)
川口　恵子　(群馬パース学園短期大学)
太田　理津子　(慶應義塾大学 [非常勤講師])

第 4 室（1602 室）　司会　**長沼　君主** (清泉女子大学)

[19]　発表 III　大規模英語学力テストにおける局所独立性の検討―長文読解問題における局所独立の成立状況について―
熊谷　龍一　(名古屋大学大学 [院生]・
学校法人河合塾嘱託研究員）

[20]　発表 IV　チェック・リスト評価法による日本語発話テストの試行結果について
庄司　惠雄　(お茶の水女子大学)
野口　裕之　(名古屋大学大学院)
春原　憲一郎　(海外技術者研修協会）

第 5 室（1603 室）　司会　**島田　勝正**　(桃山学院大学)

[21]　発表 III　JMP を活用して成績データを読む
安間　一雄　(玉川大学)
野田　昭夫　(SAS Institute Japan 株式会社)

[22]　発表 IV　オープンソースソフトウェア Moodle を基盤としたテスト分析機能付きオンラインテストサーバの開発
秋山　實　(合資会社 e ラーニングサービス)

15:15～15:25　**休　憩**

15:25～17:05　**シンポジウム**（1503 室）

言語テストとその妥当性：妥当性へのアプローチ

コーディネーター 兼 パネリスト:　**藤田　智子**　（東海大学）

パネリスト:　**村木　英治**　（東北大学）

小林　美代子（神田外語大学）

17:05～17:10　**休　憩**

17:10～17:35　**総　会**（1503 室）

議長選出

報告　**中村　洋一**（JLTA 事務局長・常磐大学）

17:35～17:45　**閉会行事**（1503 室）

挨拶　**木下　正義**（JLTA 副会長・福岡国際大学）

18:00～19:40　**懇 親 会**（れいたくキャンパスプラザ 宴会場）

司会　**法月　健**（大会運営副委員長・静岡産業大学）

挨拶　**大友　賢二**（JLTA 名誉会長・常磐大学・筑波大学名誉教授）

展示協賛企業

株式会社　桐原書店

財団法人　国際ビジネスコミュニケーション協会

株式会社教育測定研究所

発　表　要　旨　(ABSTRACTS)

基　調　講　演

コミュニケーション能力論の諸問題

柳瀬　陽介　（広島大学）

妥当性、とりわけ構成概念妥当性を検討する際には、実際のデータによる実証(empiricism)だけでなく、どのように言語や言語使用を構想するかという思索(speculation)も必要である。今回の講演では、日本でしばしば「コミュニケーション能力」として総括されている構成概念の思索に関する諸問題ー(1)翻訳語の問題、(2)方法論の問題、(3)「知識」概念の問題、(4)「過程」理論の問題ーについて報告する。

(1)翻訳語の問題：およそ思索を行う場合には、正確な言葉遣いが必要であるが、日本語で「コミュニケーション能力」に関して語る場合、しばしば諸概念に一貫した訳語が充てられることなく、多くの概念がただ「能力」と訳されているように思われる。例えば、コミュニケーション能力諸概念については、「*competence* は the language *faculty* が経験を通じて成長したものであるが、それは *ability* の要素を含む知識ともいえ、それが *capacity* によって活用されると *capability* としての communicative *competence* の表れとなり、それは *proficiency* として測定される」といった関係があると考えられるが、もし仮にこれらの概念をすべて「能力」と訳してしまったら、その関係も日本語では「能力は言語能力が経験を通じて成長したものであるが、それは能力の要素を含む知識ともいえ、それが能力によって活用されると能力としてのコミュニケーション能力の表れとなり、それは能力として測定される」といった意味不明の文になってしまう。今回は諸概念の簡単な整理と、それに基づくそれぞれの翻訳語の提示を試みる。

(2)方法論の問題：しかし、そもそも「コミュニケーション能力論」などという思弁は、研究の名前に値するのであろうか。言語に関する研究をあくまでも自然科学として推進しようとする Chomsky (2000)は、コミュニケーションといった問題を総じて論ずる考察を「何でも学」(a study of everything)であると批判している。Chomsky が言うように「コミュニケーション能力論」が自然科学として成立しない論考であるのなら、そういった論考はどのような規範にしたがって進められるべきなのであろうか。今回は応用言語学者の見解を若干まとめた後、それと、20 世紀（分析）哲学の根幹に関わる考察を行なった Wittgenstein 哲学との親近性を報告する。

(3)「知識」概念の問題：Communicative competence および、その下位項目の各種の competence に関しては、Taylor (1988)の批判により、その用法が Chomsky による原義から著しく離れているとされ、以後、communicative competence は(communicative) language ability、各種 competence は knowledge と呼びかえられることが多くなった。だがこの言い換えによって問題は解決したわけではなく、「知識」(knowledge)概念の二義性に関し

ては問題が残されたままである。すなわち、「知識」概念には、Chomsky が念頭においている合理論的意味合いと、応用言語学者が念頭においている経験論的意味合いがある。今回は、前者を「能力」(competence)、後者を「慣習」(convention)と呼び対比させた上で、コミュニケーション能力論はこの両義をどのように扱うべきかを論じる。

(4)「過程」理論の問題：コミュニケーションを論ずる場合、概念を二分化するのではなく三分化することが提唱されて久しい。Taylor (1988)は「能力」(competence)、「実力」(proficiency)、「パフォーマンス」(performance)という用語で、Lyons (1996)は「言語システム」(language-system)、「パフォーマンス」、「テクスト」(text)という用語で、McNamara (1996)は「知識」、「パフォーマンス」、「実際の使用」(actual use)という用語で表現しているが、「原因」、「過程」、「結果」という三区分であることには変わりはない。ここで一番問題となるのは過程、およびそれに関する理論であり、McNamara はこの領域を「パンドラの箱」とまで表現した。この「パンドラの箱」に関する構成概念の一つが Bachman(1990)、Bachman and Palmer (1996)の方略的能力(strategic competence)であるが、発表者はコミュニケーションに関する一般理論である関連性理論(Relevance Theory, Sperber and Wilson 1995)が過程理論として成立すると考える。今回は過程理論として考える関連性理論の概要を報告する。

シ ン ポ ジ ウ ム

言語テストとその妥当性：妥当性へのアプローチ

コーディネーター 兼 パネリスト:　**藤田　智子**　（東海大学）

パネリスト:　**村木　英治**　（東北大学）

小林　美代子（神田外語大学）

Messick の妥当性概念と言語テスト

村木　英治　（東北大学）

妥当性とは、心理測定などで使われるテストが我々の測定したい心理特性をどの程度正しく測定しているかを示す概念である。この妥当性の種類として、多くの心理教育測定の教科書では内容妥当性、基準関連妥当性、構成概念妥当性などを挙げている。しかしテストの妥当性は、そのスコアの使用目的の適切さに関連するものであり、テストの使用目的は多岐にわたるため、妥当性の名称の数はそれに従って増えていくことになる。実際、先に述べた各妥当性の名称の他に表面的妥当性、因子的妥当性、本質的妥当性、弁別的妥当性などの名称がよく使われる。妥当性検証のプロセスとは、あるテストスコアが特定の使われ方をした場合の妥当性の検証であり、それは科学的研究の仮説検証の手続きをも含む。Messick はこの点に注目し、妥当性の本質は統合化された単一の概念であるといった。彼にとって妥当性の確認は科学的な探求行為である。このシンポジウムにおいて、Messick のこの単一性の概念としての妥当性についての考え方を紹介し、それが言語テストにどう関わってくるのかを述べていきたいと思う。

TOEIC®と KEPT の比較　－　妥当性を求めて

小林　美代子　（神田外語大学）

近年、言語テストの妥当性研究の中で、Consequential Validity の重要性が認識されてきている。これは、テストが及ぼす社会全体への影響を鑑み、テスト作成者も利用者も、テストが測定しようとしている能力を明確に掌握し、本来の目的に沿った利用をする責任があることを意味する。ここでは、現在日本で広く普及している TOEIC®と、神田外語大学独自で開発した英語能力試験 KEPT の内容・形式等を比較し、日本の大学生の英語能力を測定するための適切なテストとは何かという観点から、テストの妥当性を検証する。また、妥当性を検証するためのさまざまなアプローチについても言及し、今後の妥当性研究の発展を促したい。

プレースメントテストとその使用の妥当性検証

藤田　智子　（東海大学）

妥当性の検証（Validation）は、これを完全に証明するという性質のものではない。「いかにそのテストとその使用が妥当か議論する為のサポートとなる有効な証拠を収集する」というたいへんに根気のいる作業である。しかしながら、そのテストを使用する者の社会的責任として、妥当性を検証することは、欠くことのできない義務である。そこで、この妥当性検証を効率よく迅速に進めるために、そのテストのための妥当性検証フレームワークを作成することを推奨したい。実証例として、ある大学の英語プログラムのプレースメントテストとその使用の妥当性検証フレームワーク（テスト実施手順、分析、単一の妥当性の関係をあらわした表）を作成し、それに従ってプレースメントテストとその使用の妥当性をテスト改訂前と後を比較しながら検証した。その結果、改訂版プレースメントテストでは、より正当な妥当性を議論できる証拠を収集できた。しかし、まだ改善しなければならない点はあり、今後も改訂を行いつつ妥当性検証を続け、さらに有効な証拠を収集する努力をしていかなければならない。

研　究　発　表

[1] Approaches to the Validation of Language Tests for College Admissions

S. J.　Ross　(Kwansei Gakuin University)

As Japan's population of high school graduates continues to decline, universities and colleges are increasingly considering innovative alternatives to conventional testing for college admissions. Like never before, opportunities for using alternative assessments in 'high stakes' language testing have become increasingly frequent. The validity consequences of alternative admissions, however, have not been widely examined.

This paper presents a comparative post-matriculation analysis of undergraduates admitted via conventional competitive exams and by the recommendation of high school teachers. The comparisons are based on eventual growth in EFL proficiency and grade point average changes over eight required foreign language courses. The validity of admissions decisions, it will be argued, can be based on three criteria after admissions are made: **a**) absolute differences in proficiency growth accruing from language instruction, as well as the maintenance of grade point averages; **b**) the impact of language proficiency and achievement motivation on success in subsequent English-medium content courses; **c**) continued self-selection of undergraduates into English-medium content courses. For comparisons of long term growth in criterion **a**, latent growth curves are fit for each subset of the 938 undergraduates sampled. The issue of the relative impact of proficiency versus achievement in criterion **b** is addressed with the use of covariance structure analysis.

Results of the growth and impact analyses suggest that admissions based solely on the competitive examinations may yield undergraduates with higher starting proficiency which may not convert itself into subsequent achievement. Recommended students, in contrast, show strong growth slopes in proficiency and maintain steady levels of achievement.

[2] Applying Rasch Measurement to Entrance Exams

Eton Churchill (Kanagawa University)

David Aline (Kanagawa University)

While one of the most important milestones faced by Japanese learners of English is the university entrance exam, analysis on these exams is seldom performed and even less frequently reported. In this presentation, we share how we are using Rasch Measurement Theory in a post hoc analysis to evaluate exam reliability and item fit, discrimination and difficulty. We first present the persons map of examinee ability on item difficulty of our *zenki* (February) 2003 Entrance Exam (3,968 examinees) and discuss the overall difficulty and reliability of our exam. We then compare the performance of specific item types (e.g., reading comprehension, contextualized vocabulary matching and isolated grammar completion) focusing on comparative difficulty, item fit and discrimination. We demonstrate how Rasch can be used to identify problematic items, discuss factors that may be affecting item performance and present our recommendations for future test writers. To supplement our discussion, we draw on analysis of our 2003 *koki* (March) Entrance Exam (1,892 examinees). Finally, we share how, within the constraints imposed by the exam writing process (e.g., no piloting, no item banking), we are using our analysis to make better informed decisions about item types and overall test difficulty. In conclusion, we suggest that post hoc Rasch analysis can be used to incrementally improve the validity of entrance exams in Japan.

[3] Reasons for “Saying No” to Peer Assessment of EFL Presentations

Hidetoshi Saito (Ibaraki University)

A number of studies that have been conducted in both general education (Falchikov & Goldfinch, 2000)and psychology (Fletcher & Baldry, 1999) including L2 fields (e.g., Patri, 2002; Saito & Fujita, 2004) in general lend support for psychometric validity of peer assessment. Concerning student reactions, studies reported both positive and negative reactions to peer assessment and peer review of writing (e.g., Amores, 1997; Cheng & Warren, 1997). Hardly known is what leads students to feeling negative about peer assessment, however.

The present study examines factors that influence student negative reactions to peer assessment of EFL presentations. One hundred one university students in Japan received instructions on presentation skills, gave presentations, assessed peer performance, received feedbacks, and filled in a questionnaire on which students indicated their attitude towards peer assessment. Based on the results of the questionnaire, sixteen students who reacted negatively to the questionnaire were interviewed by two trained graduate students.

Interview data were all transcribed and qualitatively analyzed. Two critical concepts emerged from the analysis: students as poor disciples, who committed rating errors, knowing their own biases and looseness in rating, and the teacher as the master, who are knowledgeable and only ones whom they can trust as far as grading is concerned. Although students understood the benefits of assessing peer performance, various external (e.g., traditional roles) and internal attributes (e.g., lack of experience) seem to mitigate their acceptance of assessing peers.

[4] Aspiration Factors Affecting TOEIC Score Gains: Influences of the Peer Group

Yoko Kozaki (Part-time Lecturer, Mukogawa Women's Univ.)

S. J. Ross (Kwansei GakuinUniv.)

Sumie Matsuno (Part-time Lecturer, Nagoya Univ.)

As growing importance has been placed on English proficiency for career development, career-related aspiration may be becoming one factor that influences college learners' motive to pursue study of language. In this study, we investigated not only the influence of learners' own aspiration but also that of their peers' normative attributes on this construct on the L2 proficiency gains. With learners' own aspiration being conceptualized as an individual difference factor and their perceptions of their peers' normative aspiration as a contextual factor, we looked into how these two factors interact with each other, especially in ways to influence learners' L2 proficiency gains. As such, we generated the following two hypotheses:

> Hypothesis 1: Learners' aspiration directly and uniquely predicts their L2 proficiency gains.
> Hypothesis 2: Learners' perceptions of their peers' aspiration influence their own aspiration, which in turn lead to their L2 proficiency gains.

The 915 participants were students of three colleges in Japan, which showed different growth in L2 proficiency. The L2 proficiency was measured by TOEIC/Bridge Test in a pre-instruction, post-instruction design. The aspiration constructs were operationalized by the participants' responses to survey items. Structural equation models were used to examine the significance of the latent variable paths specified for the above hypotheses and to test the model fit.

[5] Can-do-statement における評定尺度の等間隔性について

脇田　貴文　(名古屋大学大学院 [院生])

野口　裕之　(名古屋大学大学院)

Can-do-statement では，「読む」「書く」「話す」「聞く」の 4 技能についてそれぞれ 15 項目に対する回答（「1. 全くできない」から「7. 問題なくできる」）の素点を合計したものを個人の当該能力を表す指標として通常扱っている。この処理は，回答者の回答を間隔尺度水準であると見なすため可能となる。しかし，厳密には，評定尺度法により得られたデータは順序尺度水準である。したがって，より精度の高い測定を実現しようとする場合，評定尺度の等間隔性について確認することは重要である。

そこで，Can-do-statement において各カテゴリ間の間隔を求め，現在の 7 件法による測定が適切であるか否かを検討する。また，ここで用いられている「全くできない」「ある程度できる」「問題なくできる」という評定尺度表現が適しているのかについて検討する。

本研究では，脇田（印刷中）で提案した評定尺度法における間隔の評価方法を示した後に，上述の検討を行う。また，Can-do-statement に対して多値型 IRT 分析を行うことで得られる知見についても言及する。

[6] 日本語 Can-do-statements に対する IRT 多値型モデルの適用

野口　裕之（名古屋大学大学院）
熊谷　龍一（名古屋大学大学院 [院生]・
学校法人河合塾嘱託研究員）
脇田　貴文（名古屋大学大学院 [院生]）
和田　晃子（早稲田大学大学院 [院生]）

日本語 Can-do-statements は、日本語能力試験の妥当性を検証するための外的基準として、開発された自己評価方式による、日本語能力の評価尺度である。「読む」「書く」「話す」「聞く」の４技能につき各 15 項目、全部で 60 項目の質問から構成される。野口ほか（2003）では、この日本語 Can-do-statements に対して、項目応答理論の２値型モデルを適用して、１）ラッシュ・モデルと２パラメタ・ロジスティック・モデルの比較、２）尺度に含まれる項目難易度の分布の検討、３）テスト情報量による測定精度の評価、４）IRT 特性尺度値と従来の単純加算得点との関係の検討、６）中国語母語話者集団と韓国語母語話者集団の間で DIF 項目を同定した結果を示した。本研究は、7 段階評定尺度データの情報を２値化することなく、そのまま利用して IRT 尺度を構成した場合に、尺度の現実的有用性にどの程度効果をもたらすかについて検討する。具体的には、１）２パラメタ・ロジスティック・モデルと段階反応モデルおよび部分得点モデルとの比較、２）テスト情報量による尺度値推定精度の比較、３）異なるモデルの推定尺度値間の比較、４）IRCCC による項目内用の評価、５）IRT 多値型モデルによる DIF の検討、などである。野口ほか（2003）の結果と合わせて、最終的に日本語 Can-do-statements の尺度構成に用いるのに最適な IRT モデルを決定する。

[7] 対称性を持たせて共通項目法により段階反応モデルの項目母数を等化する方法

服部　環　(筑波大学)

Samejima（1969）の段階反応モデルは多値的に採点した項目得点、小問得点を合計した大問得点などに適用する項目反応モデルの 1 つである。段階反応モデルの項目母数を共通項目法によって等化する方法として、項目母数の記述統計量を用いる Mean & Mean 法、Mean & Sigma 法、また、Baker（1993a）のテスト特性曲線を利用する方法などがある。また、Baker（1993b）は計算事例を示していないが、項目反応カテゴリ特性曲線を利用する方法を提案している。ところで、冊子 1 の尺度値を冊子 2 の尺度値へ等化するための係数の推定値と、冊子 2 の尺度値を冊子 1 の尺度値へ等化するための係数の推定値との間に一意の関係があるとき、推定値は対称性を持つが、Baker（1993a,1993b）の方法による等化係数の推定値は対称性を持たない。そこで本稿は、Baker（1993a,1993b）の方法を等化係数の推定値が対称性を持つように変容し、さらに項目特性曲線を用いて等化係数を推定する方法を提案して、既存の方法を含めて推定値の相違を数値実験を通して検討した。

[8] スピーキングテストの並存的妥当性(Concurrent Validity)の検証

―直接テスト SST と半直接テスト T-SST における検証―

金子　恵美子 (株式会社アルク)

本研究報告では、直接スピーキングテストのＳＳＴと、電話で行う半直接スピーキングテスト、Ｔ－ＳＳＴの並存的妥当性の検証（concurrent validation）を報告する。

㈱アルクでは、1997 年にＡＣＴＦＬと共同でＳＳＴ(Standard Speaking Test)というインタビュー形式のスピーキングテストを開発し、実績を重ねている。2004 年秋には、独立行政法人通信総合研究所（現情報通信研究機構）を中心に構築された「ＳＳＴコーパス」（収容語数 100 万語を超える日本人英語学習者コーパス）が完成し、公開される。

一方、アルクではＳＳＴで蓄積されたノウハウを元に、半直接テストＴ－ＳＳＴが開発の最終段階を迎えている。ＳＳＴ，Ｔ－ＳＳＴとも、クライテリオン準拠で口頭運用能力を９段階で判定し、高い相関を持つ（相関係数 0.955）テストであるが、更に詳細な妥当性検証を行うため、Ｔ－ＳＳＴ２６件の書き起こしテキストにＳＳＴコーパスと同様のタグ付けを行い、口頭運用能力別に、ポーズ、フィラー、繰り返し、自己訂正の多寡、使用される語彙レベルに焦点を絞り、ＳＳＴとＴ－ＳＳＴの比較・検討を行った。その結果を報告したい。

[9] L1 と L2 のメンタルレキシコン：英単語仕分け課題の結果にみられる semantic clustering の特徴

折田　充　(熊本大学)

１．研究目的

本研究は、L1 と Ｌ２のメンタルレキシコン内において単語がどのように結びついているかという semantic clustering の問題を扱う。メンタルレキシコンの構造（特にメンタルレキシコン内の単語間のネットワーク構造）を調べるためには，単語連想テスト (word association tests) を用いることが多い (e.g., Kruse, Pankhurst, & Sharwood Smith, 1987; Schur, 2003; Wilks & Meara, 2002; Wolter, 1999)。しかし，個々の刺激語に対する反応語の表出結果の分析を基本とする単語連想テストは、メンタルレキシコン内で単語が意味的結びつきのもとにどのような clusters を形成しているかという問題に答えるには，必ずしも適していない。本研究は、簡単な英単語仕分け課題を用いて、L1 と Ｌ２のメンタルレキシコン内の semantic clustering の特徴を明らかにすることを目的とする。

２．方法

(1) 被験者 : 日本人英語話者（上級レベル）(NNSs) 28 名および英語母国語話者 (NSs) 28 名

(2) 課題 :『JACET List of 8000 Basic Words』Level 1 より無作為に抽出した 50 個の英単語をそれぞれカードに印刷し、それらを被験者に自身が考える意味のまとまりごとにグループ (cluster) 分けしてもらった。作成するグループの数や大きさに制限は加えず、どの単語ともグループを形成しないと被験者が思う単語については一つのままでよいとした。制限時間は 20 分とし、早く終了しても可とした。

３．結果と考察

被験者が作成した clusters の数や cluster ごとの単語数に両群間で明確に有意な差は見られず、結果について両群とも被験者間の個人差が大きかった。また、両群ともに、1～10 個の単語と cluster を形成する単語が圧倒的に多かった。なお、NSs ではいずれの単語とも cluster を形成しない単語が少なく、一方 NNSs では多数の単語と cluster を形成する単語が多いことが判明した。

参考文献

Kruse, H., Pankhurst, J., & Sharwood Smith, M. (1987). A multiple word association probe in second language acquisition research. *Studies in Second Language Acquisition, 9*, 141-154.

Schur, E. (2003). *An exploration of the structural properties of L2 vocabulary networks: A graph theoretical approach.* Unpublished PhD thesis, University of Wales Swansea.

Wilks, C., & Meara, P. (2002). Untangling word webs: Graph theory and the notion of density in second language word association networks. *Second Language Research, 18*, 303-324.

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model. *Studies in Second Language Acquisition, 23*, 41-69.

[10] ＥＡＰ語彙の広さと深さの関係

島田勝正　(桃山学院大学)

Academic Word List (AWL)は、570 語で構成される English for Academic Purposes (EAP)の代表的な語彙表であり、350 万余語の Academic Corpus から作成されている。本研究の目的は、AWL の学術語彙に関して、その広さと深さの関係を調べることにある。AWL の分布に基づき、名詞９語、動詞９語、形容詞５語、合計２３語を選定し、英単語の意味を４つの選択肢から選ぶ English Japanese Test (EJ)２版(Form X, Form Y)を作成した。さらに、この 23 語を刺激語とする Word Association Test (WAT)２版(Form X, Form Y)を作成した。WAT は、paradigmatic relation (PR)の代表として synonym を、syntagmatic relation (SR)の代表として collocate を用い、４つの選択肢からそれぞれ２語選ぶ、recognition test の形式を採用した。この２種類のテストを大学生 91 名に実施し、平均点、信頼性係数、相関係数を版別、頻度別、EJ の成績別（EJ２版の合算点により、上位・下位群に分類）に分析した。版、頻度、EJ の成績により、多少の変異は観察されるが、EJ と WAT の PR の相関は、EJ と WAT の SR の相関より高く、WAT において、PR の平均点は SR のそれよりも高いという結果を得た。

[11] 速読力の測定におけるテキスト構造と項目タイプとの関連性の検証

長沼　君主 (清泉女子大学)

和田　朋子 (東京外国語大学大学院 [院生])

長沼・和田(2002)ではリーディング能力の直接的な能力指標の１つとしての速読力の妥当性を検証し、物語文－論説文といったテキストタイプより、人文系－科学系といったトピックタイプの方が速読力に影響を及ぼすことを指摘した。長沼・和田(2003)では人文系－科学系に加えて、テキストの長さと難易度の影響を調べた結果、前回と同様に人文系テキストにおいて有意に速読力が高く、難易度があがるにつれ、速読力が低下する傾向が観察された。また、テキストの長さに関しては、長いテキストで速読力が高くなることも確認された。ただし、テキストの難易度としては語彙などのミクロ構造しか参照していなかった。

そこで本研究では論説文タイプのテキストにおけるマクロ構造に着目し、速読力の測定における理解確認問題の項目タイプとの関係を検証した。情報の関連の緊密さを軸にテキスト構造を分類した Meyer(1985)をベースとした Kobayashi(1995)の４つのテキストタイプの分類を採用し、概要、部分、比較、参照といった４つの項目タイプとの交互作用から、速読力との関連を検証した。今後さらにテキスト構造によって保持される情報と理解確認に適切な項目タイプとの関連を調べて行きたい。

[12] A Comparison of Summarization and Free Recall as L2 Reading Test Tasks

Yasuyo Sawaki (University of California, Los Angeles)

The use of extended-response reading comprehension tasks has recently been proposed for language assessment purposes. Potential advantages of this approach include the usefulness of the detailed learner responses for diagnostic purposes and the absence of the effects of reader-question interaction on reading performance, which complicates interpretation of test results. The use of such tasks remains infrequent in assessment, however, due to the lack of empirical validity evidence. To address this issue the present study compared two extended-response reading comprehension task types—summarization and free recall—regarding the reliability of ratings and the structural relationships among the constructs these task types tap into.

One hundred and eighty-eight university-level learners of Japanese as a foreign language in the United States read two e-mail messages written in Japanese and completed summaries and free recalls of each passage in English in two web-based testing sessions. Two independent raters rated each summary or recall protocol by using analytic rating scales. A confirmatory factor analysis (CFA) of the summary and recall ratings suggested that the recall tasks tapped into two highly correlated and yet distinct constructs, comprehension and integration of main ideas vs. comprehension of details, whereas the summary tasks assessed comprehension and integration of main ideas only. Moreover, the CFA analysis and a generalizability analysis agreed that the summarization task was associated with lower dependability of ratings. The raters' verbal protocols collected during the rating sessions suggested that this might be due to the increased judgment of appropriateness of learner responses required of the raters for rating the summarization protocols.

[13] Group Oral Testing: Amount of Floor Time and Score Variance

Miyoko Kobayashi (Kanda Univ. of International Studies)

Alistair Van Moere (Lancaster University)

Performance testing through group discussion is a promising test paradigm in that it is time- and cost-effective. Moreover, it reflects real-world and classroom tasks, providing rich samples of learners' natural language use in interaction. However, the quantity and quality of such samples largely depends on task conditions such as rater characteristics, rating scales, candidate characteristics and interlocutors as well as the tasks themselves (Skehan 1998). Most research so far conducted on oral assessment has examined interview format or pair-tasks, and very few studies have investigated group discussions. The current study is part of a wider series of research projects which investigate a range of variables related to this test format. The study examines the performance of approximately 120 Japanese university students within about 40 separate small-group discussion tests, and explores whether there is any interaction between the number of turns an individual takes, the length of turns, the learner's relation within the test group, the overall language proficiency level of the test group, and the group gender configuration. All performances were video-recorded and transcribed. They were then analysed both quantitatively and qualitatively and were compared with the raters' holistic and analytic grades. The data showed that the number of turns taken had no impact as a main effect, but the number of words spoken did have an impact. This impact was particularly apparent in the Communicative Skills and Fluency rating categories. These findings will have practical implications for the future design and use of group discussion as an examination format.

[14] Application of Toulmin Argument Structure to the Case of TOEIC Validity

Mark Chapman (Hokkaido University)

This paper will discuss a new approach to assessing the validity of TOEIC. At the Language Testing Research Colloquium 2004, Lyle Bachman proposed the use of Toulmin argument structure for the validation of language tests. This presentation will apply the framework described by Bachman to the TOEIC. This new approach to language test validity and utilization employs features of argument structure (backing, warrants, data, claims, inferences and rebuttals) to investigate the issues of validity, test use, consequences and fairness.

The presenter will suggest that the TOEIC lacks validity as a test of the ability to communicate in English. The usefulness and relevance of TOEIC scores in inferring the suitability of Japanese corporate employees for positions requiring English proficiency will also be questioned. The consequences of corporations being excessively reliant on TOEIC scores at the expense of more direct testing is another issue that will be addressed. The aim of this presentation is to apply Bachman's theoretical framework to a real test and language testing situation (the use of TOEIC in Japanese companies). Bachman's use of argument structure will be presented as a step forward from Messick's view of validity, in that Bachman's model provides for a more thorough examination of test utilization. It is hoped that this new model for assessing test validity will come to be seen as more practical and useful than Messick's theoretical framework.

[15] A FACETS Analysis of a Performance Test Measuring EFL Pronunciation

Hiroko Yoshida (Osaka Jogakuin College)

The present study examined the factors affecting pronunciation performance scores. A total of 60 Japanese EFL college students read aloud two different materials (a prose type reading and a dialog type reading) that are designed for diagnosing learners' pronunciation problems, and audiotaped their readings. Their performances were rated by three L1 English instructors and three L1 Japanese instructors. The assessment items used were 15 items of three aspects of the sound system of General American English (segmentals, suprasegmentals, and paralinguistic features). Segmental features included vowel, diphthong, consonant, consonant cluster, and aspiration. Suprasegmentals consisted of word stress, sentence stress, rhythm, intonation, and weak form. Paralinguistic features included loudness, rate/tempo, smoothness, energy, and clarity. The data were analyzed using Multifaceted Rasch Analysis (McNamara, 1996). The questions asked include: (1) To what degree are the assessment items difficult? (2) Are there any differences between L1 Japanese judges and L1 English judges in terms of rater severity? (3) To what degree are judges consistent in assessing performances? (4) Are there any differences in task difficulty between two different types of reading material for examining pronunciation? (5) Are there any unique bias patterns specific to examinees, items, and tasks? This study demonstrates the usefulness of Multifaceted Rasch Analysis in assessing pronunciation performance that has been neglected in the EFL classroom. It is hoped that the findings of this study will be incorporated into a full-scale development of an instrument for evaluating pronunciation for English instructors.

[16] The Effects of Task Types on Listening Test Performance: A Retrospective Study

Yo In'nami (Graduate student, University of Tsukuba)

The increasing interest in outcome-based approaches to assessment in language testing (e.g., Brindley, 1998; McKay, 2000) has heightened the need for more research on fair testing by which more valid inferences can be drawn. Among many variables related to test performance, of particular interest are the effects of task types (Chapelle, 1998; Kunnan, 2000). Despite determined efforts to investigate how task types affect test performance, most studies (Berne, 1992; Brindley & Slatyer, 2002; Kobayashi, 1995; Rodriguez, 2003; Shohamy, 1984; Wolf, 1993) take only quantitative approaches to this issue. Although a few studies use qualitative methods, they focus on independent task types (multiple-choice tasks, Buck, 1991; matching tasks, Ross, 1997; cloze tasks, Storey, 1997; multiple-choice tasks, Wu, 1998; gap-filling tasks, Yamashita, 2003) and do not compare multiple task types directly. In addition, these studies use different definitions of proficiency levels (in-house measures of proficiency, Storey; impressionistic judgment, Wu; standard test scores, Yamashita) and thus it seems hard to generalize the findings beyond their context.

The purpose of the present research is to examine qualitatively the effects of three task types (i.e., multiple-choice, open-ended, and summary gap-filling tasks) on listening test performance, taking the two limitations of the previous studies into account. The present study specifically addresses how task types affect the cognitive processes of taking listening tests and how their effects vary in accordance with the listening proficiency levels of the Common European Framework (Council of Europe, 2001).

[17] 多肢選択式聴解試験の項目様式効果：TOEIC リスニングテスト Part3 と Part4

柳川　浩三（神奈川県立伊勢原高校・早稲田大学大学院 [院生] ）

英語による口頭コミュニケーション能力の育成が国家的緊急の課題である一方、英語コミュニケーション能力の基底をなす英語リスニング力の測定は、今まで十分に関心が払われてきたとは言い難い。現に、多肢選択式聴解試験の項目様式(item format)の違いがテストパフォーマンスに及ぼす影響については研究が少なく、解明しなければならない点が多い。本発表では、TOEIC リスニングテストの Part3 及び Part4 で採用する項目様式とそれ以外の項目様式を比較検証し、問題文と選択肢の提示時期による項目様式の相違がテストパフォーマンスに与える影響を検証する。

項目様式が異なる①本文を聞く前に問題文と選択肢が提示される様式②本文を聞く前に選択肢が提示され、本文を聞いた後にはじめて問題文が提示される様式③本文を聞く前に問題文が提示され、本文を聞き終えた後にはじめて選択肢が提示される様式の 3 種類のテストを、過去の TOEIC 公開テストを基に作成し、約 250 名のボランティア参加者に無作為に割り当て受験させた。結果の分析にあたっては、項目困難度や項目弁別度等の項目特性が項目様式によりどのように変動し、各項目様式のテスト特性（情報量）と受験者能力がどのような関係にあるかを検証する。また、各様式と受験者能力との間に交互作用があるかどうかを共分散分析(ANCOVA)を用いて検証する。こうした検証を通じて、受験者能力に応じてどの項目様式が最も適切か、さらに、TOEIC リスニングテストの Part3 と Part4 の項目様式の妥当性についても考察を加える。

尚、本発表は、平成 14 年度国際ビジネスコミュニケーション協会 TOEIC 運営委員会助成研究の一部である。

[18] TOEIC IP テスト受験者の業務経験と自己評定の関連性の研究

伊東　田恵　　（豊田工業大学）
川口　恵子　　（群馬パース学園短期大学）
太田　理津子　（慶應義塾大学 [非常勤講師] ）

「自己評定」による言語能力測定はある程度信頼できるとされている。しかし、学習者がある言語タスクに対する能力を自己評定する場合、そのタスクを経験したことがあるかないかが、自己評定に影響を及ぼす可能性があると指摘されている。本研究では、TOEIC のＩＰ（Institutional Program）テスト受験者 8,000 名に仕事に関連した言語タスク 65 項目からなる can-do statements のアンケートを用い、１（全くできない）から５（問題なくできる）の 5 段階尺度で能力を自己評価してもらった。その際に当該業務の経験があるかないかも尋ねた結果、経験の有無で、そのタスクの自己評価点が変わることが判明した。経験があると、経験がない場合より一般に高めに自己評価する傾向が見られたのである。今回は言語の習熟度、タスクの難易度、および、タスクの内容の専門性という観点から、経験要因と自己評定の関係の検証を行い、「経験」が自己評定に与える影響をより明確にする。

[19]　大規模英語学力テストにおける局所独立性の検討―長文読解問題における局所独立の成立状況について―

熊谷　龍一 (名古屋大学大学院院生・学校法人河合塾嘱託研究員）

項目応答理論によりテスト分析を行う場合には, 項目間に局所独立の仮定が成り立っている必要がある。局所独立の仮定とは、「ある項目に対する解答が、その他の項目への解答とは独立に生ずるもの」というものである。受験者の特性値や項目パラメタ値を推定する際に計算される尤度関数において、通常この仮定が重要な意味を持つ。

ところで実際の英語学力テストにおいては、数十語から数百語からなる１つの文章（passage）に対して複数の項目が出題されていることがある。複数の項目が１つの文章で関係づけられていることから、とりわけこのような問題に対しては項目間で局所独立の仮定が成り立っているのかを慎重に検討しなければならない。

本研究では、Yen（1984）による Q3 統計量を一つの指標として、英語学力テストにおける局所独立の仮定の成立状況を検討することを目的とする。また、本研究で取り扱うデータは実際に実施された大規模学力テストのものであり、このような大規模データにおける項目応答理論の分析例を示す。

[20] チェック・リスト評価法による日本語発話テストの試行結果について

庄司　惠雄　（お茶の水女子大学）
野口　裕之　（名古屋大学大学院）
春原　憲一郎（海外技術者研修協会）

財団法人海外技術者研修協会では、2004年度に実施する「実地研修中の研修生日本語能力及び日本語学習に関する縦断的調査プロジェクト」において、紙筆テスト及び発話テストによって研修生の日本語能力を測定する計画である。本発表では、本実施に先立ち昨年度に実施した発話テストのパイロット研究に関する結果を報告する。

パイロット研究に使用した発話テストは、録音済みの音声テープから提示する課題に反応させる方式で、Kuo, J.ほか(1997)を参照し、質問に応答するなど４類の課題で構成した。各課題は相互に独立で、先行課題が後続課題の解答に関係を持たないよう設計した。課題提示と解答発話の関係は、課題提示→課題遂行の一方的関係のみで、相互のインターアクションはない。

評価はチェック・リスト法と査定基準参照による程度評価法を併用し、同協会所属教師12 名がランダムに２名ずつ６組のペアに分かれて評定を行った。その結果、評定者間相関係数では、チェック・リスト得点が 0.924、査定基準得点は 0.629、総合得点で 0.880 と高い値を示した。また、全課題の評定結果の和で得点表示した信頼性係数の値をα係数で推定したところ 0.7967 であった。チェック・リストによる評定に比較し、査定基準参照による程度評定では判定に「ゆれ」を生じる部分があることが判明したため、相関係数の低かった発話標本について詳細に検討した結果、改善すべき評定項目が明らかになった。

[21]　JMP を活用して成績データを読む

安間　一雄 (玉川大学)

野田　昭夫 (SAS Institute Japan 株式会社)

近年言語テスティングにおいては学習者・学習環境の多様性の追求という理論的要請に応えるべく様々な統計的手法が提案され、この専門化の傾向は言語テスト研究者一般の学習負担を増大させ研究理解を困難にせしめている。統計ソフト JMP(R) はその卓越したユーザーインターフェースで基礎的および先端的統計解析の扱いを容易にするものである。特に新たに取り込まれた項目反応理論は多くの研究者にとって有益な研究環境を提供する。この発表では第 2 言語学習において一般的な分析事例を取り上げ、本ソフトによるデータ中心の解決方法を紹介することで言語テスト普及の一助としたい。

[22] オープンソースソフトウェア Moodle を基盤としたテスト分析機能付きオンラインテストサーバの開発

秋山　實　(合資会社ｅラーニングサービス)

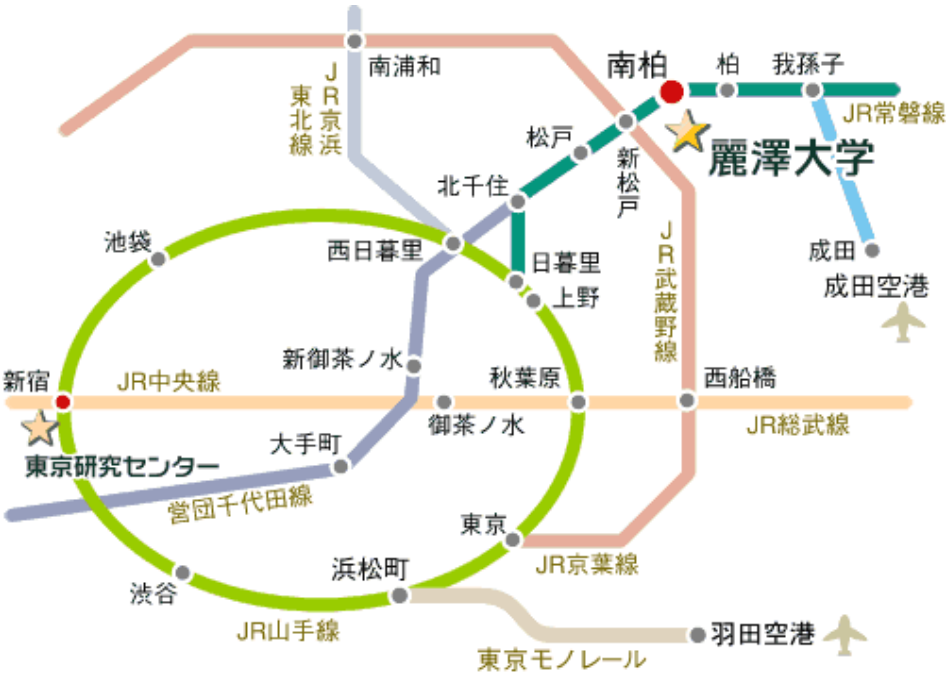
大友・中村（1996）による TDAP ver.1.0 は、後に大友・中村・秋山（2002）によって処理データ数の制約と処理スピードが改善され、Windows へ移植されたことにより、現在普及しているパソコンで手軽に利用できるようになった。しかし、運用上のもう一つの課題であったデータの入力は依然として、手作業で行わざるをえなかった。

本稿では、オープンソースソフトウェアとして広く公開されている Moodle を基盤として、その小テストモジュールの機能を利用しながら、TDAP のテスト分析機能を組込み、オンラインテストサーバとすることによって、データの手入力を不要とし、運用上の最後の課題を解決する試みについて報告する。

交　通　案　内

東京駅から

東京駅よりＪＲ山手線乗車→上野駅でＪＲ常磐線(快速電車)乗車→松戸駅にて JR 常磐線(各停)柏・我孫子・取手行きに乗り換え→南柏駅下車。



南柏駅から

東口の１番乗り場から東武バスに乗車し約５分。「麗澤大学前」で下車。

お　知　ら　せ

１．　受付は、 1 号棟 5 階中央廊下で行います。

２．　大会参加費は、会員が 1000 円、非会員が 3000 円です。

３．　昼食は、学園内にあるキャンパスプラザ・レストランまんりょうをお勧めします。ランチセットは通常 2-3 種類あり、¥1,000 です。予約なしでも可能ですが、お席に着くまでの待ち時間と調理のお時間を節約するために、ぜひ各自で直接、お席とメニューのご予約を事前にお電話でされることをお勧め致します。大会会場から、歩いて約 5 分です。http://www.kiu.ne.jp/hj/plaza/rest/ph/ 予約：04-7173-3558 (フロント)

４．　懇親会費は 5000 円です。会費は、学会当日、受付でお支払い下さい。会場は、れいたくキャンパスプラザ 宴会場 です。

５．宿泊の斡旋等は行いませんので、各自でご手配ください。下記に、麗澤大学近郊のホテルを紹介いたしますので、参考になさって下さい。

ホテルサンガーデン柏　http://www.gardenhotels.co.jp/kashiwa/accomo.html
〒277-0005 千葉県柏市柏 4-3-1　Tel : 04-7166-3111　Fax : 04-7166-3194
JR 常磐線「柏駅」東口より徒歩 2 分
シングル 8,400 円より（税別）

ザ・クレスト・ホテル柏　http://www.cresthotel.co.jp/kashiwa/01top/01_top.html
〒277-0842　千葉県柏市末広町 14-1　Tel: 04-7146-1111　Fax: 04-7146-1121
JR 常磐線「柏駅」西口より徒歩 2 分
シングル 8,000 円より（サービス料と税は別途）

柏フェニックスホテル　http://www.k-phoenixhotel.com/
〒277-0005 千葉県柏市柏 5 丁目 2 番 5 号 TEL 04-7164-6421　Fax: 04-7164-6493
JR 常磐線「柏駅」東口より徒歩 5 分
シングル 6,500 円より（税込み）

柏プラザホテル
〒277-0852 千葉県柏市旭町 1-5-3　Tel: 04-7147-1111　Fax: 04-7147-1117
JR 常磐線「柏駅」南口／東口より　徒歩 2-3 分
シングル 5,985 円より　（税込み）

れいたくキャンパスプラザ　http://www.kiu.ne.jp/hj/plaza/index.html
〒277-8656　千葉県柏市光ヶ丘 2-1-1　TEL:04-7173-3558　FAX:04-7173-3550
シングルＡ 5,500 円　シングルＢ 7,000 円