JART2016

日本テスト学会

共催　JSPS 科研費 基盤研究(A) 15H01772

「大規模型 e テスティング・システムとその運営モデルの開発」

後援　電気通信大学

第 14 回大会　発表論文抄録集

２０１６年９月８日（木）・９日（金）

於　電気通信大学

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　1　2016/08/29　13:59:34

協賛

本大会開催にあたり、下記の企業・団体の皆様より、多大なご支援をいただきました。心より感謝の意を表します。

株式会社 旺文社

株式会社 金子書房

株式会社 北大路書房

株式会社 教育測定研究所

一般財団法人 統計質保証推進協会

ナショナル・コンピュータシステムズ・ジャパン株式会社

公益財団法人 日本英語検定協会

公益財団法人 日本人事試験研究センター

日本データパシフィック株式会社

株式会社 ニホン・ミック

プロメトリック株式会社

株式会社 ベネッセホールディングス

株式会社 リクルートマネジメントソリューションズ

50 音順

日本テスト学会第 14 回大会実行委員会
委員長　植野真臣

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 2 2016/08/29 13:59:34

大会ご参加の皆様へ

1. 総会について

１日目の 13 時 30 分からシンポジウム会場（B 棟 202 教室）にて行います。日本テスト学会会員の皆様はご出席ください。

2. 懇親会について

１日目の 18 時から学生会館 3 階のレストラン ROYAL にて行います。当日参加ご希望の方は、大会受付にて懇親会費（5,000 円）をお支払いください。事前にお支払いいただいている場合には、懇親会受付にて、名札のご提示をお願いいたします。

3. 休憩室について

１日目は B 棟 101 教室、２日目は B 棟 201 教室に飲み物、お菓子類をご用意いたしています。

4. クロークについて

大会受付にて荷物をお預かりいたします。貴重品は各自お持ちください。ご利用時間は、１日目 9 時 00 分～18 時 30 分、２日目 8 時 00 分～19 時 00 分となります。

5. 喫煙所について

館内禁煙となります。屋外に喫煙所がありますので、大会スタッフにお尋ねください。

6. 大会発表賞推薦について

大会受付にて日本テスト学会発行「日本テスト学会第 14 回大会　大会発表推薦用紙」を配布いたします。ご記入の上、会場内の投票箱へお入れくださいますようお願いいたします。

7. 抄録集について

日本テスト学会第 14 回大会の抄録集の PDF ファイルを、大会 Web サイトにてダウンロードできます。ダウンロードには、以下のパスワードを使用してください。
パスワード：jart16uec

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　3　2016/08/29　13:59:34

発表者の方へ

1. 発表時間は、質疑応答を含め原則として 20 分です。

2. 発表者はセッション開始の 5 分前までに、発表会場にご参集ください。

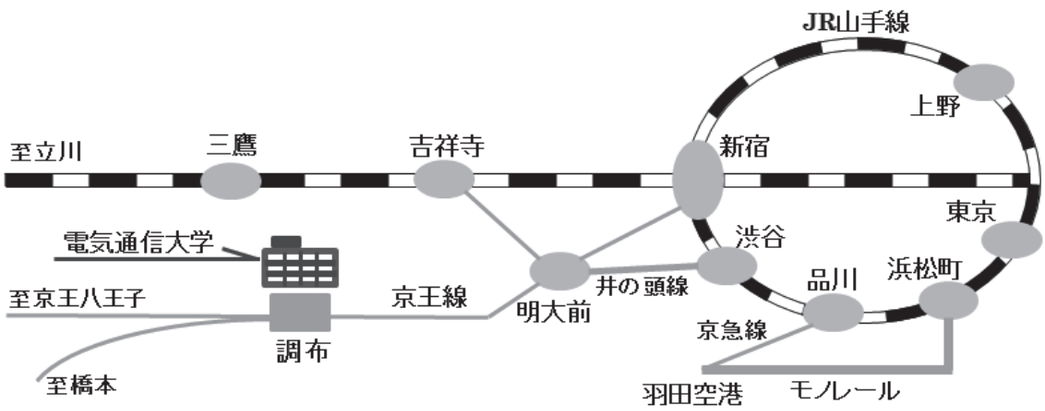
3. 会場には、プロジェクターと PC を用意しています。PC の OS は Windows 8 で、Microsoft Office 2016 がインストールされています。プレゼンテーションの資料は USB 接続可能なメディア（USB フラシュメモリ、外付け HDD 等）にてご持参ください。SD カード、メモリスッティック等のメディアをご利用の際は、USB 接続可能なアダプタのご用意をお願いします。CD、DVD などの光学メディアには対応しかねる場合がありますのでご注意ください。

4. ご自身でパソコン等の機器を御用いいただく場合は、プロジェクターとの接続確認を、セッションが始まる前までに必ず行ってください。

5. 補足資料を配布する場合は、50 部程度を目安に各自でご用意いただき、会場にいるスタッフにお渡しください。事前の預かり及び大会事務局でのコピーはいたしかねますのでご了承ください。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 4 2016/08/29 13:59:34

会場までのアクセス



主要駅から電気通信大学最寄駅（調布駅）まで

(1) JR **新宿駅から**

新宿駅　→　京王線（特急・準特急約 15 分、急行約 20 分）→　調布駅

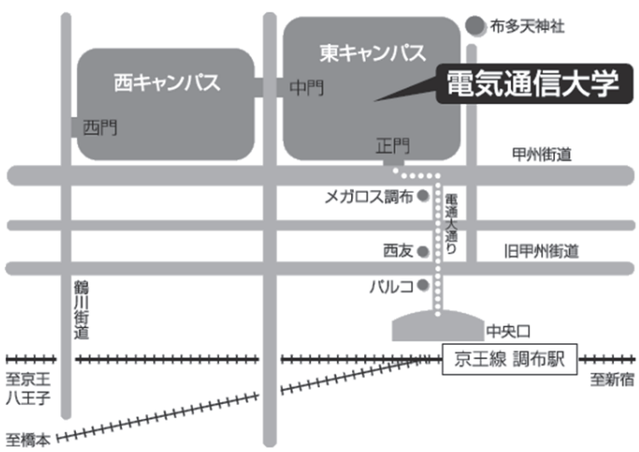
(2) JR **渋谷駅から**

渋谷駅　→ 井の頭線（急行約 7 分）→ 明大前駅→　京王線（特急・準特急約 9 分）→　調布駅

(3) **羽田空港　国内線ターミナル（京急急行）から**

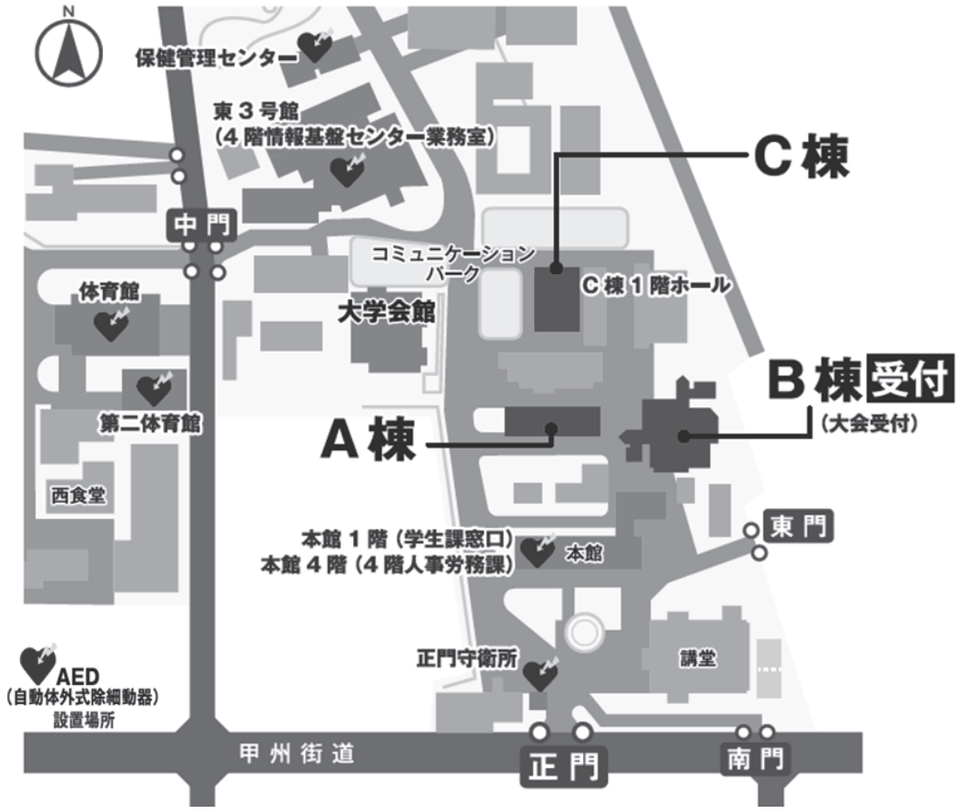
羽田空港　→ 京急空港線（快特約 16 分）→　品川駅　→ JR 山手線外回り（約 20 分）→　新宿駅　→ 京王線（特急・準特急 15 分）→　調布駅

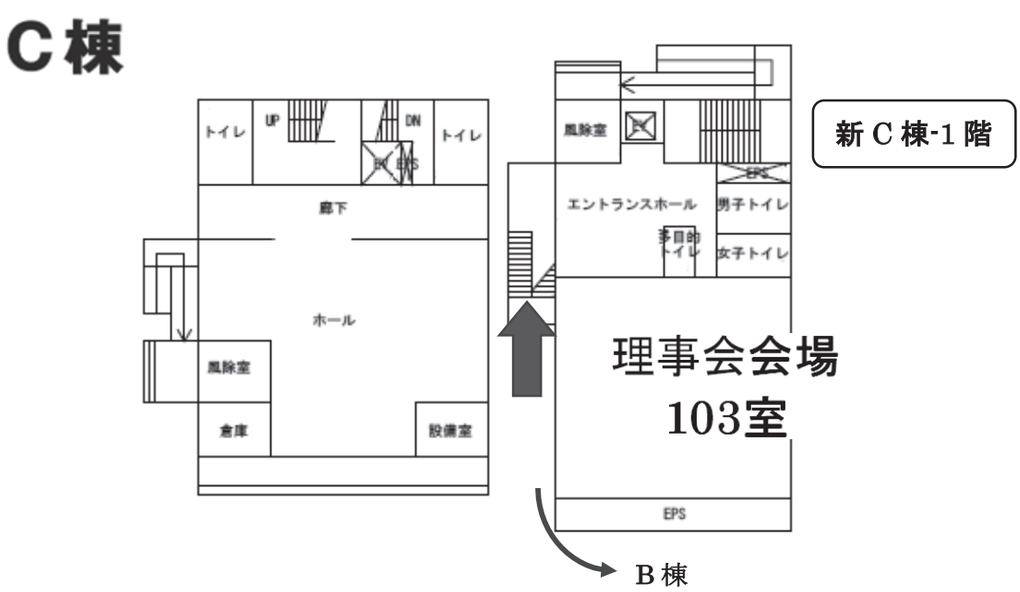
調布駅中央口から電気通信大学まで（徒歩約 5 分）



230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　5　2016/08/29　13:59:35

会場案内図

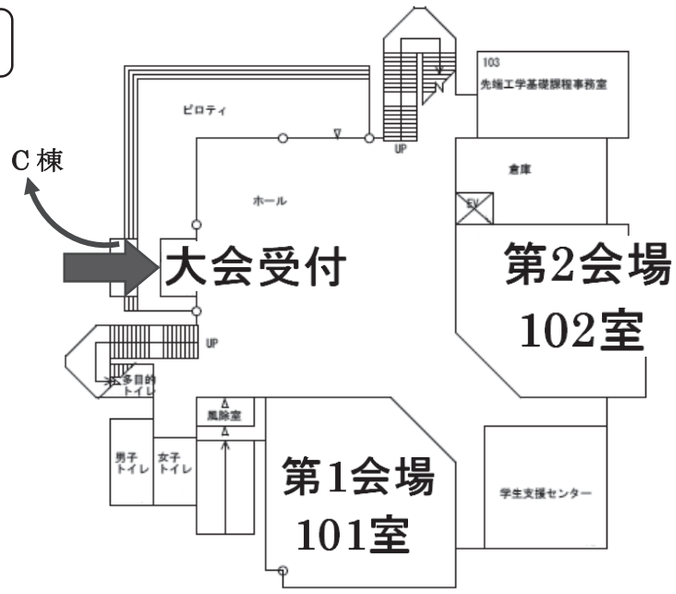




230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 6 2016/08/29 13:59:35

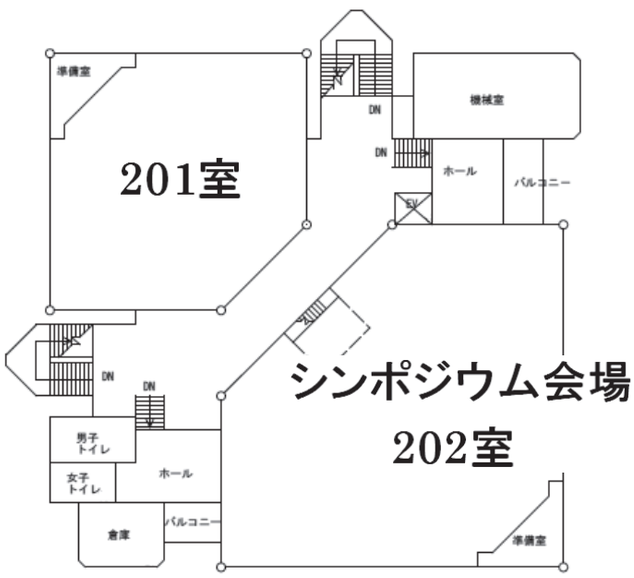
B棟

B棟-1階



※1日目大会本部102教室、
2日目大会本部201教室

B棟-2階



230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 7 2016/08/29 13:59:35

大会日程

９月８日（木）	
時間	シンポジウム会場（B202）
9:00 -	受付
10:30 - 12:00	シンポジウム１ 「e テスティング」
12:00 - 13:30	昼休み
13:30 - 14:15	総会
14:30 -16:00	基調講演
16:20 -17:50	シンポジウム２ 「入試改革」
18:00 -	懇親会

９月９日（金）		
時間	第１会場（B101）	第２会場（B102）
8:00	受付	
9:00 - 10:20	一般セッション１ 「大学入試」	一般セッション２ 「ジェネリックスキル」
10:30 - 11:50	一般セッション３ 「記述式・論述式テスト」	一般セッション４ 「初年次教育」
11:50 - 13:00	昼休み	
13:00 - 14:40	企画セッション１ 「e ポートフォリオ」	一般セッション５ 「テスト統計」
14:50 - 16:20	企画セッション２ 「21 世紀スキル」	一般セッション６ 「問題・項目作成」
16:30 - 18:30	企画セッション３ 「認知診断モデル」	一般セッション７ 「e テスティング」

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 8 2016/08/29 13:59:35

公開シンポジウム１：
大規模型ｅテスティングの運用モデルの確立を目指して
9 月 8 日(木) 10:30 - 12:00 シンポジウム会場（B202）

企画：植野　真臣　（電気通信大学）

近年、テストの国際標準 ISO 規格(ISO/IEC 2007)で定義されるｅテスティングが世界的に普及しつつある。しかし、同一テストの測定精度の等質性を保証しなければならないｅテスティングでは、大規模なアイテムバンク（テスト項目のデータベース）と高度なコンピュータ技術が必要であり、その運用方法については未だ確立されていない。また、アイテムバンク方式が普及している西洋とは違い、日本ではこれからそれを構築する組織が多い。そのため、容易に導入可能なｅテスティング・システムの開発とアイテムバンクの継続的構築も考慮した運営モデルの開発が急務である。本シンポジウムでは、すでに大規模型ｅテスティングの実施を進めてきた方々にそれぞれの組織での運用方法についてご発表いただき、大規模型ｅテスティングの運用モデルを明らかにしていく。

1. 国家試験『IT パスポート試験』におけるｅテスティング …… 22

谷澤　昭紀　(独立行政法人 情報処理推進機構 IPA)
本多　康弘　(独立行政法人 情報処理推進機構 IPA)

2. GTEC CBT の運用を支える ICT と測定理論 …… 26

込山　智之　（一般財団法人進学基準研究機構 CEES）
野澤　雄樹　（ベネッセ教育総合研究所）

3. 大規模英語テストにおけるアイテムプール拡張と管理・運用について …… 28

仲村　圭太　（公益財団法人 日本英語検定協会）

4. 医療系大学間共用試験の特色 …… 30

仁田　善雄、齋藤　宣彦、石田　達樹、島谷　一芳、江藤　一洋
(公益社団法人 医療系大学間共用試験実施評価機構)

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 9 2016/08/29 13:59:36

基調講演：

認知科学を基礎とした新たなアセスメント研究動向

9 月 8 日(木) 14:30 - 16:00 シンポジウム会場（B202）

認知科学にもとづいた、学習の、学習のための、学習としてのアセスメント
― ETS による研究イニシアチブ …… 32

講演：ETS 首席研究員 Paul Deane 博士

この基調講演では、ETS 首席研究員の Paul Deane 博士をお招きし、氏が携わる CBAL（Cognitively Based Assessment of, for, and as Learning）イニシアチブを通して、認知科学を基礎とした次世代型のアセスメントについてお話いただく。

CBAL イニシアチブは、米国における現行の K-12 のアセスメントを、「学習への貢献」を強く意識した次世代型アセスメント・システムへと変革するための研究プロジェクトである。その特徴は三つある。一つ目は、社会的文脈における学習と習熟を考慮した言語の「認知モデル」を構築していることである。認知モデルに基づいてアセスメントを設計し、評価を行うことで、学習の到達状況に即したきめ細かなフィードバック、指導が可能になることが期待される。二つ目は、相互に関連づけられた形成的評価と総括的評価を年間にわたって複数回行う「アセスメント・システム」を構想していることである。三つ目は、コンピュータ上での試験実施や機械採点など、最新の技術を積極的に活用する点である。

本講演では、認知科学に基づくフレームワークを、どのようにして具体的なアセスメントタスクに落とし込み、測定し、評価するのかといった CBAL の具体的な手法について、これまでの研究成果を交えながら紹介していただく予定である。

司会：加藤　健太郎　（ベネッセ教育総合研究所）

備考：同時通訳有り

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 10 2016/08/29 13:59:36

公開シンポジウム２：
大学入試改革の概要と課題
９月８日(木) 16:20 – 17:50　シンポジウム会場（B202）

高大接続システム改革と「大学入学希望者テスト」の導入 …… 36

講演：鈴木　寛
（東京大学公共政策大学院教授 兼 慶應義塾大学政策・メディア研究科教授）

指定討論者
徳永　健伸　（東京工業大学）
石岡　恒憲　（大学入試センター）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 11 2016/08/29 13:59:36

一般セッション１：大学入試

９月９日（金）9:00 – 10:20　第１会場（B101）

座長：橋本　貴充　（独立行政法人大学入試センター）

1. 国立大学における個別学力試験の解答形式に関する研究（１） …… 40

○宮本　友弘　（東北大学）
庄司　強　（東北大学）
田中　光晴　（東北大学）
石上　正敏　（東北大学）
倉元　直樹　（東北大学）

2. 国立大学における個別学力試験の解答形式に関する研究（２） …… 42

○庄司　強　（東北大学）
宮本　友弘　（東北大学）
田中　光晴　（東北大学）
石上　正敏　（東北大学）
倉元　直樹　（東北大学）

3. センター試験の成績未利用行動の分析（1）－理科②の成績未利用者の分析－ · 44

○橋本 貴充　（独立行政法人　大学入試センター）
内田 照久　（独立行政法人　大学入試センター）
鈴木 規夫　（独立行政法人　大学入試センター）

4. 大学・大学院入試に向けた英語スピーキングテストの尺度化事例
－受験者特性と評価者属性を考慮したモデルによる検討－ …… 48

○光永　悠彦　（島根大学）
神澤　克徳　（京都工芸繊維大学）
坪田　康　（京都工芸繊維大学）
羽藤　由美　（京都工芸繊維大学）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　12　2016/08/29　13:59:36

一般セッション２：ジェネリックスキル

９月９日（金）9:00 – 10:20　第２会場（B102）

座長：木村　哲夫　（新潟青陵大学）

1. 論述力を測定する混合フォーマットテスト開発の試み …… 52

○野澤　雄樹　（ベネッセ教育総合研究所）
伊藤　素江　（ベネッセ教育総合研究所）
須永　正巳　（ベネッセ教育総合研究所）
堂下　雄輝　（ベネッセ教育総合研究所）
村田　維沙　（ベネッセ教育総合研究所）

2. 「実社会・実生活に生きる多様な力を測るテスト」の多面的な分析 …… 56

○岡村　彩音　（岡山大学大学院教育学研究科）
○志水　黎　（岡山大学大学院教育学研究科）
山田　剛史　（岡山大学大学院教育学研究科）
長谷川　康代　（(株)ベネッセコーポレーション）
牧野　直道　（(株)ベネッセコーポレーション）
岡下　博美　（(株)ベネッセコーポレーション）

3. クリティカルシンキングの適応型テストの尺度開発 …… 58

○若山　昇　（帝京大学/CRET）
宮澤　芳光　（東京学芸大学）
梶谷　真司　（東京大学）
植野　真臣　（電気通信大学）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　13　2016/08/29　13:59:36

一般セッション 3：記述式・論述式テスト

9 月 9 日（金）10:30 – 11:50　第 1 会場（B101）

座長：倉元　直樹　（東北大学）

1. 評価者と課題の多様な特性を考慮した項目反応モデル …… 62

○宇都　雅輝（電気通信大学大学院情報理工学研究科）
植野　真臣（電気通信大学大学院情報理工学研究科）

2. ピアアセスメントの精度を最適化する評価者選択手法…… 66

○Nguyen Duc Thien（電気通信大学大学院情報理工学研究科）
宇都雅輝（電気通信大学大学院情報理工学研究科）
植野真臣（電気通信大学大学院情報理工学研究科）

3. IRT による理系記述式テストデータの分析 (2)

――記述式テスト IRT 化検討の背景―― …… 70

○倉元　直樹（東北大学）
泉　毅（教育測定研究所）

4. IRT による理系記述式テストデータの分析 (3)

――項目パラメタ・能力パラメタによる IRT 適用課題の導出―― …… 72

○泉　毅（教育測定研究所）
倉元　直樹（東北大学）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　14　2016/08/29　13:59:37

一般セッション４：初年次教育

９月９日（金）10:30 – 11:50　第２会場（B102）

座長：大森　拓哉　（多摩大学経営情報学部）

1. オンライン英会話グループ学習を用いた英語学習態度改善の取り組みと測定 · 76

森　一将　（文教大学経営学部）
遠山　道子　（文教大学経営学部）
新谷　真由　（東京電機大学理工学部共通教育群）
大森　拓哉　（多摩大学経営情報学部）

2. 工学系大学の数学基礎教育における e テスティングの取り組み …… 78

○亀田　真澄　（山陽小野田市立山口東京理科大学）
宇田川　暢　（山口県立大学）

3. 大学１年生 1000 人を対象とした毎週オンライン IRT テスティングシステムとその運用 …… 82

○廣瀬　英雄　（広島工業大学）

4. Co-creating ルーブリックの自己制御学習および自己評価力への影響分析 …… 86

○山本　美紀　（電気通信大学）
宇都　雅輝　（電気通信大学）
西山　悠　（電気通信大学）
川野　秀一　（電気通信大学）
植野　真臣　（電気通信大学）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　15　2016/08/29　13:59:37

企画セッション１：ｅポートフォリオ

９月９日（金）13:00 – 14:40　第１会場（B101）

企画：森本　康彦　（東京学芸大学）

【総説】

1. ｅポートフォリオ／学習記録データを活用したアクティブ・ラーニングの学習評価 …… 90

森本　康彦　（東京学芸大学）

【初等中等教育セッション】

2. 汎用的な資質・能力を育成するためのパフォーマンス評価 …… 92

鄭　谷心　（東京学芸大学）

3. 項目反応理論に基づくアクティブ・ラーニングの学習評価 …… 96

宮澤　芳光　（東京学芸大学）

【高等教育セッション】

4. 大学の授業におけるｅポートフォリオを活用した学習評価 …… 98

田中　洋一　（仁愛女子短期大学）

5. ｅポートフォリオを用いた大学教育の質保証 …… 100

宮崎　誠　（畿央大学）

【パネルディスカッション】

登壇者全員によるパネルディスカッションを行う．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　16　2016/08/29　13:59:37

一般セッション５：テスト統計

９月９日（金）13:00 – 14:40　第２会場（B102）

座長：柴山　直　（東北大学大学院教育学研究科）

1. 複数国において実施された性格特性検査における因子構造 …… 102

○酒匂　志野　（株式会社リクルートマネジメントソリューションズ）
坂本　佑太朗　（株式会社リクルートマネジメントソリューションズ）
今城　志保　（株式会社リクルートマネジメントソリューションズ）
分寺　杏介　（東京大学大学院）

2. 複数国において実施された性格特性検査における DIF の検出 …… 106

○坂本　佑太朗　（株式会社リクルートマネジメントソリューションズ）
酒匂　志野　（株式会社リクルートマネジメントソリューションズ）
今城　志保　（株式会社リクルートマネジメントソリューションズ）

3. 連続変数を含む補償的認知診断モデルの開発 …… 108

○山口　一大　（東京大学大学院教育学研究科）
星野　崇宏　（慶應義塾大学大学院経済学研究科）

4. カテゴリ項目尺度における希薄化の修正について …… 112

○岡本　安晴　（日本女子大学人間社会学部心理学科）

5. 対応づけ得点の信頼性係数について …… 116

○佐藤　喜一　（新潟大学教育戦略統括室）
柴山　直　（東北大学大学院教育学研究科）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　17　2016/08/29　13:59:37

企画セッション２：
21 世紀型スキルの学習環境と評価のありかた
９月９日（金）14:50 – 16:20　第１会場（B101）

企画・司会：益川　弘如　（静岡大学）

1. 21 世紀型スキルが喚起する学習評価のあり方に関する議論 …… 120
白水　始　（東京大学）

2. CoREF における評価の取り組み …… 121
齊藤　萌木　（東京大学）

3. 資質・能力を育む学習環境における縦断的評価 …… 121
益川　弘如　（静岡大学）

指定討論者：南風原　朝和　（東京大学）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　18　2016/08/29　13:59:38

一般セッション６：問題・項目作成

９月９日（金）14:50 – 16:20　第２会場（B102）

座長：石井秀宗（名古屋大学大学院教育発達科学研究科・教育学部）

1. The Impact of LAMS Project Training on Test Development Skills of Trainees and Change in their Perceived Self-Efficacy …… 122

○Etenesh Mekonnen
（Mathematics and Science Improvement Center, Ministry of Education, Ethiopia）
Hidetoki Ishii
（Psychology and Human Developmental Science, Graduate School of Education and Human Development, Nagoya University）

2. Quality Difference between LAMS and Conventional Mathematics Question Items in Item Difficulty and Discrimination Poweron Grade Seven Students in Bahir Dar City, Ethiopian …… 126

○Bimerew Kerie
（Mathematics and Science Improvement Center, Ministry of Education, Ethiopia）
Hidetoki Ishii
（Psychology and Human Developmental Science, Graduate School of Education and Human Development, Nagoya University）

3. エビデンスに基づいた多肢選択式問題作成ルールの策定に向けて …… 130

○上松　慮生　（プロメトリック株式会社）

4. MMQ 形式の問題項目の採点方法について …… 132

○荒井　清佳　（大学入試センター研究開発部）
宮埜　寿夫　（大学入試センター研究開発部）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　19　2016/08/29　13:59:38

企画セッション 3：認知診断モデル

9 月 9 日（金）16:30 – 18:30 第 1 会場（B101）

企画・司会：服部 環（法政大学）

受験者の学習改善を目的としたとき，テストの合計点もしくは 1 次元性を仮定した項目反応モデルに基づく特性値をフィードバックするだけでは，十分とは言えない。そのため，近年では，個々の項目に正答するために必要な知識や理解力の関係を積極的に利用して受験者の認知的な診断を行う測定モデルが注目されている。そこで，本邦において，そのような認知診断を行うモデルを運用し，また，認知的な診断を行うために有用な情報や測定技術を創出する研究に携わっている方々が最新の研究成果を発表する場を設けることに意義があると考え，本セッションを企画した。

1. 認知モデルによる誤り要因分析 …… 134

大森 拓哉 （多摩大学）
森 一将 （文教大学）

2. 認知的診断のための測定論的情報を抽出する IRT 分析 …… 136

坂本佑太朗 （株式会社リクルートマネジメントソリューションズ）

3. IRT モデルにもとづく学力評価ルーブリックの作成手法の試み
—小学校算数・面積学習領域を対象として— …… 140

佐藤 誠子 （石巻専修大学）
柴山 直 （東北大学）

4. 認知診断モデルによる学習診断の教育実践での利用可能性 …… 144

鈴木 雅之 （横浜国立大学）

5. 外国語学習支援に向けた認知診断テストの開発 …… 146

孫 媛 （国立情報学研究所）
島田 めぐみ（国立情報学研究所）
谷部 弘子 （東京学芸大学留学生センター）

指定討論者：村木 英治 （東北大学）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 20 2016/08/29 13:59:38

一般セッション７：ｅテスティング

９月９日（金）16:30 – 18:30　第２会場（B102）

座長：藤田　彩子（株式会社リクルートキャリア）

1. 適応型テストの回答データを用いた新作項目の項目特性値推定 …… 150

○杉山　剛　（株式会社リクルートキャリア測定技術研究所）
仁田　光彦　（株式会社リクルートキャリア測定技術研究所）

2. 大規模 CBT 実施における標準手法
－Asia ATP のホワイトペーパーに基づく検討－ …… 154

○齋藤　康寛　（Asia ATP／株式会社イー・コミュニケーションズ）

3. 日本の初等中等教育における CBT 導入の意義についての考察
～理論、技術、運用の議論以外に何が必要なのか　～ …… 156

○劉　東岳　（学研ホールディングス、学研教育総合研究所）

4. 上方カメラを用いたタブレット端末におけるタッチジェスチャによる
替え玉判別 …… 160

○中村　祐文　（成蹊大学大学院理工学研究科）
渡邉　悠二　（元成蹊大学大学院理工学研究科）
小方　博之　（成蹊大学理工学部）

5. オンライン受験における聴覚的不正行為防止を目的としたマスキング音が
存在する時の集中力への影響 …… 162

三浦　晋矢　（成蹊大学理工学部）
窪田　瑞穂　（成蹊大学理工学部）
○張　成龍　（成蹊大学大学院理工学研究科）
小方　博之　（成蹊大学理工学部）

6. 能力検査受検にスマートフォンを使用することの影響とその要因 …… 164

○園田　友樹　（株式会社リクルートキャリア）
藤田　彩子　（株式会社リクルートキャリア）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　21　2016/08/29　13:59:38

公開シンポジウム１：

大規模型ｅテスティングの運用モデルの確立を目指して

９月８日（木）　10:30 - 12:00　シンポジウム会場（B202）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　23　2016/08/29　13:59:39

国家試験『ITパスポート試験』におけるｅテスティング
（導入フェーズから現在に至るまでのあらましのご紹介）

独立行政法人情報処理推進機構（IPA）IT人材育成本部情報処理技術者試験センター
○谷澤　昭紀，本多　康弘

１．はじめに

現在，国家試験として唯一CBT（Computer Based Testing）方式で実施しているITパスポート試験（ｉパス）は，平成21年春から平成23年秋までは年2回開催するペーパー方式であった。今では全国約110箇所の会場で毎日のように開催する，年間応募者数8万人規模のCBT試験に成長した。

第12回大会（2014年）では，IPAは“今後の日本のｅテスティングの普及，展開においてもリードしていきたい”と表明した。今回の第14回大会では，CBT導入プロジェクトにおける検討等について，本学会の場で情報共有を図り，日本のｅテスティングの普及，展開に寄与したい。

２．CBT導入の背景

CBT導入のきっかけは，平成19年7月「経済産業省産業構造審議会人材育成ワーキンググループ報告書」に示された，“民間活力の活用”であった。試験実施業務そのものを民間事業者に開放し，民間事業者のノウハウを生かして，安定的かつ効率的な方法で試験を実施し，受験者の利便性を向上させることを目指した。その後，IPAは，平成19年12月「新試験制度審議委員会報告書」の中で，平成21年度に創設するｉパスにおいて，“平成23年度を目処に本格的なCBT方式の導入を目指す”ことを公表した。CBT方式では，“年間を通じて頻繁に試験を実施することを目指す”，“受験者がその場で試験結果を確認できるようにする”といった“受験者の利便性の向上”も狙いとした。なお，“国家試験として初めての試みにつき入念な準備が必要”とし，“ｉパス開始当初はペーパー方式で試験を実施”，“採点はｉパス開始当初は素点方式とし，準備が整った段階で統計的手法（IRT）を導入する”などを具体的に表明し，以降本格的な導入プロジェクトがスタートした。

３．導入プロジェクト

(1) プロジェクト推進体制構築と役割分担

CBT導入に向け，IPAでは2つのワーキンググループ（問題作成WG，試験実施WG）を設置し，それぞれが深く専門的に検討できる体制を整備した。なお，各WGにはテスト理論の専門家，CBT事業の専門家を招き，助言を受けながら密な関係をもって推進した。

＜問題作成WG：アイテムバンクやIRTといったテスト設計を中心に検討＞

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　22　2016/08/29　13:59:39

ａ．アイテムバンク設計，アイテムバンク登録基準の検討，IRT の導入
ｂ．問題セット組立て・問題セット間の等質化，問題セットの配信方法
ｃ．パソコン上の試験画面の検討，結果データ収集・分析・評価の PDCA サイクル検討
ｄ．アイテムを非公開とする議論
ｅ．アイテム作成量の検討と作成スケジュール計画と作成体制増強
ここで，ｂ.「問題セット組立て・問題セットの等質化」については，第 12 回大会の抄録で紹介しているので参照いただきたい。

＜試験実施 WG：試験会場や試験運営スタッフ，会場試験運営業務を中心に検討＞
ａ．試験会場の設置基準検討と設置スケジュール及び会場確保
ｂ．試験会場の業務運用や試験運営スタッフの確保及び教育
ｃ．試験制度変更に関する全国説明会の企画及び開催
ここで，ｂ.「試験会場の業務運用」については，第 12 回大会の抄録で紹介しているので参照いただきたい。

(2) 全国規模によるリハーサル試験の企画と開催
試験会場手配，会場要員教育，試験システムの導入等が完了すると，最後はこれらが本番運用に耐えられるものか，気付いていない課題が残されていないか等を確認する目的で，できるだけ本番さながらの環境（試験会場，受験者数など）を想定し，リハーサル試験を企画，開催した。
リハーサル試験参加者には，本番と同様の CBT 環境，試験時間で試験問題を解いてもらい，最後にアンケートに回答してもらった。

＜第 1 回リハーサル試験＞
ａ．平成 23 年 1 月 17 日～平成 23 年 3 月 27 日（約 2.5 か月間）
ｂ．全国 47 都道府県／100 会場以上で実施
ｃ．一人 1 回の無料体験，合否判定なし，試験結果を正答数（正答率%）で表示，試験問題は非公表
リハーサル試験実施期間中の 3 月 11 日に東日本大震災が発生し，やむなく 15 会場以上が中止となった。なお，それでもリハーサル試験は最後まで継続し，結果として 16 項目の改善箇所を抽出することができた（表 1)。

改善箇所	改善理由	重要度（利用者視線）	改善後の仕様（案）
試験問題の初期文字サイズ	文字の初期表示サイズが小さく、見にくいという意見が多数	大	現状の140％を初期サイズとして（案）、あわせて拡大率の上限も拡張か。
解答一覧の表示機能	「後で見直す」にチェックした問題だけを一覧で表示させたいという意見が多数	中	一覧表示機能を追加
試験問題以外の文字サイズ	弱視の方でも受験できるようにも文字倍率変更機能があるが、ログイン画面や同意事項など、試験問題以外の文字サイズが小さいので結局受験できない。	大	文字のポイント数を大きくする

表 1　改善項目（一部）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　23　2016/08/29　13:59:39

＜第 2 回リハーサル試験＞

ａ．平成 23 年 6 月 22 日～平成 23 年 8 月 21 日（約 2 か月間）

ｂ．全国 14 都道府県／23 会場で実施

ｃ．一人 1 回の無料体験，合否判定なし，試験結果を正答数（正答率%）で表示，試験問題は非公表

第 1 回リハーサル試験から抽出された 16 項目の改善箇所についての確認と本番稼働前の最終点検の目的で実施した。当初の計画では第 2 回リハーサル試験は想定していなかったが，国家試験としての安全性や信頼性の確保を最優先とした。第 1 回及び第 2 回リハーサル試験を通して，すべての試験会場とすべての試験運営スタッフが参加した訓練が実施できたこと，また多くの利用者の声を直接聞くことと，改善箇所として取り入れることができたことは大変有意義であった。

(3) リハーサル試験のもう一つの目的

IRT モデルの CBT では，あらかじめ a 値，b 値，項目情報量といった項目特性値が評価されていなければならない。そのため，項目特性値を評価するための被験者データをいかに取得するかについて検討した。法政大学／服部環先生，名古屋大学／野口裕之先生，電気通信大学／植野真臣先生にご支援いただき，被験者データ取得にかかわるコスト、容易性等で判断し，リハーサル試験の被験者データを使用することとした。ここで，リハーサル試験参加者にもう一つの目的を知らせるべきか議論したが、知らせた場合，故意に解答を間違えるなど，項目特性値に悪影響を与えることが想定されたことから，知らせることはしなかった。

４．利用者メリットに配慮したサービスの継続的な導入

ｉパスのような能力認定試験は，受験時点における受験者の能力を評価する。そのため，試験結果を踏まえて苦手分野をブラッシュアップすることで受験者の能力はさらに向上する。IPA は，CBT の使いやすさや教育面に配慮するなど，CBT 開始以来，次のような付加サービスを導入していることをご紹介したい。

＜申込内容変更期間の短縮＞

申込内容を変更できる期間を，当初の受験日 7 日前から 3 日前まで短縮することで，受験者の都合による受験機会を逃すといった事態にできるだけ対応できるようにした。

＜“強み・弱み”の可視化（バウチャー購入者に対する付加サービス機能）＞

ｉパスの成績を大きく 14 の知識領域にグループ化し，“自組織の強み・弱み”をレーダチャートにより可視化できるようにした（図 1）。

https://www3.jitec.ipa.go.jp/JitesCbt/html/examination/poc_report.html

＜疑似体験ソフトウェアの配布＞

CBT を事前に自宅で体験してもらうべく疑似体験ソフトウェアの配布を開始した。本ソフトウェ

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 24 2016/08/29 13:59:40

アには，年 2 回（各 100 問）公開している問題が登録されており，本番と同様に問題を解きながら，CBT 画面を体験することができる。

https://www3.jitec.ipa.go.jp/JitesCbt/html/guidance/trial_examapp.html

なお，本ソフトウェアは i パス CD としても配布しているのでご希望があればお問い合わせいただきたい（図 2）。



図 1　レーダチャート



図 2　配布用 i パス CD

５．終わりに

i パスは，平成 25 年 6 月に閣議決定された「世界最先端 IT 国家創造宣言」において，“情報モラルや情報セキュリティに関する知識を含め，国民全体の情報の利活用力の向上を図る”とし，その進捗状況や達成度を表す指標（KPI）に“IT パスポート試験の延べ合格者数”が示された。このようにわが国の IT 政策の一環として活用され，期待されている国家試験として，CBT のメリットを最大限に生かしつつ，さらなる e テスティングの普及のためにも IPA は継続して貢献していきたい。

参考文献

1. 植野真臣，永岡慶三（2007）e テスティング，培風館
2. ISO/IEC (2007) ISO/IEC 23988, Information technology - A code of practice for the use of information technology (IT) in the delivery of assessments.
3. Pokpong Songmuang and Maomi Ueno: Bees algorithm for construction of multiple test forms in e-testing. IEEE Transactions on Learning Technologies 4, 209-221, IEEE Computer Society (2011)
4. Takatoshi Ishii, Pokpong Songmuang, Maomi Ueno: Maximum Clique Algorithm and its approximation for Uniform Test Form Assembly. IEEE Transactions on Learning Technologies 7, 88-95, IEEE Computer Society (2014)

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　25　2016/08/29　13:59:40

GTEC CBT の運用を支える ICT と測定理論

○込山智之*　　○野澤雄樹**
* 一般財団法人進学基準研究機構（CEES）　　** ベネッセ教育総合研究所

GTEC CBT の概要

グローバル社会で活躍できる人材の育成を目的として，大学教育およびその入り口である大学入試が大きく変わり始めている。文部科学省は，全国の大学において英語 4 技能（「聞くこと」「話すこと」「読むこと」「書くこと」）を測定する外部検定試験の活用を促す提言を 2012 年に行っている。

ベネッセコーポレーションでは，1998 年より中高生の英語運用力を測定する GTEC for STUDENTS を提供してきている。GTEC for STUDENTS は紙ベースでリーディング，リスニング，ライティングの 3 技能，タブレットでスピーキングの測定を行うテストであり，2015 年度には全国の約 1,500 校の高校で採用され，約 81 万人が受験した。テストの結果は，学校における指導改善や，生徒個人の学習の動機づけのために活用されている。

GTEC CBT は，GTEC for STUDENTS の運用を通じて得られた知見に基づいて開発されたコンピュータ型テストであり，4 技能を約 175 分で測定する。時代の変化に伴う高校卒業後の進路選択機会の拡大や，「大学の一般入試でも，高い英語力を測定したい」という大学のニーズに対応した検定試験となっている。GTEC CBT は，学習指導要領に定められた語彙・表現を基本としながらも，その範囲に留まらない出題を行い，大学レベルの授業を理解し，専門性が高い文献を読み，レポートや論文を書くことができる「アカデミックな英語力」を測定する。テスト問題および得点尺度は，ヨーロッパ全体で言語能力を示す際に用いられるガイドライン CEFR（Common European Framework of Reference for Languages）に準拠して作成されており，日本の高校生に対応する A2 レベルから，海外大学進学希望者に対応する B2～C1 レベルまでを測定できるように設計されている。

大学入学後に求められる英語力を測定することができるハイステイクスな検定として，一般財団法人進学基準研究機構（CEES）とベネッセコーポレーションの共催で運営・実施を行っている。

GTEC CBT を支える ICT

GTEC CBT は，その名の CBT （Computer-Based Testing）が示すように，コンピュータの画面に問題を表示し，キーボードやマウスで解答を入力する形式の試験である。特にスピーキングテストでは，CBT を活用することで面接試験官の確保という人的リソースに依存せず，大規模な同時実施が可能となる。

全国 100 を数える拠点において，年 3 回・同時刻に数千人が一斉受験をするという大規模な実

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　26　　2016/08/29　13:59:41

施・運営を行っている。各拠点で使用する PC 等の設備は，OS やネットワークなど一定の基準を満たしている必要があるが，会場により多少の違いが発生する。そのため，環境にできる限り依存しない形で，システムのセキュリティを担保し，かつ，ヒューマンエラーが発生してもリカバリーできる CBT システムにすることで，テストの安定実施を実現している。

また，4 つの技能を限られた時間で精度よく測定するためには，テストの効率を上げる必要がある。先述のように，GTEC CBT は CEFR の A2～C1 という広い範囲を測定しているので，従来の固定問題セットでは高い測定精度を確保できない。そのため，GTEC CBT のリスニングテストとリーディングテストにおいては適応型テスト（Computerized Adaptive Testing, CAT）を導入し，受験者の英語力に応じて適切な問題を出題する形式で実施している。CAT を活用した場合，受験者の正解・不正解の状況に基づいて情報量の高い問題が優先的に出題されるため，固定の問題セットの場合と比べて短時間で精度の高い測定が可能となる。また，各受験者の受験履歴を管理し，過去に出題された問題が再度出題されないように制限するなど，GTEC CBT では ICT を柱とした厳格なテストの運用が行われている。

GTEC CBT を支える測定理論

GTEC CBT では，技能ごとに等化を行い，受験回に拠らず同じ解釈が可能な尺度得点を報告している。そのため，各受験回のスコアを比較することで，能力の伸びを直接的に把握することができる。測定精度の面では，リスニングテストとリーディングテストでは受験者の能力に応じて問題を選択する CAT を採用し，どの受験者に対しても一定水準の測定精度を確保できるように出題を管理している。固定版であるスピーキングテストとライティングテストでは，測定精度が版によって大きく異ならないように事前検証がなされている。これらを可能にしているのが，項目反応理論（Item Response Theory, 以下 IRT）であり，GTEC CBT の運用を支えるもう 1 つの柱である。

IRT は非常に強力なツールであるが，さまざまな統計学的仮定を置いているため，それらの仮定が本当に成り立っているかを常にモニタリングしなければならない。例えば，GTEC CBT では，1 回目のフィールドテストで統計的な基準を満たした項目を選別し，2 回目のフィールドテストで項目パラメータを推定している。この項目パラメータは本試験のスコアリングに使用されることになるが，これが可能なのは，フィールドテストと本試験で，母集団レベルでは項目パラメータが変動しないという仮定があるからである。しかし実際には，さまざまな要因（テスト形式の違い，モチベーションの違い，受験者の性質の違い，採点基準の時間的変化）によって，母集団レベルで項目パラメータが変動してしまうことがある。GTEC CBT では，本試験のデータを実施後すぐに分析することで，この仮定が成り立っているかどうかを検証している。もし仮定が成り立っていない場合，その影響を慎重に検討したうえで，項目パラメータの再推定など，さまざまな対応を取ることになる。これらの作業の各段階では，2 人以上の専門家による相互チェックが行われ，ミスの発生を防止する体制になっている。

このほかにも，項目プールの管理や効率の良いフィールドテストの設計など，測定理論に基づいた中長期的なサポートが行われている。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 27 2016/08/29 13:59:41

大規模英語テストにおけるアイテムプール拡張と管理・運用について

仲村 圭太
公益財団法人　日本英語検定協会

1. 現状

日本の大学入試は高大接続を焦点に改革が進みつつある。特に英語に関してはグローバル人材の育成の観点から中央教育審議会高大接続部会においても大学入試における外部試験の活用が検討されている。またテスティング及び言語テスティングの観点(植野他 2008, バックマン他 2000)においては、各種利害関係者のため事前の予備テストを行い、出題する問題の質や難易度等に関する情報を取得し、アイテムプールを作成し、試験後の質の振り返りなどを行うことが望ましいとされている。

2. 目的

今後英語テストの結果が様々な用途へ使用されるのに伴い、アイテムプールの拡張と管理・運用を適正に進めることがこれまで以上に求められると考えられる。そのため、今回は日本英語検定協会が問題開発・実施運営を行っている英語テスト（英検・TEAP）をベースにして、現状の問題作成及びアイテムプールの管理・運用に関して現状の共有と課題の洗い出しを行った。

3. 方法

a) 優良アイテムを作成するための工夫、b) アイテムバンクの具体的な作成方法、c) テストの構成方法、d) テスト問題の管理、e) テスト実施方法、f) テスト実施後の評価、等のポイントから現状の運用を記述し、更なる改善点を洗い出した。

4. 結果概要

a) 優良アイテムを作成するための工夫

テストが測定する構成概念（アカデミック・ジェネラル）の設定、言語テストに求められる目標言語使用領域であるTLU（各技能を用いる場面等）の設定、各タスクで求められるスキルを明確化し、制作チームと共有している。また各テストで求められる単語数、難易度、語彙レベル、正解選択肢・誤答選択肢の作成方法などを示した設計図を作成し、活用している。

b) アイテムバンクの具体的な作成方法

年に複数回、予備テストを行い、新規アイテムの質を確認しながらアイテムバンクを拡張させている。具体的には既存アイテムプール内のパラメータ既知アイテムを用いて、新作アイテムのパラメータを

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　28　2016/08/29　13:59:42

推定している。また予備テストの結果、作成側で事前に想定した難易度及び回答パターンが得られているかを確認している。作成チームが想定した以外の回答が得られたり、正解選択肢よりも多く選ばれた選択肢があったりした場合には、制作チームにフィードバックを行い、内容の再確認及び必要に応じて改善を行っている。

c) テストの構成方法

各テストには固定のタスク構成（難易度・出題数等）があり、その構成にしたがって毎回出題している。同一セット内での問題トピックやターゲット単語・表現に重複がないように注意している。問題が非公開のテストにおいては全てパラメータ既知アイテムを用いて、事前にTCCを作成し年間で実施セット間での難易度の比較・確認を行っている。

d) テスト問題の管理

実施問題セットを作成する際に、DBに各小問の出題年度、回次、日程情報を入力している。同一受験者が年間に複数回受験することも多くあるので、DB上でのアイテムの使用履歴の確認しながらのセット作成を行っている。

e) テスト実施方法

R,L,Wは受験教室内での一斉実施を行っている、Sは各受験者が面接委員と直接対話を行う形式がメインのため、個々の受験者が個別の面接室内で面接試験を受験している。面接室においては各受験者の音声をICレコーダーやタブレット等で録音している。問題が非公開のテストにおいては、事前に各会場・教室に配布した問題冊子の数と試験直後に回収する問題冊子の数を比較して確認している。

f) テスト実施後の評価

毎回実施されたテストセットは項目レベルで古典的分析とIRT分析を両方行い、質を確認している。既知パラメータアイテムの場合、当初のパラメータとのズレの度合いを示す指標を活用している。新規アイテムの場合、モデル適合度指標を活用している。採点者を伴う技能の場合、事前のオンライントレーニングを経た採点者に対し、共通回答を全採点者が採点し協会が指定した結果とある一定以上の一致度が確保されているかを確認している。

5. まとめ

上記観点から、英検協会が問題開発・実施運営を行っている英語テストに関して、現状を記載した。その結果、構成概念の違い、測定技能、問題公開/非公開といった要素の違いはあれ、概ね同様のプロセスで運営を行っていた。その他当日、各観点の課題点も共有する。

6. 参考文献

[1] 植野真臣・永岡慶三　(2008)「eテスティング」培風館

[2] L.F.バックマン・A.S.パーマー　(2000)「言語テスト作成法」大修館書店

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　29　2016/08/29　13:59:42

医療系大学間共用試験の特色

○仁田　善雄，齋藤　宣彦，石田　達樹，島谷　一芳，江藤　一洋
（ 医療系大学間共用試験実施評価機構）

1.はじめに

医学系・歯学系の大学は，2005年12月から臨床実習開始前の共用試験を正式に運用している．共用試験実施評価機構は参加大学が会員となっている組織である．良質な問題を参加大学の教員が作成し，参加大学が試験を実施し，問題を作成する人と異なる人により問題評価・成績評価を行う試験である．Peer reviewが基本となる．この試験の目的は臨床実習開始前の到達度評価であり，臨床実習において必要となる知識，技能，態度を評価するものである．知識については，コンピュータを用いた客観試験（Computer Based Testing CBT)で評価し，基本的診療技能・態度については，客観的臨床能力試験（Objective Structured Clinical Examination OSCE)で評価している．以下に共用試験CBTの特色について詳細に示すこととする．

2.共用試験 CBT の内容と特色

2-1. アイテムバンク

共用試験は，アイテムバンクを作成し，それに基づき出題している．アイテムバンク構築において，毎年参加大学の教員が多くの試験項目を作成することから，他の試験に比べると新作の問題数が多く，良質のアイテムバンクの構築と多数の新陳代謝が可能となる．新作問題の評価にはトライアルを4年間実施し，十分な問題数を確保した後，本試験ではプール問題と新作問題を同時に出題して新作問題の項目特性値の評価を行う方式をとっている．

2-2. 問題作成

試験問題の作成はある特定の問題作成の専門家に依頼する方式ではなく，各大学の教員に作問を依頼する方式をとっている．毎年異なる領域の作問を依頼することから，作問者は該当する領域を専門とする教員となることが多い．また，教員の異動も多く（特に臨床系），毎年新たな作問者が参加することになる．このような状況のため，問題作成のためのルール作りと毎年の講習会が必要になってくる．作問マニュアルと問題作成のためのツールを配布し，問題作成講習会（大学独自のブラッシュアップのやり方の説明も含む）を開催している．さらに大学から要望があれば，問題作成のワークショップの講師派遣（歯学系）も行っている．

参加大学は，上記の講習会後，ツールの操作説明とマニュアルの内容説明のための独自の説明会を開催している．また，作成された問題のブラッシュアップも行っている．その後，大学は，ブラッシュアップした問題を機構に送る．機構ではさらに，参加大学から選ばれた委員による集中ブラッシュアップ（グループ作業）がなされる．このように，一つの問題に対して多くの教員が数回検討して採択の有無を決定している．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　30　2016/08/29　13:59:43

2-3. 試験実施

試験環境：テスト会場の場所の決定と設営，試験監督，試験システムの運用は参加大学が行う．したがって，大学により試験環境やコンピュータシステムの違いが生じる．

試験の流れ：機構から試験システム用プログラムなどを大学に配布し，大学側は学内のサーバーにインストールし，外部と遮断されたネットワークにつなげて CBT を実施する．サーバーに解答データを集めて，試験結果を機構に郵送する．機構が解答処理を行い，個人成績表などの結果を大学に返却するシステムとなっている．

試験システムの事前確認ツール：試験システムの動作確認用のツールを配布し，各大学のネットワークシステムの動作検証を行っている．さらにこの試験システムの事前体験可能なツールを配布し、受験生に実施してもらっている．このツールの施行は，大学の試験運営のトライアルにもなることから、受験者と実施主体の両方に十分な事前体験が可能となる．

モニター制度：試験の信頼性の観点から他大学のモニターが機構から派遣され，試験の実施状況を把握し，試験が問題なく実施されたかどうかを判断している．

2-4. 試験方式

臨床実習開始時期が大学により異なることから同一時期での単一フォームの試験を行うことが不可能である．そのため，受験生ごとに異なる問題セットで行う試験となっている．再試験も行うことから，同一個人に異なる問題セットが出るようにシステム制御を行っている．項目反応理論を用いて，問題の評価，成績評価を行っている．

2-5. 成績などのフィードバック

参加大学への成績返却，試験データの解析結果の提供，テスト理論についての専門家による講演会の開催，テスト理論の説明文書・用語説明・パンフレットの提供が行われている．受験生個人については，個人成績表の提供と試験全体の成績などを網羅した冊子の提供を行っている．

3. まとめ

共用試験は，会員である参加大学の多くの教員の努力により成り立っている．お互いに評価し合い，良好な試験にするために改善を繰り返している．今後，教員の負担軽減も含めたより精度の高い適切な試験を展開していくことが検討されている．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 31 2016/08/29 13:59:43

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 32 2016/08/29 13:59:44

基調講演：

認知科学を基礎とした新たなアセスメント研究動向

9 月 8 日（木） 14:30 - 16:00　シンポジウム会場（B202）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　33　2016/08/29　13:59:44

認知科学にもとづいた、学習の、学習のための、学習としてのアセスメントーETS による研究イニシアチブ

基調講演

ポール・ディーン（博士）

Educational Testing Service 首席研究員

Educational Testing Service（ETS）は、2007 年から、指導と学習をより効果的に支援するようなアセスメント・システムを構築するための方法を探る、長期的な研究イニシアチブに取り組んできた。このイニシアチブは、それが柱とする原則の頭文字をとって CBAL（**C**ognitively **B**ased **A**ssessment of, for, and as **L**earning ；認知科学にもとづいた、学習の、学習のための、学習としてのアセスメント）と呼ばれている。

CBAL イニシアチブは、近年の認知科学、学習科学における研究の知見を活用した「認知科学にもとづくアセスメント」の開発に取り組んできた。CBAL は、アセスメントにおける 3 つの目的をバランスよく達成することを目指している。3 つの目的とは、「学習のアセスメント」すなわち、学習者たちが何を知っていて、何ができるのかを特定すること、「学習のためのアセスメント」すなわち、指導上の次のステップを特定すること、そして、「学習としてのアセスメント」すなわち、アセスメントのタスクそのものを学習経験として機能させることである。

こうした目的の実現に向けて、CBAL の研究者たちはエビデンス・センタード・デザイン（evidence-centered design）を積極的に採用してアセスメントの開発を進めてきた。エビデンス・センタード・デザインでは、領域の分析（domain analysis）とエビデンスに関する議論（evidentiary argument）を構築することによりアセスメント設計を進めていく。これまでに、数学、科学、英語（国語）の 3 つの領域で並行して研究が行われてきた。

CBAL のアセスメント設計に対するアプローチには、以下の特徴がある。(1)充実した領域および学習者モデルの開発（「ラーニング・プログレッションズ」をともなう「コンピテンシー・モデル」）、(2)測定の質を損なうことなく、複雑なパフォーマンス・タスクにおけるキー・ステップを模した「シナリオ型アセスメント」、(3)学習者の認知に関するより豊富なエビデンスを集めることができるデジタル技術の使用、(4)より大規模なアセスメント・システムにおいてこれらの要素の実現を可能にする、革新的な心理測定学的分析である。本講演では、こうした特徴的な点について、英語（国語）における枠組みに関する筆者の研究を用いながら、CBAL 研究イニシアチブがどのようにこれらのテーマを扱ってきたのかを中心に論じる。CBAL の数学と科学に関するプロジェクトについての情報や出版物は、http://www.ets.org/cbal/ を参照されたい。

コンピテンシー・モデルとラーニング・プログレッションズ

CBAL における英語領域のモデル開発は、読解に関する先行研究（O'Reilly & Sheehan, 2009a, b; Sheehan & O'Reilly, 2011; Sabatini, O'Reilly & Deane, 2013）、論述に関する先行研究（Deane, Odendahl, Quinlan, Fowles, Welsh & Bivens-Tatum, 2008; Deane, Quinlan, & Kostin, 2011; Deane, Sabatini, & Fowles, 2012）、また、討論や研究といった特定の分野に関するスキルに焦点を当てた読解と論述の統合に

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 32 2016/08/29 13:59:44

関する先行研究（Deane, Sabatini, Feng, Sparks, Song, Fowles, O'Reilly, Jueds, Krovetz, & Foley, 2015; O'Reilly, Deane & Sabatini, 2015; Deane & Song, 2015; Sparks & Deane, 2015）の綿密なレビューにもとづいている。このレビューにより、英語（国語）スキルの基盤となる少数の「鍵となる活動（key practice）」を特定するに至った。これらは、独立したものとしての読解や論述のスキルを超えて、ストーリーを理解し他者に伝える、説明文から知識を構築し共有する、議論を組み立てる、研究プロジェクトを実行する、文学的・修辞的分析を行う、といった能力に焦点を当てている。それぞれの「鍵となる活動」では、特定のテキストジャンルや思考様式、問題解決様式に関連する統合的な読解と論述のスキルに焦点を当てている。また、それぞれの「鍵となる活動」は、一連のラーニング・プログレッションズと関連付けられている。ラーニング・プログレッションズとは、高等教育や就業に当たって十分なレベルのパフォーマンスを発揮できる状態に到達するためのスキルを、学習者がどのように習熟させていくかという仮説を記述したものである。初期（2011 年頃）の CBAL コンピテンシー・モデルとラーニング・プログレッションズについては、http://elalp.cbalwiki.ets.org/ を参照されたい。特定の「鍵となる活動」のラーニング・プログレッションズは改訂が重ねられており、上記の研究報告の一部に掲載されている。

シナリオ型アセスメント

学校における従来のアセスメントは二極化している。一方では多くの学校が多肢選択形式を多用し、幅広いスキルを測るために互いに出来る限り無関係で独立した項目を多数集めたアセスメントが行われている。それに対し、他の学校では、総合的なパフォーマンス型のアセスメントが行われ、独力でエッセイを書くといった複雑なタスクが課せられている。また、過去 20 年ほどの間に、シミュレーションのような、技術支援型のパフォーマンス・タスクへの関心も高まってきた。そのような中で、CBAL イニシアチブは、シナリオ型アセスメント（scenario-based assessments）と呼ばれる、両極の中間に位置する形式のアセスメントに注目してきた。シナリオ型アセスメントの典型としてシミュレーション型パフォーマンス・タスクがあるが、学習者にある複雑なタスクが与えられ、熟達者がタスクを完了するまでの過程を学習者が一段階ずつ辿るように構成されている。この過程の各段階で、ひとつあるいは複数の項目が出題され、タスクのその部分を遂行するための準備がどの程度できているかが測定される。シナリオ型アセスメントは、各下位タスクを遂行するのに必要なスキルを測定すると同時に、そのタスクを自力で行うために習得しなければならない方略の手本を示しているのである。シナリオ型アセスメントの事例は、Sheehan & O'Reilly (2012), Sabatini, O'Reilly, Halderman & Bruce (2014a, b), and Deane, Sabatini, Feng, Sparks, Song, Fowles, O'Reilly, Jueds, Krovetz, & Foley (2015)などの論文で取り上げられている。シナリオ型アセスメントが適切に設計されれば、コンピテンシー・モデルとラーニング・プログレッションズで明確化された一連の「鍵となる活動」と、シナリオ型アセスメントにおける一連の活動とを直接結びつけることができる（Deane & Song, 2015; Bennett, Deane, & van Rijn, 2016）。

エビデンス収集における革新

学習者の反応の根底にある認知についての理解を深めるために、CBAL の研究チームは、学習者のパフォーマンスについてのエビデンス収集にデジタル媒体を活用している。特に論述についてのエビデンスを収集するにあたり、学習者が書いたものを自然言語処理技術を使って自動評価したり、キー入力の記録を利用して書く過程の自動評価を行ったりする革新的な方法を採用している。一連の研究によれば、これらの方法を使うことによって、学習者が複雑なパフォーマンスをする際の活動の特徴をより細かく把握することが可能になり、シナリオ型アセスメントを補完できることが示

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 33 2016/08/29 13:59:45

唆されている（Almond, Deane, Quinlan & Wagner, 2012; Deane, 2013, 2014; Deane & Quinlan, 2011; Deane, Wililams, & Trapani, 2013; Deane & Zhang, 2015; Zhang & Deane, 2015）。特にキー入力の記録は、学習者が複雑なパフォーマンス・タスクにおいてどのように時間管理するかについての詳細を明らかにし、作文や転写のスキルにおける流暢さと効率についてのエビデンスを提供してくれる。

ラーニング・プログレッションズとシナリオ型アセスメントの心理測定学

CBAL のアセスメント設計の特性から、心理測定学上重要な問題が提起されている。例えば、以下のような問題である。シナリオ型アセスメントの設計の仕方には、従来のアセスメントと比べてどのような違いがあるのか（Bennett, 2015; Sabatini, Bennett, & Deane, 2011）。項目の相対的難易度からラーニング・プログレッションズを復元することで、アセスメント設計に使われたラーニング・プログレッションズの妥当性が示せるのか（van Rijn, Graf, and Deane, 2014; Bennett, Deane, & van Rijn, 2016）。また、シナリオ型アセスメントで使われるシナリオの構造が学習者のパフォーマンスにどのような影響を与えるのか（Zhang, Bennett, Deane, & van Rijn, under review）などである。

結論

本講演で論じた革新は、学習に関する理論とより密接に関係しうるさまざまな新しいタイプのアセスメントの基盤となる。また、高等教育・就業へのレディネスを達成するために学ぶ必要のあるパフォーマンスを、アセスメントの中でより現実に近い形で経験することを可能にしてくれる。本講演では、現在の研究の概要を示すとともに、こうしたアプローチが実験的なものから実用段階のアセスメントに適用されていく中で生じる重要な問題や懸念点を明らかにする。

参考文献

R. Almond, P. Deane, T. Quinlan, & M. Wagner (2012) A Preliminary Analysis of Keystroke Log Data from a Timed Writing Task. ETS Research Report No. RR-12-23.

R. E. Bennett, (2015) The Changing Nature of Educational Assessment. Review of Research in Education, Vol. 39, No.1, pp. 370–407

R. Bennett, P. Deane & P. van Rijn (2016) From Cognitive-Domain Theory to Assessment Practice. Educational Psychologist 51(1):1-26

P. Deane (2013) On the Relation Between Automated Essay Scoring and Modern Views of the Writing Construct. Assessing Writing, Vol. 18, No. 1, pp. 7–24.

P. Deane (2014a) Using Writing Process and Product Features to Assess Writing Quality and Explore How Those Features Relate to Other Literacy Tasks. ETS Research Report No. RR-14-03

P. Deane, N. Odendahl, T. Quinlan, M. Fowles, C. Welsh, & J. Bivens-Tatum (2008). Cognitive Models of Writing: Writing Proficiency as a Complex Integrated Skill. Princeton, NJ: ETS Research Report No. RR-08-55.

P. Deane, J. Sabatini, G. Feng, J. Sparks, Y. Song, M. Fowles, T. O'Reilly, K. Jueds, R. Krovetz, & C. Foley (2015) Key Practices in the English Language Arts (ELA): Linking Learning Theory, Assessment, and Instruction. Princeton, New Jersey: ETS Research Report No. RR-15-17

P. Deane, J. Sabatini, & M. Fowles (2012). Rethinking K–12 Writing Assessment to Support Best Instructional Practices. Chapter in C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Editors), International Advances in Writing Research: Cultures, Places, Measures. Parlor Press.

P. Deane & Y. Song (2014) A Case Study in Principled Assessment Design: Designing Assessments to Measure and Support the Development of Argumentative Reading and Writing Skills. Psicología Educativa, Vol. 20, No. 2, pp. 99–108

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 34 2016/08/29 13:59:46

P. Deane, & Y. Song (2015) The Key Practice, Discuss and Debate Ideas: Conceptual Framework, Literature Review, and Provisional Learning Progressions for Argumentation. Princeton, New Jersey: ETS Research Report No. RR-15-33.
P. Deane, F. Williams, V. Weng, & C. S. Trapani (2013) Automated Essay Scoring in Innovative Assessments of Writing from Sources. Journal of Writing Assessment, Vol. 6, No. 1
P. Deane, & M. Zhang (2015) Exploring the Feasibility of Using Writing Process Features to Assess Text Production Skills. ETS Research Report No. RR-15-26
T. O'Reilly, P. Deane, & J. Sabatini (2015) Building and Sharing Knowledge Key Practice: What Do You Know, What Don't You Know, What Did You Learn? Princeton, New Jersey: ETS Research Report No. RR-15-24.
T. O'Reilly & K. M. Sheehan (2009a). Cognitively Based Assessment of, for, and as Learning: A 21st–Century Approach for Assessing Reading Competency. Princeton, NJ: ETS Research Memorandum No. RM-09-04.
T. O'Reilly & K. M. Sheehan (2009). Cognitively Based Assessment of, for, and as Learning: A Framework for Assessing Reading Competency. ETS Research Report No. RR-09-26.
J. Sabatini, T. O'Reilly, L. Halderman, & K. Bruce (2014) Integrating Scenario-Based and Component Reading Skill Measures to Understand the Reading Behavior of Struggling Readers. Learning Disabilities Research and Practice, Vol. 29, No. 1, pp. 36–43.
J. Sabatini, T. O'Reilly, & P. Deane. (2013). Preliminary Reading Literacy Assessment Framework: Foundation and Rationale for Assessment and System Design. ETS Research Report No. RR-13-30.
K. M. Sheehan & T. O'Reilly (2011). The CBAL Reading Assessment: An Approach for Balancing Measurement and Learning Goals. Princeton, NJ: ETS Research Report No. RR-11-21.
K. M. Sheehan & T. O'Reilly (2012) The Case for Scenario-Based Assessments of Reading Competency. Chapter in J. Sabatini, E. Albro, & T. O'Reilly (Editors), Reaching an Understanding: Innovation in How We View Reading Assessment. R&L Education.
J. R. Sparks, & P. Deane (2015) Cognitively Based Assessment of Research and Inquiry Skills: Defining a Key Practice in the English Language Arts. Princeton, New Jersey: ETS Research Report No. RR-15-35
P. W. van Rijn, E. A. Graf, & P. Deane (2014) Empirical Recovery of Argumentation Learning Progressions in Scenario-Based Assessments of English Language Arts. Psicología Educativa, Vol. 20, No. 2, pp. 109–115
M. Zhang, & P. Deane (2015) Process Features in Writing: Internal Structure and Incremental Value Over Product Features. ETS Research Report No. RR-15-27
Zhang, M, Bennet, R., Deane, P., and van Rijn, P. (under review). The Validity of Scenario-Based Assessment: Empirical Results.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 35 2016/08/29 13:59:47

公開シンポジウム 2 :

大学入試改革の概要と課題

9 月 8 日（木） 16:20 – 17:50　シンポジウム会場（B202）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　37　2016/08/29　13:59:47

高大接続システム改革と「大学入学希望者テスト」の導入

鈴木寛
東京大学公共政策大学院教授　兼
慶應義塾大学政策・メディア研究科教授

高大接続システム改革の背景・目的

現在、文部科学省は、高等学校教育、大学入学者選抜、大学教育を一体で改革する高大接続システム改革を構想・準備中である。2016 年 3 月 31 日には、高大接続システム改革会議が最終報告をとりまとめられ、改革の骨格は固まったが、その実施をいかに円滑に行っていくか、精力的な議論が現在行われている。

まず、改革の背景を述べる。報告書の冒頭では、「これからの時代に我が国で学ぶ子供たちは、明治以来の近代教育が支えてきた社会とは質的に異なる社会で生活をし、仕事をしていくことになる。国際的には、グローバル化・多極化の進展、新興国・地域の勃興、産業構造や就業構造の転換、国内では、生産年齢人口の急減、労働生産性の低迷、地方創生への対応等、新たな時代に向けて、国内外に大きな社会変動が起こっている。

このような社会変動の中では、これから我が国や世界がどのような産業構造が形成され、どのような社会が実現されていくか、誰も予見できない。確実に言えるのは、先行き不透明な時代であるからこそ、多様な人々と協力しながら主体性を持って人生を切り開いていく力が重要になるということである。また、知識の量だけでなく、混とんとして状況の中に問題を発見し、答えを生み出し、新たな価値を創造していくための資質や能力が重要になるということである。・・・(中略)・・・・これからの時代に向けた教育改革を進めるにあたり、身に付けるべき力として特に重視すべきは、①十分な知識・技能　②それらを基盤にして答えが一つに定まらない問題に自ら解を見出していく思考力・判断力・表現力等の能力、そして③これらの基になる主体性を持って多様な人々と協働して学ぶ態度である」と述べている。

現に、子供たちの６５％は大学卒業後、今は存在していない職業に就く（キャシー・デビッドソン NY 市立大学教授）、今後１０年から２０年程度で、約 47%の仕事が自動化される可能性が高い（マイケル・オズボーン・オックスフォード大学教授）などの発言が注目されているが、これまでの大量生産・流通・消費文明を支えていた教育は大転換が迫られており、①②③の学力の三要素をいかに高等学校、大学で習得していくかが今回の改革の主眼である。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　36　2016/08/29　13:59:48

こうした議論は我が国だけで行われているわけではない。2016 年 5 月 14・15 日、筆者も議長代行として参加した倉敷 G７教育大臣会合でとりまとめられた倉敷宣言でも、世界は、貧困、若者の失業や社会経済的不平等、最近の国境を超える移民・難民のかってないほどの増大、一部の若者の間での暴力的な過激化・急進化、環境・気候変動問題など地球規模課題など、極めて根深い課題や新たな課題に対峙している。一方で、人工知能や IOE、ロボット、再生医療はじめ科学技術テクノロジーの発展は目覚ましく、それらを活用した社会イノベーションのポテンシャルも極めて高くなっている。こうした時代の大きな変化を踏まえ、これから求められる教育として、予測困難な変化の激しい世界を生きる次世代が、自らの将来を作り出すことができるようにすべきとの認識のもと、与えられた課題に効率的に解答する力にとどまらず、自ら新たな問いを立ててその解決を目指し、他者と協働しながら新たな価値を生み出していくための力を身につけていく必要があるとの認識が共有された。加えて、何を知っているか（知識に関するもの）、知っていることをどう使うか（コンピテンシーに関するもの）、どのように社会・世界とかかわり、よりよい人生を送るか。といった視点を教育実践の基盤としていくことも共有された。

また、OECD においても、教育２０３０プロジェクトが 2015 年から進行中であり、改めて、これからの教育に何が求められるのかを再検討し、そのモデルのづくりを行っている最中だが、同様の問題意識のもと、これから求められるコンピテンシーとして、Knowledge、Skill、とともに Attitude & Value が挙げられている。

高大接続システム改革の概要

高等学校教育改革

高校段階における改革としては、高等学校学習指導要領の改訂を行い、これまで暗記科目と目されてきた社会科に、歴史総合（仮称）、地理総合（仮称）、公共（仮称）を新たに設置し、思考力、判断力、表現力を重視した内容としていくほか、数理探求（仮称）や、情報活用能力を育成する新科目などを設置し、さらに、学力の三要素をバランスよく育んでいくこととしている。

また、学習・指導方法の改善として、アクティブ・ラーニングの視点からの学習・指導方法の改善、教員の養成・採用・研修の見直しを行うこととしている。

さらに、多面的な評価の推進を行うこととしており、学習評価の改善を行うこととしており、多様な学習成果を測定するツールの充実として「高等学校基礎学力テスト（仮称）の導入を行い、基礎学力の定着度合いを把握し、指導の工夫に生かす仕組みを導入することとしている。このテストは、平成 31 年度から 34 年度まで試行実施し、平成 35 年度から新学習指導要領に対応して、本格実施することとしている。

大学入学者選抜改革

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 37 2016/08/29 13:59:48

学力の三要素の育成に向けて、高等学校における指導の在り方を本質的な改善を促し、また、大学教育の質的転換を大きく加速し、改革の好循環をもたらすものとなるよう、個別大学の入学者選抜と大学入学者選抜における共通テストの双方について改革を進めることとしている。

まず、各大学において、アドミッション・ポリシーとカリキュラムポリシーとディプロマ・ポリシーを明確化し、その方針にもとづき、入学希望者に求める能力と評価方法との関係を明確化し、それに基づく入学者選抜を実施するよう改善がなされていく。

共通テストとして、大学入学希望者学力評価テスト（仮称）を導入し、知識・技能はもとより「思考力・判断力・表現力」の十分な評価を行うようにする。同時に、個別大学においては、大学入学希望者学力評価テスト（仮称）に加えて、高校時代の学習・活動歴、調査書、入学希望理由書、学修計画書や面接など多様な評価方法を工夫しつつ、高校生活における活動実績をしっかりと評価することによって、「主体性を持って多様な人々と協働して学ぶ態度」についての評価を重視することとしている。

すでに、国立大学協会は、2015 年 9 月に、国立大学入学定員の 3 割まで AO・推薦枠を広げることを合意した。すでに、東京大学「推薦入試」や京都大学「特色入試」の導入、大阪大学「世界適塾入試」導入・拡大、東北大学「AO 入試」の拡大などが実施ないし決定している。

大学教育改革

卒業認定・学位授与（DP）、教育課程編成・実施（CP）、入学者受け入れ（AP）の三つの方針の一体的な策定・公表の制度化が平成２９年４月実施される。すでにその策定・運用に関する「参考指針」が平成２８年３月に作成されており、各大学において、育成を目指す人材像や具体的な教育活動を明確化することとしている。三つの方針に基づく大学教育の質的転換促進や、内部質保証を重視した評価に改善することとしている。大学においては、大学教育を受けるに足る入学者の選抜をしっかり行ったうえで、多様な入学者とそれに合わせた教育プログラムへの転換を図っていく。

記述式の導入と実現への課題

現在の大学入試センター試験に代えて、大学入学希望者学力評価テストを平成 32 年度から実施し、平成 36 年度から、新学習指導要領に対応して本格実施することとしている。新テストは、（選抜という観点もさることながら）、大学教育を受けるための必要な能力について把握することを主たる目的とし、知識・技能を十分有しているかの評価も行いつつ「思考力・判断力・表現力」を中心に評価することとしている。

マークシート式問題については、正解が一つに限られない問題、正解を選択させるのでは

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 38 2016/08/29 13:59:49

なく、数値や記号等を直接マークさせる問題などの、思考力・判断力・表現力を重視した作問へ改善していく。また、現在よりも多くの情報（問ごとの解答状況も合わせて提供するなど）を各大学に提供する。

今回の新テストの最大の特徴は、英語四技能と並んで記述式の導入にある。平成 32 年度から条件付き記述、平成 36 年度から記述式の導入を行うこととしている。この背景には、スマートフォンの使用時間が増え（男子高校生３．８時間、女子高校生５．５時間　デジタルアーツ株「第８回未成年者と保護者のスマートフォンやネット利活用における意識調査より）、さらに、高校生の読書力が低く（高校生の不読者比率が４８．７％、高校生の新聞不読率が 4 割、全国学校図書館協議会・毎日新聞社「第６０回学校読書調査２０１４」より）、国立大学でも、その受験者の約 6 割が記述力を入学段階で評価されていないなどの理由から、文章を書く基本的なスキルの欠如した学生が多くなっていることが挙げられる（大学側の 8 割、高校側の 7 割が指摘ベネッセ調べ）。

記述力は思考力・判断力・表現力の育成に直結するものでもあると考えられるし、今後、どのような分野においても主体性を持って活動するために重要となる、複数の情報を統合し構造化して新しい考えをまとめる思考・判断の能力や、その過程を表現する能力の評価のため、記述式を導入していくこととしている。英・仏・独・などの大学入学資格試験でも記述式は多く、また、記述式導入により高等学校教育を生徒の能動的な学習をより重視したものに改善されることを狙いとしている。

記述式問題が増えることについては、文部科学省の委託調査によれば、調査した５００名の大学一年生の約 6 割が賛成で、その主な理由として、①記述式の方が能力を正確に評価してもらえる、②大学での勉強や研究を進める上で必要な力が身につきそう、③表現できる力が身に付きそう　となっている。

一方で、約 7 割が正確な採点をしてくれるかどうか不安だとも回答しているし、約５５万人強の受験生が一斉に受験することが予想され、大量の記述式の採点をどのくらい期間でおこなえるかが目下の課題となっている。採点期間を短縮するための方策として、出題形式の工夫による採点工程の工夫（キーワード、字数制限などの客観的な条件を問題に設定する）と、採点の工夫を助ける技術の導入（クラスタリングの改善により、答案の分類・並び替えの精度を向上し、採点工程を工夫する。OCR 機能を向上させるための研究開発を行うなど）などが考えられるが、その両面からのさらなる検討と技術開発と安定性の実証などが急務となっている。

平成 36 年度の本格的記述式導入に向け、採点期間と採点のバラつきなどの課題を解決していくために、いかなる体制で、いかなるプロセスで、いかなる技術を磨き、いかなる準備をしていくことが必要なのか、テスト学会会員の知的貢献が強く期待されている。

（鈴木寛：suzukan@pp.u-tokyo.ac.jp）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　39　2016/08/29　13:59:49

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 40 2016/08/29 13:59:50

一般セッション１：大学入試

９月９日（金）　9:00 – 10:20　第１会場（B101）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　41　2016/08/29　13:59:50

国立大学における個別学力試験の解答形式に関する研究(1)

○宮本友弘・庄司強・田中光晴・石上正敏・倉元直樹
(東北大学)

1. 問題と目的

センター試験に替わる新テストには記述式問題の導入が検討されている。その理由として、高大接続システム改革会議(2016)によれば、思考力・判断力・表現力の発揮が期待できるという記述式問題のメリットが挙げられている。また、“現状において、大学によっては……知識に偏重した選択式問題が中心で記述式問題を実施していない場合もある”との指摘もなされている。

こうした大学入試問題の解答形式に関する研究としては、古くは、大学入試センターが、国語、英語、数学、社会、理科のそれぞれにおいて、多肢選択式あるいはマークシートの問題を記述式の問題に変換し、両者の識別力や困難度、測定される能力の比較を行っている(鈴木・山田・池田・赤木, 1991 ; 石塚・前川, 1991 ; 豊田・山村・藤芳, 1991 ; 石塚・平・清水, 1992 ; 山田・鈴木・豊田・清水, 1992)。その後も、同様の解答形式の比較研究が、とくに数学分野において精力的になされている(村上他, 2007 ; 村上他, 2008 ; 安野他, 2013)。

しかしながら、そもそも大学入試問題の解答形式の実態がどのようなものであるかを調べた研究はほとんどない。深沢(1999)が「全国大学入試問題正解・英語〈国公立大編〉」(旺文社)を資料にして、公表されている65の国立大学の、前期日程の1996年度と1997年度の英語の問題形式の分布を比較するにとどまっている。

そこで、本プロジェクト研究では、国立大学の一般入試個別学力試験問題を収集して解答形式を分析し、現在、国立大学の一般入試が抱えている問題を析出することを目的とした。ここではその第1段階として、国語、数学、英語を検討する。

2. 方法

分析対象　大学院大学 4 校を除く国立大学 82 校のうち、「2016年版　大学入試シリーズ」(教学社)に、国語、数学、英語のいずれかの科目の2015年度一般入試個別学力試験問題(前期日程、後期日程)が収載された75校と、大学に直接依頼して問題を送ってもらった1校、計76校が分析対象となった。そのうち、前期日程は全76校が実施し、後期日程は42校のみが実施していた。各科目の大学数、問題数の詳細は、表1に示す通りである。

分析方法　『テスト・スタンダード』(日本テスト学会, 2007)及び高大接続システム会議「最終報告」等に基づき、表2に示すカテゴリーを作成し、解答の最小単位（枝問）を分類した。

表1　分析対象となった科目別の大学数と問題数

		国語	数学(文)	数学(理)	英語	英語(L)
前期日程	大学数	52	48	69	71	10
	問題数	1,891	571	1,553	2,428	124
後期日程	大学数	3	5	35	20	1
	問題数	97	48	591	395	5
合計	大学数	52	48	70	71	10
	問題数	1,988	619	2,144	2,823	129

注）文：文系　理：理系　L：リスニング

表2　解答形式の分類カテゴリー

	カテゴリー
客観式	A1:○×式　A2:多肢選択式　A3:複数選択式 A4:組み合わせ式　A5:並べ替え式 A6:その他　A7:分類不能
記述式	B1:穴埋め式　B2:短答式　B3:短文　B4:長文 B5:英文和訳　B6:和文英訳　B7:英文日本語要約　B8:英作文　B9:小論文　B10:数式 B11:図・絵等
その他	C1:コンピュータ式　C2:その他

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　40　2016/08/29　13:59:50

3. 結果と考察

ここでは、大学を込みにして科目数について分析する。各科目の客観式と記述式の割合をみると（表 3）、前期・後期日程ともに国語、数学（文）、数学（理）では、ほとんどが記述式であった。英語と英語（L）（前期日程のみ）は客観式と記述式がほぼ同程度であった。英語（L）の後期日程はわずか 5 問であったことを踏まえると、いずれの教科も前期日程と後期日程はほぼ同様の傾向にあると考えられる。

前期・後期日程を込みにして、客観式の内訳をみると（表 4）、全教科ともに多肢選択肢がほとんどを占めた。一方、記述式の内訳をみると（表 5）、数学（文）、数学（理）はともに数式（数式の展開など、数式で解答する記述式問題）がほとんどを占めた。国語と英語（L）は短答式、長文（40 文字程度以上）、短文（40 文字程度以下）が相対的に多かった。英語は数式、図・絵等以外のすべてのカテゴリーに分布した。

以上から、国立大学の一般入試個別試験の解答形式は、国語、数学では記述式が中心であり、英語に一部客観式が含まれていることが示唆された。

表 3　各科目の客観式と記述式の割合（%）

		国語	数学（文）	数学（理）	英語	英語（L）
前期日程	*N*	1,891	571	1,553	2,428	124
	客観	12.0		0.1	47.2	50.0
	記述	88.0	100.0	99.9	52.8	50.0
後期日程	*N*	97	48	591	395	5
	客観	4.1			50.1	
	記述	95.9	100.0	100.0	49.9	100.0

表 4　客観式の下位分類の割合（%）

	国語	数学（理）	英語	英語（L）
N	230	1	1,344	62
○×式	6.1		7.4	
多肢選択式	88.3	100.0	82.6	87.1
複数選択式	4.3		1.3	
並べ替え式	0.9		6.3	6.5
その他(客観式)	0.4		2.3	6.5

表 5　記述式の下位分類の割合（%）

	国語	数学（文）	数学（理）	英語	英語（L）
N	1,758	619	2,143	1,479	67
穴埋め式	0.1			8.8	
短答式	41.9		1.3	7.9	32.8
短文	25.1	0.2	0.0	11.0	34.3
長文	32.3			20.1	16.4
英文和訳				21.4	3.0
和文英訳				9.9	3.0
英文日本語要約				5.9	4.5
英作文				14.4	6.0
小論文	0.6			0.5	
数式		96.9	96.6		
図・絵等	0.1	2.9	2.1		

引用文献

深沢清治（1999）．大学入試問題の比較研究 ： 国立大学入試問題形式の推移について（1996-97） 広島外国語教育研究，2，93-100.

石塚智一・前川眞一（1991）．英語の試験問題の出題形式に関する比較研究　研究紀要（独立行政法人大学入試センター），20，47-72.

石塚智一・平直樹・清水留三郎（1992）．社会の試験問題の出題形式に関する比較研究 研究紀要（独立行政法人大学入試センター），21，1-33.

高大接続システム改革会議（2016）．最終報告 http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2016/06/02/1369232_01_2.pdf （2016 年 8 月 3 日）

村上隆・三宅正武・藤村宣之（2007）．マークシート形式と記述形式による数学の「学力」 大学入試研究ジャーナル，17，175-182.

村上隆・三宅正武・藤村宣之（2008）．マークシート形式と記述形式による数学の「学力」（2） 大学入試研究ジャーナル，18，163-170.

日本テスト学会（編）（2007）．テスト・スタンダード－日本のテストの将来に向けて　金子書房

鈴木規夫・山田文康・池田輝政・赤木愛知（1991）．国語の試験問題の出題形式に関する比較研究　研究紀要（独立行政法人大学入試センター），20，1-45.

豊田秀樹・山村滋・藤芳衛（1991）．数学の試験問題の出題形式と設問過程に関する比較研究　研究紀要（独立行政法人大学入試センター），20，73-91.

山田文康・鈴木規夫・豊田秀樹・清水留三郎（1992）．理科の試験問題の出題形式に関する比較研究　研究紀要（独立行政法人大学入試センター），21，35-57.

安野史子・浪川幸彦・森田康夫・三宅正武・西辻正副・倉元直樹…村上隆（2013）．解答形式とパフォーマンスに関する実証的研究　大学入試研究ジャーナル，23，143-150.

連絡先：宮本友弘（30000188） tomohiro@tohoku.ac.jp

謝辞：本研究は JSPS 科研費 16H02051 の助成を受けた。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　41　2016/08/29　13:59:51

国立大学における個別学力試験の解答形式に関する研究(2)

○庄司強・宮本友弘・田中光晴・石上正敏・倉元直樹
(東北大学)

1. 問題と目的

本研究プロジェクトの目的は、国立大学の一般入試個別学力試験問題を収集して解答形式を分析することにより、現在、国立大学の一般入試が抱えている問題を析出することである。解答形式にかかわる入試改革の制度設計の根本的問題点が国立大学にも存在するならば、それを特定して改善することが個別大学の責務と考えられている状況だからである。そこで、研究(2)では、大学を単位としての分析を行う。すなわち、大学の入試問題に対する考え方や大学の特性が解答形式の分布に現れているか否かを検討することを目的とする。

高校側は、大学入試問題に大学からのメッセージや大学の特性を読み取ろうとする傾向がある。例えば、高梨(2011)は、東大は全大学の模範的な立ち位置が要求されるので分野に偏りなく出題されるが、京都大学は東大と同じ趣旨では存在理由に説得力がなくなるので少し違う出題の方向性が存在するはずである、と洞察している。こうした解釈のもと、大学の特性に応じた進路指導や受験対策がなされる。その意味において、大学入試は教育の一貫であり、大学入試問題は大切な教材である(倉元, 2001)。また、大学側からすれば、大学が求める学生像をもっとも体現しているのが入試問題であるという考え方がある(中畝, 2011)。この見方は高校側の見解と軌を一にしている。一方、大学入試問題の内容に本質的な教育的意義を感じていない大学もある。そのような大学では作題・採点に対する負担感が強い。例えば、過去の入試問題を再利用することによって、負担軽減を図ることを目的にして、岐阜大学が主幹となり共同提案された「入試過去問活用宣言」(以下、活用宣言)がある。現在、国立大学では28校が参加している。日本テスト学会(2007)が編纂した『テスト・スタンダード』によれば、公開された問題の再利用は、とくにハイステークスなテストでは項目の性質を根本的に損なうもので避けるべきとされている。

文部科学省(2016)は、「平成28年度国立大学法人運営費交付金における3つの重点支援枠」に基づき、大学を次の3つに分類した。①主として、地域に貢献する取組とともに、専門分野の特性に配慮しつつ、強み・特色のある分野で世界・全国的な教育研究を推進する取組を中核とする国立大学55校(以下、地域貢献)、②主として、専門分野の特性に配慮しつつ強み・特色のある分野で地域というより世界・全国的な教育研究を推進する取組を中核とする国立大学15校(以下、教育研究)、③主として、卓越した成果を創出している海外大学と伍して、全学的に卓越した教育研究、社会実装を推進する取組を中核とする国立大学16校(以下、卓越教育研究)である。これらの分類には、人的・財政的資源の配分を目的に応じて傾斜させるしくみが付随しており、結果的に入試に対する負担感も大学の立ち位置によって異なっていることが想像できる。

2. 方法

分析対象　研究(1)と同じく国立大学76校が分析対象となった。表1に、活用宣言への参加・不参加及び3分類ごとの各科目の個別学力試験での実施状況(大学数)を示した。

分析方法　研究(1)と同じカテゴリーを使って、解答の最小単位(枝問)を分類した。各カテゴリーの度数が1以上であれば、当該大学で実施とカウントとした。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　42　2016/08/29　13:59:52

表１　科目別個別学力試験の実施大学数

	N	国語	数学（文）	数学（理）	英語	英語（L）
〈活用宣言〉						
参加	27	19	16	27	25	4
不参加	49	33	32	43	46	6
〈3 分類〉						
地域貢献	51	35	32	49	48	5
教育研究	10	4	3	7	8	2
卓越教育研究	15	13	13	14	15	3

3. 結果と考察

全体では（表 2）、いずれの科目においても、個別試験を実施した大学は、前期・後期日程にかかわらず記述式を実施した。また、国語と英語では客観式を実施した大学が半数以上みられた。客観式（表 3）では、いずれの科目も多肢選択式を実施した大学がもっとも多かったが、国語、英語では、一定数の大学が他の形式も実施した。記述式（表 4）では、国語は短答式、短文、長文、数学（文、理）は数式、図絵等に集中したが、英語は数式、図・絵等以外の各形式に一定数ずつ分布した。活用宣言参加別（表 5）、3 分類別（表 6）の客観式と記述式の実施状況は全体と同様の傾向にあった。

以上から、国立大学の場合、高大接続システム改革会議（2016）による“知識に偏重した選択式問題が中心で記述式問題を実施していない場合もある”という指摘には当たらない。

表２　客観・記述の実施大学数

		国語	数学（文）	数学（理）	英語	英語（L）
前期日程	N	52	48	69	71	10
	客観	32		1	57	4
	記述	52	48	69	71	8
後期日程	N	3	5	35	20	1
	客観	2			12	
	記述	3	5	35	19	1

表３　客観式の下位分類別の実施大学数

	国語	数学（理）	英語	英語（L）
N	33	1	57	4
○×式	2		12	
多肢選択式	32	1	53	3
複数選択式	8		7	
並べ替え式	2		20	1
その他（客観式）	1		4	1

表４　記述式の下位分類別の実施大学数

	国語	数学（文）	数学（理）	英語	英語（L）
N	52	48	70	71	8
穴埋め式	1			12	
短答式	50		1	26	2
短文	49	1	1	42	5
長文	52			56	4
英文和訳				60	1
和文英訳				32	1
英文日本語要約				31	2
英作文				56	4
小論文	5			5	
数式		48	70		
図・絵等	1	14	30		

表５　活用宣言参加別の客観・記述の実施大学数

		国語	数学（文）	数学（理）	英語	英語（L）
参加	N	19	16	27	25	4
	客観	13		1	22	2
	記述	19	16	27	25	2
不参加	N	33	32	43	46	6
	客観	20			35	2
	記述	33	32	43	46	6

表６　3 分類別の客観・記述の実施大学数

		国語	数学（文）	数学（理）	英語	英語（L）
地域貢献	N	35	32	49	48	5
	客観	25		1	39	1
	記述	35	32	49	48	4
教育研究	N	4	4	7	8	2
	客観	1			4	2
	記述	4	4	7	8	1
卓越した教育研究	N	13	13	14	15	3
	客観	7			14	1
	記述	13	13	14	15	3

引用文献

倉元直樹(2001). 日本の大学入試に何が欠けているか　西岡和雄（編）　教育が危ない 2　ゆとりを奪った「ゆとり教育」(pp.164－194)　日本経済新聞社

文部科学省(2016). 高等教育局主要事項－平成28年度概算要求－ http://www.mext.go.jp/component/b_menu/other/__icsFiles/afieldfile/2015/08/27/1361291_1.pdf （2016 年 8 月 4 日）

中畝菜穂子(2011). 入試問題を用いた高大連携―新潟大学ヴァーチャル入試体験　東北大学高等教育開発推進センター(編)　高大接続パラダイム転換と再構築(pp.65－75)　東北大学出版会

高梨誠之(2011). メッセージとしての大学入試問題　東北大学高等教育開発推進センター(編)　高大接続パラダイム転換と再構築(pp.183－198)　東北大学出版会

連絡先：宮本友弘(30000188) tomohiro@tohoku.ac.jp

謝辞：本研究は JSPS 科研費 16H02051 の助成を受けた。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　43　2016/08/29　13:59:52

センター試験の成績未利用行動の分析 (1)
—理科②の成績未利用者の分析—

○橋本 貴充，内田 照久，鈴木 規夫
独立行政法人 大学入試センター

1 はじめに

大学入試センター試験（以下，「センター試験」とする。）は毎年約 50 万人が受験する全国規模の試験であるが，その受験者像は一様ではない。内田 他（2016）は，センター試験の成績を私立大学への出願のみに利用する受験者，および出願に全く利用しない受験者を「新参入層」とし，センター試験の成績を国公立大学の出願に利用する旧来の受験者層（中核層）と得点の分布に異なる特徴があることを明らかにした。センター試験を旧来と異なる方法で利用する受験者は，理科，地理歴史，公民など，科目選択の余地のある教科で，何らかの影響を及ぼしていると考えられる。そこで，センター試験の特定の科目を，解答しても出願に利用しない受験者は，成績を出願に利用する一般の受験者と，どのように異なるのか，理科②の成績を対象として分析する。

2 方法

本報告では，理科②の物理，化学，生物，地学の 4 科目を分析対象とした。分析を行った年度のセンター試験は，学習指導要領改定に伴い，旧教育課程履修者に対する経過措置科目が出題されたため，経過措置科目である旧課程の物理 I，化学 I，生物 I，地学 I も分析対象とした。ただし，追再試験の成績は分析から除外した。なお，この年度のセンター試験の理科②では得点調整が行われたが，本報告では調整前の成績を分析した。

理科②の各科目の成績について，(1)当該科目の成績を少なくとも 1 つの大学・短大への出願に利用（以下，「利用」と略記），(2)他の科目を用いて大学・短大に出願したが当該科目の成績は出願に未利用（以下，「未利用」と略記），(3)全ての科目を出願に未利用（つまり大学・短大への出願を全く行っていない，以下「未出願」と略記）の 3 つに分類し，科目ごとに比較した。

3 結果

3.1 各科目の未利用・未出願数およびその比率

分析対象 8 科目の未利用・未出願者の数は図 1 のとおりである。科目ごとに受験者の総数が異なるため，受験者数で除した未利用・未出願率は図 2 のようになった。未利用・未出願率の合計が，物理では 14.0%であるのに対し，生物では 26.4%に達した。生物および地学では，未利用と未出願の合計が 2 割を超えており，この 2 科目の未利用率は特に高いことがわかる。地学と旧地学 I に関しては，大学などに出願したものの地学（および旧地学 I）の成績を利用しなかった「未

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 44 2016/08/29 13:59:53

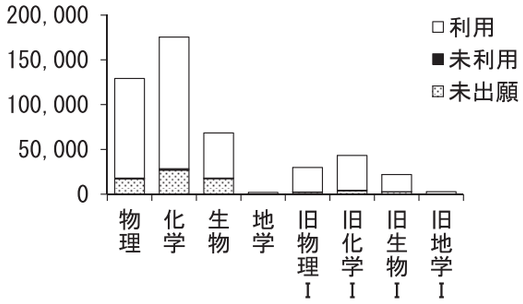


図 1: 各科目の未利用・未出願数

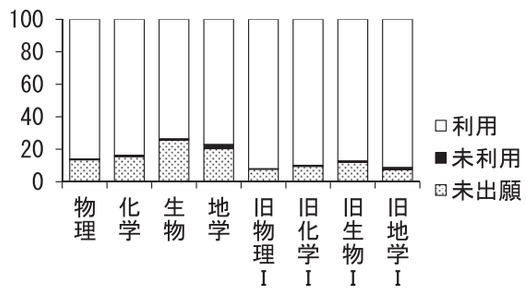


図 2: 各科目の未利用・未出願率(%)

利用」の比率が，他の科目に比べて高いという特徴も見られた。

3.2　解答科目順別の未利用・未出願率

各科目の成績を，1 科目受験，2 科目受験の第 1 解答科目，2 科目受験の第 2 解答科目に分けた場合の未利用・未出願の比率は，図 3 および図 4 のようになった。図 3 より，新課程科目のうち，地学を除く 3 科目では 1 科目受験の場合に未利用・未出願率が最も高く，地学では逆に 1 科目受験の場合には未利用率が最も低い結果となった。大学などに出願したが当該科目の成績を利用しなかった「未利用」の割合に着目すると，物理および化学では第 2 解答科目として解答した場合に最も高くなったのに対し，生物および地学では 1 科目受験の場合に最も高くなるという結果となった。

図 4 より，旧課程科目では未利用・未出願の比率は全体として低い傾向にあるものの，旧地学Ⅰを 1 科目受験した場合が特に低かった。つまり，理科②で旧地学Ⅰを 1 科目だけ受験した者のほとんどは，旧地学Ⅰを大学受験のために解答したといえる。また，旧地学Ⅰを第 2 解答科目として解答した場合の「未利用」が特に多いという結果も見られた。

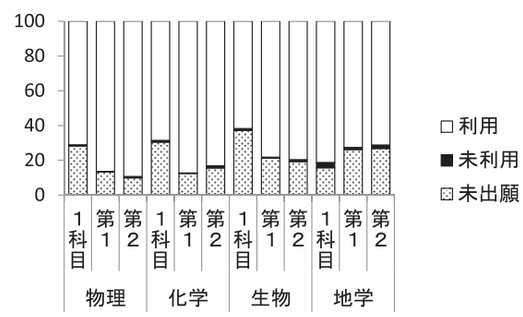


図 3: 解答科目順別の未利用・未出願率(%)
（新課程科目）

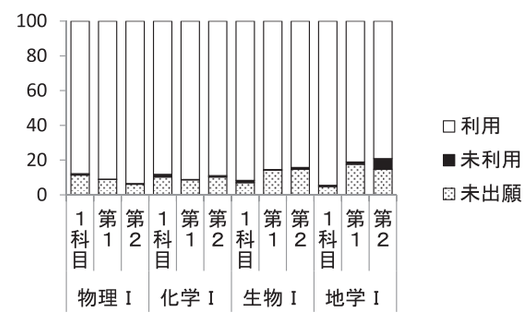


図 4: 解答科目順別の未利用・未出願率(%)
（旧課程科目）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　45　2016/08/29　13:59:54

3.3　得点の比較

それぞれの科目の成績を出願に利用した場合としなかった場合の得点の度数分布は，図 5 のようになった。いずれの科目でも未利用・未出願の受験者の得点は利用者に比べて全体的に低かった。未利用・未出願者が相対的に少ない旧物理 I では影響が小さいのに対し，未利用・未出願者が多い生物はその影響を大きく受けていることがわかった。

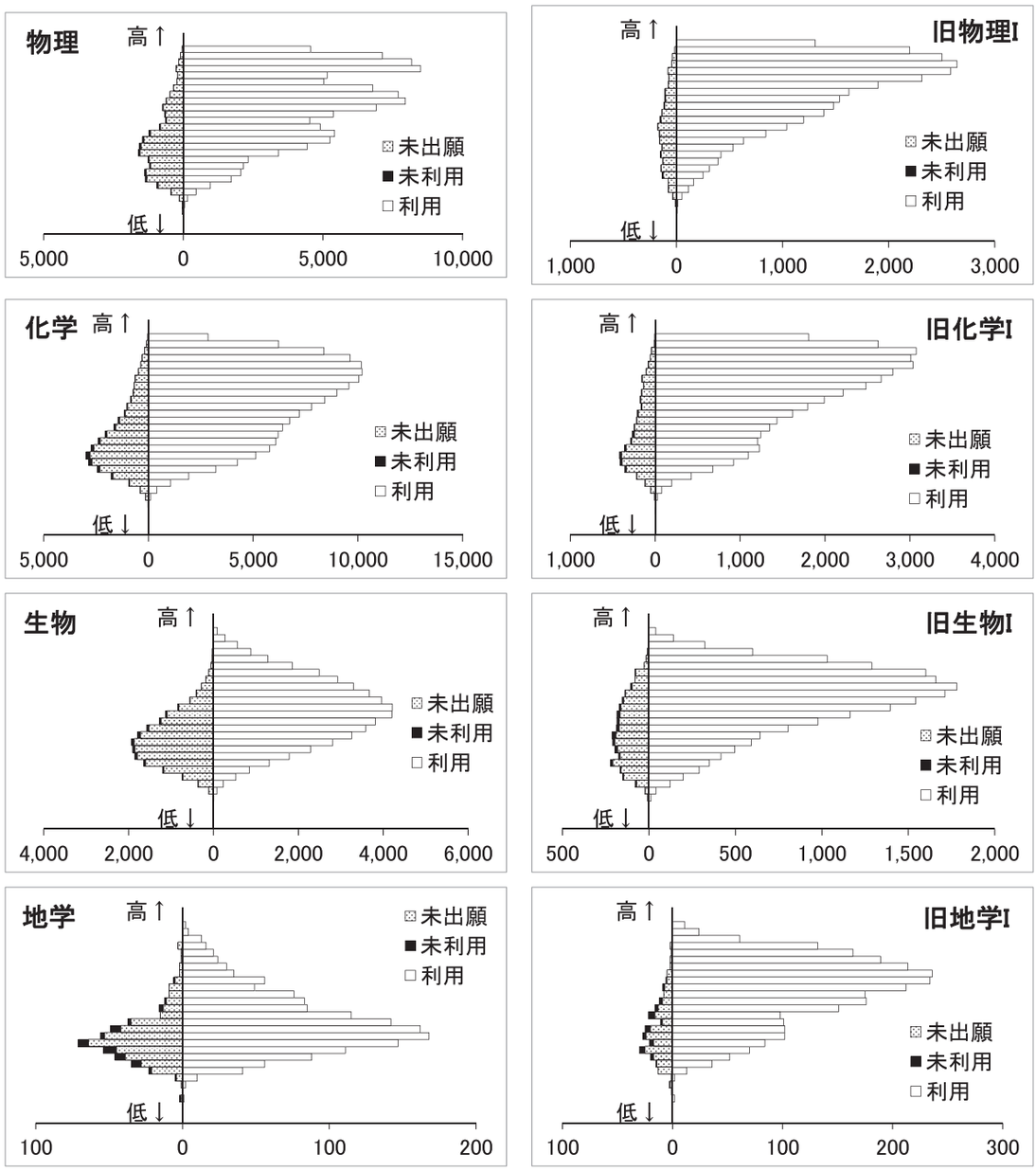


図 5: 各科目の利用状況別の得点の度数分布

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　46　2016/08/29　13:59:54

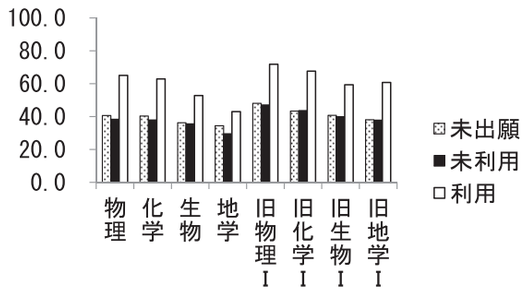


図 6: 各科目の利用状況別の得点の平均点

平均点は図 6 のとおりである。出願に当該科目を利用しなかった「未利用」の平均点に注目すると，「利用」より「未出願」に近い。当該科目を出願に利用しないという行動は，他の科目を出願に利用しているか否かにかかわらず，同様に扱ってよいと考えられる。

4　考察

平成 27 年度の理科②の科目について，成績の利用状況を科目間で比較したところ，旧課程科目では未利用率が低く，新課程科目の地学と生物で高いことがわかった。解答科目順で分けて検討した結果，未利用率は当該科目を 1 科目のみで受験した場合に高いことがわかった。成績を利用しなかった受験者の中には，全ての成績を利用しない「未出願」のほかに，当該科目のみ出願に利用しない「未利用」の者が少数ながら存在したが，平均点は「未出願」に近いものであった。新卒者の 2 割が，センター試験を受験しながら，その成績を全く出願に利用していない（内田・橋本・鈴木, 2014）という現状もあるため，成績を出願に利用しない受験者の行動が受験者全体にどのような影響をどの程度及ぼすのか，今後詳細な検討が必要である。

引用文献

内田照久，橋本貴充，鈴木規夫 (2014). 18 歳人口減少期のセンター試験の出願状況の年次推移と地域特性－志願者の 2 層構造化と出願行動の地域特徴－．日本テスト学会誌, **10**, 47－68.

内田照久，中村裕行，橋本貴充，鈴木規夫，荒井克弘 (2016). センター試験の受験目的の多様化と学力分布の層別特性．大学入試センター研究紀要, **45**, 1－11.

(hashimot@rd.dnc.ac.jp)

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　47　2016/08/29　13:59:54

大学・大学院入試に向けた英語スピーキングテストの尺度化事例

—受験者特性と評価者属性を考慮したモデルによる検討—

○光永　悠彦*　　神澤　克徳**　　坪田　康**　　羽藤　由美**
*島根大学　　**京都工芸繊維大学

1　背景と目的

高大接続に対する大学側の取り組みとして入試方法の改革の必要性が叫ばれている．またグローバル化の潮流によって，大学入学段階で実践的英語能力が学生に求められる．しかし，大学入試においては「リーディング」や「リスニング」という 2 技能を測定するにとどまっている場合が多い．また大学院課程では，学術的な情報交換やプレゼンテーションなどで英語によるスピーキングの技能が求められることが多い．そのため，大学入試や大学院入試へスピーキング能力を問うテストを導入することが意義深い．

スピーキング能力の測定は，パフォーマンステストとして実施される．すなわち受験者にいくつかの設問項目を提示し，その反応を解答として収集する．その後，評価者（rater）が受験者ごとに解答を評価（rating）し，その結果を記録した上で，評価結果を用いて受験者の成績（スコア）を算出する．このようなパフォーマンステストのデータは，一般に多相の（many-facet）データといわれる．具体的には，受験者ごとの各項目に対する反応を複数の評価者によって評価した結果である．この場合，受験者・項目・評価者の 3 相データということができる．これに加え，項目反応理論（item response theory, IRT）に代表される確率モデルを多相データに当てはめることで，各層におけるパラメタ（受験者の能力値，項目の識別力及び困難度など，評価者の評価の辛さ（rater severity）など）を推定することができる．

本論では，大学生の英語スピーキング能力を測定するためのパフォーマンステストデータを用いて，多相データに対する分析を行う．本論で使用するのは，京都工芸繊維大学において 1 年次生全員を対象として行った英語スピーキングテストで得られた多相データである．評価観点として 2 つの尺度（Task Achievement 及び Task Delivery，以下 TA，TD と略記）が測定されており，TA は各項目が要求する課題を英語でどの程度達成できたかの指標，TD はその際に言語をどの程度効果的かつ流暢に用いたかの指標である（Hato, Kanzawa, Mitsunaga, Tsubota and Underhill, 2016）．

本論では，項目パラメタを検討することに加え，以下の 2 点を明らかにすることを目的とした．

評価者の属性の違い　本論で用いられた評価者は日本の大学で英語を教える英語のネイティブスピーカー（NS）と，フィリピンで英語を教える「リンガフランカとしての英語（English as a Lingua Franca）」話者（ELF）であった．両者の評価者特性について検討する．

受験者特性との関連性　各受験者のリーディングとリスニングに関する属性が，スピーキングテ

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　48　　2016/08/29　13:59:55

ストのスコアにどれだけ影響を及ぼすかを推定する. 受験者特性尺度として TOEIC IP を用いる.

2　方法

2.1 データ

2015 年 12 月に京都工芸繊維大学において実施した英語スピーキングテストのデータを用いた.

デザイン　異なる項目セットを 2 種類（Ver.1 と Ver.2）用意し，受験者を 2 群に分けて割り付けた. また, 両方の項目セットを同時に受験するモニター受験者を用意した（共通受験者デザイン）.

受験者　受験者は京都工芸繊維大学の 1 年次生で，共通受験者は 2 年次以上とした．1 年次生のうち Ver.1 を受験した群は N=302，Ver.2 を受験した群は N=272 であった．また共通受験者は N=69 であった．ただし，Ver.1 及び Ver.2 でそれぞれ 1 名，採点不能の解答が項目全体の半数以上となったため，除外された.

項目　それぞれの項目セットについて，英語により課題を提示し，英語により課題への解答をさせる形式の 9 項目とした．全体で 18 項目について項目特性が推定されることになる.

評価者，評価項目及び評価方法　各受験者の各項目に対する解答を，事前に評価者トレーニングを受けた NS 及び ELF が併行して評価した．評価観点は TA 及び TD であった．評価データ上では評価者の区別はせず，その評価が NS か ELF かのみを区別し，評価者属性とした．いずれも 6 件法だったが，最下カテゴリが少数であったため，最下カテゴリを一つ上のカテゴリと併合して 5 件法の評価データとした.

受験者特性　共通受験者以外の受験者に対して，TOEIC IP をスピーキングテストと同日に課し，リーディングとリスニングのそれぞれのスコアを記録した．ただし，Ver.1 受験者で 4 名，Ver.2 受験者で 9 名が未受験であった.

2.2 モデル

以下 2 種類のモデルを適用した.

MFRM　Linacre(1994)の Many facet Rasch model（以下「MFRM」と表記）を用いて受験者のスピーキング能力値$\boldsymbol{\theta}_{MFRM} = \{\theta_1, \dots, \theta_i, \dots \theta_N\}'$を推定した上で，受験者の特性が$\boldsymbol{\theta}_{MFRM}$の説明にどれだけ寄与しているかを探るため，受験者の TOEIC IP のスコアを説明変数，$\boldsymbol{\theta}_{MFRM}$を目的変数とした重回帰分析を行い，標準偏回帰係数$b_s(s = 1,2)$を受験者特性の影響力指標とした（TOEIC IP 未受験者は重回帰分析から除外した）．本論で用いたモデルを式 1 に示した.

$$P_{rj_k}(\theta_i) = \frac{\exp\left(a_j(\theta_i - R_r - D_{j_k})\right)}{1 + \exp\left(a_j(\theta_i - R_r - D_{j_k})\right)} \tag{1}$$

ここで$D_{j_k} = d_j + \tau_{j_k}$とおき Partial Credit Model と仮定した. θ_i（$= \boldsymbol{\theta}_{MFRM}$）は受験者$i(= 1,2, \dots, N)$の能力値，$a_j$は項目$j$の識別力，$R_r$は評価者$r(= 1,2, \dots, R)$の評価の辛さ，$d_j$は項目$j(= 1,2, \dots, J)$の困難度，$\tau_{j_k}$は項目$j$における評価カテゴリ$k(= 1,2, \dots, K)$の閾値を表す.

宇佐美(2010)のモデル　受験者の特性と評価者の特性を同時に考慮した Generalized Partial Credit Model の拡張として提案された宇佐美(2010)のモデルを用いて，評価データと TOEIC IP のデータを同時分析した．推定手法は MCMC（Markov Chain Monte Carlo）法とした.

$$P_{rj_k}(\theta_i) = \frac{\exp(\alpha_{jr}(k(\theta_i - \delta_{jr}) - \sum_{m=0}^{k} \tau_{rj_m}))}{\sum_{l=0}^{K-1} \exp(\alpha_{jr}(l(\theta_i - \delta_{jr}) - \sum_{m=0}^{l} \tau_{rj_m}))} \tag{2}$$

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　49　2016/08/29　13:59:56

ここで$\alpha_{jr}=\alpha_j\alpha_r, \delta_{jr}=\delta_j\delta_r, \tau_{rj_m}=\tau_r\tau_{j_m}$とし，$\alpha_j$は項目$j$における識別力，$\alpha_r$は評価者$r$における評価点決定の際の識別力の高さ（採点者の一貫性），δ_jは項目jにおける困難度，δ_rは評価者rにおける評価の辛さ，τ_{j_m}は項目jにおける評価カテゴリmの閾値，τ_rは評価者rが下した評価点の平均的散らばりの程度を表す（ただし$\tau_{rj_0}=0$及び$\sum_k^K \tau_{rj_k}=0$）．またS項目の受験者特性をそれぞれ標準化したデータz_{is}を用いて式 3 のようにθ_iを再母数化する．

$$\theta_i=\sum_{s=1}^{S}\beta_s z_{is}+\theta_i^* \tag{3}$$

ここでβ_sは受験者特性の影響力指標（標準偏回帰係数），θ_i^*は説明変数から説明されない残差を表す．ただしパラメタの識別性より「全 rater でα_r及びτ_rの積は 1，δ_rの和は 0」「全受験者についてθ_iの和は 0，二乗和は 1，θ_i^*の和は 0，$z_{is}+\theta_i^*$の和は 0」という制約をおく．

2.3 分析ソフトウェア及び推定の詳細

MFRM のパラメタ推定には Facets を用いた．また宇佐美(2010)の推定は R を用いた．MCMC 法を用いたため，推定の安定性をチェックするため 3 chains にわたって推定を行い，Gelman-Rubin の$\hat{R}$を各パラメタについて求めた．

3　パラメタ推定結果

MCMC 法による推定は，$\hat{R}$が 1.2 を下回っており，TA, TD ともに収束しているとみなされた．

図 1 に TA について，MFRM のa_jと宇佐美(2010)のα_j及びd_jとδ_jを項目ごとに比較した結果を示した．また図 2 に TD について比較した結果を示した．また表 1 にR_rとδ_r及びb_sとβ_sを示した．項目パラメタについて，推定手法間で似たような傾向となる項目とそうでない項目が混在した．

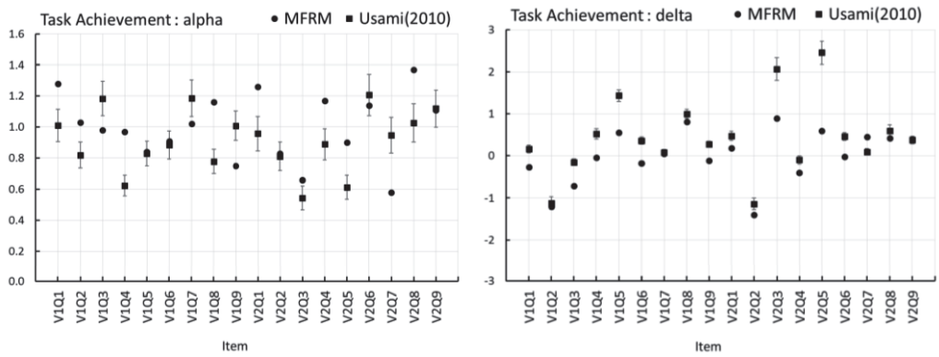


図 1：　TA 尺度の項目パラメタの比較結果．左は識別力，右は困難度．

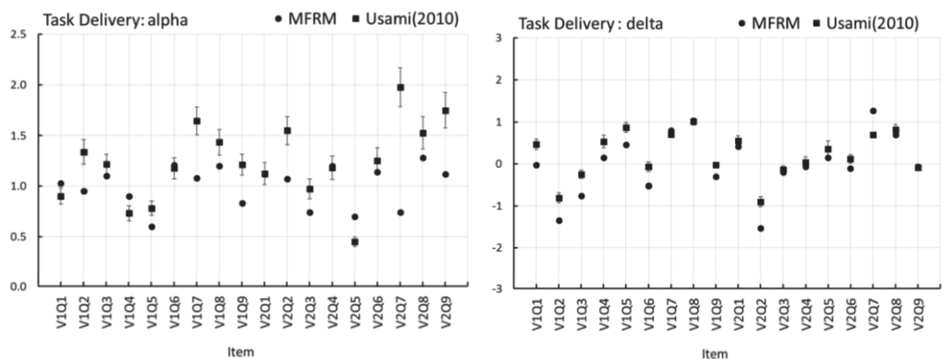


図 2：　TD 尺度の項目パラメタの比較結果．左は識別力，右は困難度．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　50　2016/08/29　13:59:56

表 1：　評価の辛さ及び受験者特性の影響力指標の推定結果．推定方法，TA，TD 別に記した．

	TA		TD	
	R_r	δ_r	R_r	δ_r
NS	-0.12	-0.30	-0.36	-0.46
ELF	0.12	0.30	0.36	0.46

	TA		TD	
	b_s	β_s	b_s	β_s
Listening	0.32	0.43	0.34	0.43
Reading	0.28	0.31	0.27	0.31

4　考察

評価者の属性の差については，TD に関しては ELF が NS より厳しい傾向があった．一般的に ELF 使用者は文脈に依存した言語処理をしがちであるといわれる．本結果から，採点評価においては言語面を厳しく評価しがちであることを前提に，採点や評価者トレーニングの方法を考える必要があるかもしれない．また，TA は，NS と ELF の違いによって特徴が現れない「普遍的な」尺度である可能性が指摘できる．

受験者特性として英語の他技能（リーディング，リスニング）を用いたが，分析手法がいずれの場合も，リスニングの影響がリーディングよりも大きかった．これは，リスニング能力とスピーキング能力の相関が高いことを示唆するとともに，本テストの項目の一部が英語の対話を聴取し，内容を理解した上でなければ解答しにくいものであったことが影響しているともいえる．

また，MFRM による受験者特性の影響力指標は 2 段階推定を経て推定されているが，宇佐美(2010)のモデルは同時に推定されており，推定の方法論的により望ましい推定手法であるといえる．加えて，MFRM では評価者別の識別力の大きさを推定していないが，宇佐美(2010)では推定することができる．さらに，MFRM による方法は FACET という専用プログラムを推定に用いており，交互作用項の指定などで細かいモデルの指定はできない．それに対し宇佐美(2010)の方法で用いられた MCMC 法は，収束の判定などに困難があるものの，モデルを柔軟に設定できる．ただし，確度の高い結果を得るためには，MCMC 法は長い計算時間を必要とする場合がある．

今後，評価者個人の影響や局所独立の仮定を考慮したモデルを導入し，より実践的な推定を行う手続きを見出すことが課題である．また MCMC 法を用いた推定結果をどのように実際のテストに用いるかを検討する必要があるだろう．

引用文献

Hato, Y., Kanzawa, K., Mitsunaga, H., Tsubota, Y. & Underhill,N. (2016). Developing a computer-based speaking test of English as a Lingua Franca. *50th international IATEFL Conference and Exhibition*, Birmingham, UK.

Linacre, J.M. (1994). *Many-facet Rasch measurement (2nd ed.)*. Chicago, IL: MESA Press.

宇佐美慧 (2010). 採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル―MCMC アルゴリズムによる推定―，教育心理学研究，**58**, 163-175.

謝辞

本研究は JSPS 科学研究費補助金（科研費）16H03448 の助成を受けた．また，R による MCMC 法の推定についてスクリプトを提供いただいた宇佐美慧氏に，ここに記して感謝申し上げます．

（連絡先：mitsunaga-h@soc.shimane-u.ac.jp）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　51　2016/08/29　13:59:57

一般セッション２：ジェネリックスキル

９月９日（金）9:00 – 10:20　第２会場（B102）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　53　2016/08/29　13:59:57

論述力を測定する混合フォーマットテスト開発の試み

○野澤雄樹　伊藤素江　須永正巳　堂下雄輝　村田維沙
ベネッセ教育総合研究所

はじめに

与えられた資料から情報を適切に抽出し，それらを根拠として論理的な文章を書く力は，大学生が習得すべき汎用的な能力の一つである。このような能力は，その性質から，論述式の問題で測定することが自然であるように思われる。しかし，一般に論述式の問題は解答に要する時間が長く，採点のコストが大きいため，出題数を多くすることが難しい。そのため，十分な信頼性を確保することができず，正確な測定ができないという問題がある。

この問題に対する一つの解決策は，論述式の問題のみでテストを構成するのではなく，解答に要する時間が比較的短く採点コストが低い多肢選択式の問題と組み合わせて出題することである。多肢選択式の問題は出題数を多めに設定できるため，内容の偏りを減らし，信頼性を高めるのに貢献する。複数の解答形式の問題を組み合わせたテストは混合フォーマット（mixed-format）と呼ばれ，Advanced Placement Exams のような重要性が高いテストにおいても用いられている（例えば，Reshetar & Melican, 2010）。

本研究の目的

本研究では，論述力を「目的に応じてインプットした情報を考察し、考察した内容を論理的・効果的に言語で表現する能力」と定義し，混合フォーマットのテストによってどの程度正確に論述力を測定できるかを検討する。ここで論述力は単なるライティング・スキルではなく，読解力や思考力も含めた総合的な能力である。そのため，論述式の問題と組み合わせて出題する多肢選択式の問題は，語彙力のような基礎的な能力を測定する問題から，読解力や思考力のような高次の認知能力を測定する問題まで，幅広いものが考えられる。本研究の目的は，さまざまな能力を測定する多肢選択式の問題の中で，どのような内容の問題がもっとも論述力の測定に寄与するのか調べることである。

方法

データの収集

首都圏の大学に通う学生 500 人に対し，都内のテストセンターで複数のテストをコンピューターで実施した。テストの一覧を**表 1** に示す。500 人の受検者のうちランダムに選ばれた 167 人が論述テストを受検した。データの収集は 2015 年 12 月に行った。

テストの内容

論述テストでは，与えられた資料から情報を抽出・吟味し，それを根拠として，目的に合致した論理的な文章を書くことが求められる。特に字数の制限はないが，600 字程度の解答を想定している。

批判的思考テストは，批判的思考力を構成する「情報の明確化」「情報の分析」「推論」

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　52　2016/08/29　13:59:58

（楠見・子安・道田，2011）という三つのスキルを測定する多肢選択式のテストである。

汎用読解テストおよび教科型読解テストはどちらも読解力を測定するテストであるが，前者が連続・非連続テキストの両方を含むのに対し，後者は評論文の論の展開や趣旨として適切なものを選ばせる従来の国語科で出題されるものに近い内容である。

短答式・語彙読解テストは，論理的な文章を書くために必要な「語彙力」と「読解力」を測定するテストである。このテストは多肢選択式ではなく，1単語～数単語程度の短答を解答する形式であるが，コンピュータである程度自動的に採点できるため，本研究で検討するテストとして含めた。

語彙テストは，芝（1978）が開発した多肢選択式の語彙問題で構成されるテストである。

以降では，論述テスト以外のテストをまとめて客観テストと呼ぶ。その他，受検者の在学する大学・学部・学科の入試偏差値を適宜分析に加えた。

表1　実施したテスト一覧

テスト名	問題形式	問題数	各問題の採点	得点	試験時間（分）	受検者数
論述	論述	3	5段階評価	評定値の合計	110	167
批判的思考	多肢選択	40	正誤	正答数	60	500
汎用読解	多肢選択	25	正誤	正答数	40	500
教科型読解	多肢選択	17	正誤	正答数	60	498*
短答式・語彙読解	短答	61	正誤	正答数	40	500
語彙	多肢選択	30	正誤	正答数	15	500

* 無解答が多かった2名を分析から除外した

研究の枠組み

本研究では，論述テストが論述力を適切に測定していると仮定し，古典的テスト理論に基づいて以下のようにモデル化する。

$$X_{論} = T_{論} + E_{論} \quad (1)$$

ここで$X_{論}$，$T_{論}$，$E_{論}$はそれぞれ論述テストの観測得点，真値，測定誤差であり，$T_{論}$と$E_{論}$は無相関であることを仮定する。この枠組みの下で，論述力を正確に測定することと，$T_{論}$を正確に推定することは同じ意味をもつ。論述テストの問題数を増やし，測定誤差の分散を減らすことが，論述力を正確に測定する直接的な方法となる。しかし，論述テストの問題数を増やすことは容易ではないため，代わりに客観テストの問題を組み入れることで，測定の正確さを上げることを試みる。

評価指標

客観テストpの観測得点をX_pで表し，同様に古典的テスト理論に基づいてモデル化する。

$$X_p = T_p + E_p \quad (2)$$

このとき，複合得点（composite score）を以下のように定義する。

$$X_{\mathrm{comp}} = w_1 X_{論} + w_2 X_p \quad (3)$$

特に明示しないが，$X_{論}$およびX_pは標準化されているものとする。w_1とw_2はそれぞれ論述テストと客観テストに対する重みである。

$T_{論}$とE_p，T_pと$E_{論}$が無相関であることを仮定すると，論述テストの真値と複合得点の相関係数は次式のように書ける。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　53　2016/08/29　13:59:58

$$\mathrm{cor}\left(T_{論}, X_{\mathrm{comp}}\right) = \frac{w_1\sqrt{\mathrm{rel}(X_{論})}+w_2\mathrm{cor}\left(T_{論},T_p\right)\sqrt{\mathrm{rel}(X_p)}}{\sqrt{w_1^2+w_2^2+2w_1w_2\mathrm{cor}\left(T_{論},T_p\right)\sqrt{\mathrm{rel}(X_{論})}\sqrt{\mathrm{rel}(X_p)}}} \quad (4)$$

ここで cor は相関係数，rel は信頼性係数を表すものとする。本研究では，信頼性係数が観測得点と真値の相関係数の 2 乗であることの類推として，(4)式を 2 乗したものを測定の正確さの指標（評価指標）として用いる。

結果

論述テストと客観テストの関係

各テストに対して一般化可能性理論に基づく分析を行い，問題数を変化させたときの一般化可能性係数，および客観テストと論述テストの間の相関係数を**表 2** に示した。

希薄化の修正後に論述テストと最も相関が高かったのは批判的思考テストであり，汎用読解，教科型読解，短答式・語彙読解，語彙の順だった。

表 2　各客観テストの一般化可能性係数と論述テストとの相関

	一般化可能性係数					論述テストとの相関	
	元	20	30	40	80	希薄化修正前	希薄化修正後
批判的思考	.780	.640	.727	.780	.876	.593	.860
汎用読解	.755	.711	.787	.831	.908	.531	.782
教科型読解	.691	.725	.798	.840	.913	.445	.685
短答式・語彙読解	.828	.611	.702	.759	.863	.459	.646
語彙	.844	.783	.844	.878	.935	.440	.613

論述テストの一般化可能性係数は，元の問題数（3 問）のとき.610 であり，1 問のとき.343，2 問のとき.511，4 問のとき.676 だった。

因子分析の結果

R の psych パッケージ（Revelle, 2016）に含まれる fa.parallel 関数を使って平行分析を行った結果，2 因子が適当であると判断された。同パッケージに含まれる fa 関数で最尤法に基づく因子分析を行い，回転前の因子負荷をプロットしたものを**図 1** に示した。図から，批判的思考テストおよび汎用読解テストが測定している能力は，論述テストが測定している能力と近いことが示唆された。一方，短答式・語彙読解テストおよび語彙テストが測定している能力は，偏差値との位置関係から，教科学力に近いことが示唆された。

評価指標を使った分析

論述テストと客観テストの重みを設定し，希薄化修正後の相関係数および一般化可能性係数を(4)式に代入して 2 乗することで，さまざまな条件における評価指標を求めることができる。**図 2** は論述テストの問題数を 2 問，批判的思考テストの問題数を 20 問に設定した場合（パタン A）と，論述テストを 1 問，批判的思考テストの問題数を 40 問にした場合（パタン B）の評価指標を示したものである（$w_1 = 1 - w_2$に設定した）。パタン A の評価指標は批判的思考テストに対する重み（w_2）が 0.5 のときに 0.659 で最大になった。パタン B の評価指標はw_2が 0.7 のときに 0.652 で最大になった。この結果から，重みを適切に決めることで，パタン A とパタン B で同程度の評価指標の値が得らえることがわかった。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　54　2016/08/29　13:59:59

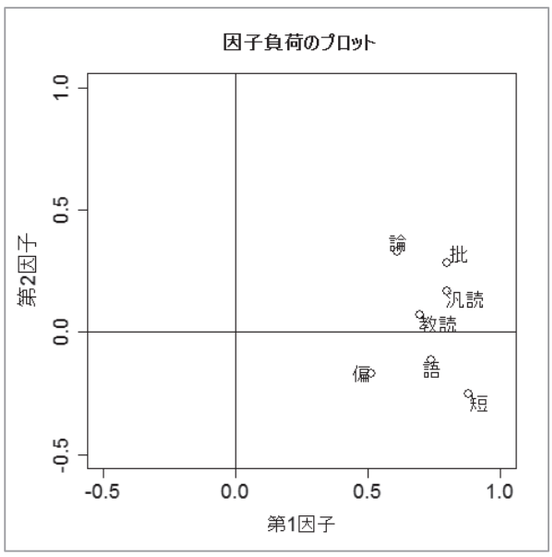


図１　因子負荷のプロット

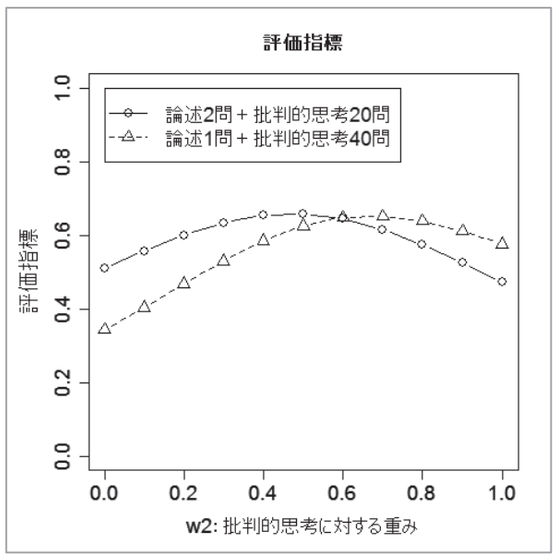


図２ 批判的思考テストと組み合わせたときの評価指標

考察

希薄化修正後の相関係数および因子分析の結果から，批判的思考テストが最も論述テストと近い能力を測定していることが示唆された。評価指標を用いた分析では，論述テストの問題を１問増やすのと同程度の効果が，批判的思考テストの問題を 20 問増やすことで得られることがわかった。批判的思考テストの問題を 20 問解くのに要する時間は 30 分程度であり，論述テストの問題を１問解くのに要する時間と同程度である。そのため，テスト時間を短縮することにはならないものの，採点のコストを半分程度に減らすことができるので，論述テストと批判的思考テストを組み合わせて出題するメリットは大きいと思われる。

この研究で使用した論述テストおよび客観テストは，妥当性が確立されているわけではない。そのため，今回の結果は，あくまで開発中のテストで見られた結果であり，批判的思考力と論述力の関係を一般的に示したものではないことに注意する必要がある。

引用文献

楠見　孝・子安増生・道田泰司（編）（2011）．批判的思考力を育む―学士力と社会人基礎力の基盤形成　有斐閣

Reshetar, R., & Melican, G. (2010, April). *Design and evaluation of mixed-format large scale assessments for the Advanced Placement Program (AP).* Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.

Revelle, W. (2016). psych: Procedures for personality and psychological research [Computer software]. http://CRAN.R-project.org/package=psych. (R package version 1.6.6)

芝 祐順（1978）．語彙理解尺度作成の試み　東京大学教育学部紀要　第 17 巻　pp.47-58.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　55　2016/08/29　13:59:59

「実社会・実生活に生きる多様な力を測るテスト」の多面的な分析

○岡村彩音*，○志水 黎*，山田剛史*，長谷川康代**，牧野直道**，岡下博美**
岡山大学大学院教育学研究科*，(株)ベネッセコーポレーション**

問題と目的

ベネッセコーポレーションが，2013 年度から 2015 年度の 3 カ年にわたって実施した文部科学省調査研究「高等学校における多様な学習成果の評価手法の調査研究」事業(岡下博美 牧野直道 長谷川康代(研究代表))では，高等教育を通じて生徒が身に付けるべき，「社会・職業への移行に必要な資質・能力」のうち，論理的思考力・批判的思考力，問題発見・解決力，メタ認知や，実践力(人間関係形成力・社会参画力)のもととなる認識，態度を適切に評価するテストの開発を試みた。このテストは選択式問題や，記述式問題，論述式問題，質問紙など多様な手法を組み合わせ，高い信頼性・妥当性を持つ評価手法の開発が追及された(ベネッセコーポレーション，2016)。本研究では，文部科学省調査研究で開発された「実社会・実生活に生きる多様な力を測るテスト」(以下，「評価テスト」と略す）の 2014 年度のデータを用いて，評価テストの各項目についての古典的テスト理論と項目反応理論による統計的な分析を行い，また，様々な心理特性についての質問紙調査との関連を検討することを目的とした。

方法

2014 年度の「評価テスト」は，高校生 1991 名，大学生 231 名，社会人 156 名の計 2378 名を対象に実施された。うち，527 名は去年も 2013 年度の評価テストを受けた共通受験者である。2014 年度の評価テストは，多肢選択式問題，態度測定質問，論述式問題の 3 つを集団質問形式で実施された。本研究では高校生 1991 人の多肢選択式問題と態度測定質問のデータについて扱うこととする。

多肢選択式問題は，3 つの力（論理的思考力，問題解決力，人間関係形成力）で構成されている。各力の定義、項目数、平均点、標準偏差、α係数をまとめた（表 1)。全 58 項目のうち 14 項目が 2013 年度実施のテストからの流用問題である。

表 1　多肢選択式問題を構成する 3 つの力の定義

多肢選択式問題を構成する力	定義	項目数	平均	標準偏差	α係数
論理的思考力	必要な情報を正しく取り出し，情報の正確さを分析・評価して判断するなど，物事を論理的に考察して説明する力	19	9.69	3.39	0.71
問題解決力	潜在的な情報を探るなど，与えられた情報の本質をとらえて批判的に分析し，問題を発見・解決したり，新しいアイデアを生み出したりする力	19	11.71	4.35	0.82
人間関係形成力	多様な他者の考えや価値観を理解し，他者と効果的なコミュニケーションをとり，意見の対立を解消するための解決策を導き出す力のもととなる，認識・方略・態度	20	12.56	3.02	0.61
テスト全体		58	33.98	9.43	0.88

態度測定質問紙は，論理的思考態度 15 項目，コスモポリタニズム尺度 2 項目，問題解決態度尺度 19 項目，人間関係形成態度尺度 5 項目，自己理解・管理態度尺度 14 項目，Big five 尺度 19 項目，対人関係コントロール尺度 6 項目，主体的意欲・態度尺度 6 項目， 社会への参画態度尺度 10 項目から構成された。各項目は，それぞれが尋ねる態度に「どの程度あてはまるか」「どの程度身に着けているか」を 5 件法で問うものであった(1:あてはまらない or 身につけていない-5:あてはまる or 身につけている)。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 56 2016/08/29 14:00:00

結果と考察

1. 統計的分析による項目の評価

分析には高校生 1991 人のデータを用いた。テストの信頼性（内的一貫性）を確認するため多肢選択式問題の全 58 項目を用いてクロンバックのα係数を求めたところ，0.88 という十分に高い結果が得られた。

古典的テスト理論による分析（以下，CTT による分析と略す）では正答率と点相列相関係数の値で評価した。CTT の評価が高かった項目は，登場人物の言うセリフを提示された状況から考え答える問題などであった。CTT の評価が低かった項目は，登場人物がとった手立てがどの要素を満たしているか選ぶ問題，5 つの文のうち筆者の主張が書かれている文を選ぶ問題などであった。

スクリープロットの結果から 58 項目全体で一次元性が確認されたため，項目反応理論（以下，IRT と略す）を適用できると判断した。本研究では 2 パラメタロジスティックモデルを用いて IRT 分析を行った。項目パラメタの推定値により項目の評価を行ったところ，例示された文章の論理構造と同じ構造のものを選ぶ問題，主張を裏付ける理由として適切なものを選ぶ問題，登場人物の主張を支える理由を記述する問題といった問題が良い問題と判断された。一方，5 つの文の中から敬語の使いかたとして最上の敬意を示すものを選ぶ問題，5 つの文のうち筆者の主張が書かれている文を選ぶ問題といった問題が悪い問題と判断された。

また，CTT と IRT の両方の視点から評価を行ったところ，提示された条件を読み取りグラフを描く問題，主張として提示された文章に対して，それを裏付ける文章を選ぶ問題，ある事象が起こる確率を答える問題などが良い問題として判断された。一方，5 つの文で成り立っている文章の中から，筆者の伝えたいことが書かれているものを選ぶ問題（この問題は受験者の能力値に関係なく正答した受験者と間違えた受験者が存在した。），5 つの文の中から敬語の使いかたとして最上の敬意を示すものを選ぶ問題，登場人物がとった行動が提示された条件にどれだけ当てはまっているか答える問題などが悪い問題として判断された。

2. 質問紙調査との関連

評価テストで測定している能力と態度との間に関連があるかどうかを確認した。評価テストを構成する 3 つの力（論理的思考力，問題解決力，人間関係形成力）については，各力ごとに素点を算出しテスト得点とした。評価テスト合計点は 3 つの力のテスト得点の合計である。態度測定質問紙調査については，下位尺度を構成する項目得点の和を尺度得点とした(表 2)。態度測定質問紙の尺度と評価テスト得点の相関を見たところ，「論理的思考態度」，「主体的意欲・態度」，「問題解決態度」の 3 つの尺度得点とテスト合計得点との間に弱い正の相関が見られた。しかし，他の態度質問紙との相関係数の値は低かった。このことから、評価テストで測定している能力と態度は独立したものである可能性がある。

表 2　テストと質問紙調査との関連（N = 1991）

テスト	質問紙調査										
	論理的思考態度	コスモポリタニズム	問題解決態度	人間関係形成態度	自己理解・管理態度	性格（開放性）	性格（誠実性）	性格（調和性）	対人関係コントロール	主体的意欲・態度	社会への参画態度
論理的思考	0.17**	-0.07**	0.22**	-0.01	0.10**	0.07**	0.08**	0.02	0.04*	0.22**	0.05*
問題解決	0.18**	-0.05**	0.22**	0.01	0.09**	0.06**	0.07**	-0.01	0.05*	0.24**	0.03
人間関係	0.17**	-0.01	0.21**	0.05*	0.11**	0.05**	0.10**	0.07**	0.06**	0.25**	0.09**
テスト合計	0.20**	-0.05**	0.24**	0.02	0.11**	0.07**	0.10**	0.02	0.06**	0.27**	0.06**

*p<.05　　**p<.01

3. 今後の課題と展望

本研究では，ベネッセコーポレーションによって開発された「評価テスト」(2014 年度実施）について，主に統計的な観点からテスト項目の評価を行った。本研究で扱った「評価テスト」のような新しいテストに対する研究の蓄積はまだ少なく、本研究における検討は、今後の統計的なテスト評価の最初のステップとなるだろう。また，テスト項目の良し悪しは，統計的な，量的な観点のみからなされるのではない。例えば，テストの文章表現やテスト形式，選択肢の長さや形式といった，質的な観点からも評価することができるだろう。そうした他の観点からの評価も行うことで，開発された評価テストの問題の吟味をより多面的に行うことができるだろう。

引用文献

株式会社ベネッセコーポレーション(2016)　社会・職業への移行に必要な資質・能力の評価手法の開発と高校の指導の質向上へ生かす方法の調査研究，平成 25 年～平成 27 年文部科学省高等学校における「多様な学習成果の評価手法に関する調査研究」事業最終報告書

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　57　2016/08/29　14:00:01

クリティカルシンキングの適応型テストの尺度開発

若山 昇　帝京大学/CRET
宮澤 芳光　東京学芸大学
梶谷 真司　東京大学
植野 真臣　電気通信大学

概　要

現代社会はいろいろな情報があふれており，情報を十分に吟味して体系的に理解し，意思決定を的確に行うためには，クリティカルシンキング(CT)は重要となる。本研究の目的は，CT の適応型テストの開発のため，尺度の開発を試みるものである。大学生 736 人のデータを分析した結果，CT における 3 尺度（分析，推論，読解）は，相関，散布図，情報量，固有値などから，概ね独立的であり妥当であると考えられた。

キーワード：クリティカルシンキング，適応型テスト，項目反応理論，尺度開発，試験

1.　はじめに

現代社会はいろいろな情報があふれており，情報を十分に吟味して体系的に理解し，意思決定を的確に行う必要がある。このため社会人基礎力(経済産業省 2006)，学士力(文部科学省 2008)，21 世紀型スキル(Griffin ほか 2011)において謳われているように，クリティカルシンキング(CT)は必要不可欠となる。CT とは，先入観に囚われず，論理的に考え，合理的な決定を導き出す能力と意思である(若山 2009)。CT の重要性の高まりとともに，学生・社会人の CT 能力を測定する必要性が高まっている。CT の能力を評価する方法は，いまだ確立されているとは言えず，海外のテストを活用する例(平山ほか 2011)や，国内においても能力評価の研究が存在する(楠見ほか 2010)。しかし，CT の領域は広く，解答するには読解時間に加えて，考える時間が十分に必要になるため時間がかかってしまう。

一方，既存のテストでは，同じテストを受検した者同士でしか，能力を比較することができないが，項目反応理論(IRT)では，異なる試験問題でも同一の尺度上で受検者を評価することができる。また，IRT に基づき解答履歴から能力を逐次的に推定し，受検者ごとに的確な問題をプール(アイテムバンク)から瞬時に選別して出題できる適応型テスト(CAT)という手法がある。実際に，IRT を用いた CAT は英語の検定試験をはじめ，情報処理技術者試験や医療系大学間共用試験，人材測定テスト SPI などでも実施されるようになっている。CT においても短時間で効率的に能力を測定するために CAT の技術開発が必要である。そのために尺度の開発が不可欠となる。本研究の目的は，CT の適応型テストの開発のために，尺度の開発を試みるものである。

2.　方　法

CT の能力評価は，行動観察や口頭試問などがあるが本研究では多肢選択型試験で測定できると想定し，尺度開発のためにペーパー試験を実施した。2014 年 7 月～2016 年 12 月に，東京及び近郊の大学生 736 人を対象とした。18 歳から概ね 22 歳であり留学生及び重複・欠損値のある回答は対象から除外している。倫理的配慮として，成績に無関係で教育・研究目的以外には使用せず個人名は特定されないことを説明し，学生の了解を得た。

大学教員でかつ教育に関する研究を行っている博士号を有する研究者 4 人で議論し，尺度と問題を設定した。尺度は①分析的思考力，②論理・推論能力，③読解・理解能力の 3 尺度とした。

①分析的思考力について

情報や問題を的確に把握し理解するための分解・解析する能力である。いわゆる公務員試験や法科大学院の適性試験では分析，数的処理といわれる分野がこれに含まれる。具体的には，順列・組合せ，確率，順序・手順，位置・対応，嘘つき・暗号問題などが含まれるが，数学の知識の有無の影響のないよう配慮した。例えば，変数 x,y の 2 つを使わないと解けない問題は避けた。

②論理・推論能力について

論理展開の妥当性・整合性の有無及びその背景を的確に把握する能力である。いわゆる公務員試験や適性試験では論理，推論といわれる分野がこれに含まれる。具体的には，集合・論理，命題，演繹・帰納，逆・裏・対偶，論理の前提や飛躍・省略，論理構造，因果・相関，第 3 変数などがこれに含まれる。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　58　2016/08/29　14:00:02

③読解・理解能力について

情報を有機的に結合し活用して，内容の本質を的確に理解する能力である。PISA の応用力はこれに含まれる。難解すぎる文章や長い文章は，国語力が左右するので避けた。さらに受検者の慣れ/不慣れや既存の教科の知識量による影響を避けるべく，論理展開・構造が複雑すぎないものとし，概ね 500～1000 文字程度の文章とした。

設問は，独自で作問したものに加えて，公務員試験など既存の問題の中から CT の定義に沿ってこれらの尺度能力を測る問題で構成した。各設問例(イメージ)は以下のとおり。

① 分析的思考力の設問例

コインを 3 枚同時に投げた。今，その 3 枚は表・表・表ではなく，裏・裏・裏でもないことが分かった。このとき 3 枚のうち 2 枚が裏である確率は，次のどれか。 なお，コインの表裏の出る確率はそれぞれ 50％である。

(1) $\frac{1}{8}$ (2) $\frac{1}{4}$ (3) $\frac{1}{3}$ (4) $\frac{3}{8}$ (5) $\frac{1}{2}$ (6) $\frac{5}{8}$ (7) $\frac{2}{3}$ (8) $\frac{3}{4}$ (9) $\frac{7}{8}$ (0) これらのいずれでもない。

② 論理・推論能力の設問例

「人は死ぬ」「ソクラテスは人である」この 2 つの命題を合わせることから「ソクラテスは死ぬ」という 100％確実な結論になる。以下の設問において，結論「 ？ 」に最適なものを(1)～(5)の中から選べ。

「新幹線に乗ると遅れないで着く」「あの人は，遅れないで着いた」したがって「 ？ 」

(1) あの人は，新幹線に乗った (2) 遅れるのは新幹線に乗らないからだ (3) 遅れないためには新幹線に乗る
(4) 新幹線は，時速 200kmを超えている (5) 何も入りえない

③ 読解・理解能力の設問例

次の文章の要旨として最も妥当なのはどれか。

通貨を統合すると何が起きるか。通貨を統合した国とは戦争が起こせないという事である。単純な話である。もし経済的に同一の国の中で戦争になれば，経済的には「国内紛争」となり，その「国」の国力を激減させる。経済的に見て敗者も勝者もなく，搾取側も被搾取側も存在しない。戦後処理が国内で負担されるのであるから，その「国」にとって基本的なメリットはない。自分で自分の首を締めることになるからだ。確かに，通貨統合の象徴である中央銀行の設立は，市民の意識を大きく変えることになる。明治維新後の日本銀行設立，アメリカ独立後の FRB 設立，ユーロの中央銀行設立以降，国内の紛争は激減している。

そもそも面倒な通貨統合を成し遂げてまで目指そうとする基軸通貨とは，何だろうか。ユーロは通貨統合を成し遂げて世界の基軸通貨になることを目指している。現在，世界の基軸通貨はドルである。では，アメリカのドルが基軸通貨になると，アメリカにとって何が好都合なのであろうか。通貨であるドルが，アメリカ国外でも流通することでドル通貨圏が拡大する。端的に言えば，これによりアメリカの中央銀行(FRB)が世界経済の中央銀行になることに近づくことができる。つまりアメリカが世界経済を左右する事が可能となる。では，世界の経済が一つになり，完璧に融合されたときの状態はどうなるか。 世界全体が「拡大アメリカ合衆国」となり，戦争が消え平和が訪れる，と考える有識者もいる。ある意味で真実だろう。しかし，世界の人が本当にそれを望んでいるだろうか。

大切な事は，欧米にできた事は，アジアでも充分にできると言うことである。それどころか現状のままでは，ドルとユーロの2大通貨がアジアを分断統治することになろう。アジアが通貨覇権競争において草刈場となりつつある。西にあるトルコはユーロの勢力範囲に，東にあるニッポン，フィリピン，太平洋の島々はドルの勢力範囲になりつつあると見るのは，早合点であろうか。10 年後にはインドを含めたアジアは人口，経済ともに世界 1 位になるだろう。今こそアジアの力の結集を検討するべきではないだろうか。アジアの英知が，いやその英知の有無までもが，今まさに試されようとしているのだから。

(1) アジアの諸国は，ユーロで見られたような通貨統合を，検討するべきである。
(2) アメリカのドルを基軸通貨とする世界経済は安定しており，アメリカ中心の世界が戦争のない世界平和をもたらす。
(3) 通貨を統合すると，統合した国同士では戦争は起こせない。なぜなら，経済的には「国内紛争」となり，相手から搾取する意味がなくなるからである。
(4) 基軸通貨になると，その国の世界的な影響力が拡大する。なぜなら，その国の中央銀行が世界の中央銀行のような力を持つからである。
(5) ドルとユーロの2大通貨の勢力がアジア地域に及んでいる。西のトルコ，東のニッポンなどをみれば，明らかである。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 59 2016/08/29 14:00:03

各尺度内ではそれぞれ 5 問を共通問題として，各尺度で 15 問の試験を 5 組作成した。試験時間は 1 問 4〜5 分として，60〜75 分に設定し，1 組に 50 人程度が解答した。大学の授業では試験を1組ずつ行い合計 632 人のデータを得た。別途週末に 55 人ずつ(3 人が重複)2 日で 107 人が集まり①分析と③読解の両方に設問，加えて②推論の共通問題に解答してもらった(表 1)。IRTPRO3，SPSS20 を用いて相関，散布図，情報量，固有値などの分析を行い尺度の妥当性を検討した。

3. 結　果

個人の能力における各尺度間の相関(表 2)と尺度間の散布図(図 1a〜1c)は以下のとおり。各尺度毎の困難度・識別力の散布図を以下に示す(図 2a〜2c)。さらに，各尺度のテスト情報量(I)と標準誤差(S.E.)は以下のとおり(表 3，図 3a〜3c)。また，各尺度の固有値プロットを図 4a〜4c に示す。

表1　尺度毎の試験問題と受検者の概要　　○,◎：おおむね50人以上

		①分析的思考力						②論理・推論能力						③読解・理解能力					
	試験	1A	1B	1C	1D	1E	1共通	2A	2B	2C	2D	2E	2共通	3A	3B	3C	3D	3E	3共通
被検者	問題数	10	10	10	10	10	5	10	10	10	10	10	5	10	10	10	10	10	5
授業 632人								○	○	○	○	○	◎						
													◎		○				◎
				○			◎												
週末 55人			○	○	○		◎						◎		○	○			◎
週末 55人		○		○		○	◎						◎	○			○	○	◎

授業と2つの週末において3人のダブりがある

表2　個人能力の尺度間相関

	Θ1	Θ2	Θ3
Θ1		.568**	.570**
Θ2			.423**
Θ3			

**: $p < .01$

表3　尺度のテスト情報量と標準誤差

	情報量(max)	標準誤差	困難度
①分析的思考力	11.65	0.29	-0.20
②論理・推論能力	17.44	0.24	0.20
③読解・理解能力	10.15	0.31	-0.70

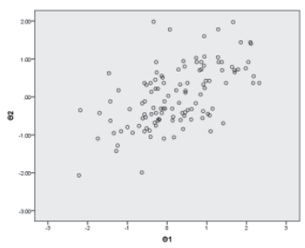


図 1a ①分析的思考力と②論理・推論能力

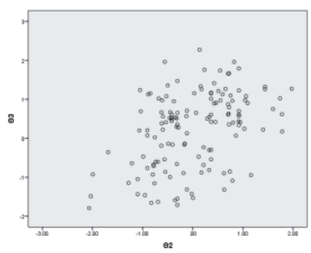


図 1b ②論理・推論能力と③読解・理解能力

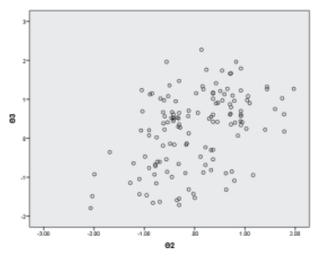


図 1c ③読解・理解能力と①分析的思考力

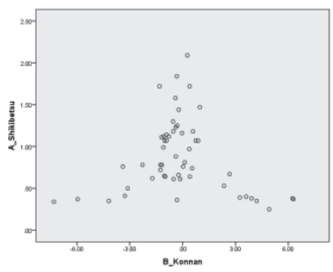


図 2a ①分析的思考力の困難度と識別力

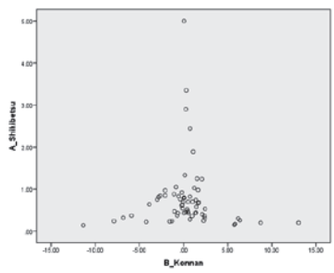


図 2b ②論理・推論能力の困難度と識別力

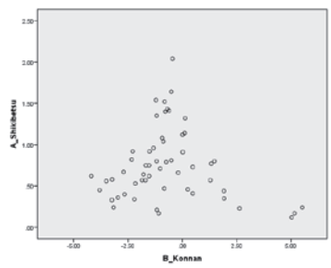


図 2c ③読解・理解能力の困難度と識別力

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　60　　2016/08/29　14:00:04

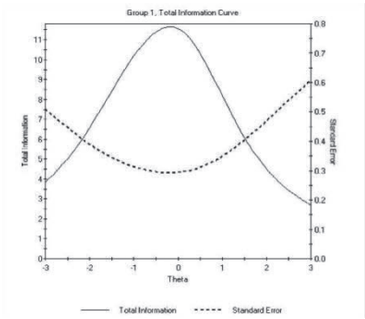


図 3a ①分析的思考力の情報量と標準誤差

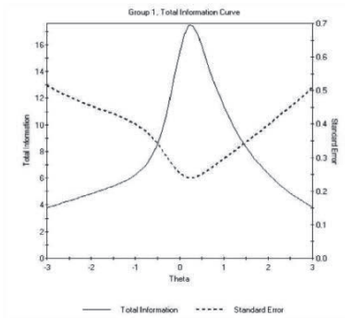


図 3b ②論理・推論能力の情報量と標準誤差

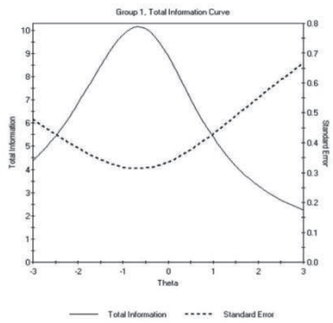


図 3c ③読解・理解能力の情報量と標準誤差

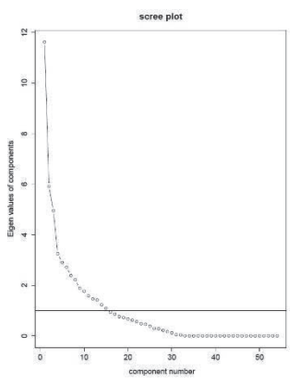


図 4a ①分析的思考力の固有値プロット

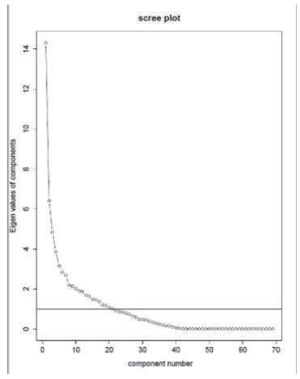


図 4b ②論理・推論能力の固有値プロット

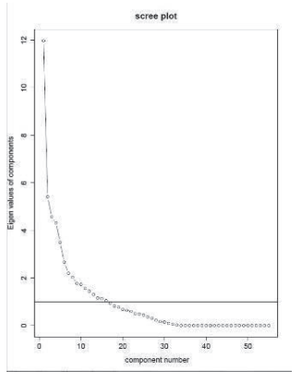


図 4c ③読解・理解能力の固有値プロット

4. 考 察

各尺度(①分析的思考力,②論理・推論能力,③読解・理解能力)の相関は，.423～.568 となり，想定よりも小さくなった。散布図を見ても各図で 2 元性以上は確認されず,偏りもそれほど大きくはない。困難度・識別力の散布図は，想定通りであり困難度 0 付近を中心に左右対称で正規分布のような形状が確認された。固有値プロットでは，どの尺度も第 1 と第 2 因子間で，大きく半減しており 1 因子の可能性が示唆された。被検者数が必ずしも十分に大きくはないこともありテスト情報量(I)と標準誤差(S.E.)は限界があるものの概ね妥当な範囲におさまった。また，②推論ではグラフには若干の歪みが見られたが，授業による被検者数が他の尺度より多く，誤差は小さく(0.24)となった。これらから,各尺度は概ね独立的であり，概ね妥当であることが示唆された。

5. おわりに

本研究により，想定した 3 尺度は概ね独立しており，妥当であることが示唆された。今後は被検者を増やすと共に，識別力の極端に低いものを排除して検証を行うことが必要となろう。また，今回は CT の能力評価において 3 尺度の妥当性を検討したが，今後は 3 次元尺度の最適性を探究することが望まれよう。なお，本抄録は若山ほか(2016)に分析結果を加えてまとめたものである。本研究は科研費（A）15H01772，（C）15K01088 及び CRET(教育テスト研究センター)の助成を受けている。

<参考文献>

Robert L. Linn (原著，編)，池田央ら(訳): “教育測定学” 上下巻，学習評価研究所 (1992)
若山昇:“大学におけるクリティカルシンキング演習授業の効果”，大学教育学会, Vol.31, No.1, pp.145-153 (2009)
若山昇ほか:“クリティカルシンキングの適応型テストのアイテムバンク構築”，教育システム情報学会 第 41 回全国大会講演論文集 (2016 予定)

Email: Wakayama.class@pobox.com

会員番号 30000224

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 61 2016/08/29 14:00:04

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 62 2016/08/29 14:00:04

一般セッション３：記述式・論述式テスト

９月９日（金）10:30 – 11:50　第１会場（B101）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　63　2016/08/29　14:00:04

評価者と課題の多様な特性を考慮した項目反応モデル

○宇都雅輝 *1，植野真臣 *1

電気通信大学大学院情報理工学研究科 *1

1　はじめに

近年，論理的思考力や問題解決力といった高次の能力を測定するニーズが高まっており，これを実現する手法の一つとしてパフォーマンス評価が注目されている [1, 2]．パフォーマンス評価は，これまでにも，大学入試における論述式テストや外国語のリスニング試験，人事考課における面接試験，学習場面でのレポート評価など，様々な形式で活用されてきた．また，将来の導入が検討されている「大学入学希望者学力評価テスト」でも採用が検討されており，パフォーマンス評価の重要性は今後ますます増加すると考えられる．

パフォーマンス評価では，受験者にいくつかの課題を与え，そのパフォーマンスを複数の評価者に採点させることで受験者の能力を測定する．しかし，この評価法では，得られる評点が評価者と課題の特性に強く依存することが知られており，これにより能力測定の精度が低下する問題がある [1, 3, 4]．

具体的には，次の評価者特性が能力測定のバイアス要因となることが知られている（e.g., [1, 4]）．

- **評価の厳しさ（甘さ）：** 全ての受験者に対して低い（または高い）評点を与える特性．
- **尺度範囲の制限：** いくつかの評点に評価が集中する特性．評価尺度の平均値付近に評点が集中する特性は中心化傾向と呼ばれる．
- **評価の一貫性：** 類似したパフォーマンスに対して評価が一貫している程度を表す特性．一貫性の低い評価者は，類似したパフォーマンスに対して評点がばらつく．

また，課題特性としては以下が想定されることが多い（e.g., [3, 5, 6]）．

- **困難度：** 得られる評点が全体として低く（高く）なる特性．
- **識別力：** 関心下の能力の大小がパフォーマンスの質に反映される程度を表す特性．識別力の低い課題では，同程度の能力を持つ受験者でもパフォーマンスの質がばらつく．

したがって，パフォーマンス評価では，このような評価者と課題のバイアスをできる限り取り除いて関心下の能力を精度よく測定することが重要課題となる．

これらのバイアスを排除した能力測定を実現するためには，(1) 十分な評価者数と課題数を用意すること，(2) 評価者のトレーニングを十分に行うこと，(3) 厳密な評価基準を提示すること，(4) 関心下の能力が明確に表出される課題とすること，などが必要となる [1]．しかし，実際のパフォーマンス評価では，人的・時間的・経済的な制約から，これらの条件を満たす理想的な評価を実施することは難しい場合も多い [1]．

この問題を解決する手法の一つとして，評価者と課題の特性パラメータを付与した項目反応モデルが提案されてきた（e.g., [3, 5, 6]）．しかし，これらのモデルは，上記のバイアス要因の一部しか考慮しないため，評価者と課題の多様な特性が想定される場合には，測定誤差を十分に取り除くことができず，能力測定精度が低下する．

例えば，多相ラッシュモデル (Many-Facet Rasch Model: MFRM) を拡張した hybrid MFRM[4] では，評価者の厳しさと尺度範囲の制限，課題の困難度を考慮した能力推定が可能であるが，評価者の一貫性と課題の識別力については全ての評価者と課題で一定としており，能力測定精度はこれらの特性に依存しないと仮定する．しかし，一般に，これらの特性は評価者間・課題間で異なり，一貫性が低い評価者や識別力の低い課題では能力測定誤差が増加する．すなわち，このモデルでは，評価者一貫性と課題識別力の低さに由来する測定誤差が能力推定値に過度に反映され，能力測定精度が低下する点が問題となる．

一方，評価者一貫性と課題識別力を考慮することで，測定誤差が大きい評価者や課題の影響を小さくして能力推定を行うモデルが提案されている [3, 6]．しかし，これらのモデルでは，各評点の使用割合が全ての評価者について一定であると仮定するため，尺度範囲の制限の特性が表現できない．そのため，尺度範囲の制限の生起が想定される場合，次の理由により能力測定精度が低下する．(1) 尺度範囲の制限が強い評価者は，幅広い能力の受験者に同一評点を付与し，それらの受験者の能力差を区別しないため，同一評点が与られる能力区間では能力測定誤差が増加する傾向がある．しかし，既存モデルでは，この特性を表現できないため，尺度範囲の制限が強い評価者の測定誤差が部分的に過小評価される．結果として，現実の誤差の影響が能力推定値に過度に反映され，能力測定の精度が低下する．(2) 既存モデルでは，各評点を一定割合かつ受験者の能力に応じて使い分ける評価者を一貫性が高いと捉える．そのため，尺度範囲の制限が強い評価者は，一貫性が低く推定され，能力測定誤差も能力尺度全域で大きいと解釈される傾向がある．しかし，実際には，尺度範囲の制限が強くとも，受験者の能力に相関した評点を与える評価者は一貫性が高いと解釈

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　62　2016/08/29　14:00:05

でき，評点の使用傾向が変化する能力値周辺では能力差の識別精度も高いといえる．既存モデルでは，能力測定精度改善に有効なこのような評価者の影響を能力推定に十分に反映できず，能力測定精度が低下する．

以上の問題を解決するために，本研究では，評価者と課題の多様な特性を考慮できる新たな項目反応モデルを提案する．具体的には，hybrid MFRM に対して評価者の一貫性と課題の識別力パラメータを付与したモデルを提案する．提案モデルは，以下の特徴により既存モデルより高精度な能力測定が期待できる．(1) 尺度範囲の制限が強い評価者でも，受験者の能力に相関した評価を行っていれば，一貫性を高く推定できる．(2) 尺度範囲の制限が認められる評価者に対して，評点の使用傾向の偏りを適切に表現でき，能力尺度の一部において能力測定誤差が増加する特徴も表現できる．(3) 一貫性が低い評価者と識別力の低い課題が能力推定に与える影響を小さくできる．(4) 上記 (1)〜(3) を考慮した能力推定を行うことができる．

2　評価者と課題の特性を考慮した項目反応モデル

本研究では，受験者 j（$j=1,\cdots,J$）の課題 i（$i=1,\cdots,I$）に対する評価者 r（$r=1,\cdots,R$）の評価カテゴリ k（$k=1,...,K$）で構成される「受験者」×「課題」×「評価者」の 3 相データ $\boldsymbol{U}$ を扱う．このような 3 相データから，課題と評価者の影響を同時に考慮した受験者の能力推定を行うために，評価者と課題の特性パラメータを付与した項目反応モデルが多数提案されてきた．

例えば，課題の困難度と評価者の厳しさ，尺度範囲の制限を考慮できる項目反応モデルとして，MFRM を拡張した以下の hybrid MFRM が知られている [4]．

$$P_{irk}(\theta_j)=\frac{\exp\sum_{m=1}^{k}\left[\theta_j-\beta_i-\beta_r-d_{rm}\right]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\left[\theta_j-\beta_i-\beta_r-d_{rm}\right]}, \tag{1}$$

ここで，θ_j は受験者 j の能力，β_i は課題 i の困難度，β_r は評価者 r の厳しさ，d_{rk} は評価者 r の評点 $k-1$ に対する評点 k への厳しさを表す．このモデルでは，d_{rk} により尺度範囲の制限を表現できる．なお，モデルの識別性のために，$\beta_{r=1}=0$, $d_{r1}=0$，$\sum_{k=2}^{K}d_{rk}=0$ を所与とする．

また，宇佐美 [6] は，一般化部分採点モデル (Generalized Parcial Credit Model: GPCM) に評価者パラメータを付与した以下のモデルを提案している．

$$P_{irk}(\theta_j)=\frac{\exp\sum_{m=1}^{k}\left[\alpha_i\alpha_r(\theta_j-\beta_i-\beta_r-d_rd_{im})\right]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\left[\alpha_i\alpha_r(\theta_j-\beta_i-\beta_r-d_rd_{im})\right]} \tag{2}$$

ここで，α_i は課題 i の識別力，α_r は評価者 r の一貫性，d_r は評価者 r の中心化傾向，d_{ik} は課題 i においてカテゴリ $k-1$ から k に遷移する難しさを表す．ただし，パラメータの識別性のために，$\Pi_r\alpha_r=1$，$\sum_r\beta_r=0$，$\Pi_rd_r=1$，$d_{i1}=0$，$\sum_{k=2}^{K}d_{ik}=0;\forall i$ を仮定する．このモデルでは，$\theta=0$ を中心とした反応分布全体の分散を表す d_r で中心化傾向を表現するが，尺度範囲の制限を必ずしも表現できるわけではない．

また，Uto and Ueno[3] は，評価者の一貫性と課題の識別力の特性を考慮できるモデルとして，段階反応モデル (Graded Response Model: GRM) を拡張した以下のモデルを提案している．

$$P_{irk}(\theta_j)=P^{*}_{irk-1}(\theta_j)-P^{*}_{irk}(\theta_j) \tag{3}$$

$$P^{*}_{irk}(\theta_j)=\left[1+\exp(-\alpha_i\alpha_r(\theta_j-b_{ik}-\varepsilon_r))\right]^{-1}$$

ただし，$P^{*}_{ir0}(\theta_j)=1$，$P^{*}_{irK}(\theta_j)=0$ とする．ここで，ε_r は評価者 r の厳しさ，b_{ik} は課題 i において k より大きい得点を得る困難度を表す．ただし，$b_{i1}<b_{i2}<\cdots<b_{iK-1}$ とする．また，パラメータの識別性のために，$\alpha_{r=1}=1$，$\varepsilon_1=0$ を仮定する．

しかし，1 章で述べたように，評価者と課題の多様な特性が想定される場合，これらのモデルでは評価者と課題のバイアスを十分に取り除くことができず，能力測定精度が低下する問題が残る．

3　提案モデル

上記の問題を解決するために，本研究では，評価者と課題の多様な特性を考慮した以下の項目反応モデルを提案する．

$$P_{irk}(\theta_j)=\frac{\exp\sum_{m=1}^{k}\left[\alpha_r\alpha_i(\theta_j-\beta_i-\beta_r-d_{rm})\right]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\left[\alpha_r\alpha_i(\theta_j-\beta_i-\beta_r-d_{rm})\right]}, \tag{4}$$

ただし，パラメータの識別性のために，$\alpha_{r=1}=1$，$\beta_{r=1}=0$，$d_{r1}=0$，$\sum_{k=2}^{K}d_{rk}=0$ を仮定する．

ここで，提案モデルのパラメータと能力測定精度の解釈を説明するために，表 1 の評価者特性パラメータを持つ 6 名の評価者について，提案モデルの項目反応曲線と標準誤差関数を図 4 に示す．標準誤差関数 $SE_{ir}(\theta)$ は，能力 θ を持つ受験者の課題 i へのパフォーマンスに対する評価者 r の測定誤差を表し，フィッシャー情報量 $I_{ir}(\theta)$ を用いて次のように定義される．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　63　2016/08/29　14:00:07

表 1: 異なる特性を持つ 6 名の評価者の特性パラメータ

	α_r	β_r	d_{rk}		α_r	β_r	d_{rk}
Rater 1	0.5	0.0	{0, -1.5, -0.5, 0.5, 1.5}	Rater 4	2.0	0.0	{0, -2.5, 0.5, 0.8, 1.2}
Rater 2	2.0	0.0	{0, -1.5, -0.5, 0.5, 1.5}	Rater 5	2.0	0.0	{0, -2.0, -1.5, 1.5, 2.0}
Rater 3	2.0	1.0	{0, -1.5, -0.5, 0.5, 1.5}	Rater 6	2.0	0.0	{0, -1.2, -0.8, -0.5, 2.5}

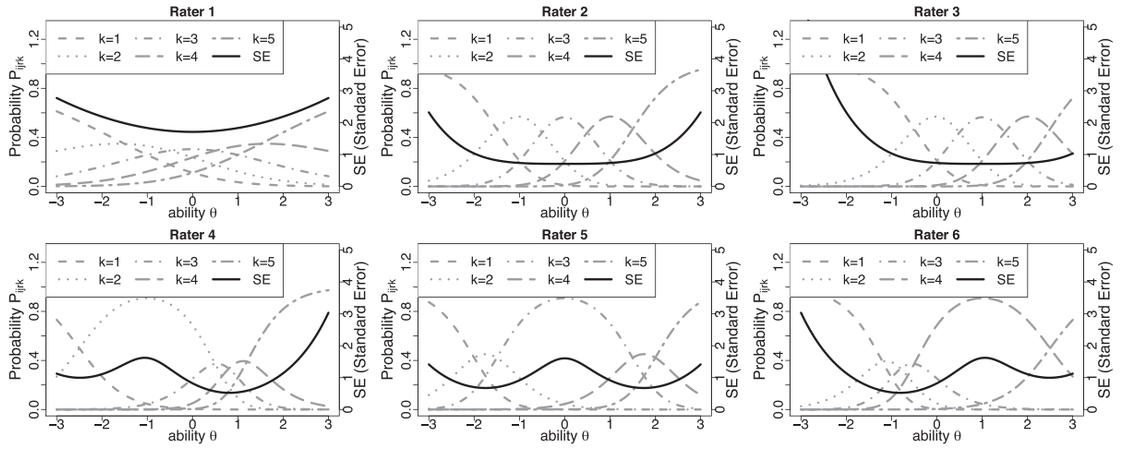


図 1: 異なる 6 名の評価者に対する提案モデルの反応曲線

$$SE_{ir}(\theta) = I_{ir}(\theta)^{-\frac{1}{2}} \ ; \ where \ \ I_{ir}(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\sum_k \ln P_{irk}(\theta)\right)^2 \middle| \theta\right] \tag{5}$$

また，図 4 は，横軸が能力値 θ，第 1 縦軸が反応確率 $P_{irk}(\theta)$，第 2 縦軸が標準誤差 $SE_{ir}(\theta)$ を示し，グレーの破線が各評点に対応する項目反応曲線を，黒の実線が標準誤差関数を表す．

図 4 より，評価の一貫性 α_r が低い Rater1 の分布では，一貫性が高い Rater2 の分布に比べて，評点間での反応確率の差異が小さいことがわかる．これは，同一の能力の受験者に対して評価が一貫しないことを表しており，結果として，Rater1 の標準誤差は能力尺度全域で Rater2 より大きく推定されている．

また，Rater2 に比べて，評価の厳しさ ε_r が大きい Rater3 では，反応曲線が全体として右に移動している．これは，Rater3 から高い評点を得るためには，Rater2 から同じ評点を得るより高い能力が必要であること意味しており，ε_r が評価の厳しさを表現していることがわかる．

尺度範囲の制限は d_{rk} により表現される．d_{rk}（$k \geq 2$）は評点 $k-1$ と評点 k の反応曲線の交点に対応する能力値を表す．$d_{rk+1} - d_{rk}$ が正に大きい値をとることで，評点 k への反応確率が能力尺度広域で高くなり，尺度範囲の制限が表現される．Rater4，5，6 は，Rater2 に比べて尺度範囲の制限が強い評価者である．Rater2 の分布が能力尺度広域に対して均一な標準誤差を示すのに対し，これらの評価者の分布は，同一評点を与えやすい能力区間で標準誤差が大きく推定されている．例えば，Rater4 は，能力が平均以下の受験者に評点 2 を与える傾向が強く，それらの受験者間の能力差を識別しないため，その能力区間で標準誤差が低くなっている．このことから，提案モデルが，尺度範囲の制限に由来する測定誤差の不均一を適切に表現することがわかる．

また，課題の識別力 α_i は評価者の一貫性 α_r と，困難度 β_i は評価者の厳しさ β_r と同様に解釈できる．

次節では，実データに提案モデルと既存モデルを適用し，提案モデルの性能を示す．

4 実データ適用

本研究では，実データとして，34 名の大学生と大学院生に 4 つのライティング課題を与え，全ての成果物を全ての被験者に 5 段階で評価させたデータを用いる．

このデータを用いて，提案モデル，hybrid MFRM，式 (3) のモデル（Uto2016 と呼ぶ），式 (2) のモデル（Usami2010 と呼ぶ）のパラメータを推定した．パラメータ推定は，マルコフ連鎖モンテカルロ法を用いた EAP (Expected A Posteriori) 推定により実行した．

図 2 と図 3 に，評点の使用傾向が類似しており評価の一貫性が異なる 2 名の評価者について，提案モデルと hybrid MFRM の反応曲線を示した．図より，提案モデルでは，評価の一貫性の差異が分布に反映されているのに対し，一貫性を考慮しない hybrid MFRM では，これらの評価者の差異が表現されていないことがわかる．同様に，課題の識別力の差異も hybrid MFRM では表現されない．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 64 2016/08/29 14:00:08

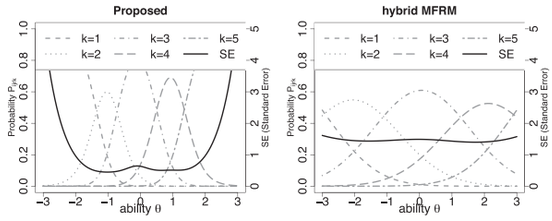


図 2: 一貫性が高い評価者の反応曲線

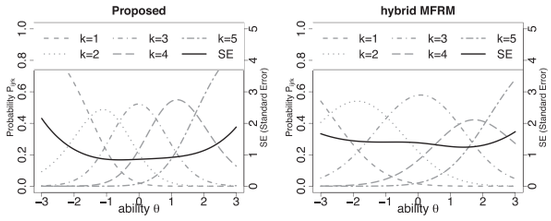


図 3: 平均的な一貫性の評価者の反応曲線

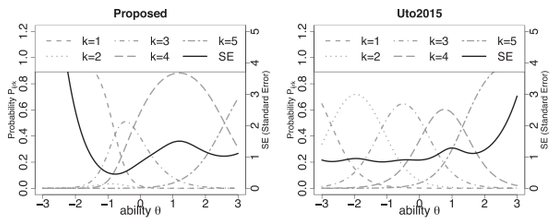


図 4: 尺度範囲の制限が強い評価者の反応曲線

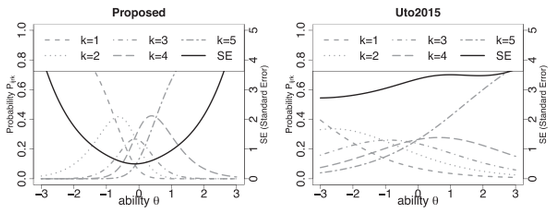


図 5: 尺度範囲の制限が強く一貫性が高い評価者の反応曲線

また，図 4 に，尺度範囲の制限が強い評価者について，提案モデルと Uto2016 の反応曲線を示した．提案モデルの分布から，この評価者は，能力が高い受験者に評点 4 を与える傾向が強く，$\theta = -1.0 \sim 0.0$ 程度の範囲では評点 1,3,4 を使い分けていることがわかる．提案モデルでは，この特性が分布に反映され，能力が高い区間で測定誤差が大きく評価されている．一方で，Uto2016 では，全ての評価者が各評点を一定の割合で使用するとみなすため，この特性が分布に反映されず，尺度範囲の制限に伴う標準誤差の不均一も表現されていない．Uto2016 のこの性質については，Usami2010 でも同様の傾向が見られた．

さらに，尺度範囲の制限は強いが能力に強く相関した評価を行う評価者に対する推定結果を分析するために，そのような特性を持つ評価者データをシミュレーションで生成し，実データに加えて再度パラメータ推定を行った．図 5 に，この評価者に対する提案モデルと Uto2016 の反応曲線を示す．図 5 より，提案モデルでは，この評価者の一貫性が高く推定されていることがわかる．これに対し，Uto2016 では，一貫性が低く推定されている．これは，この評価者による各評点の使用割合が共通の基準を逸脱しているためと解釈できる．一方，Usami2010 では，中心化傾向を表すパラメータ d_r により尺度範囲の制限の一部を表現できるため，必ずしも一貫性の過小評価が起こるとは限らない．しかし，Usami2010 において，尺度範囲の制限が強い評価者の一貫性を大きく推定すると，図 4 の Uto2015 の分布のように，実際には能力測定精度が低い能力区間であるにも関わらず測定精度を過大に評価する傾向がより顕著になり，能力測定精度の低下を引き起こすと考えられる．

以上より，提案モデルでは，既存モデルでは捉えきれない評価者と課題の多様な特性を適切に表現しており，測定誤差の評価も妥当になされていることがことがわかる．提案モデルでは，これらの特性と誤差評価を反映した受験者の能力推定を行うため，高精度な能力測定が可能となる．

5 むすび

紙面の都合上詳細は割愛したが，実データを用いて，提案モデルと複数の既存モデルとの性能比較を行った．情報量基準（周辺尤度，DIC，WAIC）を用いたモデル比較では，すべての情報量基準において提案モデルが最も高いスコアを示した．また，能力測定精度の観点でも，提案モデルは他モデルより優位に高い性能を示した．

参考文献

[1] 宇都雅輝, 植野真臣. パフォーマンス評価のため項目反応モデルの比較と展望. 日本テスト学会誌, Vol. 12, No. 1, pp. 55–75, 2016.

[2] 宇佐美慧. 論述式テストの測定論的問題再考-主要な論点の整理とその現実的解決のために-. 日本テスト学会, 2015.

[3] Masaki Uto and Maomi Ueno. Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, Vol. 9, No. 2, pp. 157–170, April-June 2016.

[4] Thomas Eckes. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments (Language Testing and Evaluation)*. Peter Lang Pub Inc, 2015.

[5] Maomi Ueno and Toshio Okamoto. Item response theory for peer assessment. In *Proc. IEEE International Conference on Advanced Learning Technologies*, pp. 554–558, 2008.

[6] 宇佐美慧. 採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル：MCMC アルゴリズムに基づく推定. 教育心理学研究, Vol. 58, No. 2, pp. 163–175, 2010.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 65 2016/08/29 14:00:09

ピアアセスメントの精度を最適化する評価者選択手法

○ Nguyen Duc Thien[*1]，宇都雅輝[*1]，植野真臣[*1]

電気通信大学大学院情報理工学研究科[*1]

1　はじめに

近年，社会構成主義における学習評価法として，学習者同士による学習成果物の相互評価法を表すピアアセスメントが様々な場面で活用されている [1, 2, 3, 4]．しかし，ピアアセスメントでは，学習者の能力測定精度が評価者の甘さ・厳しさなどの特性に強く依存することが知られている [1, 2]．この問題を解決する手法の一つとして，評価者特性パラメータを付加した項目反応モデルが提案されてきた [1, 2]．これらのモデルにより，評価者のバイアスを考慮した学習者の能力推定を行うことができる [2]．

しかし，学習者を複数の小グループに分割してピアアセスメントを行う場合，これらのモデルを用いても，グループ内の評価者特性分布に学習者の能力測定精度が依存する問題が残る．例えば，信頼性が低い評価者が多数存在するグループでは，そのグループ内の学習者に対して高精度な能力測定は期待できない [5]．このようなピアアセスメントの能力測定精度を改善するためには，グループ構成の最適化が必要といえる．

これまでにも，学習場面におけるグループ構成手法がいくつか提案されてきた [6, 7, 8]．例えば，Lin ら [8] は，学習効果や学習成果の質を最大化することを目的としたグループ構成法を，遺伝的アルゴリズムを用いた手法として提案している．また，Srba and Bielikova [6] や Kardan and Sadeghi [7] は，協調学習の効果を高めるために，年齢や国籍，協調性，言語能力などの特性が類似したグループを構成する手法を提案している．しかし，これらのグループ構成手法は，学習効果の改善を目標としており，能力測定精度が最適化される保証はない．

一方，Nguyen ら [3] は全ての学習者に同等な能力測定精度を与えることを目指し，複数回のピアアセスメントを通して各学習者が出来る限り多様な評価者に評価されるようにグループを構成する手法を，ハミング距離を最大化する整理計画問題として提案している．この手法では，各学習者が多様な評価者に評価されるため，学習者間での能力測定精度の差異を小さくできる．しかし，この手法でも，能力測定精度が最大化される保証はない．

以上の問題を解決するために，本研究では，評価者特性を考慮した項目反応モデルを用いて，ピアアセスメントの精度を最適化するグループ構成手法を開発する．具体的には，各学習者に対する評価者の能力測定精度を表すフィッシャー情報量を最大化するようなグループ構成を，整数計画問題として定式化する．これにより，最適なグループが構成できれば，各学習者に対する能力測定精度を最適化できると期待できる．しかし，シミュレーション実験と実データ実験の結果，提案手法を用いてもランダムにグループ構成を行う場合と比べて，能力測定精度の優位な改善は認められなかった．これは，グループ内の学習者同士でのみ評価を行うという制約が存在すると，情報量最大化を目的関数としても必ずしも全ての学習者に高い情報量の評価者が割り当てられるとは限らないことを意味する．

そこで，本研究では，この条件を緩和し，各学習者に対してグループ外からも数名の評価者を割り当てるグループ外評価者選択手法を提案する．提案手法は，各学習者に対するグループ外評価者のフィシャー情報量の下限を最大化する整理計画問題として定式化する．シミュレーション実験と実データ実験から，グループ外の情報量が高い評価者を採用することが能力測定精度の改善に不可欠であることを示す．

2　ピアアセスメント

本研究では，複数の課題 i（$i = 1, \ldots, N$）を学習者 j（$j = 1, \ldots, J$）に与え，その成果物をピアアセスメントで評価することを考える．このとき，ピアアセスメントは，複数の小グループ g（$g = 1, \ldots, G$）に分割して行い，グループ構成は課題ごとに再構成すると仮定する．

したがって，x_{igj} を，課題 i において学習者 j がグループ g に属している場合に 1，そうでない場合に 0 をとるダミー変数とするとき，課題 i におけるグループ構成は次のベクトルで表現できる．

$$\boldsymbol{x}_i = \{x_{igj} \mid x_{igr} \in \{0, 1\}\}(g = 1, \cdots G, j = 1, \cdots J) \tag{1}$$

また，得られる評価データ $\boldsymbol{U}$ は，学習者 j の課題 i に対する評価者 r（$r = 1, \ldots, J$）の評価カテゴリー k（$k = 1, \ldots, K$）で構成され，以下のように定義できる．

$$\boldsymbol{U} = \{u_{ijr} \mid u_{ijr} \in \{\text{-}1, 1, \ldots, K\}\}(i = 1, \cdots N, j = 1, \cdots J, r = 1, \cdots J) \tag{2}$$

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　66　2016/08/29　14:00:10

ここで，$u_{ijr} = -1$ は欠測データを表す．同一グループ g に学習者 j と r が同時に存在しない場合，すなわち，$x_{igj} = 1$ かつ $x_{igr} = 0$ のとき，$u_{ijr} = -1$ となる．

本研究では，このようなピアアセスメントの精度を改善するために，評価者特性を考慮した項目反応モデルを用いる．

3　ピアアセスメントにおける項目反応理論

本研究では，Uto and Ueno[1] が提案した以下の項目反応モデルを用いる．

$$P_{ijrk} = P_{ijrk}(u_{ijr} = k \mid \theta_j) = P^*_{ijr,k-1} - P^*_{ijrk} \tag{3}$$
$$P^*_{irk}(\theta_j) = \left[1 + \exp(-\alpha_i\alpha_r(\theta_j - b_{ik} - \varepsilon_r))\right]^{-1}$$

ただし，$P^*_{ir0}(\theta_j) = 1$，$P^*_{irK}(\theta_j) = 0$ とする．ここで，α_i は課題 i の識別力，b_{ik} は課題 i において k より大きい得点を得る困難度，α_r は評価者 r の一貫性，ε_r は評価者 r の厳しさ，θ_j は学習者 j の能力を表す．ただし，$b_{i1} < b_{i2} < \cdots < b_{iK-1}$ とする．また，パラメータの識別性のために，$\alpha_{r=1} = 1$，$\varepsilon_{r=1} = 0$ を仮定する．

一般に，項目反応理論では，能力測定精度をフィシャー情報量により評価する．Uto and Ueno[1] のモデルでは，学習者 j に対して課題 i における評価者 r が与えるフィシャー情報量 $I_{ir}(\theta_j)$ を，次のように定義できる．

$$I_{ir}(\theta_j) = {\alpha_i}^2{\alpha_r}^2 \sum_{k=1}^{K} \left[{P^*}_{ijr,k-1}{Q^*}_{ijr,k-1} - {P^*}_{ijrk}{Q^*}_{ijrk}\right]^2 \left[P^*_{ijr,k-1} - P^*_{ijrk}\right]^{-1} \tag{4}$$

ここで，${Q^*}_{ijrk} = 1 - {P^*}_{ijrk}$ とする．

能力 θ_j を所与として $I_{ir}(\theta_j)$ が大きくなる評価者は，受験者 j の能力を精度よく評価できる評価者 r であると解釈できる．したがって，全ての学習者に対してフィッシャ情報量が最大になる評価者を割り当てられれば，能力推定精度を最適化できるといえる．

4　項目反応理論を用いたグループ構成手法

本研究では，評価者特性を考慮した項目反応モデルを用いて，学習者に対するフィッシャー情報量を最大化するようなグループ構成手法を開発する．

4.1　グループ構成手法

グループ構成手法は以下の整数計画問題として定式化する．

$$\begin{aligned}
&\text{Maximize } \textstyle\sum_{j=1}^{J}\sum_{g}^{G} x_{igj}\left[\sum_{r}^{\mathbf{J}\setminus j} I_{ir}(\theta_j)x_{igr}\right] \\
&\text{Subject to} \\
&\quad \underline{J}_g \le \textstyle\sum_j^J x_{igj} \le \bar{J}_g,\ \forall g = 1,\ldots,G \\
&\quad \textstyle\sum_g^G x_{igj} = 1,\ \forall j = 1,\ldots,J \\
&\quad J/G - 1 < \underline{J}_g \le J/G,\ \forall g = 1,\ldots,G \\
&\quad J/G \le \bar{J}_g \le J/G + 1,\ \forall g = 1,\ldots,G
\end{aligned} \tag{5}$$

ここで，一つ目の制約式は，各グループに配分される学習者数を表す．二つ目の制約式は，学習者が一つのグループのみに属することを制約する．残りの制約式は，各グループに所属する学習者数を制約する．目的関数は，これらの制約を満たしつつ，学習者への情報量を最大化する関数として定義される．

以上の整理計画モデルを解くことにより，課題 i において学習者に対するフィッシャー情報量を最大化するグループ構成が得られると期待できる．

4.2　シミュレーション実験

本グループ構成手法の性能を評価するために，ここでは，以下のシミュレーション実験を行った．

1. 学習者数 $J = R = 30$，課題数 $N \in \{3, 4\}$ について，式（3）のモデルの全てパラメータをランダムに生成した．
2. 課題 $i = 1$ に対し，グループ数 $G \in \{3, 5, 6, 7, 10\}$ について，ランダムにグループを構成した．
3. 課題 $i \in \{2, \ldots, N\}$ に対して，次の各手法を用いてグループ構成を行った．（1）提案手法（Proposed），（2）Nguyen ら [3] によるハミング距離最大化手法（Hamming），（3）ランダム構成手法（Random）．ここで，提案手法でグループ構成を行う場合，手順 (1) で生成した真のパラメータを所与とした．
4. 手順 (2) - (3) により得られたグループと，手順 (1) で生成したパラメータを所与とし，各課題に対する評価データを生成した．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　67　2016/08/29　14:00:11

N	G	グループ構成手法		
		Random	Hamming	Proposed
3	3	.266(.059)	.258(.021)	.261(.037)
	5	.418(.057)	.397(.078)	.406(.069)
	6	.425(.070)	.408(.051)	.418(.067)
	7	.481(.090)	.478(.073)	.468(.090)
	10	.533(.084)	.489(.065)	.529(.075)
4	3	.223(.047)	.229(.036)	.246(.037)
	5	.311(.029)	.311(.057)	.324(.067)
	6	.334(.082)	.348(.082)	.372(.077)
	7	.350(.054)	.356(.048)	.386(.069)
	10	.493(.115)	.479(.123)	.492(.101)

表 1　シミュレーションによるグループ構成の能力測定精度結果

N	G	グループ構成手法		
		Random	Hamming	Proposed
4	3	.221(.039)	.209(.019)	.266(.009)
	5	.255(.037)	.276(.033)	.290(.025)
	6	.319(.031)	.287(.027)	.302(.034)
	7	.346(.031)	.336(.039)	.383(.029)
	10	.405(.065)	.382(.052)	.341(.036)

表 2　実データを用いたグループ構成の能力測定精度結果

選択評価者数		$R_u = 1$						$R_u = 2$						$R_u = 3$					
グループ構成手法		Random		Hamming		Proposed		Random		Hamming		Proposed		Random		Hamming		Proposed	
グループ外評価者選択手法		RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED
N =3	G = 3	.254(.057)	.245(.049)	.243(.021)	.237(.030)	.252(.039)	.239(.040)	.252(.055)	.236(.047)	.252(.028)	.226(.023)	.246(.036)	.225(.037)	.245(.055)	.230(.050)	.242(.021)	.219(.026)	.236(.039)	.213(.041)
	G = 5	.392(.042)	.357(.050)	.383(.076)	.346(.072)	.389(.072)	.347(.070)	.379(.045)	.328(.044)	.360(.081)	.317(.075)	.363(.079)	.321(.069)	.367(.053)	.301(.043)	.358(.058)	.304(.066)	.358(.085)	.299(.072)
	G = 6	.404(.076)	.336(.068)	.387(.045)	.326(.054)	.394(.061)	.331(.046)	.382(.075)	.310(.066)	.371(.038)	.298(.046)	.367(.063)	.294(.065)	.340(.065)	.296(.058)	.361(.032)	.281(.049)	.358(.050)	.278(.065)
	G = 7	.428(.080)	.391(.068)	.447(.073)	.406(.065)	.446(.096)	.367(.084)	.411(.072)	.355(.063)	.412(.062)	.366(.061)	.430(.082)	.336(.071)	.392(.073)	.328(.064)	.411(.085)	.333(.054)	.392(.093)	.310(.067)
	G = 10	.470(.082)	.406(.060)	.425(.060)	.391(.062)	.481(.079)	.397(.055)	.437(.080)	.346(.061)	.415(.064)	.331(.055)	.416(.071)	.354(.045)	.423(.086)	.318(.063)	.383(.063)	.305(.055)	.408(.078)	.337(.047)
N = 4	G = 3	.218(.042)	.216(.043)	.223(.032)	.211(.030)	.234(.040)	.227(.034)	.217(.037)	.206(.038)	.220(.033)	.205(.035)	.233(.036)	.217(.042)	.209(.043)	.204(.035)	.217(.037)	.200(.033)	.227(.041)	.206(.040)
	G = 5	.305(.034)	.264(.036)	.299(.047)	.265(.054)	.300(.049)	.265(.059)	.273(.046)	.248(.035)	.279(.043)	.239(.054)	.288(.075)	.239(.055)	.276(.032)	.234(.040)	.275(.052)	.231(.061)	.282(.064)	.219(.049)
	G = 6	.315(.071)	.282(.075)	.328(.088)	.274(.056)	.353(.086)	.293(.088)	.311(.080)	.263(.075)	.325(.095)	.256(.056)	.328(.079)	.266(.087)	.291(.067)	.246(.074)	.296(.077)	.238(.056)	.307(.081)	.243(.075)
	G = 7	.324(.066)	.272(.049)	.327(.041)	.281(.035)	.350(.062)	.290(.053)	.306(.041)	.249(.033)	.293(.039)	.239(.033)	.325(.053)	.260(.044)	.292(.020)	.236(.038)	.272(.052)	.222(.043)	.313(.028)	.241(.051)
	G = 10	.443(.101)	.365(.078)	.415(.122)	.354(.086)	.436(.124)	.359(.085)	.415(.086)	.317(.069)	.389(.083)	.314(.070)	.432(.111)	.309(.075)	.388(.103)	.297(.072)	.405(.087)	.272(.067)	.378(.081)	.272(.059)

表 3　シミュレーションによるグループ外評価者選択の能力測定精度結果

選択評価者数		$R_u = 1$						$R_u = 2$						$R_u = 3$					
グループ構成手法		Random		Hamming		Proposed		Random		Hamming		Proposed		Random		Hamming		Proposed	
グループ外評価者選択手法		RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED	RANDOM	PROPOSED
N = 4	G = 3	.208(.034)	.209(.031)	.207(.017)	.191(.026)	.247(.016)	.260(.011)	.198(.033)	.177(.024)	.194(.017)	.172(.027)	.233(.015)	.219(.012)	.189(.037)	.159(.026)	.189(.023)	.155(.021)	.212(.025)	.181(.009)
	G = 5	.249(.030)	.254(.029)	.264(.034)	.257(.031)	.272(.029)	.275(.021)	.236(.030)	.210(.020)	.255(.036)	.217(.023)	.261(.026)	.240(.013)	.229(.024)	.196(.020)	.223(.034)	.202(.022)	.239(.027)	.218(.008)
	G = 6	.291(.028)	.299(.038)	.273(.028)	.269(.034)	.292(.037)	.292(.023)	.278(.031)	.252(.020)	.254(.026)	.222(.023)	.273(.042)	.225(.016)	.252(.031)	.214(.006)	.246(.022)	.209(.022)	.236(.023)	.196(.009)
	G = 7	.316(.042)	.327(.027)	.305(.037)	.321(.034)	.336(.022)	.373(.025)	.300(.032)	.259(.022)	.292(.030)	.258(.029)	.329(.035)	.277(.014)	.277(.026)	.219(.019)	.275(.023)	.229(.026)	.298(.039)	.222(.012)
	G = 10	.368(.069)	.385(.054)	.361(.047)	.354(.042)	.326(.031)	.350(.027)	.324(.028)	.278(.027)	.332(.048)	.271(.019)	.312(.046)	.232(.014)	.309(.036)	.238(.019)	.297(.048)	.245(.022)	.263(.042)	.215(.010)

表 4　実データを用いたグループ外評価者選択の能力測定精度結果

5. 評価データを用いて学習者の能力測定を行った．ここで，パラメータ推定には EAP（Expected A Posteriori）推定法を用いた．
6. 能力推定値と手順 (1) で生成した真値との平均平方 2 乗誤差値 (以降，RMSE と呼ぶ) を計算した．
7. 以上の手順 1 - 6 を，10 回に繰り返し，RMSE の平均と標準偏差を計算した．

以上の実験結果を表 1 に示す．表 1 より，提案手法の能力測定精度は，ランダム構成手法とハミング距離最大化手法に比べて，必ずしも改善しないことが確認できる．これは，グループ内の学習者同士でのみ評価を行うという制約が存在すると，情報量最大化を目的関数としても必ずしも全ての学習者に高い情報量の評価者が割り当てられるとは限らないことを意味する．

そこで，本研究では，この条件を緩和し，各学習者に対してグループ外からも評価者を割り当てるグループ外評価者選択手法を提案する．

5　項目反応理論を用いたグループ外評価者選択手法

5.1　グループ外評価者選択手法

グループ外評価者選択手法は，グループ g に属する学習者 j に対し，フィッシャー情報量の下限を最大化するように，異なるグループ $g' \in \mathbf{G} \backslash g$ に属する評価者を R_u 人割り当てる整数計画問題として定式化する．

$$
\begin{aligned}
&\text{Maximize } \textstyle\sum_j^J y_{ij} \\
&\text{Subject to} \\
&\qquad \textstyle\sum_{g'}^{\mathbf{G}\backslash g} \sum_{r=1}^{J} I_{ir}(\theta_j) w_{ijg'r} \geq y_{ij},\ \forall x_{igj} = 1 \\
&\qquad \textstyle\sum_{g'}^{\mathbf{G}\backslash g} \sum_{r=1}^{J} w_{ijg'r} = R_u \\
&\qquad \textstyle\sum_{g'}^{\mathbf{G}\backslash g} \sum_{j=1}^{J} w_{ijg'r} \leq J_u
\end{aligned}
\tag{6}
$$

ここで，$w_{ijg'r}$ は，課題 i において，グループ g' に所属している評価者 r が学習者 j に割り当てられた場合に 1，そうでない場合に 0 をとる変数，R_u は一人の学習者に対して割り当てられるグループ外評価者の数を表し，J_u は一人の評価者が評価できるグループ外学習者数の上限を表す．一つ目の制約式は，学習者 j に対するグ

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　68　2016/08/29　14:00:12

ループ外評価者のフィッシャー情報量が y_{ij} 以上であることを制約する．また，2 つ目の制約式は一人の学習者に対する評価者数を制約し，3 つ目の制約式は一人の評価者に対する評価負担を制約する．目的関数は，これらの制約上での情報量の下限 y_{ij} の総和最大化として定義される．

この整数計画問題を解くことにより，グループ外から情報量の高い評価者を割り当てることができる．

5.2　シミュレーション実験

提案手法の有効性を示すために，4.2 節で行ったシミュレーションと同様の実験を行った．ただし，本実験では，手順 3 でグループを構成した後，全ての学習者に対してグループ外評価者を $R_u \in \{1, 2, 3\}$ 名割り当てた．評価者の選択は，本研究で提案したグループ外評価者手法（PROPOSED）とランダム選択手法（RANDOM）の 2 つを用いた．本実験では，$J_u = 12$ とした．

実験結果を表 3 に示す．表 3 より，全ての場合で，ランダム選択手法と比べて，提案手法の RMSE が改善されていることがわかる．また，グループ外評価者数 R_u が増加するほど，提案手法の優位性が顕著になる傾向が確認できた．以上の結果より，グループ外から情報量の高い評価者を採用する提案手法が，能力測定精度の大幅な改善に不可欠であることが示せた．

6　実データ適用

ここでは，実際のピアアセスメントデータを用いて，提案手法の有効性を評価する．本研究では，34 名の被験者に 4 つのライティング課題を与え，全ての成果物を全ての被験者に 5 段階評価させたデータを用いる．

本実験では，完全データを用いて項目反応モデルのパラメータを推定した後，4.2 節と 5.2 節のシミュレーションと同様の実験を行った．

それぞれの実験結果を表 2 と表 4 に示す．表 2 から，シミュレーション同様，項目反応モデルを用いたグループ構成手法が，ランダム構成法とハミング距離最大化法に比べて能力測定精度を必ずしも改善できなかったことがわかる．一方で，グループ外から評価者を加えた場合，表 4 より，2 名以上の評価者を追加すると，グループ外評価者手法がランダムな選択手法に比べて能力測定精度を改善できていることが確認できる．$R_u = 1$ の場合に能力測定精度の改善が認められなかった原因としては，本研究で採用した項目反応モデルがデータに適合しておらず，モデルにより有益と解釈された評価者が実際には必ずしも適切ではなかった可能性が挙げられる．

7　むすび

本研究では，グループで行うピアアセスメントの精度を改善するために，項目反応モデルを用いたグループ構成手法を開発した．しかし，実験から，提案したグループ構成手法が，必ずしも能力測定精度が向上できるとは限らないことが示された．この問題を解決するために，グループ内のみで評価を行うという制約を緩和し，グループ外から評価者を追加割り当てするグループ外評価者選択手法を提案した．実験の結果，提案手法が能力測定精度を大幅に改善できることを示した．

実データ実験の考察で示したように，本研究で用いた項目反応モデルは本研究で扱った実データには適合していなかった可能性がある．今後は情報量基準を用いたモデル選択により最適なモデルを選択して採用する．また，より現実的な適用場面を想定した提案手法の適用方法についても今後の検討課題とする．

参考文献

[1] M. Uto and M. Ueno, “Item response theory for peer assessment,” IEEE Trans. Learn. Technol., vol.9, no.2, pp.157–170, April 2016.
[2] 宇都雅輝，植野真臣，“パフォーマンス評価のための項目反応モデルの比較と展望，” 日本テスト学会誌，vol.12，no.1，pp.55–75，2016.
[3] T. Nguyen, M. Uto, Y. Abe, and M. Ueno, “Reliable Peer Assessment for Team-project-based Learning using Item Response Theory,” Proc. 23rd Int. Conf. Comp. Educ., pp.144–153, Dec. 2015.
[4] 植野真臣，荘島宏二郎，学習評価の新潮流，朝倉書店，2010.
[5] Z. Wang and L. Yao, “The Effects of Rater Severity and Rater Distribution on Examinees’ Ability Estimation for Constructed Response Items,” ETS Research Report Series, vol.2013, no.2, pp.i–22, 2013.
[6] I. Srba and M. Bielikova, “Dynamic Group Formation as an Approach to Collaborative Learning Support,” IEEE Trans. Learn. Technol., vol.8, no.2, pp.173–186, April 2015.
[7] A.A. Kardan and H. Sadeghi, “An efficacious dynamic mathematical modelling approach for creation of best collaborative groups,” Math. Comput. Model. Dyn. Syst., vol.22, no.1, pp.39–53, 2015.
[8] Y.S. Lin, Y.C. Chang, and C.P. Chu, “Novel approach to facilitating tradeoff multi-objective grouping optimization,” IEEE Trans. Learn. Technol., vol.9, no.2, pp.107–119, April 2016.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　69　2016/08/29　14:00:13

IRTによる理系記述式テストデータの分析(2)
――記述式テストIRT化検討の背景――

○倉元直樹（東北大学）[i]・泉 毅（株式会社教育測定研究所）

従来，少数の例外を除いて，わが国のハイステークスな試験でIRTは積極的に活用されてこなかった。しかし，ここ数年の大学入試改革の議論の中でにわかにIRTに対する期待が高まってきた。

口火を切ったのは文部科学省の委託事業報告書として発表された『高等学校段階の学力を客観的に把握・活用できる新たな仕組みに関する調査研究』(北海道大学，2010) で検討された「高大接続テスト (仮称)」である。当否は別として，IRTを用いれば目標準拠型テストが実現され，それによって大学入試を巡る様々な問題の解決が可能であるかのように謳われた。推薦・AO入試が学力不問で大学入学可能なバイパスになっているとして，中央教育審議会で議論が進められていた入試改革の議論の中で提案された推薦・AO入試の学力把握措置の一つとしての位置づけであった (中央教育審議会，2008a)。同答申では「学力を客観的に把握する方法の一つとして一定の意義があると考えられる一方，高等学校教育の在り方との関係上，留意すべき点も種々あることから，高等学校及び大学関係者間の十分な協議・研究が行われることを期待する (中央教育審議会，2008b)」と位置づけられたが，「高大接続テスト (仮称)」に関する議論は深まらず，構想は立ち消えとなった。

2013年10月には第2次安倍内閣の官邸の下に置かれた首相の私的諮問機関である教育再生実行会議の第四次提言 (教育再生実行会議，2013) で，大学入試の共通試験として「達成度テスト (仮称)」の導入が提言された。同提言では「達成度テスト (仮称)」には，基礎レベルと発展レベルの二つの水準が設定されていた。そのうち，基礎レベル (仮称) は「高等学校の基礎的・共通的な学習の達成度を客観的に把握し，学校における指導改善に活かす」目的を持ったものとされ，「高等学校在学中に複数回受験できる仕組み」を検討することが提言された。発展レベル (仮称) は「大学が求める学力水準の達成度の判定に積極的に活用」することを目的としたテストであり「試験として課す教科・科目を勘案し，複数回挑戦を可能」とすることとされていた。さらに，発展レベル (仮称) においては「将来的には，試験問題データを集積しCBT（Computer Based Testing）方式で実施することや，言語運用能力，数理論理力・分析力，問題解決能力等を測る問題の開発も検討する」とされている。注釈に記載されたCBTの説明には「コンピューター〔ママ〕を利用した試験方式。数千～数万題の問題の中から，難易度が同じとなるよう問題を組み合わせて出題することにより，複数回受験しても安定した成績を示すことが可能となる（例 TOEFL，医学部共用試験）」とある。IRTに基づく運用が示唆されたものと推測される。

IRTへの期待と同時並行的に従来から一般的な解答形式である記述式問題に対する期待も高まってきた。教育再生実行会議の議論を受け，2014年3月に出された中央教育審議会高大接続特別部会の審議経過報告 (中央教育審議会高大接続特別部会，2014) には，達成度テスト (仮称) の在り方として「記述式を導入すること」「紙媒体ではなくコンピュータによる出題・回答の方式を導入すること」の専門的な検討を進めることが盛り込まれた。また，複数の試験機会の得点を比較可能とするには「IRT (項目反応理論) 等を用いた得点調整，得点表示方式」について検討が必要であるとされた。テスト内容にも新機軸が求められ，例えば「複数の教科・科目にまたがった内容に基づきその活用力や応用力を測る『合教科・科目型』や，教科の枠組みにとらわれない『総合型』の導入に向けて専門的な検討を進めるべきである」との記述が見られた。IRTという一般になじみの薄い学術用語も頻繁に行政文書に登場するようになった。「試験を複数回実施する場合，複数の試験間の得点を比較可能とすることが必要であり，IRT (項目反応理論) 等を用いた得点調整，得点表示方式」についての検討が必要である」といった具合である。

同年12月に公表の中央教育審議会答申 (中央教育審議会，2014) では，共通試験の名称が「高等学校基礎学力テスト (仮称)」，「大学入学希望者学力評価テスト (仮称)」と改称された。実質的な中身は「達成度テスト (仮称)」と本質的に同じであり，「合教科・科目型」「多肢選択方式だけでなく，記述式を導入」といったアイデアが継承されている。さらに「年複数回実施」「CBT方式での

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 70 2016/08/29 14:00:14

実施を前提に開発を行う」といった形で踏み込んだ提案がなされた。「高大接続テスト (仮称)」との最大の相違点は，大学入試における現在の共通試験である大学入試センター試験を廃止し，それにとって代わる試験として提案されたことである。

中央教育審議会の議論が答申にまとまり，大学入試改革の議論は次の段階に進んだ。答申を実施に移すことを目的に，高大接続システム改革会議に制度設計にかかわる具体的な検討が委ねられたのである。2015 年 9 月の高大接続システム改革会議「中間まとめ」 (高大接続システム会議，2015) では，IRT，CBT の導入に関する検討と同時に記述式テストの導入に関する議論が見られた。

しかし，以上のような政策的議論が IRT の原理やモデルとしての性質，制約条件をも踏まえて行われているとは考えにくい。我が国のテスト文化の中で IRT を基盤とする試験を実施し，軌道に乗せていくには，検討すべき課題は多い。例えば，村上 (2003) は受験機会が複数設定され，それぞれの成績を同一の尺度上での比較することに関して，我が国では理論的基盤と試験実施のためのバックアップ態勢が弱いと指摘している。IRT に基づくテストを実用に供するまでには，我が国のテスト慣行に合わせて数々の実証的研究に基づく議論を積み重ねていく必要がある。その後，最終報告 (高大接続システム改革会議，2016) では議論の焦点が記述式テストの共通試験への導入という方向に向かい，現在は記述式テストに対して IRT 適用が求められる状況ではない。しかし，大学入試改革に関する議論の流れが今後も継続するならば，将来的に理系記述式テストを念頭に置いた IRT の運用が議論の俎上に上ることも考えられる。

本研究では既存の理系記述式テストデータの分析を通じて我が国の大学入試における記述式テストへの IRT 適用の課題を探り，その一部にテスト理論的な観点から実証的検討を加えることとした。なお，すでに本研究のうち，生物の試験問題関する分析の一部は泉・倉元 (2014) において既報である。本研究はそれに続く成果について報告する。

文献

中央教育審議会 (2008a). 学士課程教育の構築に向けて (答申) 参考資料 6，平成 20 年 12 月 24 日 (http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2008/12/26/1217067_004.pdf, 最終閲覧日平成 28［2016］年 8 月 3 日)

中央教育審議会 (2008). 学士課程教育の構築に向けて (答申)，平成 20 年 12 月 24 日 (http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2008/12/26/1217067_001.pdf, 最終閲覧日平成 28［2016］年 8 月 3 日)

中央教育審議会高大接続特別部会 (2014). 中央教育審議会高大接続特別部会審議経過報告，平成 26 年 3 月 25 日 (http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2014/04/01/1346157_1.pdf, 最終閲覧日平成 28［2016］年 8 月 3 日)。

北海道大学 (2010). 高等学校段階の学力を客観的に把握・活用できる新たな仕組みに関する調査研究，文部科学省先導的大学改革推進委託事業調査研究報告書，平成 22 年 9 月 30 日 (http://www.mext.go.jp/a_menu/koutou/itaku/08082915/__icsFiles/afieldfile/2010/11/04/1298840_1.pdf，最終閲覧日平成 28［2016］年 8 月 3 日).

泉毅 (2016). 理系記述式テストへの IRT 適用課題の検討，東北大学大学院教育情報学教育部博士後期 3 年の課程学位請求論文.

泉毅・倉元直樹 (2014). IRT による理系記述式テストデータの分析――高校生対象の生物テストデータを用いて――，日本テスト学会第 12 回大会発表論文集，170-173.

高大接続システム改革会議 (2016). 最終報告，平成 28 年 3 月 31 日.
(http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2016/06/02/1369232_01_2.pdf 最終閲覧日平成 28［2016］年 8 月 3 日).

教育再生実行会議 (2013). 高等学校教育と大学教育の接続・大学入学者選抜の在り方について (第四次提言)，平成 25 年 10 月 31 日 (http://www.kantei.go.jp/jp/singi/kyouikusaisei/pdf/dai4_1.pdf，最終閲覧日平成 28［2016］年 8 月 3 日).

村上隆 (2003). 研究の背景と目的. 村上隆 (編) 我が国の公的機関における得点等化の導入に向けた心理・教育測定論的研究，平成12.～14 年度日本学術振興会科学研究費補助金 (特別推進研究費 {1})，課題番号 12800015，研究代表者　村上隆，研究成果報告書，1-11.

付記

本稿は泉 (2016) の「序章」に大幅に加筆修正を加えたものである。また，本稿は JSPS 科研費 15K13124, 16H02051 の助成に基づく研究成果の一部である。

[i] 連絡先：倉元直樹 (Naoki T. Kuramoto)
Email: ntkuramt@m.tohoku.ac.jp

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 71 2016/08/29 14:00:15

IRT による理系記述式テストデータの分析(3)
――項目パラメタ・能力パラメタによる IRT 適用課題の導出――

○泉 毅（ 株式会社教育測定研究所 ）・ 倉元 直樹 （ 東北大学 ）

1. 目的・背景

1.1. 目的

本稿では，倉元・泉(2016)の議論を受け理系記述式テストデータの IRT モデルとの整合性に関する問題を扱う。

本研究で対象とする理系記述式テストとは，短答式，数式を用いた解答，図や絵を用いた解答，穴埋め式などの解答形式で出題される理系分野の学力測定を目的とするテストを言う。本稿では，理系分野として高等学校の数学，理科を想定している。なお，一つのテストの中に記述式で解答を求める設問とともに，真偽式や多枝選択式などの選択式の解答形式の設問も含まれることがあるが，全体として理系記述式テストと呼ぶことにする。

本稿は，設問が形式的に局所独立の仮定を満たすことが難しい場合，逸脱がどの程度パラメタの推定に影響を及ぼすのか，実際のデータから項目パラメタ，能力パラメタを算出し，それらをもとに検討を加えることを目的とする。

1.2. 局所独立の仮定に関する技術的問題点

局所独立の仮定は，IRT を用いた分析を行う際に必要とされる重要な仮定であるが，理系記述式テストでは大問形式が主流である。大問形式での出題は長文問題のようなあるテーマに沿った出題，また前の問題が解けなければ後の問題が解けない出題等が考えられる。これらの形式では原理的に局所独立の仮定を満たすことができない。

Yen (1993) は，局所独立の仮定を満たさないことを局所依存 (LID) と呼び，LID を引き起こす原因を複数挙げている。記述式かつ大問形式のテストにおいて，LID を引き起こす原因として文脈依存と項目間の連鎖性が挙げられる。

文脈依存とは，複数の項目間の間接的な相互関係性を指す。大問形式のテストにおいては，一つのリード文のもとにある一定のテーマが与えられ，それに沿った複数の小問が出題される。文脈依存の典型的な出題形式と言える。

項目の連鎖性とは，二つの項目間のより直接的な相互関係性を表す。例えば，数式を展開させるような設問の場合，直前の問題に正解できなければその次の設問に正答できないというような構造の出題形式によくみられる。

1.3. 項目の連鎖性の分類

Yen (1993) の連鎖性の定義は「二項目間に正答率の影響があること」とされる。項目の関係性と影響の強さに関して明示的な言及はない。本稿では，現実のテスト場面に即して項目間の関係性の違いに着目して下位分類を行った。

テストに含まれる任意の二つの項目のペアを，出題順に「前の項目」「後の項目」と表現し，受験者群のそれぞれの項目への「正答，誤答」の解答パタンを考える。受験者の解答は，必ず「前の項目に正答かつ後の項目に正答」「前の項目に正答かつ後の項目に誤答」「前の項目に誤答かつ後の項目に正答」「前の項目に誤答かつ後の項目に誤答」パタンいずれかに属する。項目ペアの連鎖性の違いを 4 パタンの出現頻度の違いとして表現した。

1.3.1. 実質的同一項目

二つの項目に関する正答が同一の操作で同時に導かれる関係である。前の項目の正答情報さえあれば後の項目に必ず正答できるようなケースで，正誤情報はほぼ同一となる。モデル上の出現パタンは「前の項目に正答かつ後の項目に正答」と「前の項目に誤答かつ後の項目に誤答」の二つに限られる。「実質的同一項目による連鎖項目ペア」と呼ぶ。

1.3.2. 完全連鎖項目

前の項目の解答情報を用いて後の項目の解答を導く項目構造を指す。この場合，前の項目に不正解であると後の項目は必ず不正解となる。「完全連鎖項目ペア」とする。

完全連鎖項目ペアの場合，前の項目に正答した受験者は後の項目にも正答する可能性がある。しかし，前の項目に誤答した受験者は，後の項目には正答することが出来ない。よって，原理的には「前の項目に誤答かつ後の項目に正答」以外の三つの解答パタンの出現があり得る。

1.3.3. 部分連鎖項目

より連鎖性が弱い項目ペア「部分連鎖項目」と呼ぶ。4 パタンすべての解答パタンをとる可能性があるが，相対的に「前の項目に正答かつ後の項目に誤答」，「前の項目に誤答かつ後の項目に正答」のパタンが相対的に少ない。

1.3.4. 連鎖性がない項目

項目ペアに連鎖性がない場合は，4 パタンすべての解答パタンをとる。連鎖性がない項目ペアでも，同じリード文を共有する等，他のいくつかの項目と文脈依存性があるために，完全には局所独立を満たしていないケースもある。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 72 2016/08/29 14:00:16

2. 方法

2.1. 分析対象者

本研究では，倉元 (2003) で作成，実施されたテストデータを用いる。このテストは大学進学を目指す高校 3 年生の生徒, 約2, 900 名の参加のもとに解答を得たものであり，数学分野，物理分野，化学分野，生物分野から出題がなされた (倉元, 2003)。全項目無解答者を除いた解答データを分析の対象とした。分析対象者数を表 1 に示す。

表 1. 各テストの分析対象者数

	分析対象者数
数学	2733
物理	1776
化学	2639
生物	946

2.2. 分析モデル

項目パラメタの推定に大きな負荷がかからない単純なモデルを用いることとした。具体的には，正誤のデータ，すなわち 2 カテゴリのデータには 2 パラメータロジスティックモデル (2PLM) を, 2 カテゴリよりも多い多値データには段階反応モデル(GRM: Samejima, 1969) を採用する。

2.3. 分析方法

まず, 各教科の一次元性をスクリープロットで確認する。次に二値データならば，2PLM，多値データならば GRM によって分析し, 項目パラメタ・能力パラメタの推定を行う。

同じデータを再分類し，カテゴリ数が 2 の場合と多値の場合に分けて分析した。内容から見て合理的な部分点とするため，各作題者へのヒアリングをもとに閾値を定めた。作題者には本テストの項目得点のヒストグラムを提示し，妥当な閾値について判断を仰いだ。本稿では, 二値型項目，多値型項目と呼ぶ。また，連鎖性の構造についてもヒアリングを行い，連鎖性をもつ項目の把握を行った。

IRT への適用に際して，能力パラメタが標準正規分布する際に想定される識別力パラメタ推定値の値を 0〜2.0 とし，この値の範囲に収まるか否かという観点から検討を行った。2.0 以上の推定値を, 識別力の過大推定とみなすこととした。また，2PLM，GRM の能力パラメタ推定値と素点合計の散布図，相関係数を求め，分析法による違いを検討した。

識別力パラメタが過大推定された項目が連鎖性のある項目であれば, その項目群をテストレットとして再度 2PLM，GRM，能力パラメタに関する分析を行った。なお，分析には IRTPRO ver.2.1 (Cai, Thissen & du Toit, 2011) を用いた。

3. 結果

紙面の都合上，項目数が 4 分野の中で最も多かった物理分野の結果の一部について記載する。

3.1. 基礎統計量

物理分野のテストの構成は，大問 1 に 4 項目 (item2_1_01 〜 item2_1_04) ，大問 2 に 2 項目 (item2_2_01, item2_2_02) , 大問 3 に 5 項目 (item2_3_01 〜item2_3_05) が含まれ，計 11 項目であった。またそれぞれの項目には 3〜9 点の配点がなされ，計 50 点満点であった。

また，物理分野のテストでは 11 項目中 4 項目が多枝選択式の解答形式であった。その他の 7 項目は記述式の解答を求める設問であり，図を用いた解答，短答式，論述式が各 1 項目，数式を用いた解答が 4 項目であった。各項目の解答形式・配点・基礎統計量を表 2 に示す。

3.2. 連鎖性

物理分野は大問 1 と大問 2 に含まれる項目，すなわち Item2_1_01〜Item2_1_04，Item2_2_01，Item2_2_02 は独立した項目であることが分かった。

表 2. 各項目の解答形式・配点・基礎統計量 (物理分野)

項目	解答形式	配点	平均得点	得点率	標準偏差
Item2_1_01	多枝選択式	3	2.18	72.5%	1.34
Item2_1_02	多枝選択式	3	2.20	73.2%	1.33
Item2_1_03	多枝選択式	3	2.31	77.1%	1.26
Item2_1_04	多枝選択式	8	2.91	36.3%	2.65
Item2_2_01	図を用いた解答	3	0.06	2.0%	0.42
Item2_2_02	短答式	4	1.16	29.1%	1.63
Item2_3_01	数式を用いた解答	9	4.62	51.4%	3.80
Item2_3_02	論述式	3	0.38	12.7%	1.00
Item2_3_03	数式を用いた解答	4	0.86	21.6%	1.20
Item2_3_04	数式を用いた解答	5	0.65	13.1%	1.46
Item2_3_05	数式を用いた解答	5	0.04	0.8%	0.44

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 73 2016/08/29 14:00:17

大問 3 に関しては，Item2_3_01 と Item2_3_02，Item2_3_03 と Item2_3_04, Item2_3_04 と Item2_3_05, が部分連鎖の項目ペアとされた。

3.3. スクリープロット

一次元性を確認するため，配点をもとにした素点合計点から求めたスクリープロットを描いた。物理分野のスクリープロットを図 1. に示す。第一固有値と第二固有値に格段の差があると判断し，分析を続けた。

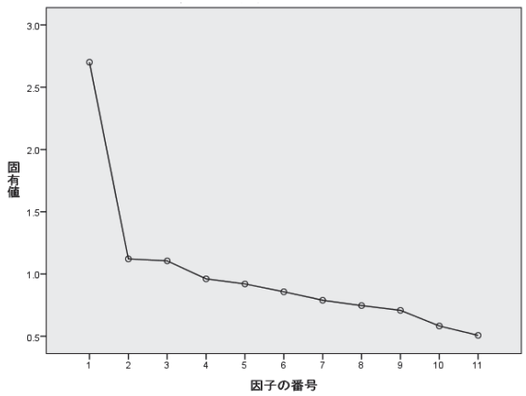


図 1. 物理分野 11 項目のスクリープロット

3.4. 項目パラメタ

二値型項目分析における，物理分野の項目パラメタの推定値を表3に示す。物理分野は item2_3_01, item2_3_03, item2_3_04，item2_3_05 の識別力パラメタが過大推定された。また，連鎖性がないが，item2_1_01，item2_1_02, item2_1_03 の識別力パラメタには，通常よりも小さい推定値が得られた。

表 3. 物理分野 11 項目（二値型）の項目パラメタ

項目	*a*	*s.e.*	*b*	*s.e.*
item2_1_01	0.17	0.06	-5.61	2.06
item2_1_02	0.35	0.07	-2.93	0.55
item2_1_03	0.09	0.07	-14.28	11.29
item2_1_04	1.11	0.11	0.25	0.09
item2_2_01	1.09	0.29	4.07	0.97
item2_2_02	1.11	0.11	0.60	0.12
item2_3_01	2.43	0.21	-0.08	0.06
item2_3_02	1.42	0.14	1.79	0.16
item2_3_03	2.51	0.23	0.40	0.08
item2_3_04	2.24	0.28	1.05	0.14
item2_3_05	3.29	2.16	2.70	0.55

各分野において一部の項目をテストレットとして扱い項目パラメタの算出を行った。テストレットを含む二値型項目分析における項目パラメタの算出結果を表 4 に示す。作題者への連鎖性の構造に関するヒアリングの結果に基づき，物理分野は部分連鎖の関係にあるとされた item2_3_01 と item2_3_02 をテストレット (testlet2_3_01_02) とし，また，同様に部分連鎖の関係にあるとされた item2_3_03, item2_3_04，item2_3_05 の 3 項目をテストレット (testlet2_3_03_04_05) とした。識別力が通常より小さく推定された item2_1_01，item2_1_02，item2_1_03 は連鎖性のない項目であった。これらの項目はテストレットとして扱うことは不適切であると判断し，テストレットとせず分析を行った。

表 4. テストレットを含む
物理分野 8 項目（二値型）の項目パラメタ

項目	*a*	*s.e.*	*b1*	*s.e.*	*b2*	*s.e.*	*b3*	*s.e.*
item2_1_01	0.17	0.06	-5.64	2.11				
item2_1_02	0.36	0.07	-2.89	0.54				
item2_1_03	0.08	0.07	-14.85	12.30				
item2_1_04	1.14	0.09	0.24	0.05				
item2_2_01	1.12	0.23	4.01	0.65				
item2_2_02	1.15	0.09	0.59	0.06				
testlet2_3_01_02	2.17	0.16	-0.14	0.04	1.58	0.07		
testlet2_3_03_04_05	2.53	0.22	0.26	0.04	1.21	0.06	3.01	0.17

その結果，二つのテストレットとした項目 (testlet2_3_01_02 と testlet2_3_03_04_05) で，再び識別力パラメタが過大推定された。また，二値型項目分析と同様，識別力パラメタが非常に低く，かつ困難度パラメタと標準誤差が非常に高い，推定結果が極めて不安定である項目がみられた (item2_1_01，item2_1_02，item2_1_03) 。

3.5. 能力パラメタ

二値型項目における能力パラメタ，多値型項目における能力パラメタ推定値の散布図を図 2 に示す。これは，2PLM と GRM の二つのモデルに基づいて各受験者の能力パラメタを推定し，その結果を散布図に表したものである。

その結果，能力パラメタ推定値が 0.5 以上，-0.5 以下の部分において，推定された能力パラメタ値の絶対値が大きいほど二つの分析による能力値の差に開きがある受験者が増加していく様子が見られる。

また，二値型項目における能力パラメタ，多値型項目における能力パラメタ推定値の相関係数の値は r=.966 であった。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 74 2016/08/29 14:00:18

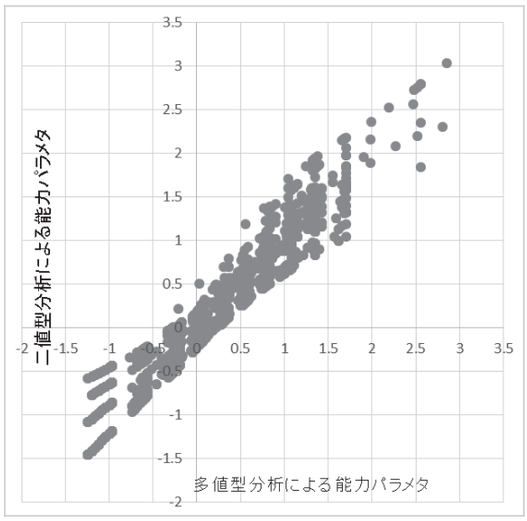


図 2. 物理二値型と多値型の能力パラメタ散布図(r=. 966)

4. 考察

物理分野においては，二値型にせよ，多値型にせよ，局所独立の仮定の逸脱は識別力パラメタの過大推定につながることが示唆された。さらにテストレット化によって局所独立を解消ないしは軽減するという方略も効果がなかった。

識別力パラメタの安定した推定という観点からは，二値型項目として分析した場合でも多値型項目として分析した場合でも，分析結果は変わらなかった。したがって，本研究の結果からは，部分点の扱いの違いが大きく識別力パラメタの推定に影響する可能性は小さいことが示唆された。

一方，他の項目との連鎖性があり，局所独立の仮定を満たさない項目については，項目パラメタの推定に悪影響を及ぼすことが実証的に示された。構造的に局所独立の仮定を逸脱する項目のペアが存在しない生物分野のテストデータに対する分析 (泉・倉元，2014) は，受容可能な範囲に識別力パラメタが収まるなど，項目パラメタの推定が比較的上手く行っていたのに対し，連鎖性がある項目を含む数学分野，物理分野，化学分野は，いずれも識別力パラメタの過大推定がみられた。しかも，識別力パラメタが過大推定された項目は，いずれも他の項目と構造的に完全連鎖，ないしは，部分連鎖という形で直接的な連鎖性がある項目であった。

能力パラメタの観点からは，どの分野においても，二値型項目分析と多値型項目分析から得られた能力パラメタ推定値の相関係数の値は.95 程度の，高い値となった。しかし，散布図を描くと，能力パラメタ推定値の絶対値が大きくなるにつれ，二値型項目分析と多値型項目分析の推定の差異が見られ，直線的な関係とは言えないことが示された。

5. IRT 適用の課題

本研究の結果からは，どの程度の強さの関係性が識別力パラメタの過大推定につながるのか，識別力パラメタの推定という観点からは，どの程度の局所独立の仮定からの逸脱であれば許されるのか，といった点に関しては，課題が残った。

また，一連の分析に用いた項目数は 6〜11 項目と一般的な客観式テストと比較すると非常に少なかった。IRT モデルによる分析の上で，項目数の確保が難しいということは，理系記述式テストの宿命とも言える本質的な構造的問題である。識別力パラメタの過大推定をはじめとして，項目パラメタの推定が不安定になるという問題は，項目の局所独立の仮定の逸脱という問題のみならず，推定に必要な項目数が確保できていないことに由来する可能性が考えられる。

少なくとも局所独立の仮定の逸脱の程度による影響の把握，また，項目数確保は，記述式テストへ IRT 適用を行うことを考えるならば，避けて通れない難題であると言えるだろう。

引用文献

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows. [Computer software]. Lincolnwood, IL: Scientific Software International

泉毅・倉元直樹 (2014). IRT による理系記述式テストデータの分析――高校生対象の生物テストデータを用いて――, 日本テスト学会第 12 回大会発表論文集, 170-173.

倉元直樹 (2003). 高校と大学の教育接続を重視した試験問題開発研究――モニター調査結果報告――, 夏目達也 (編) 高校と大学のアーティキュレーションに寄与する新しい大学入試についての実践的研究, 平成12〜14年度日本学術振興会科学研究費補助金 (基盤研究 [A]), 研究課題番号 12301014, 研究代表者 夏目達也, 研究成果報告書, 99-175.

倉元直樹・泉毅 (2016). IRTによる理系記述式テストデータの分析(2)――記述式テストIRT化検討の背景――, 日本テスト学会第14回大会発表論文集.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometorika Monograph Supplement*, **34**, 100-114.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, **30**, 187-213.

付記

本稿は JSPS 科研費 15K13124, 16H02051 の助成に基づく研究成果の一部である。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 75 2016/08/29 14:00:19

一般セッション４：初年次教育

９月９日（金）10:30 – 11:50　第２会場（B102）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　77　2016/08/29　14:00:19

オンライン英会話グループ学習を用いた英語学習態度改善の取り組みと測定

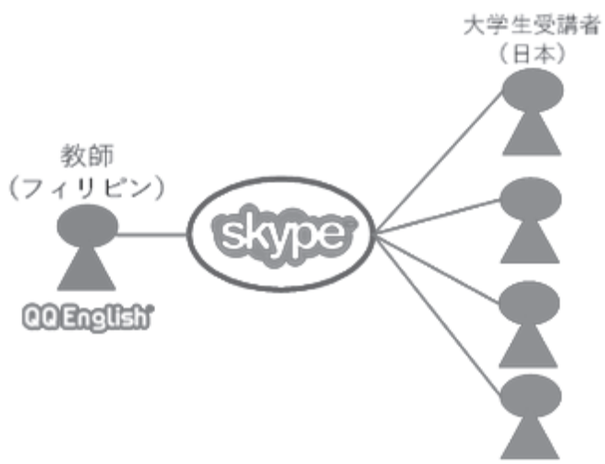
森　一将*，　遠山　道子*，新谷　真由**，大森　拓哉***
*文教大学，**東京電機大学，***多摩大学

1.　はじめに

本研究では、大学生に対し、オンライン英会話サービスを用いたグループ学習を課し、英語学習態度や、語彙・文法の運用能力の変化を計量的に測定し、これらの改善に影響を与える要因を統合的に検討する。オンライン英会話とは Skype などのインターネットツールを用い、主に日本国外の英語教師を用いて行う英会話教育サービスであり、コストが安く授業の質が比較的高いという利点がある。本研究により、英語の運用に苦手意識を抱える大学生の態度改善へのオンライン英会話グループ学習の適用可能性や、英語科目の補習としてのオンライン英会話の活用の可否を明らかにする。

2.　オンライン英会話グループ学習の概要と実験の目的

オンライン英会話では、Skype のグループ通話機能を用いると、1 人の英語教師と複数の学生受講者を結び、オンラインを介して英会話のグループ学習を行うことが可能になる。本研究ではこのオンライン英会話グループ学習（Online English Group Learning; OEGL）を用いて、大学生の英語学習態度や、語彙・文法の運用能力の変化を計量的に測定し、これらの改善に影響を与える要因を統合的に検討する。



（図 1）OEGL の概要

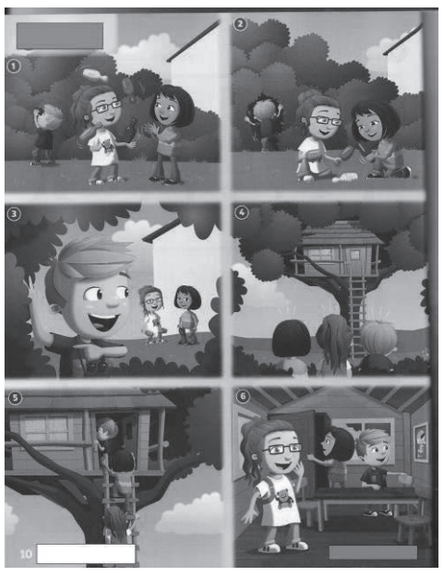
230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　76　2016/08/29　14:00:20

3. 実験の概要と手順

実験は 2016 年 7 月にある私立大学の 3 年生(13 名)を対象に以下のような手順で行われた。

① 事前に学生の英語経験(留学経験、英語資格の有無など)をアンケートで調査した。
② 実験前の学生の英語や英語学習への態度を心理尺度を用いて測定した。測定に用いた尺度は、英語スピーチへの恐怖感(Horwitz et al. 1986,Kimura, 2008)、英語学習への動機付け(Tanaka and Yamauchi,2000)、英語学習への自己効力感(Pintrich and De Groot,1990, 牧野,2023)を用いた。
③ 学生に英語スピーチテストを実施した。テストは市販の英語テキストから引用したイラスト(図 2)を用い、イラストの内容について回答させるというものである。
④ 学生に OEGL のレッスン(25 分間)を 4 回にわたって実施した。
⑤ 全レッスン終了後、②、③と同様に心理尺度による測定、スピーチテストを行った。

これらの実験後、②、③、⑤で測った心理尺度の変化やスピーチテストの評価の差異を①で調査した英語経験の違いを基に分析し、英語学習態度や、語彙・文法の運用能力の変化や改善に影響を与える要因を明らかにする。合わせて大学において OEGL を導入した場合の運用面での問題を明らかにし、英語科目の補習としての導入可能性を検討する。



(図 2) スピーチテストに用いたイラストの例

実験の結果や考察については、当日発表する。

補足：オンライン英会話サービスについて

本実験で利用したオンライン英会話サービスは株式会社 QQEnglish の協力により提供を受けたものである。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 77 2016/08/29 14:00:20

工学系大学の数学基礎教育における e テスティングの取り組み

○亀田真澄*　　宇田川暢**
*山陽小野田市立山口東京理科大学　共通教育センター
**山口県立大学　学術情報センター　情報基盤部門

【要旨】 本講演は，工学系大学の講義：微分方程式においてオンラインテスト（以下，「e-Test」という）を実行した取り組みについて紹介する。この e-Test は，学習管理システム Moodle[1]，数式処理システム Maxima[2]，数式解答評価システム STACK[3]，組版処理システム AMS-LaTeX[4]，JavaScript 数式ライブラリ MathJax[5] の構成で配信され，その実践における教育的成果・課題について報告する。
【キーワード】　数学、e-Learning、e テスティング、Moodle、STACK

1．e-Learning 環境と背景

第 1 筆者は，永年において工学系大学の数学教育に携わり，対面授業を主な学習環境として提供していた。近年において教育の場において ICT ，LMS ，BYOD などの活用をベースに，第 2 筆者との協働作業として，対面授業に連携した e-Learning 環境を構築し，e-Test を教室内外の学習で活用した。

この e-Test では，問題で係数・関数をランダムに変化させるランダム出題機能，受験終了直後にサーバが自動的に実行される自動採点機能，受講者と教師が出題問題と採点結果を振り返ることができる共有空間機能，受験時間内であれば同等な試験問題を再度受験できる反復受験機能を保持している[1]。

この e-Test で扱う数学問題の解答入力では，表計算ソフトウェア Microsoft Excel の命令コマンドと同等な，かつ数学的記号論を理解した Maxima コマンドの生成能力を必要としている。例えば，両辺が相等しくないことを示す数式を 3 組の実例で紹介する。

1/3*x+2≠1/(3*x+2),　　sin(x)^2≠sin(x^2),　　log(%e)≠cos(%pi)

ここで，%e, %pi は組み込み定数； sin(･), log(･), cos(･) は組み込み関数； (,) は括弧演算子； ^ は指数演算子； +, -, *, / は四則演算子となる Maxima の各命令コマンドであり，記載されたコマンド群に応じて数学的記号論における優先度が設定されている[3]。

本講演で扱う数学科目は，大学 2 年次に対する半期・選択科目「微分方程式論」であり，その学習においては「常微分方程式」「ラプラス変換」「逆ラプラス変換」となる 3 つの学習単元を設けて，各学習単元に対して通常の定期試験に替わる e-Test をこの 3 期分で実行した。

本抄録では，1 期：常微分方程式を基本にして議論していく。

1 http://www.moodle.org/
2 http://maxima.sourceforge.net/
3 http://stack.bham.ac.uk/
4 http://www.ams.org/publications/authors/tex/amslatex
5 https://www.mathjax.org/

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　78　　2016/08/29　14:00:21

2．試験問題とその受験履歴

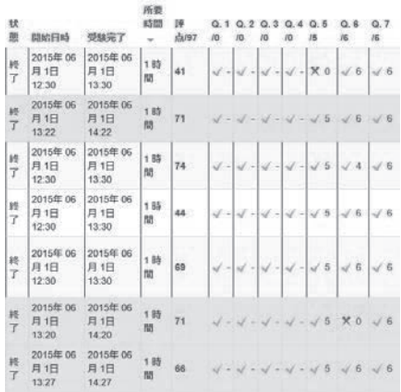
配信した e-Test は，Moodle の活動機能「小テスト（Quiz）」における問題タイプ「STACK」として配信されている。提供開始と終了（12:30〜14:30），受験残時間計測（最大試験 60 分），自動採点機能（満点 100 点），全ての受験履歴を保持（随時振り返り可能），反復受験管理（最高評点採用），評点集計機能（評点，ヒストグラム，平均・標準偏差・歪度・尖度等の基本統計処理），及び学習データ分析（ファシリティ指標，識別率，クロンバックの α 係数などの応用統計処理）などが自動化されている。

受験者は，指定時間帯において，インターネット接続されたノート PC （または大学内設置の PC，タブレット，スマートフォンなど）で稼働した Web ブラウザを用いて，好きな場所（通常教室・図書館・学生談話室・自宅など）から受験でき，試験時間内で好きな回数だけ反復して受験できる。

実際の採点済答案（問題，採点結果を含む）のスクリーンショットを紹介する。図 1 は 1 期：常微分方程式における定数係数 2 階線形同次分方程式の振り返りであり，図 2 は 1 期受験履歴である。



図 1　採点済答案イメージ



状態	開始日時	受験完了	所要時間	評点/97	Q.1 /0	Q.2 /0	Q.3 /0	Q.4 /0	Q.5 /5	Q.6 /5	Q.7 /5
終了	2015年06月1日 12:30	2015年06月1日 13:30	1時間	41	✓ -	✓ -	✓ -	✓ -	✗ 0	✓ 5	✓ 5
終了	2015年06月1日 13:22	2015年06月1日 14:22	1時間	71	✓ -	✓ -	✓ -	✓ -	✓ 5	✓ 5	✓ 5
終了	2015年06月1日 12:30	2015年06月1日 13:30	1時間	74	✓ -	✓ -	✓ -	✓ -	✓ 5	✓ 4	✓ 5
終了	2015年06月1日 12:30	2015年06月1日 13:30	1時間	44	✓ -	✓ -	✓ -	✓ -	✓ 5	✓ 5	✓ 5
終了	2015年06月1日 12:30	2015年06月1日 13:30	1時間	69	✓ -	✓ -	✓ -	✓ -	✓ 5	✓ 5	✓ 5
終了	2015年06月1日 13:20	2015年06月1日 14:20	1時間	71	✓ -	✓ -	✓ -	✓ -	✓ 5	✗ 0	✓ 5
終了	2015年06月1日 13:27	2015年06月1日 14:27	1時間	66	✓ -	✓ -	✓ -	✓ -	✓ 5	✓ 5	✓ 5

図 2　受験履歴イメージ

このとき，1 期 e-Test において受験者が 31 人，受験件数 88 件，平均受験件数 2.84 件／人（*SD*=1.95）であった。さらに同期評点において，初回受験で平均 41.23 点（*SD*=26.72），最高評点で平均 69.79 点（*SD*=23.63）であった。それ故，初回評点と最高評点の両分布の平均に対して有意差が認められた（*df*=60, *t*=6.75, *p*<0.001）。

また，同期の所要時間（秒）に関して，初回受験で平均 2,396 秒（*SD*=1,115），最高評点で平均 2,495 秒（*SD*=891）であった。それ故，初回受験と最高評点の所要時間（秒）の平均に対して有意差は認められなかった（*df*=60, *t*=0.3873, *p*>.05）。

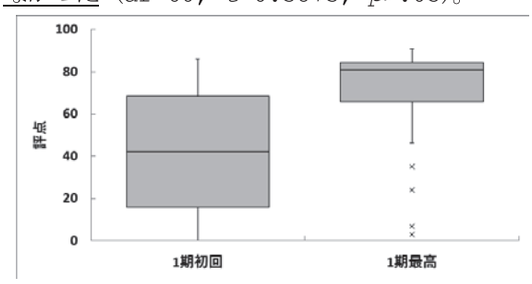


図 3　1 期回数別評点分布

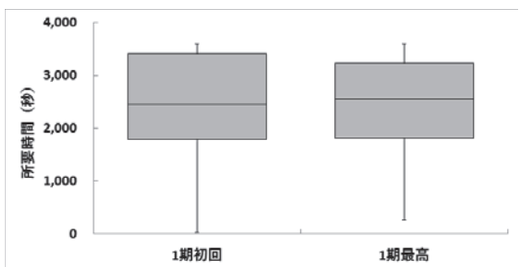


図 4　期別回数別所要時間の分布

次に，同期の最高評点分布に関して，その歪度： *skewness*=-1.63（標準正規分布の歪度は 0），その尖度： *kurtosis*=4.71（標準正規分布の尖度は 3）であった。このとき，分布の正規性に関するオム

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　79　2016/08/29　14:00:21

ニパス検定により「仮説：正規性あり」を棄却できることが認識できる（χ^2=16.24, df=2, p<0.001）。

3．Moodle による問題構造分析

1 期 e-Test は，問題が 14 題あり，各問題に対して Moodle が詳細な統計処理（問題構造分析）を自動的に実行している。すなわち各問における最高点（max），平均（M），標準偏差（SD），ファシリティ指標（F），識別率（DE）が出力される（表 1）。ただし，満点は 97 点に設定している。

表 1　1 期 e-Test の問題構造分析データ群

	問 01	問 02	問 03	問 04	問 05	問 06	問 07
MAX	5	6	6	6	6	6	4
MEAN	4.84	5.35	4.90	3.87	5.74	4.35	2.52
SD	0.90	1.82	1.99	2.42	1.00	2.09	1.29
F	96.77%	89.25%	81.72%	64.52%	95.70%	72.58%	62.90%
DE	52.41%	57.93%	68.67%	76.35%	73.12%	89.92%	61.76%
	問 08	**問 09**	**問 10**	**問 11**	**問 12**	**問 13**	**問 14**
MAX	5	6	6	6	6	6	4
MEAN	4.84	5.35	4.90	3.87	5.74	4.35	2.52
SD	0.90	1.82	1.99	2.42	1.00	2.09	1.29
F	79.57%	62.10%	70.97%	62.10%	77.42%	64.52%	59.91%
DE	68.20%	54.20%	83.64%	81.72%	91.29%	69.60%	72.42%

ここで，ファシリティ指標 F とは，設問がどれだけ難しいかを示す指標であり，$F=X_{average}/X_{max}$ （0≦F≦1）で定める。ただし，$X_{average}$ は設問の平均評点，X_{max} は設問の最大評点である。このとき一般的に 35%≦F≦65% の場合，平均的な学生に適切な設問であると判断されている[6]。1 期の e-Test では 6 問が適切な設問であり，他の設問は易しい設問と判断される。

また，識別率 DE とは，全体の評点の上位群と下位群に対して設問の正解率を識別できているかを示す指標であり，DE=（SUM_{xy}/N）/（SD_x*SD_y） （-1≦DE≦1）で定める。ただし，SUM_{xy}/N は全体評点と設問評点の共分散，SD_x と SD_y は，全体評点と設問評点のそれぞれの標準偏差である。このとき一般的に 50%≦DE の場合，適切な識別率をもつ設問と判断されている[6]。1 期 e-Test では全 14 問が適切な識別率をもつ設問と判断される。

次に，クロンバック（Kronbach）の α 係数（または内部整合性係数），単に α 係数とは，時や場所を変えても同様な結果が得られるか示す指標（再現性）であり，$\alpha=(k/(k-1)*(1-\Sigma_i S_i^2/S_x^2)$ で定める。ただし，k は設問の個数，S_i^2 は i 番目の設問評点の分散，S_x^2 は全評点の分散を示す。このとき一般的に 80%≦α の場合，問題全体の信頼性が高いと判断されている[6]。1 期の e-Test では α=91.42% より信頼性は高いと判断される。

最後に，発表時間に余裕があれば，第 1 筆者が提案する「試験等に対する信頼性と妥当性に関して可視化した幾何学的図形」を紹介したいと考えている[2][4][5]。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 80 2016/08/29 14:00:22

４．参考文献

[1] 亀田真澄，宇田川暢: 大学の数学教育に対する主体的な学びとなる学修環境作り，私立大学情報教育協会，論文誌 ICT 活用教育方法研究，第 16 巻 第 1 号，pp.36-41，2013 年

[2] 亀田真澄，宇田川暢：微分方程式におけるオンライン試験の取り組みについて，情報処理学会研究報告教育学習支援情報システム（CLE），2015-CLE-17(31)，pp.1-7，2015 年

[3] 亀田真澄，宇田川暢： 工学系大学における微分方程式 e-Learning の実践例について，東京理科大学紀要（教養編），第 48 号，pp.265-277，2016 年

[4] 村山航: 妥当性概念の展開, http://www.jartest.jp/pdf/jirei7rep1a.pdf （2015 年 09 月 09 日参照）

[5] 点群の凸包および最小包含円を求める（JavaScript 版）, http://k-ichikawa.blog.enjoy.jp/etc/HP/js/convexHull/cvx.html（2015 年 09 月 09 日参照）

[6] Phil Butcher： Brief guide to the Moodle iCMA Reports, The Open University, pp.1-2

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 81 2016/08/29 14:00:23

大学1年生1000人を対象とした 毎週オンラインIRTテスティングシステムとその運用

廣瀬 英雄（広島工業大学データサイエンスリサーチセンター）

概 要

広島工業大学では新しく開始した基礎教育のフォローアッププログラム（FP）の一環として項目反応理論（IRT）を用いた評価法を取り入れたオンラインテストシステムを組み込み2016年4月から運用している．対象学生は1年生全員（1160人）で対象科目は数学（解析基礎と線形代数）．FPは，毎週授業時での習熟度確認テスト（LCT），水準に達しない者を対象としたFPクラスでの演習（CWT）とテスト（FPT）から成っており，すべてにIRT評価が組み込まれている．全員がパソコンを持参して受験できる体制が完成し，順調に運用されている．プレースメントテスト，LCT，中間テストの関係のアナリティクスも行った．

1 Introduction

The IRT provides us the difficulties of the test items (problems) and the examinees' abilities together ([1, 2, 3, 13]), resulting in evaluating the examinees' abilities accurately and fairly. In addition, adaptive testing using the IRT selects the most appropriate items to examinees automatically, resulting in more accurate ability estimation and more efficient test procedures ([4, 5, 6, 7, 8, 14]). Thus, we have been proposing to use the adaptive online IRT systems for testing in universities, and we have decided to include the IRT testing system for our university follow-up program ([9, 10, 11, 12]) just begun since April 2016. In this paper, we first introduce our follow-up program, and then we explain the newly developed follow-up system for undergraduate students using the online IRT. Then, we illustrate this system in operation. Finally, we indicate the effectiveness of this online testing system.

2 Follow-up Program, FP

To assist students who were not necessarily taught sufficiently in high school days and junior high school days, we have decided to provide follow-up program to mathematics (*Analysis* and *Linear Algebra*), *Physics*, and *English* courses to support insufficient knowledges and skills of them. The follow-up program consists of regular classes, follow-up classes, remedial classes, and testing systems to make relations to these classes by using the portfolio system. The follow-up program includes the follow-up testing systems explained below.

Another important objectives in the systems is to educate higher level students, called peer supporters, by helping follow-up class students. As is well-known, teaching is the one of the best methods for learning.

2.1 Follow-up Program Testing Systems

Configuring the follow-up classes, we use the testing systems consisting of learning check testing (LCT), follow-up program testing (FPT), and collaborative work testing (CWT). Figure 1 shows the procedure of the follow-up program testing system to a certain subject, and Figure 2 shows the configuration of the follow-up program testing. All the students take LCT at the beginning or the end of each lecture in regular classes for about ten minutes to check if they comprehend the content of the lecture. Once students were discriminated as the candidates for the follow-up classes, they are enrolled to the follow-up program classes (FP classes) after regular classes. At the end of each follow-up class time, all the members take examinations to check if the class work can compensate the insufficient part of their knowledges and skills. In the case of mathematics courses, the exam coverage becomes the contents previously learnt or before. In the FP classes, in mathematics courses, the class work is taking the adaptive online IRT testing systems collaborated with peer supporters. The remedial classes also support the students after the FP classes.

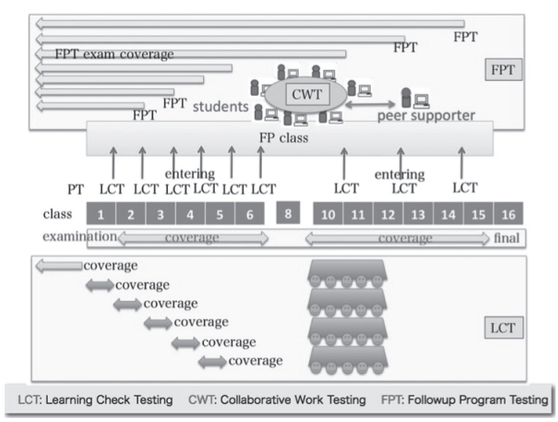


図 1: Follow-up program testing procedure.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 82 2016/08/29 14:00:23

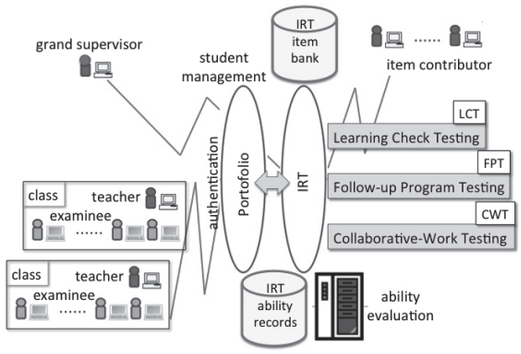


図 2: Configuration of the follow-up program testing systems.

2.2 Learning Check Testing, LCT

The learning check testing, LCT, is a kind of mini test, but the evaluation method for the LCT adopts the IRT partially in which the difficulty values are provided in advance, unlikely to the common IRT method where difficulty values and ability values are unknown simultaneously. Only six questions are prepared in each LCT. All the students in regular classes take the LCT for about ten minutes using their personal computers, that is, the test is taken online. All the questions are the same to each student, but sorted in different order. The levels of the questions are distributed from difficult one to easy one, that is, all levels are covered.

Wi-fi systems are equipped in every lecture room to assist the network connection in the campus. After a teacher in a class admits accesses to the LCT to all the attendees in the class, students can begin to take the examination. After the students finish the examination or after pre-setting testing time schedule is elapsed, the system computes the students' abilities by using the IRT, and send the scores transformed from the ability values to the portfolio system. The portfolio system sends messages to students who failed the LCT, to class teachers, department advisors, department tutors, and remedial class teachers.

The questions for the LCT are closed, and are different from the questions used to the CWT (explained later). Questions and answers are not open.

One LCT in a lecture has own unit name to be easily understood by teachers who admit access to the system, such as "special functions". Units are consisting of several sections such that "exponential, logarithmic, and trigonometric functions".

The problem items are collected from the item contributors (explained later), and installed to the item bank. The number of items in one section is more than 50 for one subject, and six items in one unit are used to the LCT. Others are provided to the FPT and the CWT. Thus, the size of the item response matrix for one subject for one LCT becomes about $1,000 \times 100 = (\#users, \#items)$.

2.3 Follow-up Program Testing, FPT

The follow-up program testing, FPT, is also a kind of mini test. Students who unfortunately failed to the LCT are automatically enrolled to the FP class, and take the CWT and the FPT. Unlikely to the LCT, the FPT adopts the adaptive online IRT testing system. The exam coverages become the contents previously learnt or before. Using the records of the LCT, the system can grasp the exam coverage to each examinee. Thus, teachers in the FP classes need not give the admission to take examination to students. The system decides whether the examinee should again attend the FP class next week or not by using the IRT ability value, and send its result to teachers, advisors, tutors, and remedial teachers.

The questions for the FPT are closed, and are different from the questions used to the CWT. Questions and answers are not open.

2.4 Collaborative Work Testing, CWT

The collaborative work testing, CWT, is mainly used in the FP classes. In the FP classes, a peer supporter is arranged for about ten FP class students, and we have provided about 40 peer supporters to analysis course and linear algebra course; more than 1,000 students are to be enrolled as freshman students. Peer supporters are assigned from junior, sophomore, senior students according to their wishes.

The questions in the CWT are chosen from some certain section; thus, students first select the sections to be taken. The CWT adopts the adaptive online IRT testing system. The exam coverage is the targeted section. The test is not considered to evaluate the students' skills exactly. Rather, the CWT is aimed at eagerness and fun in taking examinations. Thus, the number of questions is limited to five.

The scores are not used in students' evaluations, but the execution histories are sent to the portfolio system. The CWT are open to all the students for self-learning and self-study. Students can access to the system from anywhere and at anytime.

3 Roles in the Follow-up Program

We introduces the roles in the follow-up program. They are examinees (to LCT, FPT, and CWT), teachers in classes, peer supporters in the FP classes, item contributors, item managers, system manager, and grand supervisor.

Examinees:

Examinees are persons who take examinations via the internet. They are consisting of the LCT examinees, the FPT examinees, and the CWT examinees. The LCT examinees are all the Freshman students. The FPT examinees are persons who failed in the LCT. The CWT examinees are mainly the same as the examinees of the FP classes, but are not limited to. Any students can access the CWT from everywhere at anytime. All the examinees are registered to the system in advance via portfolio authentication system.

Teachers:

Teachers are responsible persons to the lecture classes. Teachers determine the unit to be tested in the LCT, and they can make triggers to start the LCT. Some teachers are also item contributors.

Peer Supporters:

In the FP classes, peer supporters assist FP class students. They make suggestions and hints to solve the problems, and give explanations if needed. Peer supporters are persons who are willing to assist the FP class students. They are collected from second to fourth grades students.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 83 2016/08/29 14:00:24

Item Contributors:
The system allows anyone who want to submit the items to contribute to the item bank. Mainly, item contributors are teachers in the regular classes, but are not limited to. The item bank is growing as the new item is registered to the system.
Item Manager:
The item managers handle the items; items are divided into the (LCT and FPT) items, and the CWT items; the CWT items are disjoint from the (LCT and FPT) items. The managers can also determine the initial difficulty values for each item.
System Manager:
The system manager usually works when system launching, when some fault happens, and when a major system update is needed. There is no duty, otherwise.
Grand Supervisor:
The grand supervisor can give access rights to teachers, item contributors, examinees, etc. The grand supervisor has special right to look at, modify, delete all the items in the item bank. He can make item calibration work when necessary.

4 Item Registration

All the problem items used in these testing systems should be installed to the item bank in advance. Item contributors (teachers in regular classes in principle) create mathematical problems, make pdf files including questions, answers, and explanations, and upload the pdf files. To do this, we have developed the item registration scheme using database system as shown in Figure 3.

Item contributors are not necessarily familiar with computer systems nor computer editing. Common mathematicians can use LaTeX language to manipulate the mathematical formula. Thus, we have decided to use pdf files as the media to be uploaded. We need not require to use LaTeX language; we only need the pdf formatted file; pdf via Word (MS) is also acceptable. In the database system, only the files in image format are saved in the database linked to the items of questions and answers in certain table in the database.

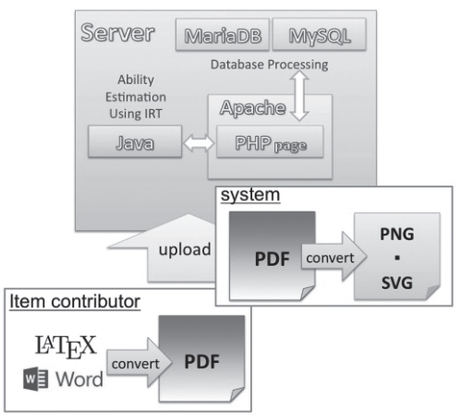


図 3: Item registration scheme.

5 Follow-up In Operation

Figures 4 and 5 show pictures of the LCT and FPT. All the students brought their PC, otherwise they use their own smartphones. The LCT also raised attendance rates from around 85% to more than 95%, and also brought vivaciousness and a tense atmosphere.

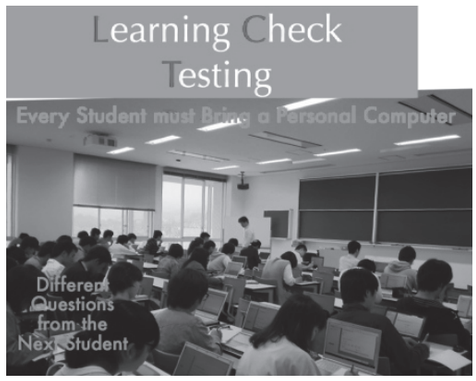


図 4: Learning check testing in operation.

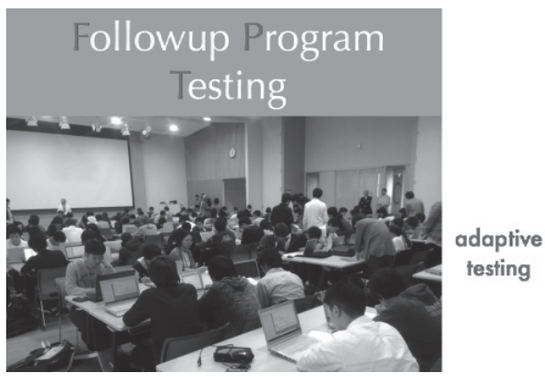


図 5: Follow-up program testing in operation.

6 Learning Analytics

To check if the accuracy of the LCT, we compared the scores among placement test, LCT, and midterm test in a class for demonstration. Figure 6 shows the relationship among them. In the figure, LCT.mean expresses the mean values from the 6 times LCT tests. These three are well correlated to each other, and we may say that the LCT is reliable although only six problem items are given in 10 minutes testing.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 84 2016/08/29 14:00:25

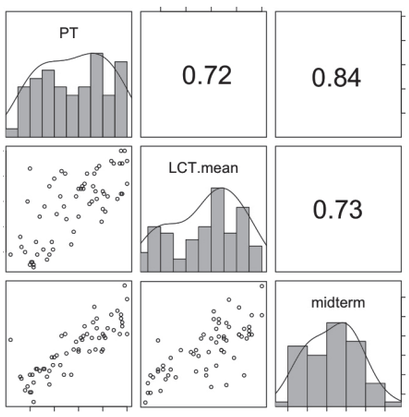


図 6: Relationship among placement test, learning check test, and midterm test.

7 Textbooks with Adaptive Online IRT System

Since April 2015, we have begun to deliver undergraduate textbooks equipped with the adaptive online IRT testing system in public. The subjects are *Linear Algebra* and *Probability and Statistics*. As commonly seen, exercises are given at the end of each chapter. A unique and advanced point is that the adaptive online IRT system combined with this book is introduced in the appendix chapter with access code to the Web. Students use these books and the online IRT testing system simultaneously.

8 Concluding Remarks

Universities in Japan are inclined to accept a wide variety of students because of the shrinkage of high school students and growth of the number of acceptance to universities, which causes various educational difficulties in universities. In response to this situation, our university has decided to enhance the follow-up program which is aimed at helping students who need basic learning and aimed at assisting teachers who have to engage in teaching a variety of students. The follow-up systems are recognized as a part of the follow-up program, and consist of the learning check testing, follow-up program testing, and collaborative work testing. We have developed the follow-up systems using the online item response theory which provides us the examinees' abilities accurately and fairly. In the system, many persons playing roles are involved, such as teachers, students, peer supporters, and item contributors. The systems have been successfully working.

参考文献

[1] R. de Ayala, The Theory and Practice of Item Response Theory. Guilford Press, 2009.

[2] R. K. Hambleton and H. Swaminathan, Item Response Theory: Principles and Applications. Springer, 1984.

[3] R. Hambleton, H. Swaminathan, and H. J. Rogers, Fundamentals of Item Response Theory. Sage Publications, 1991.

[4] H. Hirose and T. Sakumura, Test evaluation system via the web using the item response theory, in Computer and Advanced Technology in Education. pp.152-158., 2010.

[5] H. Hirose, T. Sakumura, Item response prediction for incomplete response matrix using the EM-type item response theory with application to adaptive online ability evaluation system, IEEE International Conference on Teaching, Assessment, and Learning for Engineering 2012, pp.8-12, 2012.

[6] H. Hirose, Yu Aizawa, Automatically growing dually adaptive online IRT testing system, IEEE International Conference on Teaching, Assessment, and Learning for Engineering 2014, pp.528-533, 2014.

[7] H. Hirose, Y. Tokusada, K. Noguchi, Dually adaptive online IRT testing system with application to high-school mathematics testing case, IEEE International Conference on Teaching, Assessment, and Learning for Engineering 2014, pp.447-452, 2014.

[8] H. Hirose, Y. Tokusada, A simulation study to the dually adaptive online IRT testing system, IEEE International Conference on Teaching, Assessment, and Learning for Engineering 2014, pp.97-102, December 8-10, 2014, 2014.

[9] H. Hirose, Y. Tokusada, Meticulous Learning follow-up systems for undergraduate students using the online item response theory, 5th International Conference on Learning Technologies and Learning Environments, July 10-14, 2016.

[10] H. Hirose, M. Takatou, Y. Yamauchi, T. Taniguchi, T. Honda, F. Kubo, M. Imaoka, T. Koyama, Questions and answers database construction for adaptive online IRT testing systems: analysis course and linear algebra course, 5th International Conference on Learning Technologies and Learning Environments, July 10-14, 2016.

[11] H. Hirose, Learning analytics to adaptive online IRT testing systems "Ai Arutte" harmonized with university textbooks, 5th International Conference on Learning Technologies and Learning Environments, July 10-14, 2016.

[12] Y. Tokusada, H. Hirose, Evaluation of abilities by grouping for small IRT testing systems, 5th International Conference on Learning Technologies and Learning Environments, July 10-14, 2016.

[13] W. J. D. Linden and R. K. Hambleton, Handbook of Modern Item Response Theory. Springer, 1996.

[14] T. Sakumura and H. Hirose, "Making up the complete matrix from the incomplete matrix using the EM-type IRT and its application," Transactions on Information Processing Society of Japan (TOM), 7(2), pp.17-26, 2014.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 85 2016/08/29 14:00:26

Co-creating ルーブリックの自己制御学習および自己評価力への影響分析

○山本 美紀*，宇都 雅輝*，西山 悠*，川野 秀一*，植野 真臣*
*電気通信大学大学院情報理工学研究科

1 はじめに

近年，ルーブリックは，教育評価において最も注目されている技術革新である(Andrade, 2000; Brookhart, 2003; Andrade and Du, 2005; Ross, 2006)．ルーブリックは，教師にとって，「期待することを伝達，フィードバックの提供，成績評価とその説明責任」の有効な手段として広まってきた(Andrade, 2000; Goodrich, 1997; Moskal, 2003; Popham, 1997; Andrade and Du, 2005)が，最近では，学習活動における自己評価（self-assessment）や学習者間の相互評価（peer-assessment）において，学習者に評価基準として活用させる場面も増えてきた(Hanrahan and Isaacs, 2001; Hafner and Hafner, 2003; Ross and Starling, 2005; Andrade and Du, 2007)．

自己評価とは，「学習者が自分のパフォーマンスの“価値”を自分自身で評価，判定し，学習成果を改善するために自分の強みと弱みを知ること」と定義されている（Klenowski, 1995）．これまでの研究において，自己評価が学習者の学習成果を高めることに大きな効果をもつことが実証されてきた（Schunk, 1996; Sparks, 1991; Fontana and Fernandez, 1994）．その多くは，学習者の自律性に関連し(Stallings and Tascione, 1996; Hart, 1999; Paris and Paris, 2001)，メタ認知の向上(Rivers, 2001)などが挙げられる．

自己評価の能力を高める訓練に，ルーブリックを活用した研究が行われている（Ross and Starling, 2005; Ross, 2006）．例えば，Andrade and Du（2007）は，ルーブリックを用いた自己評価が，Zimmerman の自己制御学習の枠組み（目標設定，プランニング，自己判断，自己反応）において重要な足場かけになる可能性を論及している．

一方，教育現場では「学力を単なる知識の量ととらえるのではなく，自ら学び自ら考える力」の育成が求められている．このような自己学習力を育成するためには，学習者の主体的な学習において，自己である学習者自身が自分の行動を観察し，評価し，新たな行動を起こすための自己評価力を身につけることは，不可欠であるといえる．

そこで，本論では，ルーブリックと Zimmerman の自己制御学習の枠組みにおけるメタ認知活動と学習者の学習に対する認識（学習観）と自己評価力との関係を明らかにすることによって，自己評価力を向上させるルーブリックの活用方法を実験的に検証する．

本論が着目するのは，教師によるインストラクションのためのルーブリックではなく，学習者への学習効果が期待されることから学習者と教師が共同でルーブリックを作成する実践研究（Andrade, 2005; Andrade and Do,2007; Ross and Starling, 2005; Ross, 2006）である．そして，山本ら(2015)の先行研究における学習者のみのグループでのルーブリックの共同作成である．山本ら(2015)は，ルーブリックの活用が学習観の変容を誘発し，メタ認知を含む学習方略の使用を促進するとこを示した．

学習観が，学生の教育経験によって変化することは，Morgan and Beaty(1997)によって主張されているが，その根底の知識観については，様々なモデルが存在し，Hofer and Pintrich(1997)のモデルでは，知識観の変化を巨視的な時間スケールで示されている．また，学習者の知識観が学習観に影響することは，認識論の立場から野村ら（2014）によって明らかにされている．そこで，学習者の知識観を影響要因として取り上げる．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 86 2016/08/29 14:00:27

2　方法

2.1　調査概要

大学院生博士前期課程の初年次の統計学の基礎講座においてブレンド学習が実施された．各対面講義において講座担当教師による単元内容の説明の後，e-ラーニング（SAMURAI）上で，講義範囲の課題が出題され，受講者は演習を行った．

本実験は，講座の受講生（25 名）を対象とし，e-ラーニング上での学習活動において，個人ルーブリックの作成，グループ・ルーブリックの作成（教師との共同：有／無），自己評価，相互評価を組み込んで行った．実験は，講座担当教師 2 名，TA3 名，観察者 1 名で実施した．

2.2　調査内容

(1)　対面講義および e-ラーニング（演習）スケジュール

e-ラーニング（SAMURAI）上の協調学習における自己制御学習は，Closed-end な課題（解が一意に決まる）を用いた演習を 2 回，Open-end な課題（解が一意に決まらない）を用いた演習 1 回を，図 1 に示すスケジュールで行った．それぞれの事前・事後調査を図 1 に示した．

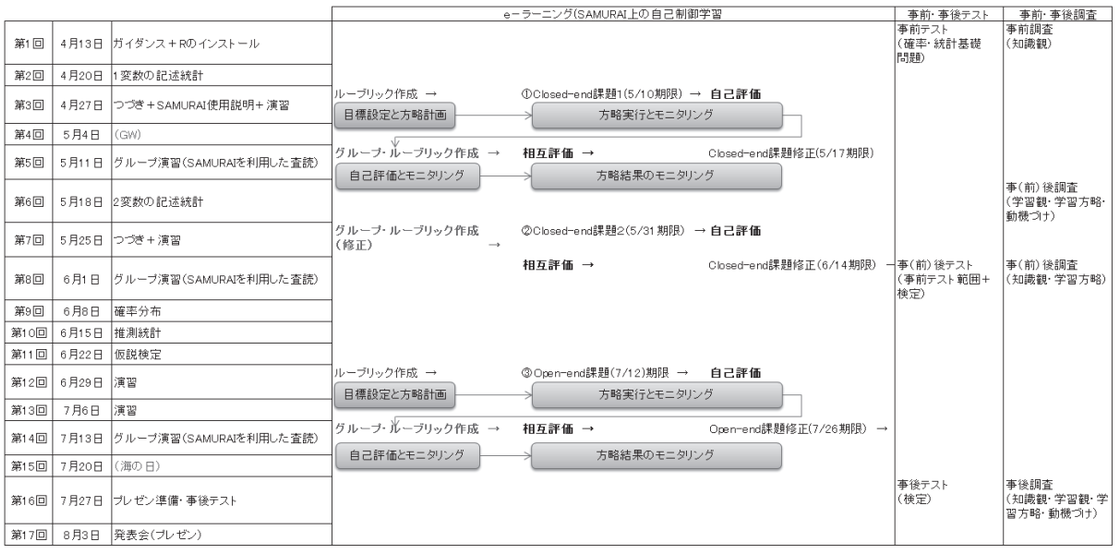


図 1　講座（実験）スケジュール

(2)　事前・事後調査

本論では，学習課題の形式，個人ルーブリックの作成とグループ・ルーブリックの作成（教師との共同：有／無）によって，これまでの評価に対する固定概念を一変させ，学習観に変化が生じると考える．これによって，学習者の学習動機にも影響を与えると考えられる．自己制御学習におけるメタ認知活動および自己評価力に影響を与える学習者個人の知識観，学習観，学習動機を調査する．知識観の質問項目には，野村ら（2014）の認識的信念尺度「知識の適用における条件性」，「知識の広い適用可能性」，「正当化された知識の実在性」，学習観の質問項目には，野村ら（2014）の授業観尺度「協同活動的授業観」，「知識電卓的授業観」を利用し，「主体的な学習観」に関する質問項目を新たに追加した．学習動機については，Pintrich and De Groot（1990）の学習動機づけ方略質問紙 MSLQ（Motivated Strategies for Learning Qestionnaire）を利用した．自己制御学習におけるメタ認知活動については，阿部ら（2010）のメタ認知尺度に関する質問項目を利用した．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　87　2016/08/29　14:00:27

2.2　仮説モデル

Hofer and Pintrich(1997)のモデルによると，今回，対象としている大学院学生は，大学教育を受け知識の多様さと不確かさを認めているが，普遍的な知識の存在を信じている段階にあって，今後，知識は相対的なものであると考え第三者的に選択することに価値を置く段階に成長していくとされる．本研究では，ルーブリックの作成に教師の参加が有／無が，学習観に最も影響を与えると考えるが，学習観への影響を与える要因として，学習者の知識観に焦点をあてる．そして，以下のような仮説モデルを立てた．

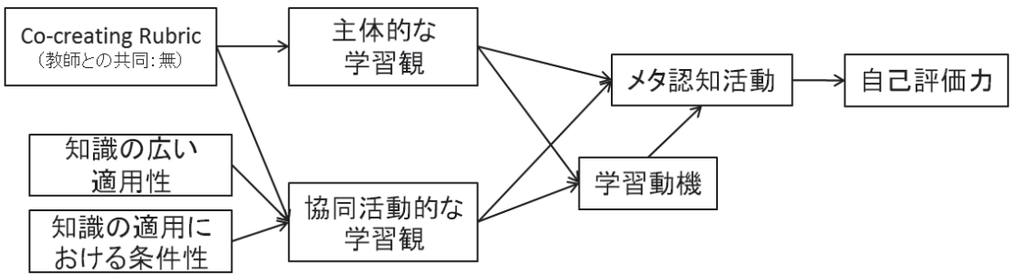


図１　ルーブリックの共同作成と知識観，学習観，学習動機，メタ認知活動，自己評価力の仮説モデル

3　分析・結果

本論では，第１回の実験（Close-end な課題）の結果データを用いた分析結果を報告する．演習レポート，質問紙を提出した学習者（22 名）のデータを分析の対象とした．分析には，Bengt and Linda Muthén によって開発された M-plus を使用した．

3.1　共分散構造分析と因果モデル

本分析では，独立変数を「ルーブリック作成方法」と「知識観（認識的信念：野村ら（2014））」とし，従属変数に学習観（「協同活動的な学習観」，「主体的な学習観」，「知識伝達的な学習観」），学習動機（「自己効力感」，「内発的価値」），メタ認知活動（「目的設定・方略計画」「方略実行」「方略結果」「点検」），そして「自己評価力」を設定した．

自己評価力は，宇都ら（2015）の項目反応モデルにおける能力推定値を相互評価点として用いて，次の式１で算出した．

自己評価力 ＝MAX - | IRT 評価 z 点－自己評価 z 点 |（式１）（ここで，MAX=10 とする）

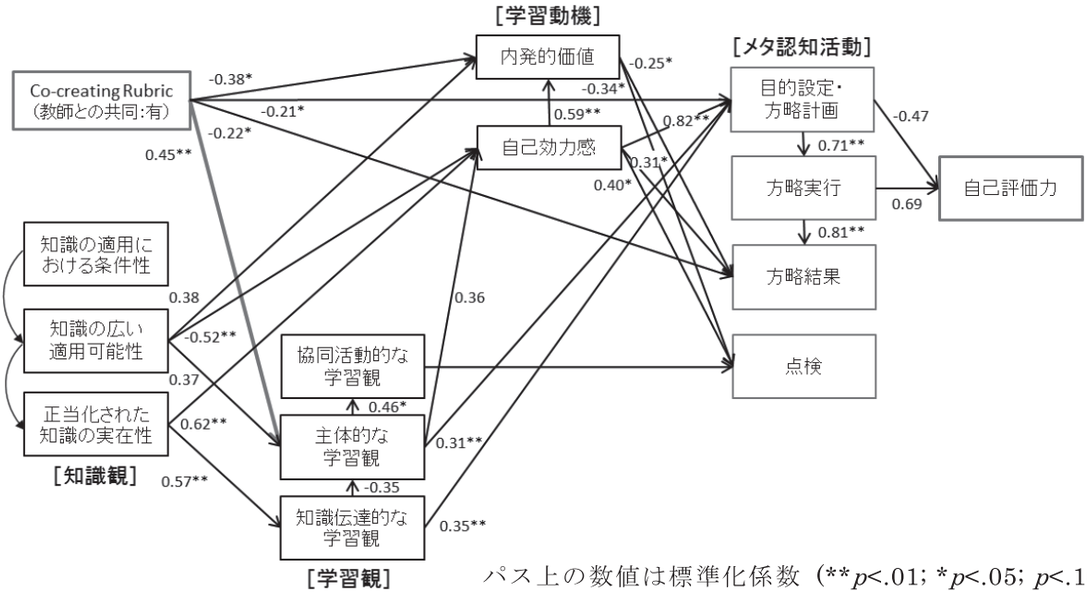
また，質的独立変数「Co-creating Rubric」は，学習者グループと教師の共同（有）の場合を 1，学習者のみのグループを 0 としてダミー変数化している（パス上の数値は学習者のみのグループに対する，学習者グループと教師の共同（有）の増分を表す）．

共分散構造分析において，有意差のあるパスを残しながら収束するまで分析を繰り返し，図２のモデルを得た．推定法は，ML を用いて解を求めた．モデルの適合度指標は，χ^2 値=62.224(df=57)(p=.296)，RMSEA=.069，CFI=.966T，TLI=.947 と良好であり，最終的なモデルとして受容した．

分析の結果，ルーブリックを学習者グループと教師の共同（有）で作成することは，学習者の主体的な学習観（すなわち，学習の主体は学習者であるという認識）を最も高めることが明らかになった．また，主体的な学習観を高めた学習者は，学習に対する自信（自己効力感）を向上させ，自己制御学習においてメタ認知活動「目的設定・方略計画」を促進させていることが分かった．さらに，学習者の自己評価力は，これらを経て「方略実行」モニタリングによって高められていることが明らかになった．

一方，知識観に注目すると，「知識の広い適用可能性」（野村ら，2014）を強く認識している学習者は，主体的な学習観が高い傾向にあることが示された．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　88　2016/08/29　14:00:28



パス上の数値は標準化係数（**p<.01; *p<.05; p<.10）

図２　因果モデル

4　考察

本論において，大学院生博士前期課程の初年次の学生にとって，学生グループでルーブリックを共同作成する際に教師との共同を行った場合とそうでない場合を比較し、学習観，学習動機，メタ認知活動，自己評価力への影響を検証した．その結果，(1) 教師との共同を行った場合，学習者の主体的な学習観を最も高める，(2) 主体的な学習観を高めた学習者は，学習に対する自信（自己効力感）を向上させる，(3)自己効力感が高いほど，自己制御学習においてメタ認知活動「目的設定・方略計画」,「方略結果」,「点検」のモニタリングを促進させる，(4)学習者の自己評価力は，「目的設定・方略計画」を経て「方略実行」モニタリングによって高められる，ことが明らかになった．

しかし，一方で，対象学生の中には，未だ「正当化された知識の実在性（野村ら，2014)」を信じ，知識伝達的な学習観を持つ学生がいることが分かった．そして，これらの学生は、主体的な学習観を低めることが示された．それでも，メタ認知活動は行われているが，自己効力感，学習に対する内発的価値は持てないでいると示唆される．

よって，これまでの学習者と教師が共同でルーブリックを作成する実践研究(Andrade and Do,2007; Ross and Starling, 2005; Ross, 2006)は学習成果を示したものであったが，ルーブリックを学習者と教師で共同作成し，それを用いて自己評価を行うことは，自己制御学習におけるメタ認知活動を促し，自己評価力を高めることができることを示した．よって，今後，自己評価力を高める訓練に用いるルーブリックの活用方法として期待できる．

引用文献

阿部真美子，井田政則(2010) 成人用メタ認知尺度の作成の試み:−Metacognitive Awareness Inventory を用いて−. 立正大学心理学研究年報（創刊), 23-34

宇都雅輝，植野真臣(2015) ピアアセスメントの低次評価者母数を持つ項目反応理論，電子報通信学会論文誌 J98-D(1), 3-16

野村亮太，丸野俊一 (2014) 授業を協同的活動の場として捉えるための認識的信念:−仮説的世界観措定仮説の検証−. 教育心理学研究 62(4), 257-272

山本 美紀，植野 真臣(2015) 構成主義的学習におけるルーブリックの活用方法が学習者に与える影響分析－目標志向性，学習観，動機づけ，学習方略，学習課題成績に着目して－．日本教育工学会論文誌, 39(2), 67-81

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　89　2016/08/29　14:00:29

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 90 2016/08/29 14:00:29

企画セッション１：ｅポートフォリオ

９月９日（金）13:00 – 14:40　第１会場（B101）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　91　2016/08/29　14:00:29

eポートフォリオ／学習記録データを活用したアクティブ・ラーニングの学習評価

○森本康彦
東京学芸大学

1. 背景

近年，学習者の継続的な学習プロセスを通して，21 世紀型スキル，学士力，社会人基礎力に代表される「テストだけでは測ることができない資質・能力」についての学習評価が強く求められるようになった．特に，学習者の主体的な学び，いわゆる「アクティブ・ラーニング」の学習評価法の確立は，文部科学省が掲げる重要課題の一つとなっている．一方，平成 28 年 7 月 29 日に文部科学省は「2020 年代に向けた教育の情報化に関する懇談会」の最終まとめを公開した[1]．これは，2020 年代の我が国の ICT を活用した教育のあり方について，目的や課題，対応策などを 3 項目でまとめたものであるが，授業・学習面と校務面の両面での ICT 活用の現状と課題のなかで「学習記録データ等の活用方法やユースケースについて具体的な方向性が示されていない」と指摘している．学習記録データとは，学習履歴，学習記録，学習成果物等の授業・学習の記録を電子的に蓄積した e ポートフォリオそのものである．e ポートフォリオを蓄積して学習評価に活用していくためには ICT の利用が必須であるが，近年の情報通信ネットワークの整備やタブレット端末の普及により，場所や時間に関係なく，どのような授業等の教育活動においても，すべての学習者が同様に e ポートフォリオを蓄積し学習評価に活用することが容易にできるようになった．

そこで，本企画セッションでは，学習者中心のアクティブ・ラーニングの学習評価に焦点を当て，初等中等教育と高等教育の教育現場におけるパフォーマンス評価や e ポートフォリオ活用の最前線について紹介する．

2. eポートフォリオ／学習記録データを活用した学習評価

学習評価とは，学習者にどれだけ学習の成果があらわれたか（あらわれているか）を測るための評価である．学習目標の達成状況を単なる知識だけでなく，多面的な観点から学習過程の全体を通して行われる[2]．また，e ポートフォリオの活用が求められる真正な学習においては，学習評価が学習のなかに埋め込まれていて切り離すことができない．つまり，学習評価すること自体が学習そのものであると言える．e ポートフォリオを活用した学習では，学習者はその学習プロセスを通して「e ポートフォリオ活動」を適宜組み合わせて繰り返し行うことで，学習の振り返りが促進され，学習が生起されていく．これは，学習評価の形成的評価を繰り返すサイクルとして捉えることができる．

図 1 は，e ポートフォリオを活用した形成的評価の繰り返しによる学習評価モデルである[3]．たとえば，ある単元の一連の授業において，本モデルに基づいた形成的評価が何度となく繰り返され，学びの改善を行いながら，学習が進んでいく．まず，教員と学習者は，学習目標と評価基準を確認し，学習指導を展開していく．学習プロセスにおいては，教員がタイミングよく最適なアセスメント方法を選択しアセスメントを行う．アセスメントの中では，学習エビデンスを e ポートフォリオとして生成

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 90 2016/08/29 14:00:29

し，その中から必要なものを抽出することで学習者の現在の学習状況を把握し，ギャップを同定する．学習者は，アセスメントを通し多くの気づきを得ることができるが，必要に応じて，自己評価を併せて行ったり，相互評価による仲間同士の相互作用や，教員評価や保護者等からの他者評価による外的フィードバックにより，さらなる学びの振り返りが促される．学習者自身や教員は，学びの改善を行い，そして，学びを継続する．やがて，ギャップが埋まり，学びの促進・成長に繋がっていく．このように，このモデルに基づいて形成的評価が実行され，ギャップが徐々になくなっていき，これが何度も繰り返されることで学習目標に近づいていくことを想定している．そして，最後の総括的評価に引き継がれていく．

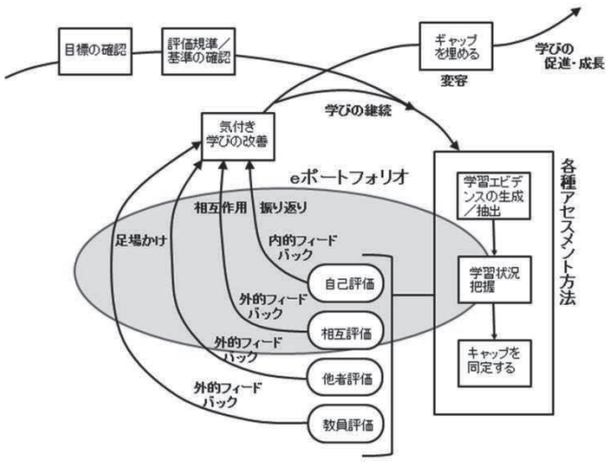


図１　学習プロセスにおけるｅポートフォリオを用いた学習評価モデル[3]

3.　企画セッションの概要

初等中等教育セッションでは，鄭谷心が，アクティブ・ラーニングの学習評価方法としてのパフォーマンス評価を紹介し，多元的・高次な学習者像をとらえるための「(紙ベースの) ポートフォリオ評価」や「ルーブリック」の理論と実践について述べる．次に，宮澤芳光が，それら実践における学習場面ごとに学習者の ｅ ポートフォリオを蓄積・活用することで項目反応理論を用いて学習者が発揮した資質・能力を評価し，資質・能力値をリアルタイムに推定する新たな学習評価法について提案する．

高等教育セッションでは，田中洋一が，ジェネリックスキルを身に付ける科目等において，どのように ｅ ポートフォリオを活用した学習評価を行うかについて実践をベースに解説し，続けて，宮崎誠が，ｅポートフォリオを活用した学習評価を受けて，どのように大学教育の質保証を実施すべきかについて言及する．

参考文献

[1] 文部科学省, “「2020年代に向けた教育の情報化に関する懇談会」最終まとめ”, 2016.
< http://www.mext.go.jp/b_menu/houdou/28/07/__icsFiles/afieldfile/2016/07/29/1375100_01_1_1.pdf>
[2] 北尾倫彦, “学びを引き出す学習評価”, 図書文化社, 2006.
[3] 森本康彦, 永田智子, 小川賀代, 山川修（編著）, “教育工学選書：教育分野におけるｅポートフォリオ”, ミネルヴァ書房, 2016.（印刷中）

E-mail: morimoto@u-gakugei.ac.jp

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　91　2016/08/29　14:00:30

汎用的な資質・能力を育成するためのパフォーマンス評価

○鄭　谷心（東京学芸大学・次世代教育研究推進機構）

1　はじめに

2016年6月23日，新しい時代に必要となる資質・能力の育成を踏まえた次期学習指導要領改訂の方向性に関する検討が中央教育審議会教育課程部会によってなされた。この検討では，評価の基準について，「何を理解しているか・何ができるか」にとどまらず，「理解していること・できることをどう使うか」へと改善することが必要と強調された。いわゆる，知識・技能だけではなく，高次の思考力・判断力・表現力やメタ認知力など，汎用的な資質・能力を評価するためのパフォーマンス評価が重視されるようになった。その背景には，OECD（経済協力開発機構）により2030年における次世代教育を開発するための事業が2015年に発足し，日本も主要メンバーとして参加していることが影響したとみられる。そこで，文部科学省は小・中学校における汎用的な資質・能力の育成のための指導モデルやアクティブ・ラーニングのモデルの開発を奨励し，パフォーマンス評価の多様な手法を生かした学校教育の改革を目指している。本稿では，これまでのパフォーマンス評価の定義や用途を整理し，それを学校教育にどのように活用するかについて検討する。

2　パフォーマンス評価の定義と適用範囲

パフォーマンス評価とは，「知識を応用・活用・総合することを要求する『真正の課題』に挑戦させ，実際の完成作品を生み出させたり，実演を行わせることによって，児童生徒の理解の様相を把握しようとする方法」である(1)。元来，パフォーマンス評価は，1980年代後半のアメリカにおいて，州政府によるトップダウンの「標準テスト」への批判とそれに基づく反省から，「真正の評価」論における代表的なものとして提案されたものである。「真正の評価」論では，「もし学ぶ価値のあるものであれば，それは評価する価値のあるものだ」(3)といったように，いずれ忘れ去ってしまう知識・理解の要素やある種の推論のパターンを教えるだけではなく，それらを総合的に使いこなすことができるような本物の学力を児童生徒に身につけさせなければならないのである。そのため，「生徒に現実の世界からの挑戦や問題を模した課題を与えること」，すなわち，「真正の課題」が必要とされている。よって，学力評価のあり方は，従来の選択回答式（客観テスト式）の設問から，多様な学習活動の自由記述式の設問や実技による評価，レポート・口頭発表等を用いた評価，およびパフォーマンス課題による評価へと移行しつつある。本稿は，これまでの学力評価法の分類をふまえてそれぞれの利点・欠点を次の表1に整理し，具体的な事例をあげながら解説する。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　92　2016/08/29　14:00:31

表１　学力評価法の利点・欠点

評価法			利点（適用範囲）		欠点	
客観テスト式	・多肢選択問題 ・正誤問題 ・順序問題 ・組み合わせ問題 ・穴埋め問題		・知識・理解の要素について標本をとるには適している。 ・ある種の推論のパターンを応用できるかについては評価できる。		高次の資質・能力の評価に十分ではない。	
	発展的な(改良された)多肢選択式問題		・従来のテスト項目に比べてより「真正」であり，正答を決定するために熟考・判断を要する。		表現が制限されている場合が多い。	
自由記述式	・知識を与えて推論させる方法		・問題を解くのに必要な知識ではなく，それを使いこなせるかどうかをみることができる。		文章力による影響が大きい。	
	・作問法		・子どもの概念理解の実態把握ができる。		同上	
	・認知的葛藤法		・子どもに認知的葛藤を起こすような問題や事象を提示して，それがどのように解消されるかを分析することにより，理解の実態を評価することができる。		・ある程度の概念が形成されないと使えない。 ・文章力による影響もある。	
	・予測－観察－説明（POE）法		・構成主義的な学習観に則った学習法 ・形成的評価に適している		文章力による影響もある。	
	視覚的描写問題	概念マップ法	文章以外で理解を表現させることができる	・論理的な思考力を育てることができる（論理的に構造化された概念マップ） ・子どもの知識の範囲や興味・関心の重点を探ることができる（連想法の場合） ・学びの過程として示すことができる	手法の用い方についての理解が必要である	あげられた言葉の数だけを評価しがちな点。
		ベン図法		・子どもの理解を教師が読み取りやすい ・個々の要素に関する知識も評価できる。		一度に多数の用語を扱えない。
		KJ 法		・情報の取り出し・分類・総合する力を評価することができる。		評価のために滅多に使用されない。
		運勢ライン法		・記述の下手な子どもも表現しやすい ・記述が上手な子どもにとっても言葉でごまかせない。		子どもが描いたラインが「正しい傾き」かの判断が難しい場合がある。
		描画法		・科学的に正しいか間違っているか，望ましい方向へ変化しているかどうかといった観点から評価できる。		絵が描けない子の場合には不適切。
実技テストの項目	検討会・面接法・口頭試問（観察や対話による評価）		・話し合う力や概念・事例・事象について質問を投げることで，子どもの理解の実態を把握できる。 ・個別的な技能を評価することが可能		さまざまな知識やスキルを総合的に使いこなすことが評価できない。	
	短文の朗読					
	実験器具の操作					
	運指練習					
	運動技能の実演					
パフォーマンス課題	・完成作品(エッセイ，観察記録，物語等) ・実演（朗読，プレゼン，彫刻，ダンス等）		・多様な知識やスキルを総合的に使いこなすことを評価することができる。 ・完成した作品そのものの特徴も評価可能。 ・チームで取り組むことによって協働的な問題解決力を評価することができる。		個別的に知識・技能を網羅することができない。	
	プロジェクト PBL (Project Based Learning)					

注：編みかけ部分以外の評価方法はすべて，パフォーマンス評価に属する。参考文献(1)(2)に基づき筆者作成。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　93　2016/08/29　14:00:31

表１には，多様な学力評価法とその利点・欠点を整理した。しかし，これらの評価法は，いずれも特定の時間や場所における学力の断片的な証拠を提供することしかできない。よって，複数の評価法を組み合わせて取り入れる必要があり，かつ，発達段階とともに変化する学習の姿を捉えることができるような評価のあり方が求められるようになっている。それは，児童生徒が成長していく姿を描くことができるポートフォリオ評価の汎用化を促している。

３　ルーブリックを活用した評価基準の設定

パフォーマンス評価やポートフォリオ評価の妥当性と信頼性をどのように保障すればよいのか。そこで最も用いられているのが，ルーブリックである。ルーブリックとは，「成功の度合いを示す数段階程度の尺度と，それぞれの評点・評語に対応するパフォーマンスの特徴を示した記述語から構成された評価基準表」である(4)。ルーブリックには，場合によってパフォーマンスの質のレベルを規定する基準だけではなく，典型的な作品や事例をもってその基準を示すものもある。執筆者は，今まで学校における数多くのルーブリックを見てきたが，いずれも以下，３種のルーブリックの１つに属しているといえる。

１つ目は，「特定課題ルーブリック」である。その名の通り，探究する力などの高次の資質・能力を評価するために，単元を貫く特定のパフォーマンス課題に対応して設定されたものである。例えば，東京学芸大学附属大泉小学校４年の UOI 単元「富浦フリータイム学習」においては，「自分にとって価値のある体験をして，特ダネを発表しよう！」というパフォーマンス課題に対応するルーブリックが作成された。そのルーブリックを取り入れた Reflection card が子どもたちの問題解決力や自己評価力を評価するために活かされる。また，アメリカでは，討論のために用いられる採点用紙（表２）があり，それは特別なジャンルに対応する包括的なルーブリックにもなっている。つまり，特定課題ルーブリックは，その課題を遂行するために必要な多様な思考力・判断力・表現力を評価する指針を示すだけでなく，短い学習活動のプロセスそのものの評価指標にもなる。

表２　討論の採点用紙としてのルーブリック

点数	アクション（聞き取りスキル・会話スキル）
2	質問に対する自分の見解を示す ・自分の見解を支える証拠を利用するか，事実に基づく情報を提供する ・類推する ・矛盾を認識する ・不適切なコメントを認識する
1	適切なコメントをするが，自分の見解を示さない。 ・他の人を議論に引き込む ・明確にするように質問をするか，議論を前に進める
−1	不適切なコメントをする
−2	邪魔をする ・注意を全く払わないか，他人の気を散らす
−3	個人攻撃をする

注：参考文献(2)に基づき筆者作成。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　94　2016/08/29　14:00:32

２つ目は，学年を越えて児童生徒の発達・成長を描き出す「長期的ルーブリック」である。それゆえに，連続的に学習記録やパフォーマンスを保存し，精選していくようなポートフォリオ評価との相性がよい。ポートフォリオに何を入れるかはその目的や見せる相手にもよるが，その中核をなすのは児童生徒によって選ばれた作品の事例とそれに基づく自己評価やふり返りの記述である。例えば，東京学芸大学次世代教育研究推進機構と附属大泉小学校との共同研究では，教師が考える児童の自己評価能力のパフォーマンスの成功の度合いを低学年・中学年・高学年の３段階に分け，４つの観点から捉えることができた(4)。それを長期的ルーブリックとして位置づけたうえで，児童の探究していく姿に照らしあわせて自己評価と相互評価をさせると，今後の指導と学習の改善につなげると期待できる。

３つ目は，「評定用ルーブック」である。これは，学校を越えてある地域社会において共通理解された学力評価の指標を指す場合が多く，主に成績付けのために開発されている。例えば，京都府乙訓地方において行われた「乙訓スタンダード」作りでは，「目標に準拠した評価」を実施するために作られた「評価計画Ⅰ」(5)はその典型的な事例である。また，上海市小中学校の学期末でパソコンによって自動生成されている「総合資質評価表」(6)も評定用ルーブリックに属していると考えられる。

４　研究の展望と課題

これからパフォーマンス評価が学校教育において活用する際，その妥当性と信頼性を保障するために，複数の教師や教育専門家によって育成すべき資質・能力の観点とカリキュラム整合性の観点からルーブリックについて検討が重ねられていく必要がある。また，評価の真正性を高めていくために，児童生徒の置かれている環境を考慮しつつ，その実態に基づいてルーブリックを絶えず修正し，再構築していく作業も大切である。これらの作業がすでに多忙化している学校教師の負担にならないように工夫を凝らし，汎用的資質・能力を評価するための有効的な方法として精緻化・構造化していくことが今後の課題となる。

謝辞

本研究は，東京学芸大学「日本における次世代対応型教育モデルの研究開発」〔文部科学省特別経費(プロジェクト分)〕の研究成果の一部であり，JSPS 科研費 JP16K17383 の助成を受けたものです。

参考文献

(1)田中耕治『教育評価』岩波書店，2008 年初版，p.149。
(2)田中耕治編著『新しい教育評価の理論と方法』日本標準，2002 年，pp.37-85。
(3)ダイアン・ハート著，田中耕治監訳『パフォーマンス評価入門――「真正の評価」論からの提案』ミネルヴァ書房，2012 年初版，pp.70-72。
(4)梶井芳明，鄭谷心，宮澤芳光，後藤由季奈「児童の自己評価能力を測定する項目の作成の試み」『日本教育工学会第 32 回全国大会講演論文集』，2016 年（掲載予定)。
(5)西岡加名恵「『乙訓スタンダード』作りの取り組み」『活用を促進する評価と授業の探究 研究成果最終報告書』2010～2012 年度科学研究費補助金基盤研究(C)，2013 年，pp.111-121。
(6)鄭谷心「上海におけるカリキュラムと評価改革の展開」田中耕治編著『グローバル化時代の教育評価改革』日本標準，2016 年，pp.16-27。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　95　　2016/08/29　14:00:33

項目反応理論に基づくアクティブ・ラーニングの学習評価

○宮澤　芳光
東京学芸大学　次世代教育研究推進機構

1　はじめに

2030 年には，世界がより複雑で不安定になり，多様化が進むと予想されており，OECD（経済協力開発機構）では，その時代に向けて，子ども達に求められるコンピテンシー（資質・能力）を再定義し，その資質・能力を育成・評価を確立する事業が進められている。我が国でも，次期学習指導要領の教育では，次世代に要求される資質・能力の育成のため，汎用的な資質・能力の育成が強調されている。資質・能力の育成には，子ども達が課題に対して主体的に学ぶ「アクティブ・ラーニング」の実践が必要である。アクティブ・ラーニングとは，多様な定義が存在するが，溝上らは，これらを包括し，「一方的な知識伝達型講義を聴くという（受動的）学習を乗り越える意味での，あらゆる能動的な学習のこと。能動的な学習には，書く・話す・発表するなどの活動への関与と，そこで生じる認知プロセスの外化を伴う」と定義している [1]。アクティブ・ラーニングの評価では，パフォーマンス課題を用いた評価が有効であることが知られている。パフォーマンス課題を用いた評価とは，学習者のパフォーマンスの能力を，完成作品および口頭発表や実技の実演によって，評価しようとデザインされた課題を用いた評価を指している [2]。パフォーマンス課題を用いた資質・能力を評価する測定試験には，批判的思考や分析的論理付け能力，問題解決能力，及び文章表現力を評価する CAE(Council for Aid to Education) の CLA(The Collegiate Learning Assessment) や大学生がいかなる知識と技能を習得したかに焦点をあてた OECD の AHELO(Assessment of Higher Education Learning Outcomes) が挙げられる。

これら測定試験では，自由記述式によるレポート等の最終成果物を用いて資質・能力を評価している。しかし，真正な学習において求められる真正な評価では，学習の結果だけでなく，学習過程が重視されるため，継続的に生成される学習成果物や学習の記録など学習エビデンスを用いて学習者のパフォーマンスを評価することが必要とされる。つまり，先の試験では，学習過程を重視した真正な評価による資質・能力の評価には不十分と言える。

そこで，本研究では，ICT を活用し資質・能力を評価する方法を開発することを目的とする。具体的には，CBT(Computer Based Testing) 上のパフォーマンス課題に学習者 (以下，受検者) が取り組むことで，各学習場面で必要な学習エビデンスを e ポートフォリオ (以下, 学習記録データ) として収集・蓄積し，課題遂行時に発揮された受検者の資質・能力をパフォーマンス課題に紐づいた評価基準を参考に項目反応理論を用いて同定する方法を提案する。筆者は，これまで，この研究に取組んできた [3][4]。本発表は，この研究内容の一部である。

2　アクティブラーニングの形式化

本章では，学習場面における学習記録データを蓄積するために，その蓄積の枠組みを定義する。学習場面とは，生徒の複数の学習活動を含み，その学習場面での成果物と，その成果物を評価するルーブリックといった学習評価も含むものとする。学習活動は，学習場面の中での生徒の具体的な活動であり，その活動の目的，主体，対象等によってデータが整理される。学習場面では，複数の学習活動が数珠つなぎのように連続しており，学習活動から生成された学習記録データで繋がっている，これは，プログラミング言語の関数をアナロジーとしており，関数が学習活動がであり，引数が学習活動において必要になるデータ，返り値

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　96　2016/08/29　14:00:34

は活動から生成された学習記録データである。学習場面で生成された成果と，その成果を評価するルーブリックを用いて，成果物を点数化し，次の章の段階反応モデルを用いて能力を推定する。

3　段階反応モデルに基づいた学習場面の評価

各学習場面で生成された成果物をルーブリックを用いて受検者の資質・能力を同定するため，テスト理論である項目反応理論を用いる。本研究では，学習場面から生成される学習記録データを一つの項目とし，その学習記録データを受検者が課題中に発揮した資質・能力を測定するための段階的な評価基準に基づいて採点したデータを反応データとする。そのため，段階的な反応データを扱うことができる段階反応モデルに基づいて受検者の資質・能力を測定する [5]。段階反応モデルとは，反応データがカテゴリ k の多値データを扱うモデルであり，受検者 j が項目 i に対してカテゴリ k と反応する確率 P_{ijk} を次式から求める。

$$P_{ijk} = P^*_{ijk-1} - P^*_{ijk} \tag{1}$$

$$\begin{cases} P^*_{ijk} = \frac{1}{1+\exp(-a_i(\theta_j - b_{ik}))} k = 1, \ldots, K-1 \\ P^*_{ij0} = 1 \\ P^*_{ijK} = 0 \end{cases}$$

ここで，a_i は項目 i の識別力パラメータ，b_{ik} は項目 i のカテゴリ k の難易度パラメータ，θ_j は受検者 j の能力パラメータを表す。ただし，$b_{i1} < b_{i2} < \cdots < b_{ik} < \cdots < b_{iK-1}$ と制約される。本研究では，受検者の学習記録データを自動採点または課題に紐づいた評価基準を参考に採点する。この評価基準によって採点された受検者の得点から段階反応モデルを用いて受検者の能力を推定する。

4　おわりに

本研究では，アクティブ・ラーニングの学習評価のために，学習活動を形式化し，段階反応モデルで能力を推定する手法を提案した。今後はシステムの開発とその評価を行っていく予定である。

謝辞

本研究は, 東京学芸大学「日本における次世代対応型教育モデルの研究開発」〔文部科学省特別経費 (プロジェクト分)〕の研究成果の一部です。

参考文献

[1] 溝上慎一. アクティブラーニングと教授学習パラダイムの転換. 東信堂, 2014.

[2] 田中耕治（編）. よくわかる教育評価　第 2 版. ミネルヴァ書房, 2010.

[3] 大橋誠, 宮澤芳光, 森本康彦. CBT における項目反応理論を用いたパフォーマンス課題の評価法の提案. 教育システム情報学会第 40 回全国大会論文集, pp. 361–362, 2015.

[4] 大橋誠, 宮澤芳光, 森本康彦. パフォーマンス課題に基づく評価のための項目反応理論を用いた CBT の提案. 日本教育工学会第 31 回全国大会論文集, pp. 361–362, 2015.

[5] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, 1969.

E-mail: miyazawa@u-gakugei.ac.jp
会員番号: 30000171

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　97　2016/08/29　14:00:35

大学の授業におけるｅポートフォリオを活用した学習評価

田中洋一
仁愛女子短期大学 生活科学学科

１．はじめに

仁愛女子短期大学では，2015 年度にディプロマポリシー，カリキュラムポリシー，アドミッションポリシーの見直しを行い，2016 年 3 月に新たな学習成果を公開した．また，生活科学学科生活情報専攻（以下，本専攻）では，「学修成果アセスメントシステムの構築―ルーブリック，ｅポートフォリオ，客観テストの連携―」というプロジェクトで平成 26 年度～平成 27 年度の学長裁量経費を受け，学習成果アセスメントシステムの設計を行った．学習評価としては，直接評価か間接評価かという縦軸，量的データか質的データかという横軸の２軸で分類すると，以下の４つのタイプに分けられる．①量的データの直接評価である客観テスト（標準テスト），②質的データの直接評価であるパフォーマンス評価（ポートフォリオ評価），③量的データの間接評価である学生調査，④質的データの間接評価であるリフレクションシート（ミニッツペーパー）．学位授与の保証としては，これら４タイプを連携することで学習成果を評価することがもとめられる．本専攻の学習成果（ディプロマポリシー）および評価（アセスメントポリシー）が明確になることにより，教員の授業目的・方法・評価の共通理解が得られ，体系的なカリキュラムが構築できる．また，カリキュラム全体のルーブリックを示すことにより，入口側の高校および出口側の企業に対して，本専攻の学びへの理解が深まり，進学及び就職のミスマッチを減少できる．

２．ｅポートフォリオの活用

本専攻の授業や学生支援では，先述の②質的データの直接評価であるパフォーマンス評価（ポートフォリオ評価）として，オープンソースのｅポートフォリオ「Mahara」を活用している．本稿では，筆者が担当する３つの授業（プログラミング科目，キャリア科目，ジェネリックスキルを身につける科目）での活用例を紹介しよう．

２−１．プログラミングⅠ

１年前期に開講する必修科目「プログラミングⅠ」ではScratch，HTML，CSS，JavaScriptを学ぶ．本授業の目的は，プログラミングの楽しさに触れると共に，問題解決能力・論理的思考力・創造力を身につけることである．

Scratchのゲーム課題は，３つの練習課題と最終課題からなる．練習課題は，オープンソースのLMS「Moodle」の説明文を読み，まずコアプログラムを完成させた後，オリジナルの改良を１つ以上加える．最終的に完成したプログラムはScratch公式サイトで公開するとともに，Maharaに埋め込み，プログラムの説明をまとめる．最終課題は，デザイン思考ミニワークのプロ

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 98 2016/08/29 14:00:35

トタイプやScratch公式サイトのゲームを参考にして，オリジナルゲームを一から制作する．各制作物は履修者全員が閲覧でき，ピアアセスメントとセルフアセスメントを行う．毎回，授業の最後には，Maharaを用いて振り返りノートをまとめている．

2−2．キャリアプランニング

本授業の目的は，「理想の自分」について考えることにより，現時点での目標を設定，行動プランを作成し，実行することである．そのため，ジェネリックスキルテスト「PROG」及びその振り返り，働く価値に関するカードを用いたワーク，名刺制作，デジタルストーリーテリング制作を通して，自己理解及び目標の設定を行う．また，教員及び社長からの講義や学外研修（テクノフェアにおける名刺交換やインタビュー等）を通して，企業研究を行う．

デジタルストーリーテリングとは，「５年後の自分への手紙」として，自分を振り返り，自分語りをした動画である．制作過程としては，ピアメンタリングを行い，お互いに幼児から今日までを振り返り，Mahara のワークシート上に簡略版の自分史を書き出した上，ナレーション原稿を作成する．次に，本専攻において在学期間中無料で貸与している Windows タブレットを用いて，ナレーションの録音，静止画や動画の編集を行い，デジタルストーリーテリングを制作する．完成した動画はファイル自体をクラウド上にアップロードして，Mahara に埋め込んだ．デジタルストーリーテリングは，制作を通じて，自分の生き方を振り返り，現在の自分を見つめ直すため，アイデンティティに関わる e ポートフォリオの一形式といえる．

2−3．生活情報論

本授業の目的は，ジェネックスキル，特に問題解決能力を身につけることである．本授業では，振り返りシート 1 枚とグループワーク用ワークシート９枚，計 10 枚の Mahara ページを１つのコレクションとして使用している．振り返りシートは２列のレイアウトで，左列に振り返りノート（授業の最後に行動や考えの変化等を毎回記述），右列に課題ノート（教科書の演習問題を解答）を配置している．ワークシートは，グループワークを構造化し，個人作業・グループ作業・クラス発表という３つの次元での思考や活動を支援するような質問と回答欄から成る．

本授業の期末課題は，「学習成果１年前期」という Mahara ページ制作である．「生活情報専攻の学習成果」というテキストブロックを左上に挿入し，本専攻各学習成果の自己評価（４件法の点数と説明文），その主張の根拠（エビデンス）となる資料を記述する．また可能な根拠資料は，授業や資格対策講座等の成果物（アーティファクト）をブロックで挿入する．つまり，この課題は，Mahara 等に収集した学習成果物をセレクションして作る，１年前期のショーケース・ポートフォリオである．

３．おわりに

最後に紹介したショーケース・ポートフォリオには，今後 Mahara のルーブリック・プラグインを用いる予定である．また，本専攻のカリキュラム・ルーブリックがまだ完成しておらず，今回は数値（４件法）で自己評価させているため個人差が大きい．ただし，エビデンスに基づき，セメスターごとの振り返りを行うことは重要である．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 99 2016/08/29 14:00:36

eポートフォリオを用いた大学教育の質保証

○宮崎 誠*1
畿央大学*1

1．はじめに

学士課程教育の維持・質の向上に向けた改革として2008年に中央教育審議会答申である「学士課程教育の構築に向けて」が出された．そこでは，「学位授与方針」（ディプロマ・ポリシー），「教育課程編成・実施の方針」（カリキュラム・ポリシー），そして「入学者受け入れの方針」（アドミッション・ポリシー）の3つのポリシーを示すことが挙げられている[1]．また，これらの指針に基づく取り組みの中で計画・実践・評価・改善(PDCA)のサイクルを確立することが重要であるとしている．これを受けて大学においては，この「学士課程答申」を指針とした教育の改善を図る取り組みが積極的に行われてきたが，学校教育法施行規則の一部改正公布（2016年3月31日付）により，3つのポリシーの策定・公表の義務化が確実となり，大学において喫緊の課題となった[2]．そのなかで，大学教育の質保証の質を保証する上で，「ディプロマ・ポリシー」の指針による取り組みにおいて，学生の学習成果をエビデンスとして示すことが有効な方略の一つであり，ｅポートフォリオが注目されている． ｅポートフォリオを利用することで，カリキュラムと学習活動との対応を体系的かつ明確にすることができ，その学習成果は学生の学習のエビデンスとして活用することができる．

本稿では，ｅポートフォリオの活用を学習の評価という観点でとらえ，ｅポートフォリオが教育の質保証に資するためのカリキュラムや授業における方略をまず述べ，次にｅポートフォリオによりどのように活用することができるか述べる．

2． 教育課程の体系化

学習の評価（アセスメント）を実施するためには，学習成果（ラーニング・アウトカム）重視の観点から，卒業までに学生がどのような能力を修得することを目指すかを，明らかにされていることが必要である．学習成果は，目標(goal)，スタンダード(standard)，期待されるもの(expectation)，コンピテンシー(competency)，目的（objective）などで示すことができる具体的な基準であり，例えば，教育課程での学習成果を策定する手法としては，教育課程における修了者像を設定し，そこから導き出すというアプローチがある．次にディプロマ・ポリシーにしたがって，教育課程を体系的に編成するために，今度は，明らかになったいくつかある学習成果のうち，科目がどれに資するかを具体的にしていく．例えば，科目の学習成果を策定する手法としては，各科目で期待される学習成果を具体的にし，その上でKJ法により課程教育の学習成果に集約するというアプローチがある．これら学習成果の評価には，パフォーマンス評価が用いられることが有効であり，ルーブリックが活用できる．ルーブリックとは，学習結果のパフォーマンスレベルの目安を数段階に分けて記述し，学習の達成度を判断する基準を示すことができる教育評価法であり，学習者は自身の学習の達成度を客観的に認識できるため，自身の学習の指針とすることができる．

3．ｅポートフォリオの活用

ｅポートフォリオを導入するためのシステムとして，商用，オープンソースを含め様々なソフトウェアがあり，その利用する主体や目的に応じて設計思想も大きく異なる[3]．そのため，ｅポートフォリオを利用する目的と学習活動を想定し，システムの機能がどのように活用できるのかを事前に検討する必要がある．例えば，学習成果物を深い学びに活用するための学習ポートフォリオでも，学習過程を形成的に振り返るのか，それとも総括的に振り返るのかで実施する学習活動も違ってくる[4]．組織が教育を評価する目的で活用するｅポートフォリオには，アセスメントポートフォリオがあり，オープンソースソフトウェアにはSakai OSPやその後継システムであるKARUTAが活用できる [5]

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 100 2016/08/29 14:00:37

[6]．また，同じくオープンソースソフトウェアであるMaharにもプラグインを導入することでルーブリックを導入したアセスメントポートフォリオへの機能強化が可能である[7]．これらシステムでは，学習成果の基準をルーブリックとして表現することができ，大学が求める学習成果の基準に学生の学習成果物をエビデンスとして整理することができる．また，学習の省察や評価，フィードバックといった活動がルーブリック上で一貫して利用できるため，学習成果に基づいた授業を支援するツールとして非常に優れている．科目によらない学習成果基準採用することで，学生自身が身につけた学習成果について科目を横断して体系的に振り返ることが可能になる．

４．教育課程の体系化によるｅポートフォリオの活用事例

熊本大学大学院教授システム学専攻では，フルオンラインで学習を行うｅラーニング大学院である．オンライン教育等のICT活用教育を実施するにあたっても，当然対面授業と同等の効果が求められることから，学習における質の保証として，修了生が身につけるべきコンピテンシーを明確にし，どの科目を履修すると身につけられるのかを公開されており，カリキュラム上の科目設計だけでなく，各科目の課題にまで，コンピテンシーと直結するように科目が設計されている[8]．このように教育課程の体系化が実現しているため，Sakai OSPを使って学習成果物を自動的にマトリクスに登録することにより，コンピテンシーに沿った体系的な振り返りが可能になり，教育の質保証として学習成果に基づいたエビデンスの提示が可能となっている[9]．

５．おわりに

本稿では，教育課程の体系化を行うことで，教育の質保証にｅポートフォリオが活用できることを述べた．これは，言い換えると，教育課程の体系化ができなければ，教育の質保証にｅポートフォリオは活用できないということである．また，ｅポートフォリオシステムにも様々な設計思想により構築されているため，ｅポートフォリオを利用する目的にあったシステムを選定しなければ，効果的な活用は難しい．しかし，Sakai OSPやその後継システムであるKARUTA，Maharaは，アセスメントポートフォリオが構築できる．今回述べたような学習成果物がｅポートフォリオに蓄積されていく学習成果物蓄積のフレームワークとも言える仕組みが確立できれば，その後，蓄積された学習成果物を使ったｅポートフォリオを活用した新しい学習活動（省察や自己評価，相互評価，他者評価，ショーケースポートフォリオの作成）にも展開可能である．

参考文献

[1]文部科学省（2008）学士課程教育の構築に向けて（答申），中央教育審議会，
http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2008/12/26/1217067_001.pdf（2016年7月30日閲覧）
[2]文部科学省（2016）学校教育法施行規則の一部を改正する省令の公布について（通知），
http://www.mext.go.jp/b_menu/hakusho/nc/1369884.htm（2016年7月30日閲覧）
[3]森本康彦（2008）ｅポートフォリオの理論と実際 教育システム情報学会，25(2):245-263
[4] Helen C. Barrett.（2009），Balancing the Two Faces of ePortfolios，
http://electronicportfolios.org/balance/（2016年7月30日閲覧）
[5]宮崎ほか（2011）ｅポートフォリオシステム評価 －Mahara と Sakai OSP －，法政大学情報メディア教育研究センター研究報告特別号 Vol.25: 12-14
[6] KARUTA，http://karutaproject.org/（2016年7月30日閲覧）
[7] Mahara - Rubric-plugin，https://github.com/eportfolio/rubric-plugin（2016年7月30日閲覧）
[8]北村ほか（2007）ｅラーニング専門家養成のためのｅラーニング大学院における質保証への取組：熊本大学大学院教授システム学専攻の事例『メディア教育研究』第3巻 2号:25-35
[9]宮崎ほか（2016）コンピテンシーに基づくカリキュラムに対応したｅポートフォリオシステムの開発，情報処理学会 情報処理学会論文誌教育とコンピュータ（TCE6）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 101 2016/08/29 14:00:37

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 102 2016/08/29 14:00:38

一般セッション５：テスト統計

９月９日（金）13:00 – 14:40　第２会場（B102）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　103　2016/08/29　14:00:38

複数国において実施された性格特性検査における因子構造

酒匂志野[1] 坂本佑太朗[1] 今城志保[1] 分寺杏介[2]
[1]株式会社リクルートマネジメントソリューションズ [2]東京大学大学院

【背景】

近年、日本企業が海外現地法人において人材の適性を見極めたいというニーズが増えており、適性検査をグローバル展開することが求められている。従来、性格特性の比較文化を行った研究は数多く見られるが、ビジネスパーソンを対象とした研究はまだ少ない。そのため本研究では、職務遂行に影響を及ぼすと考えられる性格特性を比較することとした。比較する国としては、日本企業が今後海外進出の拠点とすることがますます増えると思われるアジアに限定した。なお、通常性格特性の異文化比較では翻訳の問題が指摘されることがある。今回はすべて英語で調査を実施することによって、翻訳の問題をできる限り除いて結果を検討できるようにした。

本研究では、「対人関係性」「課題解決性」という構成概念における、複数国での測定の等価性(equivalence)について、確認的因子分析を使って検証することを目的とする。

【方法】

日本、シンガポール、マレーシア、タイのビジネスパーソンを対象に調査を実施した。

調査日：2016年2月16日～3月3日

調査方法：インターネット調査（調査会社に依頼）

調査対象者：日本、シンガポール、マレーシア、タイのビジネスパーソン各100名。国籍以外の条件をできる限り統制できるよう、属性条件を以下のように設定した（性別：男性のみ／年齢：30～49歳のみ／勤務先企業の従業員規模：500名以上のみ　など）

調査内容：

・対人関係性を表す性格4尺度：社交的でネットワーク作りが得意な性格を示す「外向性」14項目（うち反転3項目）、支配的・指示的で自己主張が強い性格を示す「統率性」15項目（うち反転2項目）、協調性がありチームワークを重んじる性格を示す「調整性」14項目（うち反転2項目）、自分の基準を大事にする性格を示す「自律性」12項目（うち反転2項目）、4件法

・課題解決性を表す性格5尺度：堅実で安定感のある性格を示す「維持」14項目（うち反転2項目）、創造的で新しいことを起こそうとすることを示す「変革」15項目（うち反転3項目）、論理的・分析的に考えることを示す「思索」14項目（うち反転2項目）、活動的で考えるよりもまず動く性格を示す「行動」14項目（うち反転2項目）、タイミングよく迷いなく意思決定を下す性格を示す「決断」14項目（うち反転3項目）、4件法

なお、反転項目は反転処理をした上で以降の分析を行った。

【結果】

基礎的な統計量を確認した上で、確認的因子分析を行った。

基礎的な統計量

全項目において国ごとに平均値と標準偏差を差出した。概ね偏らずに分布していた。外向性において、日本の平均値は1.82～2.65、標準偏差は0.653～0.798、シンガポールの平均値は1.84～2.90、標準偏差は0.579～0.692、マレーシアの平均値は1.92～3.03、標準偏差は0.592～0.760、タイの平均値は1.95～3.01、標準偏差は0.594～0.740であった。統率性において、日本の平均値は1.79～2.53、標準偏差は0.599～0.748、シンガポールの平均値は1.99～2.82、標準偏差は0.574～0.763、マレーシアの平均値は1.98～2.90、標準偏差は0.562～0.743、タイの平均値は1.99～2.90、標準偏差は0.563～0.789であった。調整性において、日本の平均値は2.11～2.96、標準偏差は0.519～0.692、シンガポールの平均値は2.16～2.96、標準偏差は0.520～0.679、マレーシアの平均値は2.15～3.09、標準偏差は0.517～0.639、タイの平均値は2.16～3.05、標準偏差は0.507～0.694であった。自律性において、日本の平均値は2.23～2.99、標準偏差は0.545～0.697、シンガポールの平均値は2.25～3.09、標準偏差は0.466～0.647、マレーシアの平均値は2.17～3.09、標準偏差は0.507～0.677、タイの平均値は

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 102 2016/08/29 14:00:39

2.15～3.16、標準偏差は 0.476～0.649 であった。

維持において、日本の平均値は 2.47～2.95、標準偏差は 0.538～0.727、シンガポールの平均値は 2.53～3.11、標準偏差は 0.480～0.677、マレーシアの平均値は 2.54～3.13、標準偏差は 0.498～0.692、タイの平均値は 2.38～3.13、標準偏差は 0.471～0.715 であった。変革において、日本の平均値は 2.04～2.92、標準偏差は 0.552～0.749、シンガポールの平均値は 1.83～2.94、標準偏差は 0.476～0.740、マレーシアの平均値は 1.80～3.11、標準偏差は 0.493～0.705、タイの平均値は 1.86～3.08、標準偏差は 0.466～0.741 であった。思索において、日本の平均値は 2.52～3.00、標準偏差は 0.550～0.718、シンガポールの平均値は 2.35～3.12、標準偏差は 0.484～0.681、マレーシアの平均値は 2.31～3.19、標準偏差は 0.468～0.748、タイの平均値は 2.28～3.17、標準偏差は 0.486～0.710 であった。行動において、日本の平均値は 2.24～2.91、標準偏差は 0.598～0.825、シンガポールの平均値は 2.18～2.97、標準偏差は 0.512～0.712、マレーシアの平均値は 2.04～3.02、標準偏差は 0.568～0.791、タイの平均値は 2.03～3.03、標準偏差は 0.517～0.722 であった。決断において、日本の平均値は 2.15～2.74、標準偏差は 0.602～0.684、シンガポールの平均値は 2.16～2.86、標準偏差は 0.552～0.747、マレーシアの平均値は 1.96～2.94、標準偏差は 0.563～0.759、タイの平均値は 1.97～2.93、標準偏差は 0.551～0.756 であった)。

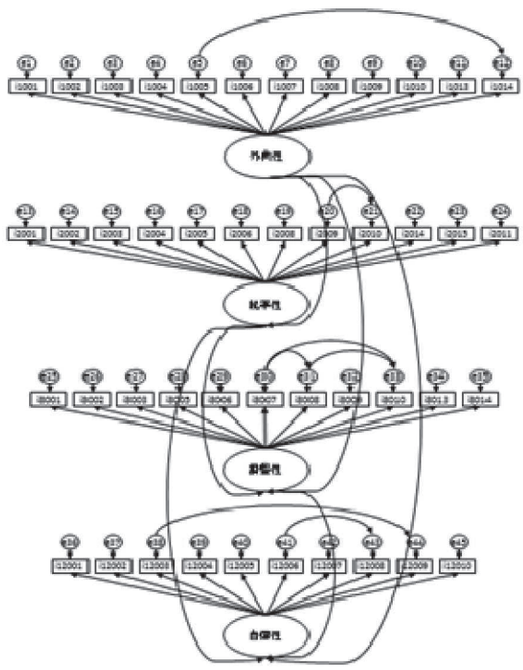
確認的因子分析（以下「CFA」）

4 母集団の CFA を行った。モデルは、配置不変を基本としている。なお、使用したソフトは Amos 22.0 であった。

① 対人関係性

調査で用いた項目をすべて投入して推定を行った。モデルに若干の修正を行い、2 度目の推定を行った。具体的には、他の因子に対して高い負荷を持っている項目をモデルから除外した。また、同じ単語を用いているなど、表現の似通っている項目については誤差間に共分散を引いた。また、パス係数が負の値になっていた 2 項目を除外した。最終的に 45 項目を投入した（図 1 参照）。パス係数が 4 母集団間で等値であるモデルとそうでないモデルを試した結果、前者は CFI=.827、RMSEA=.029、AIC=8397.094 であり、後者は CFI=.834、RMSEA=.029、AIC=8356.469 であった。最終的に後者を採用した。また、各国におけるパ

【図 1】対人関係性のモデル図



【表 1】外向性の項目へのパス係数

	日本	シンガポール	マレーシア	タイ
il001	1.000	1.000	1.000	1.000
il002	0.939	0.938	0.980	1.110
il003	0.697	1.031	0.987	1.067
il004	0.378	0.648	0.553	0.667
il005	0.813	0.883	0.913	1.038
il006	1.000	0.947	0.850	1.116
il007	0.659	0.700	0.729	0.708
il008	0.783	0.825	0.965	1.046
il009	0.710	0.950	0.847	0.959
il010	0.702	0.815	0.672	0.894
il013	0.773	0.976	0.841	0.987
il014_R	0.767	0.460	0.596	0.575

【表 2】統率性の項目へのパス係数

	日本	シンガポール	マレーシア	タイ
i2001	1.000	1.000	1.000	1.000
i2002	0.607	0.923	0.826	0.889
i2003	0.622	0.783	0.774	0.775
i2004	0.883	0.979	0.947	0.939
i2005	0.912	0.878	0.765	0.957
i2006	0.745	0.983	1.027	0.974
i2008	0.780	0.617	0.635	0.699
i2009	0.695	0.816	0.634	0.757
i2010	0.537	0.769	0.743	0.885
i2014	0.553	0.960	0.697	0.807
i2015	0.943	0.943	0.797	0.836
i2011_R	0.194	−0.074	0.104	0.052

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 103 2016/08/29 14:00:39

【表 3】調整性の項目のパス係数

	日本	シンガポール	マレーシア	タイ
i8001	1.000	1.000	1.000	1.000
i8002	0.951	1.010	0.951	1.206
i8003	1.033	1.000	0.846	0.906
i8005	1.013	1.151	0.690	1.065
i8006	1.160	1.183	0.857	1.005
i8007	0.921	0.771	0.791	0.814
i8008	0.708	0.786	0.789	0.732
i8009	0.714	0.658	0.652	0.965
i8010	0.858	1.163	0.994	1.294
i8013	0.883	0.789	0.584	0.870
i8014	0.478	0.341	0.622	0.473

【表 4】自律性の項目のパス係数

	日本	シンガポール	マレーシア	タイ
i12001	1.000	1.000	1.000	1.000
i12002	1.070	1.228	1.563	1.104
i12003	0.726	1.174	1.099	0.792
i12004	0.831	1.059	0.605	0.539
i12005	0.985	0.954	1.194	0.968
i12006	0.858	1.008	0.972	0.737
i12007	1.435	1.307	2.041	0.914
i12008	0.732	1.232	1.112	0.709
i12009	1.111	1.040	1.323	0.866
i12010	0.812	0.952	0.647	0.907

ス係数の値は表 1～4 のとおりである。

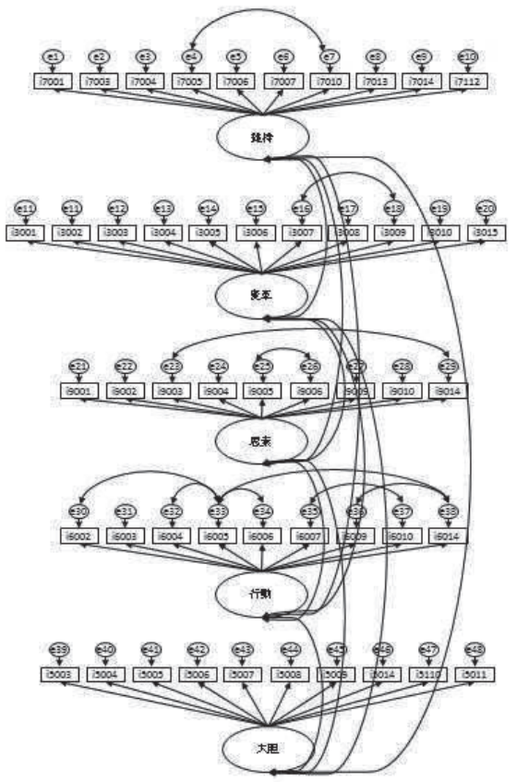
パス係数に多国間で有意差があるか見たところ、少なくともある1 カ国と別の 1 カ国の間に有意差がみられた項目は、外向性の 8 項目、統率性の 7 項目、調整性の 6 項目、自律性の 4 項目であった

なお、日本とそれ以外の国に有意差が見られたのは 20 項目、シンガポールとそれ以外の国は 21 項目、マレーシアとそれ以外の国は 10 項目、タイとそれ以外の国は 17 項目であった。

② 課題解決性

はじめに、調査で用いた項目をすべて投入して 4 母集団の配置不変 CFA を行った。モデルに若干の修正を行い、2 度目の推定を行った。具体的には、他の因子に対して高い負荷を持っている項目をモデルから除外した。また、同じ単語を用いているなど、表現の似通っている項目については誤差間に共分散を引いた。最終的に 49 項目を投入した（図 2 参照）。パス係数が 4 母集団間で等値であるモデルとそうでないモデルを試した結果、前者は CFI=.794、RMSEA=.029、AIC=9998.896 であり、後者は CFI=.800、RMSEA=.029、AIC=9986.072 であった。最終的に後者を採用した。各国におけるパス係数の値は表 5～9 のとおりである。

【図 2】課題解決性のモデル図



【表 5】維持の項目のパス係数

	日本	シンガポール	マレーシア	タイ
i7001	1	1	1	1
i7003	2.16	2.239	2.315	1.837
i7004	2.822	2.137	2.081	1.994
i7005	1.642	1.646	1.879	1.415
i7006	2.083	2.073	2.255	2.327
i7007	1.844	1.599	1.654	1.089
i7010	1.892	1.64	2.13	1.695
i7013	1.918	1.815	1.801	1.901
i7014	1.045	1.559	1.444	1.446
i7112	0.902	-0.226	0.36	-0.273

【表 6】変革の項目のパス係数

	日本	シンガポール	マレーシア	タイ
i3001	1	1	1	1
i3002	1.676	1.3	1.586	1.344
i3003	1.857	1.176	1.412	1.086
i3004	1.595	1.27	1.403	1.172
i3005	1.745	1.3	1.534	1.246
i3006	1.229	1.229	1.422	1.263
i3007	2.186	1.347	1.498	1.406
i3008	1.485	1.291	1.522	1.289
i3009	1.528	1.162	1.01	0.89
i3010	1.592	1.234	1.449	1.193
i3015	1.388	1.359	1.543	1.18

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 104 2016/08/29 14:00:40

【表 7】思索の項目のパス係数

	日本	シンガポール	マレーシア	タイ
i9001	1	1	1	1
i9002	1.039	0.929	1.352	0.879
i9003	1.287	0.824	0.777	1.099
i9004	0.907	0.8	0.945	0.923
i9005	1.268	1.168	1.421	1.244
i9006	1.603	0.975	1.082	1.164
i9009	0.676	0.87	1.173	0.846
i9010	1.211	1.249	1.512	1.559
i9014	1.336	0.755	0.904	1.037

【表 8】行動の項目のパス係数

	日本	シンガポール	マレーシア	タイ
i6002	1	1	1	1
i6003	0.64	0.631	0.287	0.72
i6004	0.74	1.121	1.218	1.122
i6005	1.061	1.338	1.668	1.425
i6006	0.955	1.133	1.251	0.861
i6007	1.048	1.067	1.401	1.098
i6009	0.858	0.965	1.288	1.143
i6010	0.835	0.939	1.573	1.307
i6014	1.013	0.979	1.537	1.346

【表 9】決断の項目のパス係数

	日本	シンガポール	マレーシア	タイ
i5003	1	1	1	1
i5004	1.209	0.784	1.019	0.872
i5005	1.414	1.118	1.101	1.07
i5006	1.231	0.927	0.83	0.901
i5007	1.391	1.015	1.115	0.997
i5008	1.159	0.92	0.859	0.988
i5009	1.083	0.891	0.881	1.073
i5014	1.044	0.955	0.903	0.869
i5110	0.545	0.838	0.537	0.725
i5011	0.69	0.313	0.251	0.143

パス係数に多国間で有意差があるか見たところ、少なくともある1 カ国と別の 1 カ国の間に有意差がみられた項目は、維持の 1 項目、変革の 3 項目、思索の 4 項目、行動の 4 項目、決断の 5 項目であった。日本とそれ以外の国に有意差が見られたのは 15 項目、シンガポールとそれ以外の国は 13 項目、マレーシアとそれ以外の国は 10 項目、タイとそれ以外の国は 9 項目であった。

【考察】

日本、シンガポール、マレーシア、タイの 4 カ国では、対人関係性の 4 因子、課題解決性の 5 因子において配置不変であることが確認された。つまり、外向性、統率性、調整性、自律性の 4 因子、維持、変革、思索、行動、大胆の 5 因子に相当する概念はいずれの国においても存在するといってもよいと思われる。ただし、それぞれの因子において、因子を代表する項目順序やその強弱が異なると考えられる。たとえば、いずれの国においても外向性を代表する項目として「友人を作るのが得意」という意味の項目が上位になるという点は共通している。一方、シンガポールやタイにおいては「会話を続けることが得意」という項目も上位になるのに対し、日本やタイではそうではないという結果であった。また、統率性を代表する項目として日本では「私はよいリーダーだと思う」という項目が上位になるが、他 3 カ国では「グループを管理するのが好きだ」という項目が上位になるという結果であった。

また、各項目の各因子へのパス係数をみると、多国間で有意差がある項目があった。対人関係性のほうが課題関係性よりも多国間で有意差がある項目が多かった。いずれの項目においても職場場面を想定した項目設計を行ったが、対人関係性の項目のほうが、職業場面であったとしても文化の影響を受けやすい可能性がある。一方、課題解決行動のほうが、商習慣が反映され、さほど国の間の差が大きくなかった可能性がある。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 105 2016/08/29 14:00:41

複数国において実施された性格特性検査における DIF の検出

坂本佑太朗・酒匂志野・今城志保
株式会社リクルートマネジメントソリューションズ

問題と目的

企業における昇進昇格等の個人や組織の将来を左右するハイステイクスな場面で用いられるテストにおいて，そのテスト項目が「測りたい特性を」（妥当性の問題）「精度よく」（信頼性の問題）測定できているかを確認することは重要である．もしそれが複数国で実施されるような場合には，「テストの公平性」の観点からも，測定の等価性を担保することは必須となる．

このような問題に対し，テスト研究においては特異項目機能（differential item functioning, DIF）の問題として取り扱われてきた．その検出方法は様々提案されているが（Su, & Wang, 2005），その中でも項目反応理論（item response theory；IRT）（Lord, 1952）による方法は 1）テスト開発において IRT が基盤となっている場合に用いやすい，2）各国における受検者の反応状況を潜在特性尺度値にもとづいて把握できることなどが利点として挙げられる．

そこで本研究では，複数国において実施された性格特性検査データにおいて，IRT を基盤とした”EasyDIF”（熊谷，2012）による DIF の検出を試みる．

分析

方法

実施対象国　日本（N=331）・シンガポール（N=313）・マレーシア（N=300）・タイ（N=300）

対象者層　各国とも 30〜49 歳の男性が対象となり，勤務先企業の従業員規模は 500 名以上である．

調査日　2016 年 2 月 16 日〜3 月 3 日

調査法　インターネット調査（調査会社に依頼）

分析対象となる項目

「対人関係性」を測定する項目群（4 件法）から，選択枝すべてに回答がない項目や EasyDIF における多母集団の段階反応モデル（graded response model；GRM）（Samejima, 1969）においてステップ母数が異常に大きく推定された項目を除外した計 45 項目を分析対象とした．

なお，「対人関係性」には下位領域として「外向性」（12 項目（内 1 項目は逆転項目）），「統率性」（11 項目），「調整性」（10 項目），「自律性」（12 項目）が設定されている．

結果

「対人関係性」を構成する「外向性」，「統率性」，「調整性」，「自律性」それぞれについて EasyDIF を用いて分析を行い，指標 K の値から「外向性」，「統率性」において 2 項目ずつ DIF 項目が検出された．その項目内容と指標 K の値を表 1，各項目の項目期待得点曲線（item expected score curve）を図 1~図 4 に示す．なお，項目名は本研究において便宜的に与えたものである．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　106　2016/08/29　14:00:42

表 1 DIF 項目と指標 K（DIF_item2 は逆転項目）

項目名	測定領域	項目内容	指標K
DIF_item1	外向性	I rarely spend time alone.	0.304
DIF_item2	外向性	I tend to be quiet around strangers.	0.473
DIF_item3	統率性	I enjoy giving instructions to other members.	0.328
DIF_item4	統率性	I like taking charge when I work in a group.	0.310

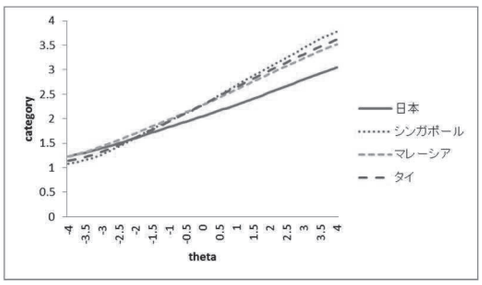


図 1 DIF_item1 の項目期待得点曲線

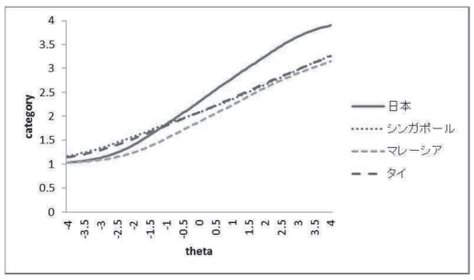


図 2 DIF_item2 の項目期待得点曲線

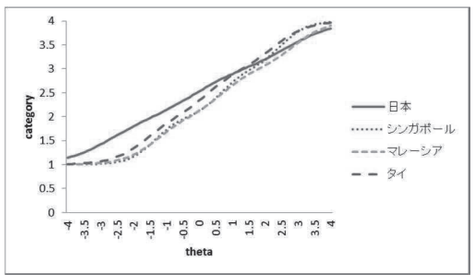


図 3 DIF_item3 の項目期待得点曲線

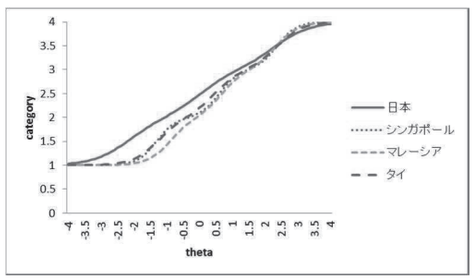


図 4 DIF_item4 の項目期待得点曲線

もっとも指標 K が大きかった DIF_item2 では，「外向性」の潜在特性尺度値θ = −1を境にして，その値が大きくなるにつれて，日本の受検者は他国の受検者よりも当該の逆転項目（”I tend to be quiet around strangers.”）に対して否定的に回答する傾向がある．言い換えれば，日本の受検者は「外向性」が高い人ほど他国の受検者よりも DIF_item2 に敏感に反応していることになる．つまり，他国と比較して日本の受検者は，見知らぬ人々の中で発言するなどの行動に対して「外向性」が反映されやすいことがわかる．

考察

本研究では，複数国において実施された性格特性検査における「対人関係性」の 4 側面について，EasyDIF を用いた DIF の検出を試みた．その結果，「外向性」「統率性」においてそれぞれ 2 項目の合計 4 項目に DIF が見られた．

今後の課題としては，以下の 2 点が指摘できる．1 つ目は，本研究で示したのが一次元性を仮定する IRT モデルの下で検出された DIF であり，厳密には「対人関係性」における因子間の関係を考慮していない点である．また，2 つ目は，同データを用いた酒匂・坂本・今城・分寺（2016）によって示された配置不変の因子構造において，各国の潜在特性尺度値の示す意味が厳密には異なる可能性があることである．そのため，DIF を引き起こす要因が，各国の文化差であるのか構成概念的なバイアスによるものであるのかの特定ができない．これらを特定するためには，最終的には研究者の定性的な判断が入らざるを得ず，これは定量的な DIF 検出法の限界であると考えられる．

230934_多摩大学_日本テスト学会発表論文抄録集_本文. indd 107 2016/08/29 14:00:42

連続変数を含む補償的認知診断モデルの開発

○山口一大[*]・星野崇宏[**]
*東京大学大学院教育学研究科，**慶應義塾大学大学院経済学研究科

1. 問題と目的

近年，認知診断モデル(Cognitive diagnostic model: CDM; Leighton & Gierl, 2007)への注目が集まり，数多くのモデルが開発されている（DiVello, Roussos, & Staout, 2007）。中でも log-linear cognitive diagnostic model(LCDM; Henson, Templin, &Willse, 2009)や Generalized DINA model(G-DINA; de la Torre, 2011)は他の CDM を包含するモデルとして一般性の高いモデルである。これらは，しかしながら，カテゴリカル潜在特性であるアトリビュートと Q 行列(Tatsuoka, 1983)によってのみ正答確率が決まるモデルであり，潜在的な連続変数を含めたモデルではない。しかしながら，このようなカテゴリカルな変数のみによるモデルでは，テストワイズネス（村山, 2006）といった，認知要素以外の正答に寄与する不純な要素を考慮できていない。また，実データへの CDM の適用論文では“全てのアトリビュートを習得している”あるいは，“１つもアトリビュートを習得していない”といった習得パタンが多く得られている（例えば，Li, Hunter, & Lei, 2016）。診断に関係する認知要素以外の不純な要素を除去することで，より純粋な認知要素による推定が可能となり，このような極端な習得パタンが多く推定される現象も低減されると期待できる。そこで，本研究では，カテゴリカルな潜在変数に加えて連続的な潜在変数を含めた CDM を開発し，実際のデータに適用することでモデルの有用性を検証する。

2. 解析モデル

本研究では，以下のモデルを提案する。

$$logit(\Pr(x_{ij}=1|\boldsymbol{\alpha}_i,\boldsymbol{q}_j,\theta_i,a_j))=\lambda_{j0}+\sum_{a=1}^{A}\lambda_{ja}\alpha_{ia}q_{ja}+\theta_i a_j$$

$logit(\Pr(\cdot))=\log\left(\frac{\Pr(\cdot)}{1-\Pr(\cdot)}\right)$という演算である。添字$i(=1,\dots,I)$は解答者，$j(=1,\dots,J)$は項目番号を意味しており，$x_{ij}=1$は解答者$i$の項目$j$への反応が 1（正答）であること意味している。$\boldsymbol{\alpha}_i=[\alpha_{i1},\dots,\alpha_{ia},\dots,\alpha_{iA}]^t$であり，各個人ごとの$A$個のアトリビュートの習得パタンを示しており，$\alpha_{ia}=\{0,1\}$である(ただし，$a=1,\dots,A$)。$\alpha_{ia}$の取る値 0，1 はそれぞれアトリビュート$a$の未習得・習得に対応する。$\boldsymbol{q}_j=\left[q_{j1},\dots,q_{ja},\dots,q_{jA}\right]^t$であり，Q 行列のある項目$j$に対して，ど

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 108 2016/08/29 14:00:43

のアトリビュートが必要であるか否かを示したベクトルである。q_{ja}が 0 ならば，アトリビュートaは項目jの正答に必要なく，1 であれば必要である。さらに，θ_iは$\boldsymbol{\alpha}_i$によらず，項目の正答に寄与する要素であり，a_jはθ_iの項目への影響を調整するパラメータである。これらのθ_i, a_jは項目反応理論における 2 パラメータロジスティックモデルでの能力値パラメータと項目識別力に対応したパラメータである。$\lambda_{j1} \ldots, \lambda_{jA}$はそれぞれの項目の正答に必要なアトリビュートを習得している場合に正答確率を変化させるパラメータである。それぞれの項目に対して，これらのパラメータはA個必要であるわけではなく，$\sum \boldsymbol{q}_j$個のみ仮定され，$\boldsymbol{q}_j$が 0 にあたるλ_{ja}は推定されない。$\lambda_{j0,}$は切片に対応し，基本的な当て推量パラメータである。

本モデルは LCDM のなかで複数が必要な項目に対する交互作用項を省略した CRUM(Rupp, Templin, & Henson, 2010)に，θ_i, a_jを加えたモデルである。θ_iは先に述べたように，$\boldsymbol{\alpha}_i$に依存しないパラメータであり，診断的な興味は無いが，正答率に寄与する個人の特性であり全般的なテスト解答方略能力などとして具体的に解釈することができる。a_jは項目への全般的なテスト方略の適用しやさを反映したパラメータと考えることができよう。

3. 方法

データには 2015 年にスタディサプリ®で収集された，高校 1 年生の数学のテストデータの中の一部を用いた。テストは必答問題 2 問，選択問題 4 問を含んでおり，分析には，必答問題 1(数学Ⅰ 式の計算)，必答問題 2(数学Ⅰ1 次不等式・集合と論証)，選択問題 1(数学Ⅰ2 次関数 最大・最小まで)の大問 3 つ(項目数 38)および，それらに解答者の中で有効である 9538 名を対象とした。

Q 行列の作成は第 1 著者が行い，教科書の領域の区分に加えて，領域横断的に必要と考えられる能力を仮定した。具体的には，「基本的演算」，「用語・記号の理解」，「不等式」，「論理と集合」，「二次方程式の性質」の 5 つのアトリビュートを仮定した。「基本的演算」は，問題の正答に必要な方程式の解を求める能力や計算能力であり，「用語・記号の理解」は問題文に出てくる数学の用語や記号の理解ができるか否かを示す能力であった。「不等式」は不等式の基本的な性質を理解しているかどうか，「論理と集合」は論理と集合の基本的な性質を理解しているかどうかを決める能力であり，最後の「二次方程式の性質」は平方完成や 2 次方程式の最大・最小，解の存在などの性質を理解しているかどうかを示す能力として定義した。

分析の際に，比較対象のモデルとして CDM の適用場面で利用されることの多い，DINA(Junker & Sijtsma, 2001)，DINO(Templin, & Henson, 2006)，CRUM，LCDM を用いた。DINA, LCDM, CRUM はいずれも補償型のモデルであり，DINO のみ非補償型のモデルである。解析には Mplus version 7.11(Muthén & Muthén, 1998-2012)を用いた。推定にはロバスト最尤推定を用いた

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 109 2016/08/29 14:00:43

4.結果・考察

モデルを比較するために，表 1 に情報量基準の一覧を示した。提案したモデルはAIC = 298110.3，BIC = 299542.9 であり他のモデルよりも低い値を示し，良好なモデルであることが示された。AIC, BIC によるモデル比較の順位はどちらも同じで, 順に LCDM, CRUM, DINA, DINO の順であった。一方，分類の精度を示す Entorpy の値では，提案モデルは.815 であり，最も低い値を示した。Entorpy の順位は CRUM, LCDM, DINA, DINO,提案モデルの順であった。

表 1．モデルごとの情報量基準

モデル	パタメータ数	対数尤度	AIC	BIC	Entropy
提案モデル	200	-148855.1	298110.3	299542.9	.815
CRUM	132	-158883.7	318031.5	318977.0	.878
DINA	107	-164159.1	328532.3	329298.7	.851
DINO	107	-165548.5	331311.1	332077.5	.830
LCDM	157	-158432.8	317179.7	318304.3	.857

図 1 に提案モデルともっとも分類の精度が良好であった CRUM モデルで推定されたアトリビュート習得パタンの割合を示した。5 桁の数字はアトリビュート習得パタンに対応し，各桁がアトリビュートに対応している。具体的には，それぞれの桁の大きい位から，「基本的演算」，「用語・記号の理解」，「不等式」，「論理と集合」，「二次方程式の性質」に対応する。例えば，「10000」は「基本演算」のみ習得している状況を意味している。図から，提案モデルでは「11111」という全てのアトリビュートを習得しているパタンや，「00000」という全くアトリビュートを習得していないパタンといった極端な習得パタンの割合が CRUM モデルよりも少なく推定された。また，「11100」のように CRUM モデルに比べて提案モデルで多く推定されたアトリビュート習得パタンもみられた。

本研究で提案したモデルは，連続的な潜在変数をモデルに組み込むことで，これまで応用場面で多く利用されてきた CDM よりも実データにおいてよい当てはまりを示し，有用である可能性が示唆された。実際のモデル適用にあたっては，テストワイズネスのような撹乱要因を除去するための工夫が必要であり，本モデルによって診断に不必要な要素を除去できたと考えられる。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 110 2016/08/29 14:00:44

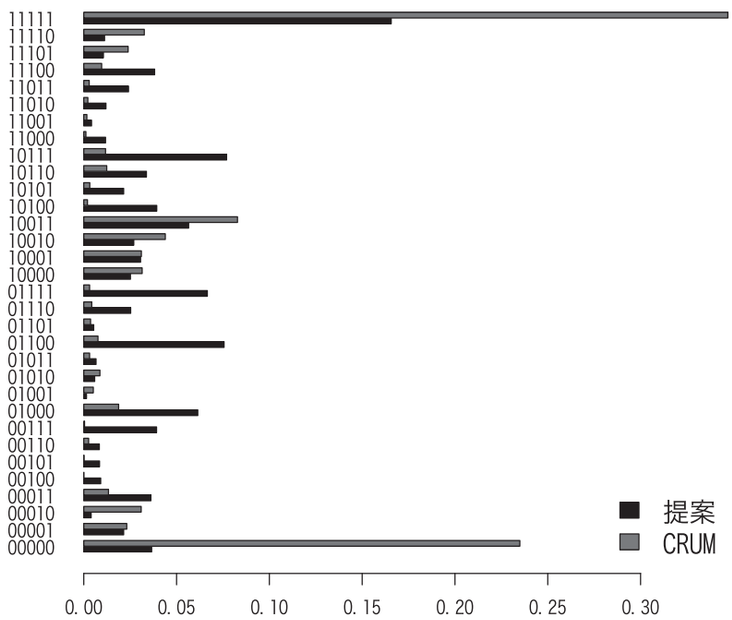


図 1 提案モデルと CRUM モデルにおける解答者のアトリビュート習得パタンの割合。5桁の数値はそれぞれの桁の大きい位から，「基本的演算」，「用語・記号の理解」，「不等式」，「論理と集合」，「二次方程式の性質」に対応する。

主要参考文献

De La Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.

Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. New York: Cambridge University Press.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York London: The Guilford Press.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. Psychometrika, 74(2), 191-210.

Email: kazz530@p.u-tokyo.ac.jp

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 111 2016/08/29 14:00:45

カテゴリ項目尺度における希薄化の修正について

○岡本安晴
日本女子大学人間社会学部

問題　測定値（項目和）間の相関係数は、それぞれの尺度の信頼性係数に応じて真値間の相関係数より小さくなる（希薄化）。真値間の相関係数を得るための修正公式があるが、この公式はテスト項目が連続量であることを想定するものである。項目が順序カテゴリ項目であるときは、注意が必要である。テスト項目が連続量である場合、修正公式は次のように導かれる（Raykov & Marcoulides, 2011, p.192）。

２つのテストの得点（連続量）をX およびY で表し、

$$\mathrm{X} = T_X + E_X \qquad \text{および} \qquad \mathrm{Y} = T_Y + E_Y \tag{1}$$

とおく。T_XとT_Yは真値、E_XとE_Yは誤差である。誤差は、他の変数とは独立である。このとき、

$$\rho_{XY} = \frac{Cov(T_X + E_X, T_Y + E_Y\ )}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(T_X, T_Y\ )}{\sqrt{Var(T_X)}\sqrt{Var(T_Y)}} \frac{\sqrt{Var(T_X)}\sqrt{Var(T_Y)}}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$
$$= \rho_{T_X T_Y}\sqrt{\rho_X}\sqrt{\rho_X}$$

が成り立つ。これより、修正式

$$\rho_{T_X T_Y} = \rho_{XY} / \left(\sqrt{\rho_X}\sqrt{\rho_Y}\right) \tag{2}$$

が導かれる。しかし、

$$\rho_X = Var(T_X)/Var(X)\text{、} \quad \rho_Y = Var(T_Y)/Var(Y)$$

は、X およびY が連続量のときのモデルから導かれるものであり、順序カテゴリ項目のモデルにおけるものではない。

順序カテゴリ項目のモデル　テスト得点X およびY が順序カテゴリ項目X_jおよびY_jの和得点であるとする。このとき、X_jおよびY_jは、項目の真値 T_{X_j}および T_{Y_j}の非線形変換

$$\begin{aligned} C_{k-1} < T_{X_j} + E_{X_j} \le C_k \quad &\rightarrow \quad X_j = k \\ D_{h-1} < T_{Y_j} + E_{Y_j} \le D_h \quad &\rightarrow \quad Y_j = h \end{aligned} \tag{3}$$

として与えられる。テスト得点 X および Yは、それらの和

$$\mathrm{X} = \sum X_j\text{、} \qquad \text{および} \qquad \mathrm{Y} = \sum Y_j$$

で与えられる。

テストの真値を

$$T_X = \sum T_{X_j}\text{、} \qquad \text{および} \qquad T_Y = \sum T_{Y_j}$$

とおくとき、真値 T_X および T_Y によって測定値 X および Y を表す回帰モデルは次式で表される。

$$X = a_X + b_X T_X + e_X \qquad \text{および} \qquad Y = a_Y + b_Y T_Y + e_Y \tag{4}$$

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　112　2016/08/29　14:00:45

このとき、

$$E[e_X] = E[e_Y] = Cov(T_X, e_X) = Cov(T_Y, e_Y) = 0$$

である。モデル（１）がデータ生成モデルであるのに対して、モデル（４）は変数間の関係を表す記述モデルであることに注意。

上のモデル（４）のもとで、テスト得点（尺度）間の相関係数

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

を計算する。

$$Cov(X,Y) = Cov(a_X + b_X T_X + e_X, a_Y + b_Y T_Y + e_Y)$$
$$= b_X b_Y Cov(T_X, T_Y) + b_X Cov(T_X, e_Y) + b_Y Cov(T_Y, e_X) + Cov(e_X, e_Y)$$

いま、仮に

$$Cov(T_X, e_Y) = Cov(T_Y, e_X) = Cov(e_X, e_Y) = 0 \qquad \text{（仮説１）}$$

であると仮定すると

$$Cor(X,Y) = \frac{b_X b_Y Cov(T_X, T_Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(T_X, T_Y)}{\sqrt{Var(T_X)Var(T_Y)}} \sqrt{\frac{b_X^2 Var(T_X)}{Var(X)}} \sqrt{\frac{b_Y^2 Var(T_Y)}{Var(Y)}}$$

$$= Cor(T_X, T_Y)\sqrt{R_{XT}^2}\sqrt{R_{YT}^2}$$

を得る。これより、次式を得る。

$$Cor(T_X, T_Y) = Cor(X,Y) \Big/ \left(\sqrt{R_{XT}^2}\sqrt{R_{YT}^2}\right) \qquad \text{（仮修正式１）}$$

ここで、R_{XT}^2およびR_{YT}^2は、モデル（４）における決定係数であり、次式で与えられる。

$$R_{XT}^2 = \frac{Var(a_X + b_X T_X)}{Var(X)} = \{Cor(X, T_X)\}^2$$

$$R_{YT}^2 = \frac{Var(a_Y + b_Y T_Y)}{Var(Y)} = \{Cor(Y, T_Y)\}^2$$

（仮説１）のもとで、修正式（２）に対応する（仮修正式１）が得られるが、この仮修正式の有用性は（仮説１）に依存している。仮説１について、シミュレーションによる検討を行ってみたところ反例が得られたので、以下に報告する。

（仮説１）の反例

順序カテゴリ項目のデータ生成モデル（３）を以下のように設定する。

テストＸおよびＹの項目数はともにＭ＝５ 項目、回答者数は ｎ＝1,000,000 人であるとする。回答者 i の X、Y、X_j、Y_j、T_{X_j}、T_{Y_j}、E_{X_j}、E_{Y_j} をそれぞれ X_i、Y_i、X_{ij}、Y_{ij}、T_{Xij}、T_{Yij}、E_{Xij}、E_{Yij} で表す。誤差E_{Xij}およびE_{Yij}は自分自身以外の変数とは独立で正規分布に従う、すなわち

$$E_{Xij} \sim N(0, \psi_{Xj}^2)、 \qquad E_{Yij} \sim N(0, \psi_{Yj}^2)$$

であるとする。カテゴリ数は、X_{ij}およびY_{ij}ともに２であるとし、カテゴリ境界値は０、すなわち

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 113 2016/08/29 14:00:46

$$C_1 = D_1 = 0$$

であるとする。真値に対して次の１因子モデルを設定する。

$$T_{Xij} = \mu_{Xj} + \lambda_{Xj} F_i 、 \quad T_{Yij} = \mu_{Yj} + \lambda_{Yj} G_i$$

ここで、

$$(F_i, G_i) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) 、 \quad \boldsymbol{\mu} = \mathbf{0} 、 \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

であり、誤差 E_{Xij} および E_{Yij} は、F_i および G_i と独立である。

シミュレーションにおけるパラメータ値を

$$\rho = 0.7 、 \quad \mu_{Xj} = \mu_{Yj} = 0 、 \quad \lambda_{Xj} = \lambda_{Yj} = 0.9 、 \quad \psi_{Xj} = \psi_{Yj} = 0.3$$

とおく。したがって、

$$X_{ij} = \begin{cases} 2 , & 0.9F_i + E_{Xij} > 0 \quad のとき \\ 1 , & 0.9F_i + E_{Xij} \leq 0 \quad のとき \end{cases}$$

$$Y_{ij} = \begin{cases} 2 , & 0.9G_i + E_{Yij} > 0 \quad のとき \\ 1 , & 0.9G_i + E_{Yij} \leq 0 \quad のとき \end{cases}$$

である。

上の条件のもとでシミュレーションを行い、以下の結果を得た。

生成されたサンプル値(F_i, G_i) の相関係数の値は

$$r_{F_i G_i} \approx 0.70005$$

であり、(X_i, Y_i) の相関係数の値は

$$r_{X_i Y_i} \approx 0.56281$$

であった。真値 F_i と G_i の標本相関係数の値 $r_{F_i G_i}$ は、モデルの母数 $\rho = 0.7$ に十分近い値である。希薄化の大きさは

$$r_{X_i Y_i} / \rho \approx 0.56281/0.7 \approx 0.804$$

となり、約 80%に希薄化されている。

希薄化の大きさと信頼性係数との関係を見るため、α 係数とR^2を求めてみた。α係数は

α(テストXのサンプル値) ≈ 0.9254、α(テストYのサンプル値) ≈ 0.9254、α(母数) ≈ 0.9254

であった。α 係数 0.9254 は、希薄化率$r_{X_i Y_i} / \rho \approx 0.804$ より大きい。修正式（２）において、信頼性係数の値が同じであるとき（$\rho_X = \rho_Y$）、$\sqrt{\rho_X}\sqrt{\rho_Y} = \rho_X$ であるので、希薄化率は信頼性係数に等しくなる。

決定係数R^2 の値は

R^2(テストXのサンプル値) ≈ 0.7436、R^2(テストYのサンプル値) ≈ 0.7439、R^2(母数) ≈ 0.7438

であった。R^2の値、約 0.744 は、希薄化率 $r_{X_i Y_i} / \rho \approx 0.804$ にα係数より近い値である。すなわち、

$$\frac{\left| r_{X_i Y_i} / \rho - R^2 \right|}{\left| r_{X_i Y_i} / \rho - \alpha \right|} \approx \frac{0.06}{0.12} = 0.5$$

しかし、R^2と$r_{X_i Y_i} / \rho$ の差 0.06 は、サンプル数が 1,000,000 であることを考えると無視できない。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 114 2016/08/29 14:00:46

（仮説１）の検証を行う。サンプルから算出された（仮説１）に関係する値は以下の通りであった。

$Cov(T_X, e_Y)$(サンプル値) ≈ −0.00036、$Cor(T_X, e_Y)$(サンプル値) ≈ −0.00032、
$Cov(T_Y, e_X)$(サンプル値) ≈ −0.00119、$Cor(T_Y, e_X)$(サンプル値) ≈ −0.00107
$Cov(e_X, e_Y)$(サンプル値) ≈ 0.20576、$Cor(T_X, e_Y)$(サンプル値) ≈ 0.16681

これらの値は、サンプル数 $n = 1{,}000{,}000$ を考えると、仮説１の反例と考えられる。特に、誤差間の共分散の値約 0.21、相関係数の値約 0.17 は、誤差間の関係が無視できないことを示す値であると考えられる。

希薄率が α 係数や R^2 と異なること（$\rho_X = \rho_Y$ のとき、$\sqrt{\rho_X}\sqrt{\rho_Y} = \rho_X$ である）、（仮説１）が成り立たないことから、希薄化の修正は、修正式（２）のような単純な式によって行うことは困難と考えられる。しかし、今日の PC の高性能化により、確率モデルを設定して相関係数などのパラメータの推定を行うことが実用的方法となっている。順序カテゴリ項目より構成されたテストの真値間の相関係数を推定するための確率モデルを以下のように設定する。

順序カテゴリ項目テストの階層的確率モデル

上で用いた順序カテゴリ項目のモデルにおいて、相関係数 ρ の分布を与える上位のモデルを設定する。すなわち、例えば β 分布を用いるなら

$$\frac{\rho + 1}{2} \sim \mathrm{Beta}(\alpha, \beta)$$

である。上式で与えられる相関係数 ρ の確率密度関数を簡単に

$$P(\rho) = P(\rho|\alpha, \beta)$$

と書くことにする。この事前分布から事後分布を求める。

尤度関数を

$$L(data|parameters) = \prod\prod P\left(X_{is} = k, Y_{it} = h|F_i, G_i, \rho, \mu_{Xj}, \mu_{Yj}, \lambda_{Xj}, \lambda_{Yj}, \psi_{Xj}, \psi_{Yj}, C_k, D_k\right)$$

と表し、事前分布を

$$P_{prior}(parameters)$$

と表す。パラメータ間の独立性と階層モデル

$$P(F_i, G_i, \rho) = P(F_i, G_i|\rho)P(\rho)$$

を仮定すれば、

$$P_{prior}(parameters) = P(F_i, G_i|\rho)P(\rho) \cdots P(D_k)$$

となる。このとき、事後分布は

$$P_{post}(parameters) \propto P_{prior}(parameters)L(data|parameters)$$

と表される。

事後分布は、MCMC により推定できるが、データとモデルとの関係において何らかの工夫が必要な場合も考えられる。Okamoto (2013)は、カテゴリが２値の場合に誤差分散に事前分布で制約を置くことの有効性を報告している。相関係数 ρ の事後分布の推定例は、別の機会に報告したい。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 115 2016/08/29 14:00:47

対応づけ得点の信頼性係数について

○佐藤喜一*
*新潟大学教育戦略統括室

柴山　直**
**東北大学大学院教育学研究科

1　はじめに

等化や予測の技術を利用すれば，形式的には任意の二つの異なるテストの対応づけ（linking）[1)]が可能である．しかしながら，意味ある対応づけを保証するには，さまざまな観点から対応づけ可能性（linkability）を吟味する必要がある．その中でも，テストの信頼性は対応づけ結果の正しい解釈のために重要な要素の一つである．

本研究の目的は，対応づけ得点の信頼性という新たな視点を対応づけ可能性分析に導入することである．佐藤・柴山（2015）では，単一グループデザインのもとでの対応づけを想定し，対応づけ得点の信頼性に関する指標をいくつか提案した（2.2 節）．本発表では，対応づけ得点の信頼性係数を新たに定式化するとともに，それらの指標を対応づけ可能性分析に応用することを試みる．

2　定式化

2.1　測定モデル

二つの異なるテスト X, Y のテスト得点が古典的テスト理論に従うと仮定する．それぞれテスト X, Y の観測得点を X, Y，真の得点を T_X, T_Y，測定誤差を E_X, E_Y とすれば，

$$X = T_X + E_X$$
$$Y = T_Y + E_Y$$

と表される．ここで，すべての変数は受験者に基づく確率変数であり，測定誤差の期待値は $\mathrm{E}(E_X) = \mathrm{E}(E_Y) = 0$ である．また，E_X は T_X, Y, T_Y, E_Y と線形独立であり，E_Y は T_Y, X, T_X, E_X と線形独立である．

いま，単一グループデザインのもとでテスト X をテスト Y に対応づけする場面を想定しよう．本研究では，対応づけ得点は共通尺度となるテストの真の得点を別のテストを通して間接的に測定した結果であると考える．この場合，対応づけ得点 $l_Y(X)$ を観測得点，テスト Y の真の得点 T_Y を $l_Y(X)$ の真の得点とみなすことができる．その差を対応づけ得点の測定誤差 $E_Y{}' = l_Y(X) - T_Y$ と定義すれば，対応づけ得点の測定モデルは，

$$l_Y(X) = T_Y + E_Y{}'$$

と記述できる．ここで，$T_Y, E_Y{}'$ は受験者に基づく確率変数であり，$l_Y(X)$ は確率変数 X の関数である．なお，どんな対応づけ方法を採用するかにより，$\mathrm{E}(E_Y{}') = 0$ とは限らないし，$E_Y{}'$ は T_Y や他の確率変数と線形独立とも限らない．

2.2　佐藤・柴山（2015）の提案指標

実際の対応づけ場面では，離散的な得点，平滑化，範囲調整，整数化などの影響により，一般には $\mathrm{E}(E_Y{}') \neq 0$ である．佐藤・柴山（2015）では，$\mathrm{E}(E_Y{}')$ を含めた信頼性を評価するため，対応づけ得点 $l_Y(X)$ とその真の得点 T_Y との距離を表す指標として，測定の MSE（mean square error）と測定の RMSE（root mean square error）を定式化した．以下に，佐藤・柴山（2015）の導出結果をリストする．ただし，任意の確率変数 U, V の相関係数を $\rho(U, V)$ と表記する．また，$\rho(Y)$ はテスト Y の信頼性係数であり，$\sqrt{\rho(Y)}$ はテスト Y の信頼性指数である．

測定誤差の期待値：

$$\mathrm{E}(E_Y{}') = \mathrm{E}[l_Y(X)] - \mathrm{E}(Y) \qquad (1)$$

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　116　2016/08/29　14:00:47

測定誤差分散：

$$\begin{aligned}&\sigma^2(E_Y{}')\\=&\sigma^2[l_Y(X)-Y]-\sigma^2(Y)[1-\rho(Y)]\end{aligned}\quad(2)$$

測定の標準誤差：$\sqrt{(2)\text{ 式}}$

測定の MSE：

$$\begin{aligned}&\mathrm{E}\{[l_Y(X)-T_Y]^2\}\\=&\sigma^2(E_Y{}')+[\mathrm{E}(E_Y{}')]^2\end{aligned}\quad(3)$$

測定の RMSE：$\sqrt{(3)\text{ 式}}$

信頼性指数：

$$\rho[l_Y(X),T_Y]=\frac{\rho[l_Y(X),Y]}{\sqrt{\rho(Y)}}\quad(4)$$

2.3　信頼性係数

テスト Y の信頼性係数は，$\sigma^2(T_Y)/\sigma^2(Y)=\sigma^2(T_Y)/[\sigma^2(T_Y)+\sigma^2(E_Y)]$ と表される．本節では，右辺の式に対応づけ得点の統計量をあてはめ，対応づけ得点 $l_Y(X)$ の信頼性係数を定義する．$l_Y(X)$ の真の得点は T_Y なので，$\sigma^2(T_Y)$ をそのまま利用する．また，$\mathrm{E}(E_Y{}')$ を含めて評価するため，測定誤差分散の代わりに測定の MSE を利用する．このとき，対応づけ得点 $l_Y(X)$ の信頼性係数は，

$$\rho[l_Y(X)]=\frac{\sigma^2(T_Y)}{\sigma^2(T_Y)+\mathrm{E}\{[l_Y(X)-T_Y]^2\}}\quad(5)$$

と表される．

(5) 式を用いると，対応づけ得点 $l_Y(X)$ の信頼性係数をテスト Y の信頼性係数と同様に評価できる．なぜなら，両者の信頼性係数が等しいとき，$\sigma^2(E_Y)=\mathrm{E}\{[l_Y(X)-T_Y]^2\}$ であり，二乗誤差基準のもとで観測得点 Y と対応づけ得点 $l_Y(X)$ の測定精度が等しいからである．なお，(5) 式の取り得る値の範囲は $0\leq\rho[l_Y(X)]\leq 1$ である．

3　応用

3.1　対応づけ可能領域の図示

(5) 式の信頼性係数がある基準値 ρ_C（例：$\rho_C=0.7$）以上であれば，対応づけ得点 $l_Y(X)$ の信頼性は十分に高いと判断できる．このとき，対応づけ得点の信頼性の観点からは，十分に意味のある対応づけ結果が得られる．本節では，このときの条件 $\rho[l_Y(X)]\geq\rho_C$ について考察する．

線形等化法や等パーセンタイル法を利用する場合，$\mathrm{E}(E_Y{}')\approx 0, \sigma[l_Y(X)]\approx\sigma(Y)$ を考慮すれば，(5) 式 $\geq\rho_C$ より，

$$\rho[l_Y(X),Y]\geq\frac{2\rho_C-1}{2\rho_C}\rho(Y)+\frac{1}{2}\quad(6)$$

という不等式が導かれる．また，(4) 式 ≤ 1 より，点 $(\rho(Y),\rho[l_Y(X),Y])$ が取り得る領域は，

$$\rho[l_Y(X),Y]\leq\sqrt{\rho(Y)}\quad(7)$$

と制限される．(6) 式と (7) 式を同時に満たす領域を図 1 に斜線で示す（$\rho_C=0.7$）．この領域が $\rho[l_Y(X)]\geq\rho_C$ を実際に満たす領域であり，対応づけ得点の信頼性の観点から対応づけ可能と考えられる領域である．

図 1 をみると，(6) 式の直線の傾きが負でないとき（$\rho_C\geq 1/2$），縦軸の値だけで対応づけ可能と判断できる領域があることがわかる．縦軸の値が点線以上のとき，実際の点 $(\rho(Y),\rho[l_Y(X),Y])$ は常に斜線の領域に含まれる．たとえば，$\rho_C=0.7$ として $\rho[l_Y(X),Y]\geq 0.786$ であれば，対応づけ得点の信頼性は $\rho[l_Y(X)]\geq 0.7$ が保証される．

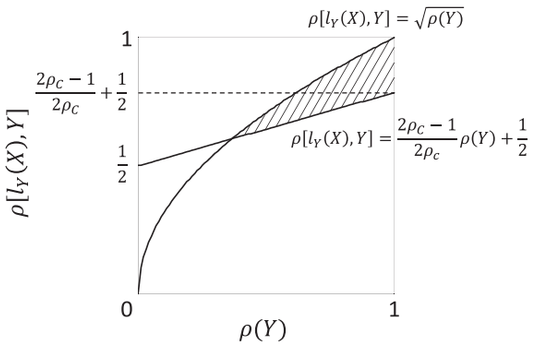


図 1: 対応づけ可能領域（境界線を含む）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　117　2016/08/29　14:00:48

3.2　α係数による推定値

対応づけ得点 $l_Y(X)$ の信頼性係数を推定するには，テスト Y の信頼性係数の推定値が必要である．本節では，信頼性係数の推定値として Cronbach の α 係数を用いた場合，(5) 式の推定値にどんな影響を与えるのかを考察する．

$\sigma^2(T_Y) = \sigma^2(Y)\rho(Y)$ より，(5) 式は，

$$\rho[l_Y(X)] = \frac{\sigma^2(Y)}{2\sigma^2(Y) + R/\rho(Y)} \qquad (8)$$

と書き換えられる．ただし，

$$R = \sigma^2[l_Y(X) - Y] - \sigma^2(Y) + [\mathrm{E}(E_Y{}')]^2$$

である．期待値と分散の推定については，一致性と不変性をもつ標本平均と不偏分散を用いれば，推定値の上限・下限の議論に影響を与えない．一方，受験者数が大きい場合，α 係数は当該テストの信頼性係数の下限値であることが証明されている．それゆえ，(8) 式の上限・下限を議論するには，$\rho(Y)$ を含む項 $R/\rho(Y)$ に着目すればよい．

R そのものは，上限・下限の議論に影響を与えない．しかしながら，R の正負によって $R/\rho(Y)$ の上限・下限が入れ替わる．α 係数を用いた信頼性係数を $\rho_\alpha(\cdot)$ とすれば，(8) 式の上限・下限の議論は以下のようにまとめられる．

- $R > 0$ すなわち $\rho[l_Y(X)] < 1/2$ のとき，$\rho_\alpha[l_Y(X)]$ は $\rho[l_Y(X)]$ の下限値となる．
- $R = 0$ すなわち $\rho[l_Y(X)] = 1/2$ のとき，$\rho_\alpha[l_Y(X)]$ は $\rho[l_Y(X)]$ に一致する．
- $R < 0$ すなわち $\rho[l_Y(X)] > 1/2$ のとき，$\rho_\alpha[l_Y(X)]$ は $\rho[l_Y(X)]$ の上限値となる．

多くの場合，信頼性係数はテストの信頼性が高いことを調べるために利用される．実用的には，3 番目の性質を利用すれば，次節に示すような対応づけ可能性分析が可能である．

3.3　対応づけ可能性のネガティブチェック

例 1 として，やや異なる構成概念を測定するテスト X, Y について，単一グループデザインのもとでテスト X をテスト Y に対応づけする場面を想定した．項目反応データは，Rasch モデルを利用して生成した．テスト X, Y の正答数得点についての基礎統計量を表 1 に示す．表 2 に，線形等化法と等パーセンタイル法を用いた場合の対応づけ得点の信頼性を示す．なお，表 2 の結果は，テスト Y の得点範囲を超えた対応づけ得点を 0 点から 50 点（満点）の範囲に制限し，さらに四捨五入して対応づけ得点を整数化したときの結果である．

いま，対応づけ得点の信頼性係数として 0.7 以上を確保したいとしよう．表 2 の信頼性係数の推定値は，信頼性係数 $\rho[l_Y(X)]$ の上限値と考えられる．ある上限値が 0.7 未満のとき，対応する真値も 0.7 未満である．したがって，どちらの等化法を利用しても対応づけ得点は信頼性不足であり，テスト X をテスト Y へ対応づけしない方がよいと判断できる．

例 2 として，同一の構成概念を測定するものの，例 1 とは項目数が半分のテスト X, Y について，テスト X をテスト Y に対応づけ（等化）する場面を想定した．項目反応データは，例 1 と同様に生成した．表 3 と表 4 に，例 1 と同様に結果を示す．

この例でも，対応づけ得点の信頼性係数として 0.7 以上を確保したいとしよう．表 4 の信頼性係数の推定値をみると，どちらの等化法を利用しても 0.7 以上の値が得られている．ただし，それらの値は $\rho[l_Y(X)]$ の上限値であり，$\rho[l_Y(X)]$ そのものが 0.7 以上であることを保証するわけではない．保守的に判断するなら，対応づけ得点の信頼性の観点からは，対応づけ可能性についての判断は保留とするのが安全である．

二つの例では，スコア間の相関係数は同じ程度でも α 係数の小さい例 2 の方が対応づけ可能性が高いという結果が得られた．これは，テスト X, Y における真の得点の相関が対応づけ可能性にとって本質的であることを示唆している．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　118　2016/08/29　14:00:48

表 1: テスト X, Y の基礎統計量（例 1）

	受験者数	項目数	平均	標準偏差	測定の標準誤差	α 係数	相関係数
テスト X	1,000	50	24.7	6.70	2.55	0.855	0.724
テスト Y		50	26.3	6.15	2.44	0.843	

表 2: 対応づけ得点 $l_Y(X)$ の信頼性（例 1）

	平均	標準偏差	測定の RMSE	信頼性係数
線形等化法	26.3	6.16	3.89	0.678
等パーセンタイル法	26.4	6.14	3.89	0.678

表 3: テスト X, Y の基礎統計量（例 2）

	受験者数	項目数	平均	標準偏差	測定の標準誤差	α 係数	相関係数
テスト X	1,000	25	11.6	3.31	1.73	0.727	0.726
テスト Y		25	10.4	2.94	1.63	0.694	

表 4: 対応づけ得点 $l_Y(X)$ の信頼性（例 2）

	平均	標準偏差	測定の RMSE	信頼性係数
線形等化法	10.4	2.97	1.48	0.731
等パーセンタイル法	10.4	2.80	1.44	0.742

4　おわりに

信頼性係数 $\rho[l_Y(X)]$ の提案をもって，対応づけ得点の信頼性を評価するための基本的な指標が一通りそろった．これらの指標により，単一グループデザインのもとで，経験則によらない定量的な対応づけ可能性分析が可能となった．本発表の成果は，対応づけ得点の性質解明という基礎研究に留まらず，新しい対応づけ技術の開発などの応用研究へと発展が見込まれる．

謝辞　本研究は JSPS 科研費 15K04052 の助成を受けたものです．

注釈

1) 本発表では，異なる二つのテスト間のスコアを結びつけることを対応づけと呼ぶ．等化や予測は，対応づけの特別な場合に位置づけられる．

参考文献

佐藤喜一・柴山　直（2015）．対応づけ可能性のネガティブチェックの試み―対応づけ得点の信頼性に着目して―．日本テスト学会第 13 回大会発表論文抄録集，44–47．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　119　2016/08/29　14:00:49

企画セッション２：

21 世紀型スキルの学習環境と評価のありかた

9 月 9 日（金）14:50 – 16:20　第 1 会場（B101）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　121　2016/08/29　14:00:49

21 世紀型スキルの学習環境と評価のありかた

企画・司会・話題提供　益川弘如*，話題提供　白水始**　齊藤萌木**
指定討論者　南風原朝和***
*静岡大学大学院教育学研究科　**東京大学大学発教育支援コンソーシアム推進機構
***東京大学大学院教育学研究科

評価とは、それがいかなる形を取るにせよ、NRC の答申本“Knowing what students know”というタイトルが示す通り、児童生徒が何を知っているかを知ろうとする営みだろう。本企画セッションは、21 世紀型スキルと呼ばれる資質・能力の具体や評価法に関する学習科学からの話題提供を通じて、21 世紀型スキルが「何を」「いかに」知るかの両面を拡張し得ることを示唆する。その示唆は、大学入試についても高校教育までの学習成果の総括的評価としてだけでなく、高大接続における形成的評価として位置付け直し、そのための学習と評価の環境を提案することにつながる可能性について議論したい。

「21 世紀型スキルが喚起する学習評価のあり方に関する議論」 白水始

21 世紀型スキルは、我々の学力観を変え、学習評価のあり方を問い直すことを促す。例えば、アイディアや知識、リソースを提供し交換してゴールを達成する「協調問題解決スキル」という 21 世紀型スキルは、「鎌倉幕府が何年にできたかを説明できる」「三角関数の微分ができる」という知識の正確な把握や問題の効率よい解決といった 20 世紀型の学習目標と二つの点で異なる。一つは、後者はゴールがはっきり定まっており、教師が答えを教えて児童生徒に覚えてもらう教え方でも対処可能なのに対し、前者はゴール自体が可変な点である。例えば、環境保全とエネルギー政策との適切なバランスに関する協調問題解決は、ここまで答えを出せば終わりというゴールが定まらない。協調問題解決に携わる中でスキル自体が伸び、求められるレベルも上がる。二つは、この違いに応じて、21 世紀型スキルでは「学習者が元々何をどこまでやでき、スキルの発揮が必然性を持つ環境の中で、どれだけスキルを駆使して伸長しているか」という形成的評価が重視される点である。

形成的評価は学習や問題解決のプロセスを捉えようとする点で、評価のコストが掛かる。しかし、「評価」が直接的には把握できない児童生徒の認知過程について、課題への児童生徒の反応の観察から、結果を解釈する営みだとすれば、プロセス評価はこの観察の窓をより多く開け、解釈を豊かにすることに寄与する。それゆえ、話題提供者らは、学習目標（学力観）の拡張だけでなく、センター入試問題の思考発話や全国学力・学習状況調査問題の協調問題解決、協調学習型授業の全発話・記述分析を通して、従来見え難かった「児童生徒が何を考えながら問題を解き、解答を導いているのか」を明らかにすることに努めてきた。これは、作問者の想定通りの認知過程がテストや授業場面で起きているかを問い直す。さらに学習目標まで若干拡張し、学問領域により真正な課題（「幕府の成立年の決め方を説明できる」等）を設定できれば、従来型の学力観や学習評価も徐々に変わる契機となる。

以上の通り、21 世紀型スキルは学力観と学習評価のあり方を豊かにすることに役立つと同時に、教育現場側が児童生徒に伸長させているスキルを妥当に捉える評価、採用側が自らの領域に真正なスキルを候補者が潜在的にも持つかを見極める評価を要請することを通して、社会全体の評価の営みを豊かにすることにもつながる。この理念を、テクノロジも使って、より低コストで実行できるようにする手立てを学会員の皆様と探りたい。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　120　2016/08/29　14:00:50

「CoREF における評価の取り組み」 齊藤萌木

学習科学の分野では、21 世紀型スキルと呼ばれる資質・能力は、課題解決の過程で引き出されるものであり、資質・能力を使って課題を解決する活動を繰り返すことをとおして資質・能力も伸長するとされる。こうした知見によれば、資質・能力を適切に評価するには、資質・能力が引き出されうる課題解決の場をデザインし、課題解決の過程において協調問題解決能力や自他の考えをまとめる力などの資質・能力を使って課題解決に取組んでいるかを評価することで、生徒のスキルの実態を把握できるということになる。

東京大学大学発教育支援コンソーシアム推進機構（CoREF）では課題解決の過程における発話・記述の外化データに基づいて資質・能力の活用を評価する 2 つの手法を提案し、教育現場と連携した実践研究によって、その有効性を検討してきた。手法は、協調的な課題解決活動の前後に、その課題について各自の答えを記述させる「前後理解比較法」と、協調的な課題解決中の発話をテキスト化して分析する「多面的対話分析法」の 2 つである。

本報告では、2 つの手法を用いた評価の実例をもとに、これらの評価手法が資質・能力の有無だけでなく、1 人ひとりの学習者の資質・能力を「その学習者が、どういう場面でどういう資質・能力を活用し、どのような解をつくりだせるか」というレベルで具体的に評価できる手法であることを示すと共に、手法の活用可能性について検討する。これまでの研究では、1 人ひとりの学習者の資質・能力について具体的な評価結果を得られるこれらの手法は、どういった課題がより多くの学習者の資質・能力を引き出すかといった授業デザイン上の指針を実証的に引き出す「形成的評価の手法」として有効性の高いものであると共に、目的に応じて学習者を多用な規準で「序列化」するに足る十分な情報を得られる点で、「評定のための手法」としても有効活用が可能であるとの示唆が得られている。

「資質・能力を育む学習環境における縦断的評価」 益川弘如

協調問題解決スキルの測定として、小学校卒業前の児童対象に、全国学力・学習状況調査問題を 2 人で解くという協調問題解決テストを実施した。そのテスト結果が、単に持っている知識を適用しただけなのか、それとも質の高い協調問題解決スキルを発揮した成果なのかについて、テスト中の対話データ記録と、児童 4 名の 3 年生から 6 年生の間に受けてきた 4 つの授業の授業中の対話データ記録を縦断的に分析した。児童の所属する学校は、3 年生まではスキルの振る舞い方を直接教授し使わせる授業環境を研修方針としていたが、4 年生以降、協調問題解決スキルを発揮したくなる授業環境へ一体的改革に取り組んでいる学校である。

分析の結果、(1) 学習者の協調問題解決スキルは個人特性よりも授業環境に依存する可能性が高い。3 年生のときの授業では各自の考えを確認し合う対話が主だったが、4 年生以降の授業では創発的に理解を深める対話に変化した。(2) 卒業前に受けた協調問題解決テストにおいても授業中と同様の協調問題解決スキルを発揮して課題を解いていた。そして、(3) 従来型の他校児童のテスト結果と対話を比較すると、スキルの発揮具合に違いが見られた。

これらより、協調問題解決テストにおいて、問題の正誤より、そこでのプロセスに大きな差が見られる可能性が高い。そこでのスキルの発揮は、過去、学習者が協調問題解決スキルを発揮する学習環境下でどれだけ繰り返し経験できたかに依存している可能性が高い。

大学入試における 21 世紀型スキル評価の妥当性を高めていくために、形成的評価と組み合わせて見ていく必要があるかもしれない。そこには、大学関係者だけでなく学習環境を構築している高校関係者も、入試の当事者として関わっていく必要があるかもしれない。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 121 2016/08/29 14:00:50

一般セッション６：問題・項目作成

９月９日（金）14:50 – 16:20　第２会場（B102）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　123　2016/08/29　14:00:51

The Impact of LAMS Project Training on Test Development Skills of Trainees and Change in their Perceived Self-Efficacy

○Etenesh Mekonnen[1], Hidetoki Ishii[2]
[1] Mathematics and Science Improvement Center, Ministry of Education, Ethiopia
[2] Psychology and Human Developmental Science, Graduate School of Education and Human Development, Nagoya University

Background

To address the issues of improvement of quality of mathematics and science education, the project for assessment and examination reform was jointly launched in October 2014 by the Government of Ethiopia and the Government of Japan (MoE-JICA, 2014).

This project, known by the name Learning Achievement in Mathematics and Science (LAMS), has done preliminary survey on the curriculum consistency on one hand and, on the other, collected baseline data concerning classroom assessment practiced in Ethiopian primary schools. The survey found that there were several cases in which a question item was judged inconsistent with the subscribed competencies. In addition, items were not scrutinized sufficiently. To fill in this gap the project launched training workshops. So far five training workshops have been conducted. However, training programs are effective only to the extent that the skills and behaviors learned and practiced during the training are actually transferred to the workplace. This is not always the case, and there are instances where training does not result in significant behavior and performance changes back on the job because of the influence of many factors (Facteau, Dobbins, Russell, Ladd & Kudisch, 1995). The question that needs to be answered in this regard therefore is "Is the LAMS training worthwhile and effective?" Effective training and measurement of its effectiveness are critical to bring about behavioral change on the trainees. Evaluation of the effectiveness of training is critical because without it there is no good way to know whether the anticipated behavioral change is attained or not. According to Kirkpatrick and Kirkpatrick (2007), when examining the impact of training programs, one should consider trainees' reactions, learning, and behavior change, along with subsequent organizational results. To this end, the objective of this study was to assess the impact of LAMS project training on test development skills of trainees and their perceived self-efficacy.

Statement of the Problem

The LAMS project aims at improving the national assessment tests like Primary School Leaving Certificate (PSLC) and Ethiopian General Secondary Education Certificate (EGSEC) Examinations. This improvement is to make the national examinations consist of "good question items" which can gauge students' learning achievement more precisely and accurately. According to the LAMS inception survey report (MoE-JICA, 2015) the development of achievement test items in Ethiopia has significant multiple problems. To circumvent these problems the LAMS project launched consistent training workshop programs for experts and teachers directly or indirectly involved in setting national and regional examinations. The progress of these training workshops was not evaluated, however. Early evaluation of the impact of the trainings enables the project to adjust its training approach

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 122 2016/08/29 14:00:52

depending on the results accomplished by the trainees. It also reminds the trainees of what they should have accomplished or should be accomplishing in the progress of the training.

As Devi and Shaik (2012) put it, results of training have to be measured in order to know its impact on business results, thereby taking measures for further improvements in the training program delivered.

Therefore, whether the LAMS training is going on the right track or not needs to be measured. If the training and workshops went as assumed, then there will be improvement in the quality of items developed by the participants in terms of their discrimination capacity, level of difficulty, and correlation. The question, how do the trainees perceive their self-efficacy in the improvement of question item writing skills, needs to be answered. The main aim of this research therefore is to answer these questions by investigating the impact of the project.

Research Design and Methodology

The study used a descriptive survey design and a mixed research method that enabled the researcher to use both quantitative and qualitative methods of data analysis. Data were collected from 1) the training participants through questionnaires (primary sources), 2) randomly selected test items developed during the second and fifth workshop (100 each) by the biology working group, and 3) item analysis results of all the test items selected and field-tested during the second and fifth workshops (240 and 479 items, respectively) (secondary sources).

Analysis

The data which were obtained from the above sources were organized in a way appropriate for responding to the research questions. The researcher used Excel software to calculate frequency counts and percentages to analyze the data from close-ended questions and data obtained from the LAMS archives. The findings about each research question were analyzed and discussed in relation to the review literature. The data from open-ended questions of the questionnaire and the documents were qualitatively analyzed and the two results were triangulated. Finally, conclusion and recommendations were drawn based on the discussion.

Major Findings

As can be seen in Fig.1, more than 90% of the respondents indicated that the training workshops impacted positively on their test development skills.

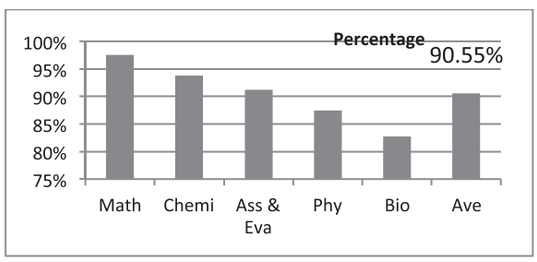


Fig.1: Average impacts of the training workshops on test development skills of the participants

On the average, 88.1% of the respondents believed that they are able to develop better test items after the training than before the training.

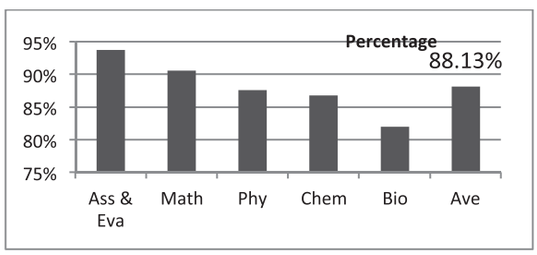


Fig.2: Average change in self-efficacy per working group

From what the participants responded to the open-ended questions about the impact of the training workshops, it was possible to

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 123 2016/08/29 14:00:52

understand that the majority of the training participants believed that the training benefited them in one way or another and they are able to cascade the training to teachers in their respective regions.

Regarding the effect of the preparation of the item pool on the teachers' and students' performance, most of the respondents said that the preparation of the item pool will help teachers to understand what good question items are and make an effort to write such types of items by using the questions in the item pool as models/examples to develop their own test items to assess their students. Students can practice on the questions in the item pool and perform well in their respective assessment tests.

Participants were asked if they have suggestions on the program's overall content, organization and coordination to improve the training. Most of the respondents indicated that the training was going in a good and well-organized manner. Many of them insisted that the LAMS project training should reach to all teachers by any means. They indicated that this is because the training given by the LAMS project improved their professional performance and qualification. They strongly suggested that the project should consider the accessibility of the item pool to both teachers and students from the remote parts of the country as they may not get chances to access the pool.

The total number of test items developed in one training workshop jumped up. It was 344 in the first workshop, then 754 (second), 1,177 (third), 1,538 (fourth) and reached 1,626 in the fifth. It was 373% rise. This indicates that the skills of the trainees significantly improved in setting test items. The number of application and higher level items also increased significantly from the second to the fifth workshop even though it relatively decreased from the fourth to fifth workshop because of the decrease in the number of participants. This again substantiates the fact that the performance of the trainees improved over time.

Table 1: Comparison of workshops 2 and 5

W	Workshop 2			Workshop 5		
G	K	A+H	Total	K	A+H	Total
M	162	88	250	336	172	508
B	86	71	157	266	115	381
C	138	44	182	210	150	360
P	92	73	165	247	130	375
T	478	276	**754**	1059	567	**1626**

K= knowledge level items
A+H= Application and higher level items

From the biology test items developed during the second and fifth training workshops, 100 items each were randomly selected and evaluated using 13 important criteria (guidelines) emphasized during the LAMS training. The results were compared to see the extent to which trainees were able to make use of important rules in item development. Out of the 555 violations of the criteria committed during the second and fifth training workshops, 352 violations were committed during the second workshop while the errors during the fifth workshop were 203, showing visible improvement in the quality of test items developed. The criteria which were most violated during the second training workshop and have shown the highest improvements in the fifth training workshop are given in Table 2.

Table 2: Most improved criteria

No.	Criteria	F*	
		WS 2	WS 5
6	Question is related to the learning outcomes	68	14
1	Task clear to students	57	22
10	Content correct	10	0

F*= Frequency of violation out of 100 items
WS 2 = Workshop 2, WS 5 = Workshop 5

Item analysis results of all field-tested items (four subjects) developed during the two training workshops (Table 3) revealed that

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 124 2016/08/29 14:00:54

there was significant improvement in all the three indices; difficulty, discrimination and I-T correlation, from the second to the fifth training workshop. On the average, difficulty, discrimination and I-T correlation have improved by 13.1%, 9.8% and 14.3%, respectively.

Table 3: Percentage of items developed in the second and fifth training workshops in the acceptable ranges of difficulty, discrimination and correlation

Working Group	Percentage of items with p value between 0.20 & 0.80		Percentage of items with D value of 0.20 & above		Percentage of items with I-T corr. value of 0.20 & above	
	WS 2	WS 5	WS 2	WS 5	WS 2	WS 5
	%	%	%	%	%	%
Math	65.0	85.0	58.3	70.8	60.0	75.8
Bio	83.3	88.3	63.3	73.3	63.3	76.7
Chem	73.3	90.0	63.3	68.3	53.3	70.8
Phy	68.3	79.2	58.3	70.0	60.0	70.8
Average	**72.5**	**85.6**	**60.8**	**70.6**	**59.2**	**73.5**

WS 2 = Workshop 2, WS 5 = Workshop 5

According to this study the trainees started to build skills of and confidence in developing good test items. Therefore, it is possible to conclude that the training is going on the right track.

Conclusion

The results of this study showed that the project brought significant impacts on test item development skills of the majority of the trainees. The improved quality of test items and the high level of self-efficacy which were seen in the results of this study are believed to increase the level of effort the training participants will make in their future performance for the improvement of learning achievement in mathematics and science in Ethiopia. As it is evident from the study, the participants' skills in developing question items that are related to the learning outcomes was improved significantly because of the LAMS training. Most of the items developed during the fifth training workshop have shown considerable improvement in their discrimination power, level of difficulty and correlation as compared with the second training workshop. In general, the data obtained from the questionnaire analysis and the data from the analysis of the test items developed during the training workshops reinforce one another and indicate the success of the project.

Therefore, it is believed that the learning achievement in mathematics and science will be improved as the result of improved quality of test items by, for example, using more diverse modes of assessment, by including items that test higher-order thinking skills, and by assessing the ability of students to apply their knowledge and skills in situations outside school as well as in a scholastic context.

Major References

Devi, V.R., & Shaik, N. (2012). Evaluating training and development effectiveness–A measurement model. *Asian Journal of Management Research*, 2 (1), 722-735.

Facteau, J.D., Dobbins, G.H., Russell, J.E.A., Ladd, R.T., & Kudisch, J.D. (1995). The influence of general perceptions of the training environment on pre-training motivation and perceived transfer of training. *Journal of Management*, 21, 1–25.

Kirkpatrick, D.L., & Kirkpatrick, J.D. (2007). *Implementing the four levels: A practical guide for effective evaluation of training programs*. California: Berrett-Koehler Publishers.

MoE-JICA. (2014). *Project for capacity development for improving learning achievement in mathematics and science education in the Federal Democratic Republic of Ethiopia: Record of discussion.* Addis Ababa.

MoE-JICA. (2015). *Project for capacity development for improving learning achievement in mathematics and science education in the Federal Democratic Republic of Ethiopia: Inception survey report*. Addis Ababa.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 125 2016/08/29 14:00:54

Quality Difference between LAMS and Conventional Mathematics Question Items in Item Difficulty and Discrimination Power on Grade Seven Students in Bahir Dar City, Ethiopia

○Bimerew Kerie[1], Hidetoki Ishii[2]
[1] Mathematics and Science Improvement Center, Ministry of Education, Ethiopia
[2] Psychology and Human Developmental Science, Graduate School of Education and Human Development, Nagoya University

Background

For the purpose of reforming the assessment and examination systems, in particular to improve the quality of mathematics and science education in Ethiopia, the Project for Capacity Development for Improving Learning Achievement in Mathematics and Science Education (LAMS) has been implemented in Ethiopia through the collaboration of JICA and Mathematics and Science Improvement Center (MSIC) of the Ministry of Education (MoE-JICA, 2014). This project has a three-year duration, October 2014–September 2017, and during this period a series of ten workshops will be organized. As of June 2016, LAMS has already implemented six workshops. The project has five working groups (Mathematics, Biology, Chemistry, Physics, and Assessment and Evaluation), whose members are invited from the 11 regional Education Bureaus, the Ministry of Education and the National Educational Assessment and Examinations Agency.

One of the main tasks of these working groups is developing multiple-choice items for the purpose of item pools for grade 7 and 8 students. LAMS has focused on quality of items and hence during the workshops it has provided special practical training on how to write good question items. Members of the working groups write items themselves, review the items mutually, and rewrite them according to the comments by other members. Some of the items are going to be field-tested. In the next LAMS workshop statistical results of the field test will be presented and analyzed in terms of item difficulty, which is a measure of the proportion of examinees who responded to the item correctly; item discrimination, which is a measure of how well the item discriminates between examinees who are knowledgeable in the content area and those who are not; and effectiveness of the distractors, a measure of how well each of the incorrect options contributes to the quality of the multiple-choice item. Finally, according to the results, the item writers rewrite the test items.

In Ethiopia, particularly for grade-seven students who are going to sit for the national primary school leaving exam soon, mathematics teachers use classroom tests as one of the assessment components. In preparing the classroom tests, particularly multiple-choice items, however, teachers do not commonly refer to general guidelines for writing multiple-choice items. They do not conduct field test or item analysis, either. Even the majority of science instructors at College of Natural Science of Addis Ababa University do not have a clear conception of test blueprint, general principles of evaluation, specific rules of test item construction or item analysis procedures (Soromessa, 2015).

Given this situation, this study intends to compare LAMS mathematics question items and conventional mathematics question items administered on grade seven students

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 126 2016/08/29 14:00:56

with respect to item difficulty and item discrimination.

Statement of the Problem

The Ministry of Education of Ethiopia emphasizes science-based education in support of the national goal to attain the status of a middle-income country by 2025 (MoFED, 2010). However, Ethiopian primary school students' mathematics achievement is low. For example, grade 8 mathematics composite mean score was 25.5% in 2011 (Woldetsadik, 2013). Quality of test items is also poor. It is even found that some of the mathematics items in the Primary School Leaving Certificate Examination (hereafter, PSLCE) are not consistent with the specific objectives of the syllabus. On average of the four regions of Ethiopia, 18.7% of the 40 items of PSLCE of 2012 were not consistent with the competencies specified in the syllabus (MoE-JICA, 2015). Furthermore, most of the PSLCE items cannot measure students' higher ability of thinking and reasoning. They usually do not undergo any item analysis before test administration. This study therefore intends to investigate quality difference between LAMS and conventional mathematics question items with respect to item difficulty and discrimination power on grade 7 students in governmental schools in Bahir Dar City, Ethiopia.

Research Design and Methodology

For the purpose of this study, 30 conventional mathematics items are taken from grade 7 four-choice mathematics items developed before the intervention of LAMS. Thirty LAMS items are selected from those items developed by the mathematics working group members at the first and second workshops. To collect the data, LAMS and conventional mathematics tests were administered for the same group of 295 grade-seven students in Teyima primary full cycle school and Atse Sertse Dingle primary school in Bahir Dar City. In the first round, LAMS test was administered at Teyima primary school and conventional test at Atse Sertse Dingle primary school. In the second round, without any intervention regarding the tests, the two tests were administered alternately in the two schools. Percentage, mean scatter diagram and Tables 1 and 2 below are used to compare LAMS and conventional items.

Table 1. Levels of item difficulty (p)

Levels of item difficulty (p)	Interpretation
$p \leq 0.25$	Difficult
$0.25 < p < 0.75$	Moderately difficult, good
$p \geq 0.75$	Easy

Source: Backhoff, Larrazolo & Rosas (2000).

Table 2. Levels of discrmination index (D)

Levels of index of discrimination (D)	Quality of items	Recommended action
$D \geq 0.40$	Very good items	Accept
$0.30 \leq D \leq 0.39$	Reasonably good items	Little or no revision is required
$0.20 \leq D \leq 0.29$	Marginal items	Needs revision
$D \leq 0.19$	Poor items	Should be rejected or completely revised

Source: Crocker & Alegina (1986).

Analysis

As Table 3 below shows, 86.7% of LAMS and 60% of conventional test items have the moderate item difficulty level ($0.25<p<0.75$). This result also indicates that 26.7% more of LAMS items have moderate item difficulty as compared to conventional test items.

Table 3. Item difficulty (p) of LAMS and conventional test items

	LAMS (%)	Conventional (%)	Difference (%)
$p < 0.25$	13.3	40.0	-26.7
$0.25 < p < 0.75$	86.7	60.0	26.7
$p \geq 0.75$	0.0	0.0	0.0

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 127 2016/08/29 14:00:56

Table 4. Index of discrimination (D) of LAMS and conventional test items

	LAMS (%)	Conventional (%)	Difference (%)
$D \geq 0.4$	50.0	10.0	40.0
$0.3 \leq D \leq 0.39$	23.3	23.3	0.0
$0.2 \leq D \leq 0.29$	16.7	30.0	-13.3
$D \leq 0.19$	10.0	36.7	-26.7

Table 5. Item total-correlation (C) of LAMS and conventional test items

	LAMS (%)	Conventional (%)	Difference (%)
$C \geq 0.4$	50.0	6.6	43.4
$0.3 \leq C \leq 0.39$	23.3	26.7	-3.4
$0.2 \leq C \leq 0.29$	16.7	36.7	-20.0
$C \leq 0.19$	10.0	30.0	-20.0

As Tables 4 and 5 show, 50.0% of LAMS question items are acceptable based on either discrimination index ($D \geq 0.4$) or item-total correlation ($C \geq 0.4$) whereas 10.0% of conventional test items are acceptable based on discrimination index and 6.6% according to item-total correlation.

Distractor analysis of LAMS and conventional items was made on the basis of the numbers of high and low scoring examinees who chose a particular distractor. Results show that 76.7% (69 distractors out of 90) of conventional items are judged plausible whereas 95.6% (86 distractors out of 90) of LAMS items are judged plausible.

Figure 1 below shows that as the difficulty value increases from 0.1 to 0.6, the values of discrimination index of conventional items oscillate between 0.0 and 0.5. However, 30% (9 items) of the conventional test items are moderately difficult ($0.25<p<0.75$) and have good discrimination power ($D>0.3$). On the other hand, as Figure 2 shows, as the difficulty value increases from 0.2 to 0.6, the discrimination index of most LAMS items increases in the range of 0.0 and 0.6. As a result, 60% (18 items) of the LAMS test items are moderately difficult ($0.25<p<0.75$) and have good discrimination power ($D>0.3$).

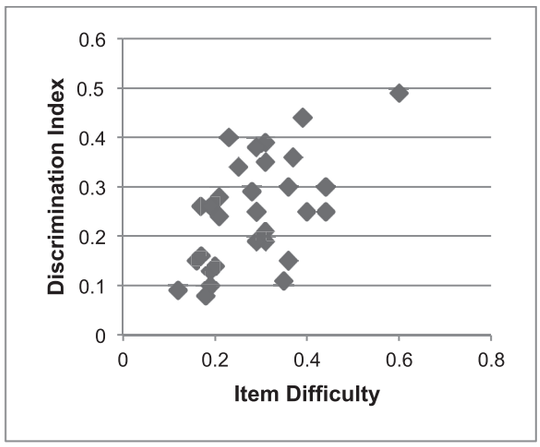


Figure 1. Relation between item difficulty and discrimination of conventional items

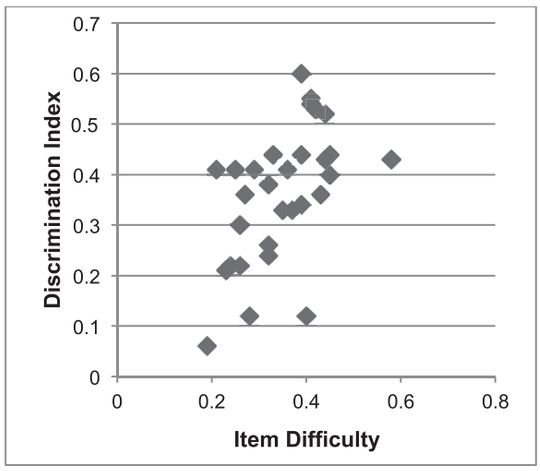


Figure 2. Relation between item difficulty and discrimination of LAMS items

Table 6. Means of item difficulty (p), index of discrimination (D) and item-total correlation (C) of LAMS and conventional mathematics question items

	Mean	
	LAMS items	Conventional items
Difficulty (p)	0.35	0.29
Index of discrimination (D)	0.40	0.25
Item-total correlation (C)	0.36	0.25

As Table 6 above shows, average level of difficulty of LAMS items is 0.35 and that of conventional items is 0.29. The mean value of item discrimination index of LAMS

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 128 2016/08/29 14:00:58

items (0.4) is greater than that of conventional items (0.25). The mean value of item-total correlation of LAMS items (0.36) is also greater than that of conventional items (0.25). The results show that, on average, conventional mathematics question items are in the part of ability range that needs improvement. By contrast, LAMS mathematics question items on average are in the part of ability range that discriminates best.

Major Findings and Conclusion

Conventional items are more difficult than LAMS items. As an indication, 86.3% of LAMS and 60% of conventional test items are moderately difficult ($0.25<p<0.75$).

The ability of LAMS test items to discriminate between knowledgeable and less knowledgeable students is better than that of conventional test items. For example, 50% of the LAMS items and 10% of the conventional items have excellent discrimination power ($D\geq0.4$).

The ability of LAMS test items to measure what the test is intended to measure is also better than that of conventional test items. For example, 50% of the LAMS test items have a strong positive relationship between scores on the item and scores on the total test items. On the other hand, only 6.6% of the conventional test items have the same level of item-total correlation.

The results of distractor analysis show that LAMS items are better composed of plausible distractors than conventional items. While 95.6% of the distractors of the LAMS items are judged plausible, 76.7% of the distractors of the conventional items are judged the same way.

Thus, LAMS mathematics item writers in the first two workshops have developed good skills to write an item that is less difficult and has good discrimination power and plausible distractors.

In general, while 30% of conventional items are "moderately difficult" ($0.25<p<0.75$) and have "good" discrimination power ($D>0.3$), 60% of LAMS items belong to the same category, thus showing their better quality in terms of the two indices.

Major References

Backhoff, E., Larrazolo, N. & Rosas, M. (2000). The level of difficulty and discrimination power of the basic knowledge and skills examination. *Revista Electrónica de Investigación Educativa*, 2(1), 1-16. Retrieved 2016, March 2 from http://redie.uabc.mx/vol2no1/contents-backhoff.html

Crocker, L. & Alegina, J. (1986). *Introduction to classical and modern test theory*. California: University of Florida. Wadsworth Group.

MoE-JICA. (2014). *Project for capacity development for improving learning achievement in mathematics and science education in the Federal Democratic Republic of Ethiopia: Record of discussion.* Addis Ababa.

MoE-JICA. (2015). *Project for capacity development for improving learning achievement in mathematics and science education in the Federal Democratic Republic of Ethiopia: Inception survey report*. Addis Ababa.

MoFED. (2010). *Growth and transformation plan*. Addis Ababa: Berhanena Selam Printing Enterprise.

Soromessa, T. (2015). The practice of student assessment: The case of College of Natural Science, Addis Ababa University, Ethiopia. *Journal of Education and Practice*, 6(25), 141-148.

Woldetsadik, G. (2013). *National learning assessment in Ethiopia: Sharing experiences and lessons*. Addis Ababa: The World Bank.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 129 2016/08/29 14:00:58

エビデンスに基づいた多肢選択式問題作成ルールの策定に向けて

上松　慮生
プロメトリック株式会社

1. はじめに

多肢選択式問題作成にあたっては、高品質な試験問題作成のための作問ルール（ガイドライン）が多数の団体・研究者から提唱され、まとめられてきた（Haladyna & Downing, 1989a; Haladyna & Rodriguez, 2013; Rodriguez, 2016）。しかし、これらのルールには必ずしも実証的なエビデンスがあるわけではなく、試験問題作成の専門家による合意によって制定されているルールも多く存在している（Haladyna & Downing, 1989b; Haladyna & Rodriguez, 2013; Rodriguez, 2016）。そういったエビデンスの無いルールの中には、実際は効果が無かったり、むしろ逆の効果があったりといったものが存在しているため、多くの分野でエビデンスに基づく施策が重視され始めており、同様のことが試験に代表される能力測定の分野でも言える（Bridgeman, 2012）。本報告では、現在提案されているガイドラインを整理するとともに、それらの実証的エビデンスについての研究を調査し、エビデンスのある試験問題作成ルールの策定に向けて、必要とされている要件を考察する。

2. 作問ルール（ガイドライン）の変遷

測定したい構成概念を過不足無く測定できる高品質な試験問題作成の指針として、『テストスタンダード』には、「1 つの質問項目は、1 つの特性のみを測定するように作成する」ことが提示されている（日本テスト学会（編）, 2007）。

一方、具体的な多肢選択式試験問題作成ルールについて検討している Haladyna & Downing（1989a）は 46 の文献をレビューすることにより 43 のルールを提唱し、Haladyna & Downing（1989b）にて、それらのルールの妥当性を検証した 96 の研究結果を整理した。ここでは、それぞれのルールを実証可能なもの（empirically testable）と実証できない・必要が無いもの（values）に分類し、実証可能と分類された 19 のルールのうち 15 について研究成果があることが確認された。

この研究を受け、Haladyna et al. (2002) は、1990 年以降に出版された 27 の文献にあたり、2 つ以上の文献で言及されている 31 の作問ルールを明らかにした。

Haladyna & Rodriguez (2013) では、Haladyna et al. (2002) の 31 項目のガイドラインを整理し、22 項目のガイドラインとして改訂し、これが現在の標準的なものとして認識されているが、Rodriguez (2016) でも指摘されているように、これらのうちエビデンスが研究されているものは決して多くない。

3. ガイドラインを逸脱することの影響

試験の品質に悪影響を与えると懸念される妥当性の脅威として構成概念と無関係な分散（Construct Irrelevant Variance : CIV）の混入が考えられるが、実際に前述のように数は限られているが、様々な研究において、ガイドラインに則らない試験問題は則ったものと比較して難易度や識別、信頼性係数に差があることが報告されており、また、その効果は受験者集団に一様ではないという研究結果もある（Haladyna et al. 2002; Rodriguez, 2016）。

ただし、こうしたエビデンスが存在してい

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　130　2016/08/29　14:00:59

る個別ルールは限定されており、多くのルールについては試験問題作成の専門家による合意を持って制定されている。こうしたエビデンスを伴わない直感や常識に従ったルールは誤っている場合もあり、科学的な根拠の検証が期待されている（Bridgeman, 2012）。

4. エビデンスに基づいた多肢選択式問題作成ルールの策定に向けて必要な要件

日本語の多肢選択式試験問題作成ルールに関するエビデンスを検証するにあたっては、以下の 3 点を考慮する必要があると考えられる。

- 言語の違いによる差異
- コンテンツの影響
- Stakes の影響

これまでの研究結果は英語母語者向けの英語による試験についてのものが大半であり、日本語母語者向けの日本語による試験における影響は別に考える必要がある。例えば、荒井（2015）は、「複雑な多肢選択式形式」は英語文献のみに掲載されていたガイドラインであり、専門家へのヒアリングでも「場合による」という回答が多かったことから、日本の文脈では妥当ではないルールと結論付けている。

難易度や識別、信頼性係数といった試験問題や試験の統計値に影響を与える要素を考えた場合、CIV よりもコンテンツのほうが大きな役割を果たしていることは容易に想像でき、例えば、社会科よりも国語（読解）といった、よりコンテンツの影響が少ない分野での実証が有効であると考えられる。

作問ルールに準拠しないことによる CIV の発生には、計測したい能力が計測できているかという妥当性の観点と、能力が無い受験者が解けるような問題になっていないかという test-wiseness の観点に分類できる。それぞれの具体的なルールとして、妥当性については、否定形を使わない；複雑な形式（type K）を使わない；ひっかけ問題を使わない等が挙げられ、test-wiseness については、ロジックで解ける問題にしない；選択肢の長さをそろえる等が挙げられる。受験者自身の生活により大きな影響を与える high-stakes な試験では、受験者が test-wiseness を可能な限り活用することが考えられるため、これに付随する CIV は high-stakes な試験での検証が望ましい。

参考文献

荒井清佳（2015）多肢選択式問題を作成する上で大切なこと – 問題作成の専門家に対する調査結果に基づいて – 日本テスト学会誌 Vol.11, No.1, 21-34.

Bridgeman, B. (2012) Why Bother with Research When We Have Common Sense? *ETS R&D Connections*, 20, 1-8.

Haladyna, T.M., Downing, S.M. (1989a) A Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*, 2(1), 37-50

Haladyna, T.M., Downing, S.M. (1989b) Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*, 2(1), 51-78

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002) A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, 15(3), 309-334.

Haladyna, T.M., & Rodriguez, M.C. (2013) *Developing and Validating Test Items*. New York, NY: Routledge.

日本テスト学会（編）(2007) テストスタンダード：日本のテストの将来に向けて. 金子書房.

Rodriguez, M.C. (2016) Selected-Response Item Development. in Lane, S., Raymond, M.R., & Haladyna, T.M. (2016) *Handbook of Test Development* (2nd ed.). New York, NY: Routledge. (pp.259-273)

（e-mail: ryo.uematsu@prometric.com）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 131 2016/08/29 14:01:00

MMQ 形式の問題項目の採点方法について

○荒井清佳，宮埜寿夫
大学入試センター　研究開発部

１．はじめに

多肢選択式は，客観式テストにおいて最も多く利用されている問題形式である。その多くは，正答選択肢が１つあるいは２つなどと，あらかじめ受験生に正答選択肢の数が示されている。それに対し，正答選択肢の数があらかじめ示されておらず「適切なものをすべて選べ」などと問う問題形式もある。このような，受験者があらかじめ用意された選択肢の中から正しいと考える選択肢をすべて選択する形式の問題は Multiple-mark 式，Multiple-multiple 式などと呼ばれ，多真偽形式（Multiple true-false form）のバリエーションの 1 つとされる（Haladyna，2004）。

このような形式（MMQ 形式）の問題については，古くから研究されており，どのように部分点を与えるか，採点方法に関する研究もなされている。しかし，どの方法が望ましいかについて一致した結論は得られていない。

本研究では MMQ 形式の問題項目の採点方法について，いくつかの方法を検討し，その特徴を比較する。

２．方法

部分点を与える方法として望ましい性質として「正答選択肢の選択が多いほど，あるいは誤答選択肢の選択が少ないほど得点が高い」ことが挙げられる。このような性質をもつ指標として，本研究では，受験者の解答パターンと正答パターンとの関連の強さや類似度を新たに取り上げた。また先行研究でよく用いられている方法を含め，次の(1)～(4)の採点方法を検討した。

以下，選択肢の総数を N，そのうち n_c 個が正答選択肢，n_w 個が誤答選択肢であるとする（$n_c + n_w = N$）。また，受験者が正答選択肢から x_c 個，誤答選択肢から x_w 個選択したとする（$x_c \leqq n_c$，$x_w \leqq n_w$）。受験者の正しい判断の数は，$i = x_c + (n_w - x_w)$ である。

１　採点方法

(1) ファイ係数…受験者が正しい判断を行った場合は正しく，誤った判断を行った場合は誤りであるとみなすと，受験者の解答は，選択肢の正誤と受験者の選択を行及び列とする 2 × 2 の分割表（表 1）にまとめられる。この表の行と列の関連の強さを表す指標であるファイ係数を用いる。

表１：選択肢の正誤と受験者の解答パターンの正誤

		受験者		計
		選択する	選択しない	
選択肢	正答	x_c	$n_c - x_c$	n_c
	誤答	x_w	$n_w - x_w$	n_w
計		$x_c + x_w$	$N-(x_c+x_w)$	N

(2) Jaccard 係数…N 個の選択肢について，受験者が選択した場合は「1」，選択しなかった場合

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　132　2016/08/29　14:01:01

は「0」として表した解答パターンと，選択肢が正答の場合を「1」，誤答の場合を「0」とした正答パターンを考える。両者の類似度を表す Jaccard 係数を用いる。なお，表 1 の表記を用いると，Jaccard 係数に基づく得点は$\frac{x_c}{n_c+x_w}$として表される。

(3) MTF（Multiple True-False）法…N 個の選択肢のそれぞれを真偽（true-false）式の項目として捉え，正しい判断（正答選択肢を選択し，誤答選択肢を選択しない）をしている選択肢の数に比例して点数を与える。i/N 点となる。

(4) NM（Negative Marking）法…正答選択肢を選択した数に応じて $1/n_c$ 点ずつ得点を与えるが，誤答選択肢を選択した場合には $1/n_c$ 点ずつ減点していく方法。

ただし，(1)と(4)の場合のように得点が 0 より小さい値になってしまう場合には 0 点とした。

２．２　採点方法の特徴の比較

例として選択肢の総数が 5 つ，そのうち正答選択肢が 3 つの場合を取り上げた。まず，すべての解答パターンについて，4 つの採点方法による得点を計算し，その性質を明らかにした。続いて，受験者数を 300 人としてシミュレーションを行った。シミュレーションにおいては，選択肢を正しく判断できる確率を項目反応モデルを用いて表すこととし，各受験者の解答パターンを乱数を用いて生成した。選択肢の困難度を変えることで，問題の難しさと，採点結果との関連を調べた。

３．結果

採点方法別に，どのような部分点になるかをみたところ（表 2），(0 点も含めて）MTF 法は 6 段階，NM 法は 4 段階に等間隔に分かれた。Jaccard 係数では 9 段階と細かく分かれた。1 点（満点）となったのは 4 つの方法とも，すべての選択肢で正しい判断をしている場合のみであった。0 点となったのは，Jaccard 係数では正答選択肢を一つも選ばない場合，MTF 法ではすべての判断が誤っている場合のみであった。

表 2：解答パターンと得点

x_c	x_w	phi	J	MTF	NM
0	0	NA	0.00	0.40	0.00
0	1	0.00	0.00	0.20	0.00
0	2	0.00	0.00	0.00	0.00
1	0	0.41	0.33	0.60	0.33
1	1	0.00	0.25	0.40	0.00
1	2	0.00	0.20	0.20	0.00
2	0	0.67	0.67	0.80	0.67
2	1	0.17	0.50	0.60	0.33
2	2	0.00	0.40	0.40	0.00
3	0	1.00	1.00	1.00	1.00
3	1	0.61	0.75	0.80	0.67
3	2	NA	0.60	0.60	0.33

シミュレーションの結果，各選択肢の困難度が高い場合には，得点が非常に低くなることが示された。その他の結果については発表当日に紹介する。

引用文献

Haladyna, T. M. (2004). Developing and Validating Multiple-choice Test Items. Psychology Press.

連絡先：sayarai@rd.dnc.ac.jp

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 133 2016/08/29 14:01:01

企画セッション 3 : 認知診断モデル

9 月 9 日（金）16:30 – 18:30　第 1 会場（B101）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　135　2016/08/29　14:01:01

認知診断モデルによる誤り要因分析

○大森 拓哉[1]、森 一将[2]
[1]多摩大学経営情報学部、[2]文教大学経営学部

1．はじめに

教授− 学習過程において教師が学習者の習得状態を把握することは教育上必要不可欠な問題である。特に，学習者がどのような誤り方をしているかを同定することは、その後の適切な指導を与える上で重要な情報となる。本稿では、いくつかの誤り同定のアプローチを概観し、定量的かつ認知的なプロセスを重視した方法を重点的に考える。

2．認知スキルの統計的分析

テストデータの分析には、項目反応理論（IRT）が用いられることが一般的である。様々なモデルが提案されているが、項目困難度パラメータの推定を中心に、そこから様々な解釈を行うことが多い。しかし項目の難しさの数値だけからではその難しさの要因や、どういった認知的プロセスにおいて誤りを犯しているかといったことを探ることは困難である。そこで困難度パラメータを更に細かく解釈するモデルが提案されている。

LLTM（Linear Logistic Test Model, Fischer (1972)）は、Rasch モデルの拡張として、困難度パラメータを各項目とそれらを解くための下位のスキルとの行列に分解し、要因の影響の大きさを測るものとして提案されている。また、Rule Space Method（RSM, Tatsuoka(1990)）は、知識や認知スキルをアトリビュートとしてまとめ、各項目とアトリビュートを対応付ける照応行列（Incident matrix,; Q-matrix）を作成するものである。

3．誤り要因の分析

学習者の誤答は，“バグ” と “スリップ” に大別されるという考えがある(Brown & Burton,1978, Brown & Van Lehn,1980,1982)。バグとは誤った考えに基づいて問題を解決してしまうものであり，例えば二桁の計算問題を筆算で考えた場合，12−9 で一の位の 2 から 9 が引けない場合，繰り下げをせずに下から上を引いてしまう(引ける方から引く、すなわち 9 – 2 としてしまう）というものである。つまり同じ位で引けない場合には常に下の数から上の数を引いてしまうという一貫した手続きをすることであり，これをバグルールという。通常のテストの分析では、項目の困難度や識別力などを主眼とすることが多いが、本稿ではこのようなバグルールといった、個人の誤り要因同定に焦点を置く。ちなみに Tatsuoka(1985) においては、項目反応理論を用いて学習者のバグルールの類型化を行っている。

4．認知構造の可視化

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 134 2016/08/29 14:01:02

誤りの同定、特に学習者個人へのフィードバックにおいては、何が問題であるかをわかりやすく明示する必要がある。そのためには認知スキルや誤り要因の構造が可視化されている方が教授側にとっても学習者にとってもわかりやすい。可視化の一つは、構造をグラフィカルモデルで表現することである。グラフの各ノードの状態を見ることにより、学習状態・理解状態を把握することができる。

5. 確率ネットワークモデルと適応型テストによる同定

確率ネットワークモデル（ベイジアンネットワークモデル）において適応型テストを実施し、知識状態を同定する研究が植野ら（1994）によって提案されている。また、大森ら（2007）においては、単純なツリーモデルにおいて適応型テストを実施することにより、学習者のバグを同定している。これらの認知構造は教師の知識を基に作成されたものであり、データから自動作成されたものではない。Mislevy (1995)では、RSM における Q-matrix からベイジアンネットワークを生成する方法を提案している。

5. ISM による認知構造学習

認知構造・教材構造を自動作成する方法として、ISM(Interpretive Structural Modeling)構造学習法がある (Warfield,1973)。ISM 構造学習法では、教材構造・知識構造を関係行列にまとめ、関係行列のべき乗の演算を操作することにより、各ノード間のリンクの生成などを行って全体構造を把握する。

6. まとめ

以上、誤り要因の分析の方法として、いくつかのモデルを概観した。特に個人にフィードバックする場合は、適応型テストである方が効率が良い。その際にはさらに項目選択のアルゴリズムなどについても考慮する必要がある。

参考文献
Brown,J.S.,Burton,R.R.(1978) Diagnostic models for procedural Bugs in Basic Mathematical Skills. Cognitive Science, 2:155-192
Brown,J.S.,VanLehn,K.(1980) Repair theory:A generative theory of bugs in procedural skills. Cognitive Science,4:379-426
Fischer, G.H. (1972). Conditional maximum-likelihood estimations of item parameters for a linear logistic test model (Research Bulletin 9). Vienna: University of Vienna, Psychological Institute.
Mislevy R.(1995) Probability-Based Inference in Cognitive Diagnosis, In P. Nichole, S. Chipman and R. Brenonn (eds.), Cognitively Diagnostic Assessment, 43/71, Hillsdale, NJ,
Tatsuoka,K. K. (1985) A probabilistic moder for diagnosing misconceptions by the pattern classification approach. *Journal of educational Statistics*,10:55-73
Tatsuoka, K. K. (1990) Toward an Integration of Item-Response Theory and Cognitive Error Diagnoses. In N. Frederiksen, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), Diagnostic Monitoring of Skill and Knowledge Acquisition. Hillsdale, NJ：Erlbaum, Pp.453-488.
植野,大西,繁桝(1994)教師の領域構造知識を利用した適応型テストの開発,日本教育工学雑誌, 18(1), 15-24
Warfield, J. N. (1973). An assault on complexity, Battelle Monograph No 3, Battelle Memorial Institute, Columbus. Ohio, USA. Warfield, J. N. 1973b. 'Intent structures'. IEEE Trans on System, Man and Cybernetics, SMC3(2), 133-140.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 135 2016/08/29 14:01:02

認知的診断のための測定論的情報を抽出する IRT 分析

○坂本　佑太朗
株式会社リクルートマネジメントソリューションズ

1. はじめに

テストデータの分析へ項目反応理論（item response theory, IRT）（Lord, 1952）を適用することが普及しつつある中で，受検者の反応データから心理測定論にもとづく認知的診断を行うための情報を抽出し，受検者に対して有益なフィードバックを行うことが求められつつある．歴史的には，Carroll & Maxwell（1979）以降，認知的診断を目的とした測定論的研究が行われてきた（Embretson, 1994）．しかしながら，わが国ではそれほど研究の蓄積は十分になく，最近では認知診断モデル（cognitive diagnostic model）（e.g. Leighton & Girel, 2007；Nichols, 1994；Rupp, Templin & Henson, 2010）や認知的 IRT モデル（cognitive IRT model）（van der Linden & Hambleton(Ed.), 1997；Embretson, 1998）を活用した研究（倉元・スコット・笠居，2003；坂本・柴山，2014；鈴木・豊田・山口・孫，2015）がなされているというのが現状であろう．

Embretson（1998）が指摘するように，認知的 IRT モデルの代表例としては線形ロジスティックテストモデル（linear logistic test model, LLTM）（Fischer, 1973）がある．IRT の場合，テスト項目が測りたい構成概念を測るように設計されている際には，項目困難度パラメータがその項目に回答するのに意図された認知的操作（cognitive operations）（認知診断モデル的にいえばアトリビュートやコンポーネント）で説明できると言われている（Embretson，1994）．LLTM はラッシュモデルにおける項目困難度パラメータを認知的操作ごとの困難度の線形和に分解する．言い換えれば，従来は推定される項目困難度パラメータに関する困難度の「要因」は分析者の定性的な判断や解釈が必要であったが，LLTM ではその「要因」を計量的に明らかにできる点に特徴がある．

しかしながら，LLTM は 1 次元性の仮定を置く心理測定モデルである．この 1 次元性の仮定について，実際には，「1 次元性を満たす」ことを確認する方法自体は多種多様であり（Hattie,1985），Embretson and Reise（2000）は「十分な 1 次元性」の判断基準は明確ではないことを指摘している．また，多くのテスト項目への回答の背後には，多次元性が内在されていると考えられている（Ackerman，Gierl & Walker，2005；Yao, Boughton, 2009）．そのため，1 次元性を仮定する IRT モデルを多次元に拡張した多次元 IRT（multidimensional IRT）（Reckase, 2009）[1]による分析が期待されているところである．最近では，Reise, Cook and Moore（2015）が 1 次元性を仮定した IRT 分析結果を報告する際，多次元 IRT を使ったテスト項目に関する精緻な分析結果も同時にするべきであると指摘している．特に，ハイステイクスな場面で用いられるテストに関わるテスト開発者/実務家こそ，テストの品質保証

[1] Embretson（1998）は，Embretson（1984），Mislevy & Verhelst（1990），Whitely（1980）を「多次元認知的 IRT モデル（multidimensional cognitive IRT model）」だと指摘しているが，一般にこれらのモデルは多次元 IRT モデルとして包含できる．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　136　2016/08/29　14:01:03

という観点から「そもそもそのテストは何を測っているのか」という基本的な問いを常に検証し続ける必要がある．そのため，多次元 IRT を用いた精緻な項目分析は，認知的な情報をテストデータから引き出し，受講者へフィードバックを行う前提として必要であると判断できる．

そこで本発表では，受検者がテスト項目に反応する際に求められる認知的操作に関する情報を得るための線形ロジスティックテストモデル（linear logistic test model, LLTM）（Fischer, 1973）を用いた分析を坂本・柴山（2014）を紹介する．また，多次元 IRT を使った分析として，テストを構成する下位領域（subscales）がもつそれ特有の測定論的な情報を計量的に明らかにする分析（坂本，2015，2016）を紹介する．本稿ではそれらのモデルと分析手法を紹介し，実際の分析例は当日示す．

2. LLTM

古典的テスト理論にもとづいて得られる項目の通過率や1次元性を仮定する IRT モデルでの項目困難度パラメータにおいて，その困難度の要因がどこにあるのかはテスト分析者の推測に頼らざるを得ない．つまり，テスト項目に反応する際に必要となる認知的操作に関する情報は，計量的に表現することができなかった．LLTM は，受検者 iの項目jに対する反応をu_{ij}，項目困難度パラメータをb_jとすると，ラッシュモデル（1PL モデル）

$$P(u_{ij}=1|\theta_i,b_j)=\frac{exp(\theta_i-b_j)}{1+exp(\theta_i-b_j)}$$

における項目困難度パラメータb_jを，

$$b_j=\sum_{k=1}^{p}w_{jl}\alpha_k+c$$

と線形和を仮定する．このとき，α_kは基本母数（basic parameters），w_{jl}はそれぞれの項目の基本母数に対する重み（行列），そしてcは基準化のための定数である．基本母数α_kの設定に当たっては，線形構造を仮定している以上，それぞれの基本母数同士が互いに独立であるという仮定を置いていることに注意が必要である．また，LLTM における重み行列$\boldsymbol{W}$が対角行列である場合が通常のラッシュモデルとして位置づけられる（Fischer, 1995）．従来，ラッシュモデルや IRT におけるロジスティックモデルなどでは項目困難度パラメータとして「項目」固有の推定値を得ていたが，LLTM では認知的操作固有の困難度パラメータを得ることができる．

つまり，テスト開発の過程で「その項目がどのような能力（認知的操作）を測るか」を定義している場合には，理論的にはそれを重み行列$\boldsymbol{W}$に落とし込み，それ固有の困難度を推定できる．言い換えれば，受検者の学習上のつまづきが，従来の項目レベルではなく測定領域ごとの困難度として計量的に示すことができる．その結果を踏まえて，今後の学習指導や受検者へのフィードバックをより詳細に行うことや，受検者集団の特徴を測定領域別に把握できる．しかしながら，その適用にあたってはいくつかの条件がある．たとえば，基本母数α_kの数は項目数mを越えてはならず（$p<m$），また$m\times(p+1)$の重み行列$\boldsymbol{W}^{+}=(\boldsymbol{W},\boldsymbol{1})$が階数（rank）$p+1$のフルランクを持つ必要がある（Fischer, 1983, 1995）．つまり，集団統計量を得ることが目的となる大規模学力テストでは，出題範囲が広く設定されるため，項目数mに対して基本母数pが多くなることが予想され，重み行列$\boldsymbol{W}$が設定しづらい．そもそも，LLTM による分析を前提に設計されるテストではあれば，その点に注意して出題範囲等を決定すればよいが，テストデータの二次分析の際には注意が必要である．また，基本母数の推定によって

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 137 2016/08/29 14:01:04

測定領域固有の困難度を計量的に示すことができるのはLLTMの特徴であるが，学力テストの場合では，重み行列$\boldsymbol{W}$の設定自体には教科の専門家によるいわば主観的な判断が必要となる．つまり，「テストの専門家」（木村，2006）としての，実際の分析を行う「教育測定（テスト理論）の専門家」とテストの測定領域についての定性的な判断を行う「教科の専門家」の協働が重要となる．

なお，発表当日にLLTMを実際のわが国で行われた学力調査データに適用した例を示す．

3．多次元IRT

多次元IRTは一般に補償型（compensatory）モデルと非補償型（noncompensatory）モデルに大別される[2]．補償型モデルは，複数の能力を測定するテストにおいて，ある能力が低い場合でも他の能力が十分高ければ当該の項目には正答しやすいという仮定を置くモデルである．つまり，数学的にはそれぞれの次元同士は和の関係にある．補償型多次元2値2PLモデル（以下，多次元2PLモデル）は数式で表現すれば

$$P(u_{ij}=1|\boldsymbol{\theta}_i,\boldsymbol{a}_j,d_j)=\frac{exp(\boldsymbol{a}_j\boldsymbol{\theta}'_i+d_j)}{1+exp(\boldsymbol{a}_j\boldsymbol{\theta}'_i+d_j)} \quad (1)$$

と表される．このとき次元数をmとすると$\boldsymbol{a}_j$は$1\times m$の項目jの識別力パラメータベクトル，$\boldsymbol{\theta}_i$は$1\times m$の受検者iの潜在特性尺度値ベクトル，d_jは困難度に関連するパラメータ（スカラー）を示している．このとき，d_jは1次元性を仮定した項目困難度パラメータb_jと同じ解釈はできないことには注意が必要である．そこで，多次元IRTの場合には多次元困難度（multidimensional difficulty, MDIFF）

$$MDIFF_j=-\frac{d_j}{\sqrt{\Sigma_{v=1}^{V}a_{jv}^2}}$$

によって同じ解釈が可能となる（Reckase, 2009）．ただし，数学的には，多次元IRTはカテゴリカル因子分析と同値であり，探索的な場合には項目識別力パラメータの単純構造への回転が可能となる（荘島，2003；Takane & De Leeuw, 1987）．

一方，非補償型モデルは当該の項目に正答するにはある能力だけが高いだけでは達成されないことをモデル化しており，数学的にはそれぞれの次元同士の積によって正答確率を定義していることに特徴がある．その代表例としては，Whitely（1980）の複合潜在特性モデル（multicomponent latent trait model，MLTM）がある[3]．MLTMは認知心理学と心理測定学とをつなぐ数理モデルとして提案されてきた背景をもっている（Whitely, 1980；Embretson, 1997）．

測定領域に関する事前の仮説がある場合，確認的な多次元IRT分析を実行することも可能である．その代表例としての双因子モデル（bifactor model）（Gibbons & Hedeker, 1992）における項目識別力パラメータは行列で表記すると

$$\boldsymbol{a}=\begin{bmatrix}a_{11} & a_{12} & 0\\ a_{21} & a_{22} & 0\\ a_{31} & 0 & a_{33}\\ a_{41} & 0 & a_{43}\end{bmatrix}$$

となる．これから明らかなように，双因子モデルは全項目へ影響を与える一般因子（general factor）と，それぞれの下位領域からの影響としてグルー

[2]最近では，岡田（2014）が補償型と非補償型を包含する多次元IRTモデルを提案している．

[3] MLTMの発展形として，項目困難度についてLLTMと同じ構造をもつ一般化複合潜在特性モデル（general multicomponent latent trait model, GLTM）（Embretson, 1984）がある．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 138 2016/08/29 14:01:05

プ因子（group factor）を定めるモデルである．このとき，一般因子を統制することによってグループ因子による影響を検討できる点がこのモデルの特徴である（Reise, et al., 2015）．

Reise, Morizot and Hays（2007），坂本（2015, 2016）はテスト全体が測定する構成概念を一般因子，テストに含まれる下位領域を「次元」とみなし（グループ因子），下位領域特有の影響が一般因子と比較してどの程度存在しているかを検証した．具体的には，IRT における項目情報量の観点からその比較を行っている．通常の IRT モデル（2PL モデル）における項目情報量は，項目jに正答する確率と誤答する確率をそれぞれ$P_j(\theta)$，$Q_j(\theta)$とすれば

$$I_j(\theta) = \frac{\{a_j P_j(\theta) Q_j(\theta)\}^2}{P_j(\theta) Q_j(\theta)} = a_j^2 P_j(\theta) Q_j(\theta) \quad (2)$$

と定義される．（2）式から，2PL モデルにおける項目情報量は，項目識別力パラメータの値によって規定されることがわかる．多次元 IRT においてはこれを多次元空間に拡張し，空間上の一点であるθ_iとv次元目のθとのなす角をα_{iv}とすると

$$I_j(\boldsymbol{\theta}) = \frac{\{\boldsymbol{a_j} P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})\}^2}{P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})} = \frac{(P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta}) \sum_{v=1}^{V} a_{jv} cos\alpha_{iv})^2}{P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta})} = P_j(\boldsymbol{\theta}) Q_j(\boldsymbol{\theta}) (\sum_{v=1}^{V} a_{jv} cos\alpha_{iv})^2 \quad (3)$$

として定義できる（Reckase & Mckinley, 1991；Reckase, 2009）．（3）式から，多次元の場合でも項目識別力パラメータの値によって項目情報量が規定されることがわかる．Reise, et al.（2007）と坂本（2015，2016）は，双因子モデルによって推定した一般因子とグループ因子の項目パラメータの 2 乗値の比較によって，それぞれの構成概念に対してもつ測定論的な情報の比較を項目レベルで行った．これを通して，たとえば一般因子を統制した後にもなお，グループ因子のほうがより多くの情報量を保有している項目については，テスト出題時の項目選択の際に注意すべきであることを示唆している．つまり，通常の IRT モデルが仮定する「1 次元」の構成概念以外の影響が強いことを意味するため，テストが測りたいものを測るという前提からそれてしまっている可能性がある．

なお，発表当日は坂本（2015，2016）の分析例を示す．

4. おわりに

本発表では，認知的 IRT モデルとしての LLTM，また多次元 IRT を使った分析を紹介した．認知的診断を行う心理測定モデルの重要性が指摘されている（Wang & Gierl, 2011）が，その前提として「そもそもそのテストは何を測っているのか」という素朴な問いは重要である．つまり，最後はテストの妥当性という観点に回帰するのである．テストの信頼性は妥当性の必要条件となるが，特にテスト開発者/実務家こそテストの妥当性検証を不断に行う必要がある．その際には，宇佐美（2015）が指摘する通り，構成概念妥当性という単一の概念に収斂するそのさまざまな側面（Messick, 1989）のどの側面について議論しているかを明らかにする必要があろう．そのような測定論的前提を踏まえてはじめて，認知診断を目的とした統計的分析を行うことができると意識し続けることが大切となる．

坂本 佑太朗（yutaro_sakamoto@recruit-ms.co.jp）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 139 2016/08/29 14:01:06

IRTモデルにもとづく学力評価ルーブリックの作成手法の試み

―小学校算数・面積学習領域を対象として―

佐藤　誠子*　　柴山　直**

*石巻専修大学人間学部　　**東北大学大学院教育学研究科

問題と目的

近年，IRTを用いて「学力」を単一の尺度値（θ）で表現する試みがなされてきている。これにより学力の経年比較や重複テスト分冊法などによる実施コストの削減が可能になる等のメリットがあるものの，教育現場へのIRTモデルの導入に際しては次の２つの問題が想定される。１点目は，IRTモデルに基づくテストの実施には使用する問題の項目プールの作成，問題非公開という条件が満たされなければならないことである。その場合，テストを学習指導に活かそうとする教育現場の立場からすれば，個々の学習者の学力特性値（θ）のみが示されても実際の学習指導改善にはつながらないとの批判が生じる。教育現場の要請に応えるには，θのレベルと実際のパフォーマンス（Can-Do）とを対応させ，指導に役立つルーブリックを作成する必要があるだろう。２点目は，the bandwidth-fidelity dilemma（Cronbach and Gleser, 1965）が生じることである。これは学力を測定する尺度の場合にもあてはまり，すべての問題をIRT尺度に合わせて作成しようとすると，近年重視されている「思考力」「問題解決力」といった高度な能力を把握しきれなくなる可能性が生じる。したがって，学力評価ルーブリックを作成する際には，IRT尺度とは別にパフォーマンス課題を用意し実際の課題解決について検討する必要があるだろう。

そこで，本研究では，学力テストとは別にパフォーマンス課題を用意し，「どの学力レベルにいる者が具体的にどのような解答のふるまいをするか，どこでつまずきやすいのか」という学力特性値（θ）のレベルと実際の解答パフォーマンスとの関連について探索的検討を行い，IRTモデルにもとづく学力評価ルーブリック作成のための手法を提案する。その際，小学校算数「面積」学習領域をとりあげる。従来の一般的なテスト問題では，面積問題は数値代入による面積量算出が主になっている。しかし，21世紀型スキルに代表されるような「思考力」が目標とされるのであれば，公式への数値代入による面積量算出だけでなく，公式の一方の変数の変動から他方の変数の変動を導くといった操作的思考（工藤，2005）も要求されるだろう。これを測定する問題として，本研究では等周長変形課題，等積変形課題（Figure1，西林，1988；工藤・白井，1991など）を用いる。これらをパフォーマンス課題として用いながら，学力レベルと実際の解答パフォーマンスとの関連について検討し，学力評価のためのルーブリック作成の手法を提案する。

方法

対象者　学校長の協力が得られた山形県内公立A，B，C小学校の第6学年児童121名である。

実施期間　2012年10月，2013年11月～2014年1月にかけて行われた。

手続き　算数学力テストおよび面積テストの2種類の問題冊子への解答を求めた。後者がパフォーマンス課題となる。配付は当該学級の担任による。所要時間は，算数学力テスト40分，面積テスト20～40分であった。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　140　　2016/08/29　14:01:06

課題構成　算数学力テスト，面積テストの 2 つの問題冊子からなる。

算数学力テスト　平成 18 年度新潟県学力テストを使用し，テスト実施時間の制約上，面積に関する問題 3 問のうち 2 問を削除した（全 23 問）。

面積テスト　等積変形課題 (2 問)，等周長変形課題 (2 問) の面積比較課題よりなる (Figure1，西林，1988；工藤・白井，1991 など参照)。いずれも，辺の長さの数値の記載がなく（あるいは不十分）底辺，高さの相対的大小を読み取り，そこから面積の大小を導出する必要がある問題である。解答はいずれも選択式による。なお，各課題で 1 問ずつ，判断の理由の記述も求めた。

各パラメタの推定　項目パラメタ（識別力，困難度）の推定には，平成 18 年度新潟県学力テストのデータ（受検者 : 小学 5 年生 21,735 名，2006 年 1 月実施）を用いた（推定方法は周辺最尤法，事前分布は標準正規分布)。受検者パラメタ（学力特性値 θ ）の推定は最尤法による。なお，これらのパラメタ推定には EasyEstimation（熊谷，2009）を使用した。

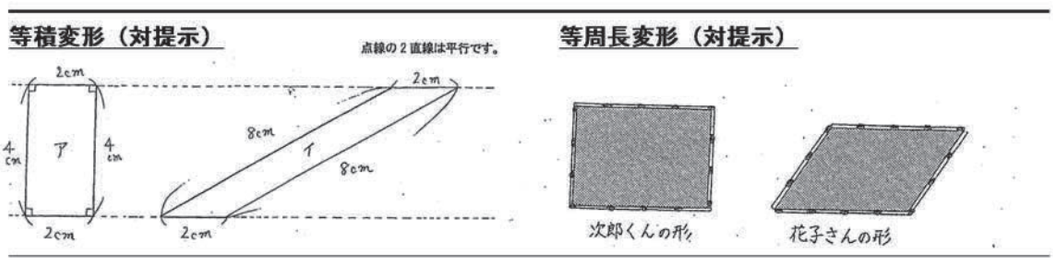


Figure1　面積比較課題で提示した図形の一例

結果と考察

算数テストの全問正答者 7 名，いずれかのテスト欠席者 2 名，回答不備 1 名を除外した 111 名が分析対象者となった。IRT モデルにより算出された学力特性値（ θ ）の平均値は 0.04，標準偏差 1.17，最小値-2.57，最大値 2.45 であった。学力レベルごとのパフォーマンス記述が目的であるため，以降，学力特性値（ θ ）について，L 群（-2.57~-0.59），M 群（-0.59~0.43），H 群（0.47~2.45）の 3 グループ（各 N=37）に分類して分析をおこなう。

学力レベルごとに，面積テストにおける各面積比較課題の平均正答率を算出した（Figure2)。全体の傾向として，等積変形課題よりも等周長変形課題の方が正答率が低く，解決が困難であることが読み取れる。また，いずれの課題でも，学力レベルが上昇するにしたがって正答率が高くなる傾向にあることがわかる。ただし，課題ごとの正答率について，学力レベルごとにみると，M 群は L 群や H 群に比して課題間の正答率の開きが大きい。面積の大小を判断する際に，一貫した判断基準を用いているわけではなく，課題によって判断の根拠を変えている可能性がある。

そこで次に，面積課題の判断理由の記述内容について分析し，各学力レベルにおける具体的なパフォーマンスの様相について質的側面から検討をおこなった。面積判断の正誤を不問にして理由記述をみると，適切属性である「高さ」への着目や公式による計算だけでなく，不適切属性である「図形の周長・斜辺」への着目や，「おしつぶす」「ひきのばす」などの「動作」への着目がみられた (Table1)。理由記述について，a.「高さへの着目」b.「公式計算」c.「周長・斜辺への着目」d.「動作への着目」e.「切り貼り変形」f.「その他」の 5 つに大別し，学力レベルごとに出現頻度

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　141　2016/08/29　14:01:07

をカウントした（Figure3）。その結果，等積変形課題，等周長変形課題とも，面積を規定する適切属性に着目した a，b の出現頻度は L 群＜M 群＜H 群であったのに対し，不適切属性である c，d の出現頻度は H 群＜M 群＝L 群であった。したがって，学力レベルが高いほど，不適切属性への着目を回避し底辺・高さや公式を利用した面積大小判断が可能になることがうかがえる。また，課題間で比較すると，H 群では a，b の頻度は課題間で差がなかったのに対し，M 群，L 群ではこれらの頻度は等周長変形の方が少なかった。むしろ，これらの群では等周長変形において c や d の出現頻度が多くなっていた。このことから，H 群では適切属性に依拠して判断する傾向にあるのに対して，L 群，M 群では面積大小の判断基準が課題によって異なり，一貫した根拠をもちえない傾向にあることがうかがえる。

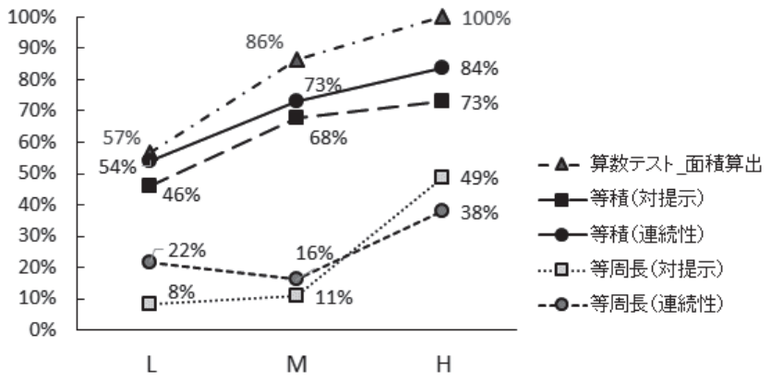


Figure2　各学力レベルにおける面積課題の正答率

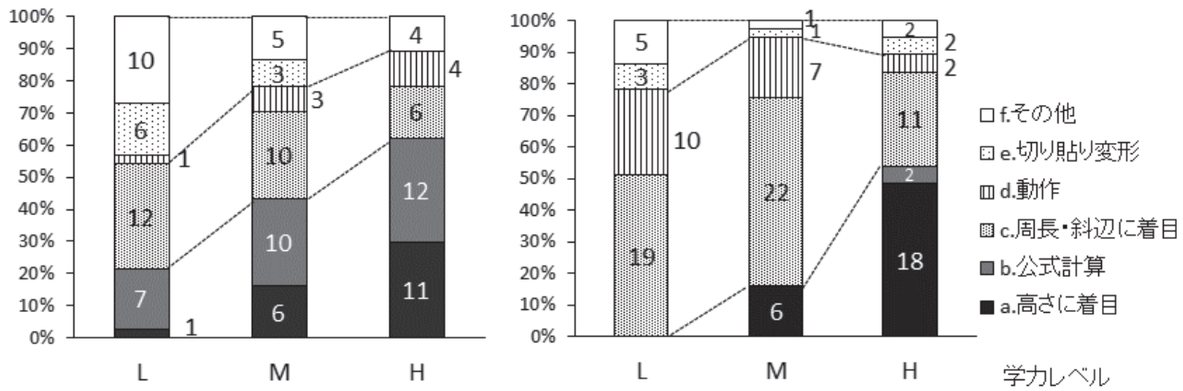
Table1　各面積課題における判断理由の分類

理由分類	等積変形	等周長変形
a.高さへの着目	・底辺も**高さ**も同じだから ・**高さ**が，アよりイの方が短いから	・花子さんの**高さ**が小さいから。
b.公式計算	・アは**4×2＝8**で，イは，高さで求めると同じく**4×2＝8**で，いっしょだから	・次郎君の形は**4×3＝12**，花子さんの形は**4×2．5＝10**だから。
c.周長・斜辺への着目	・**アは4×2＝8で，イは8×2＝16**（※底辺×斜辺）なのでイの方が大きいと思う。 ・横の長さがイの方が長いから。	・**マッチの数が同じ**だから ・**縦のマッチがどちらも3本で3×4をすると12**（※辺×辺）でどちらも同じだから。
d.動作への着目	・イはアを**ひきのばした形**になっているから，面積は同じだと思う。 ・**ただ横にのばしているだけだから**どちらも同じだと思う	・花子さんの形は，**次郎くんといっしょの形にもどりそう**だから。 ・花子さんの形は**次郎さんの形をつぶしてナナメに広がっただけ**だから
e.切り貼り変形	・まずイをよこに半分にして半分にしたイとイをくっつけると，底辺が 4 ㎝，縦が 4 ㎝でイの方が大きい	・花子さんの形の右の方を切り取って，左につけると面積が少ないから。
f.その他(無答含む)	・ななめの線は，面積をだすのに関係ないから	・平行四辺形の底辺と上にのびているマッチの所にすきまがあるから

なお，算数学力テストでも面積の問題を出題していたので，参考までにその面積算出問題の正答率を学力レベル間で比較した結果，Figure2 に示す通りとなった。この問題は，底辺，高さを同

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　142　2016/08/29　14:01:08

定してその数値を求積公式に代入することで解決可能である。L 群でも，その半数は，底辺・高さの長さが明示された場合には面積を算出することができるといえる。



等積変形課題（※数字は人数）　　等周長変形課題（※数字は人数）

Figure3　各面積課題における判断理由の分布

上記の結果に基づいて面積学習領域の学力評価ルーブリックを記述すると，次のようになる。学力レベル L 群は，底辺，高さの数値が具体的に指定された場合は公式に数値を代入して面積を算出することができる。しかし，複数の図形の面積判断の際には図形の周長に着目しがちであり，底辺と高さの相対的大小関係から面積の大小を導出するような形式的操作は困難である。次に，学力レベル M 群については，図形の底辺および高さを読み取り，公式に数値を代入して面積を算出することによって面積の大小を比較することができる。しかし，図形の底辺や高さの長さの数値が記入されていない問題など，課題によっては不適切属性（図形の周長）が面積比較の際に関与する傾向がある。また，面積の判断基準が課題内容に左右され，根拠が一貫しない傾向がある。このような学習者にとって，求積公式は「面積を算出するための手続き・道具」という認識（機械的公式観：麻柄，2009）にとどまっていると考えられる。そして，学力レベル H 群は，底辺，高さの相対的大小に着目し，そこから面積の大小を導出して判断することができる。このような学習者は，図形の周長などの不適切属性を分離させて面積を捉えることができるといえる。

課題と今後の展望

本研究では，面積学習領域を例に IRT にもとづくルーブリック作成の手法を提案したが，思考力，問題解決力などを測定するためのパフォーマンス課題の作成には，当然ながら教科内容についての教材研究が必要となる。教科学習における学習目標と照らし合わせながら作成する必要があるため，ルーブリック作成にあたっては教育現場や実践者等と共同で行うことが求められよう。今後の展望として，本研究で提案したような探索的検討を通じて指導に活かせるルーブリックが蓄積されることで，算数・数学の習熟度別指導への活用が期待できる。さらに，IRT モデルをもとに，異なる学年のルーブリックを相互に関連づけられれば，学習者個人の学年間での学力の伸びを記述することも可能になるだろう。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　143　2016/08/29　14:01:08

認知診断モデルによる学習診断の教育実践での利用可能性

鈴木 雅之
（横浜国立大学教育人間科学部）

テストを実施する目的の 1 つには，学習者の学習改善に活用することがある。この目的を達成するためには，テストによって，学習の達成度や学習上のつまずきを明確にすることが重要になる（e.g., Hattie & Timperley, 2007）。たとえば，数学の「一次関数」の単元に関するテストで成績が振るわなかったときに，方程式の解き方を習得していなかったことが原因であると特定できれば，方程式について学習をするべきだということが明確になる。しかし，多くの場合，テストの結果として学習者にフィードバックされる情報は，テスト全体の得点や正答率，受検者の中での序列であり，これらは学習の達成度やつまずきについて有用な情報を与えてくれない（池田, 2013）。

こうした中で，学習者の学習状況を診断するための方法として，認知診断モデル（cognitive diagnostic model）が注目されている（e.g., Leighton & Girel, 2007; Rupp, Templin, & Henson, 2010）。認知診断モデルによる学習診断では，アトリビュート（attribute）と呼ばれる，ある教科や単元などの学習内容を理解するのに必要とされる知識やスキルを定義し，そのアトリビュートの習得確率を推定する。たとえば，「分数の加減」に関する問題を解決するためには，「同分母の加減」「通分」「約分」の 3 つのアトリビュートが必要だとして，「分数の加減」に関する問題群の解答状況から，各アトリビュートの習得状況を知ることができる。これによって，「分数の加減」でつまずいているのは，通分の方法を知らないからなのか，そもそも分数の計算に関する知識が全くなく，同分母の加減ができていない状況であるのかを把握することができる。このように，認知診断モデルによる学習診断を行うことによって，学習上のどこにつまずきがあるかについて，より詳細な情報を得ることが可能になる。

しかし，認知診断モデルを適用した研究例はこれまでにもあるものの（e.g., Wang & Gierl, 2011），認知診断モデルによる学習診断は，教育実践の中でほとんど利用されていない。その原因の 1 つに，教育実践の中で広く利用されているテストに対して認知診断モデルを適用した事例が少ないことが挙げられる。そこで本発表では，日本の小・中学校で広く実施されている標準学力検査である「教研式標準学力検査 NRT（以下，教研式 NRT）」に対して認知診断モデルを適用した事例（鈴木・豊田・山口・孫, 2015）をもとに，認知診断テストの教育実践での利用可能性について論じる。

鈴木ほか（2015）では，中学 2 年生 948 名（男性 496 名，女性 452 名）を対象に，教研式 NRT「中学 1 年数学」を学年はじめに実施し，「計算力」と「概念理解」，「図形操作力」，「論理力」の 4 つのアトリビュートを特定し，G-DINA モデル（generalized-DINA model; de la Torre, 2011）による分析を行った。その結果，認知診断モデルを適用することで，素点に基づく学習診断で

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 144 2016/08/29 14:01:09

は推測できない学習者の特徴を知ることが可能であることが示唆された。具体的には，まず，テスト全体の得点が高い学習者であっても，未習得のアトリビュートを有する場合があることが示された。全体としてテストの成績が良いからといって，つまずきはないと判断した場合，その後の学習で問題が生じてしまう可能性があるため，つまずきを特定することは重要になる。また，テスト全体の得点が低い学習者であっても，習得しているアトリビュートを有していることがあることが示された。数学のテストで成績が振るわなかった場合に，何から学習すれば良いかを判断することは困難であるが，アトリビュートの習得状況が明らかになることで学習目標が明確になり，効率よく学習改善を試みることができる。さらに，たとえば立体の投影図から想定される立体を構成する問題の解決において，「投影図」という言葉の意味が理解できていないのか，平面図形から立体を構成する力を欠いているのかなど，問題解決できなかった原因に迫れる可能性が示された。

以上のように，認知診断モデルを適用した学習診断を行うことの利点が示される一方で，認知診断モデルによる学習診断を教育実践の中で行っていくためには課題もある。まず，アトリビュートの粒度に関する問題がある。アトリビュートの粒度の細かさによって学習診断の有用性は異なると考えられるが，どの程度の細かさが有用であるかを決定することは容易ではない。また，Q 行列（Q-matrix）を誤設定した場合には，項目パラメータの推定や診断結果に誤りが生じるなど（Rupp & Templin, 2008），Q 行列の設定に関する問題がある。さらに，認知診断モデルには様々なモデルがあり，適用する認知診断モデルによって，アトリビュート習得状況の推定結果は異なる。したがって，これらの課題を解決していくことが，認知診断モデルによる学習診断を教育実践の中で利用していく上で重要になるだろう。

引用文献

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, **76**, 179-199.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, **77**, 81-112.

池田　央 (2013). テストの過去，現在，そして未来の形を考える―その方向性と必要性―. 日本テスト学会誌, **9**, 1-14.

Leighton, J., & Gierl, K. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Rupp, A. A., Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, **68**, 78-96.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and application*. New York: Guilford.

鈴木雅之・豊田哲也・山口一大・孫媛 (2015). 認知診断モデルによる学習診断の有用性の検討―教研式標準学力検査 NRT「中学 1 年数学」への適用―. 日本テスト学会誌, **11**, 81-97.

Wang, C., & Gierl, M. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, **48**, 165-187.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　145　2016/08/29　14:01:09

外国語学習支援に向けた認知診断テストの開発

○孫 媛*・島田 めぐみ**・谷部 弘子**
*国立情報学研究所，**東京学芸大学留学生センター

1. はじめに：認知診断テストとは何か

1.1 認知診断テストの必要性

従来の学力テストでは，得点という 1 つの数値で結果が示される。だが，得点はテストの難易度に左右され，得点だけで学習者の学力や能力は判断できない。得点を偏差値に直して集団内の位置を示すこと，あるいは一定の到達基準をクリアしたか否かで「級」を与えることもあるが，正解に至らなかった原因や学習者が何を苦手としているのかは，やはりわからない。

認知診断テスト（Cognitive Diagnostic Assessment）は，習得状況そのものに焦点を当てたテストである。認知診断テストでは，注目する領域・ドメインの習得・理解に必要な認知的要素（「アトリビュート」（attribute）とよぶ）を定義し，受験者のアトリビュートの習得状況を測ることが目的となる。認知診断テストを使うことで，学習者本人や教師は，学習目標の到達度，強み弱み，躓きの原因といった形成的な診断情報を得て，学習や教育活動の改善に活かすことができる（Embretson, 1998; Gierl, Leighton, & Hunka, 2000; Nichols, 1994）。

認知診断アプローチは，test *of* learning から test *for* learning へ転換が進む中で注目されるようになった。言語教育の領域でも，従来型の能力テスト（proficiency testing）の限界が認識され，学習者の言語習得や教師の学習指導に役立つ情報の報告が重視されるようになり（Alderson, 2005），認知診断テストの活用が期待されている。

1.2 認知診断テストの作成と実施の手順

認知診断テストの作成から実施までの手順は，4 つのステップに整理できる。

1. アトリビュートの定義

教科の専門家や認知心理学者が協力して，学習領域における学習・教育上の目標を定め，アトリビュートを定義する。領域の課題分析から始め，領域の知識，解答過程のプロトコル，授業の経験，項目作成経過を分析することが有効だとされる（Buck & Tatsuoka, 1998; Leighton, Gierl, & Hunka, 2004）。

2. Q マトリックスの構成

テスト項目ごとに，アトリビュートが必要ならば 1，不要ならば 0 を与えることで，項目×アトリビュートの Q マトリックスを構成する。項目への解答パタンからアトリビュートの習得確率を推定するための認知診断モデル（Cognitive Diagnostic Model: CDM）として，多様なモデルが提案されているが，例外なく Q マトリックスを前提としている。

3. 学習者の習得状況の推定

p 個のアトリビュートの習得状況（習得＝1，未習得＝0 として 2^p 通り）×アトリビュートのマトリックスを Q マトリックスと組み合わることで，理論上の正誤パタン（正答＝1，誤答＝0）が得られる。アトリビュートの習得確率は，理論上の正誤パタンと受験者の解答パタンの関係を確率的に説明する認知診断モデルに基づいて推定される。

4. 診断結果の報告

アトリビュートに対する学習者の習得度を報告することで，学習者は学習領域における自身の

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 146 2016/08/29 14:01:10

弱点を把握できる。認知診断の結果をどのように報告すれば，学習改善を促すのか，効果的なフィードバック法の研究も重要である。

2. 日本語語彙認知診断テストの開発と実施

2.1 日本語語彙のアトリビュートの定義と項目の作成

外国語習得において語彙の重要性は高く，その獲得は母語や地域など学習者ごとの違いの影響を受けやすい。とくに中国語を母語とする学習者は，日本語能力試験の文字・語彙領域で，母語の転移と思われる誤答をすることが少なくない。そこで今回は語彙領域に注目して，日本語教育分野における認知診断テストを開発することにした。日本語能力試験の過去問題（語彙領域）を分析して基本アトリビュートを設定した後，文化庁（1978），張（2009）等を参考に母語別のアトリビュートを設定した。さらに予備調査に基づいてアトリビュートを再検討し，表 1 のように整理した。中国語母語話者を対象としたため，中国語についての母語アトリビュートを設定した。

表 1 日本語語彙領域のアトリビュート

基本アトリビュート

A：形に関わるアトリビュート
- A1：ひらがな表記の和語がわかる
- A2：漢字表記の和語がわかる
- A3：漢字表記の漢語がわかる
- A4：カタカナ表記の語がわかる
- A5：音声で提示された語がわかる

B：意味に関わるアトリビュート
- B1：基本的意味がわかる
- B2：類義語との相違がわかる
- B3：派生的な意味がわかる

C：用法に関わるアトリビュート
- C1：統語的特徴（品詞性）がわかる
- C2：共起関係がわかる

母語に関わるアトリビュート

D：中国語に関わるアトリビュート
- D1：中国語と同形，意味がほぼ同じ語がわかる
- D2：中国語と同形，意味が異なる語がわかる
- D3：中国語にないが同じ意味の漢字を含む語がわかる
- D4：中国語になく同じ意味の漢字を含まない語がわかる

日本語能力試験 N2 レベル相当の難易度を想定して，上記アトリビュートの習得状況を推定するための 73 項目（問題Ⅰ：10 項目，問題Ⅱ：32 項目，問題Ⅲ：31 項目。すべて 4 肢選択形式）を作成した。A5 の習得を確認するための問題Ⅰは，選択肢を音声で提示する。問題例を示す。

問題□　選択肢を聞く問題です。次の文の（　　）に入れるのに最もよいものを，1〜4 の中から一つ選びなさい。

（1）　来月，新しいタイプのコンピュータが（　　）されるそうだ。

（音声提示）1. 売買　　2. 商売　　3. 販売　　4. 発売

問題□　次の文の（　　）に入れるのに最もよいものを，1〜4 の中から一つ選びなさい。

（11）　約束の時間を守れないなんて，彼女は時間に（　　）人なんだね。

1. しょうがない　2. きたない　3. くだらない　4. だらしない

問題□　次の言葉の使い方として最もよいものを，1〜4 の中から一つ選びなさい。

（43）　縮む

1. 文章が長すぎたので，半分に縮んだ。
2. このページをコピーするとき，半分の大きさに縮んでください。
3. 最近，日本語学習者数が縮んできた。
4. セーターを自分で洗ったら縮んでしまったので，妹にあげた。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　147　2016/08/29　14:01:11

2.2　Q マトリックスの構成

次に，項目ごとに正答のために必要なアトリビュートを特定し，Q マトリックスを作成した。Q マトリックスのうち，上記で例示した 3 項目の部分を抜き出したのが表 2 である。たとえば，項目 1 には，アトリビュート A5（音声で提示された語がわかる），B2（類義語との相違がわかる），D3（中国語にないが同じ意味の漢字を含む語がわかる）を付与している。

表 2　日本語語彙テスト Q マトリックス（一部）

項目	アトリビュート													
	A1	A2	A3	A4	A5	B1	B2	B3	C1	C2	D1	D2	D3	D4
1	0	0	0	0	1	0	1	0	0	0	0	0	1	0
11	1	0	0	0	0	0	0	1	0	1	0	0	0	0
43	0	1	0	0	0	1	0	0	1	0	0	0	1	0

2.3　調査の実施と習得状況の推定

2013 年 3 月，中国大連の A 大学と煙台の B 大学で日本語語彙認知診断テストを実施した（A 大学で 131 名，B 大学で 103 名が受験した）。また，コンピュータでの運用を検討するために，北京の C 大学ではコンピュータによりテストを実施した（30 名）。紙媒体を使用した A 大学と B 大学の受験者 234 名において，平均 32.07，標準偏差 8.53，最大値 61 点（満点は 73 点），最小値 12 点であった。図 1 は得点分布のヒストグラムである。

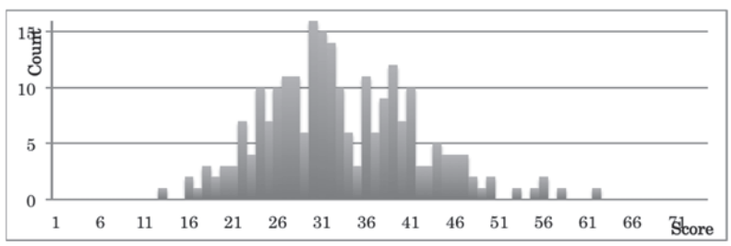


図 1　得点分布

認知診断モデルとして DINA モデル（deterministic inputs, noisy “and” gate model: Haertel, 1989；Junker & Sijtsma, 2001）を用い，各受験者についてアトリビュートごとの習得確率を推定した。得られた習得確率の平均値を表 4 に示す。

表 4　各アトリビュートの習得確率（平均）

	A1	A2	A3	A4	A5	B1	B2
Mean	0.481	0.721	0.678	0.714	0.573	0.709	0.765
Sd	0.399	0.313	0.337	0.347	0.387	0.358	0.308
	B3	C1	C2	D1	D2	D3	D4
Mean	0.583	0.685	0.717	0.698	0.589	0.700	0.730
Sd	0.329	0.327	0.303	0.292	0.335	0.327	0.347

2.4　診断結果の報告

認知診断テストでは一つの得点だけではなく，アトリビュートごとの結果を受験者や指導者にフィードバックできる。図 2 から図 5 は，同じ得点（33 点）を得た 4 名の受験者について，アト

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　148　2016/08/29　14:01:11

リビュート別に結果を示したものである。得点が同じ受験者であっても，アトリビュートごとの習得確率は大きく異なることがわかる。

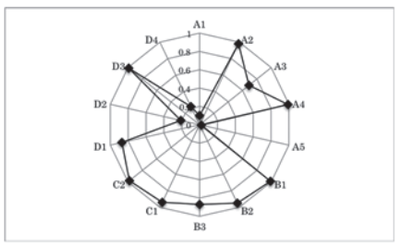


図 2 受験者 1

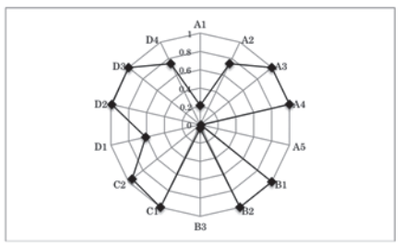


図 3 受験者 2

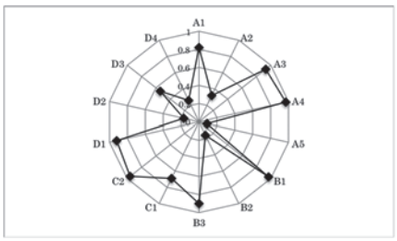


図 4 受験者 3

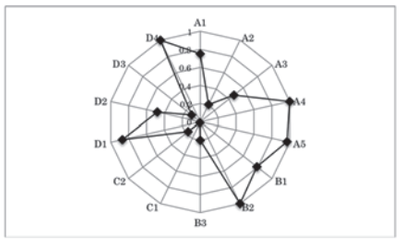


図 5 受験者 4

3. 考察と今後の展望

日本語語彙理解に必要なアトリビュートを定義し，認知診断テストのための項目を作成した。中国の大学で日本語を学習する学生に，この認知診断テストを受けてもらい，その解答データに対して DINA モデルを適用してアトリビュートの習得状況を推定した。その結果，学習者ごとの習得状況の違い，学習者の特徴を確認できた。また，診断結果を学習者本人と担当教員にフィードバックし，結果の妥当性を検証した。個々の学習者に対応した指導に繋がることが期待される。

ドメインの知識構造やアトリビュートの定義・特定は今後も研究が必要な課題であるが，ほかにも，データの蓄積および適用モデルの適合度や推定方法の検討，学習改善の促進に効果的なフィードバック方法の検討等も，取り組むべき重要な課題である。さらに，認知診断テストを LMS（Learning Management System）に実装し，学習者の習得プロフィールを即時に診断しフィードバックすると同時に，教科書や参考書などと連動させて，学習者が未習得のアトリビュートについて具体的な学習コンテンツを適切なタイミングで提示するなど，個別化学習システムを構築していくことを考えている。

Web 上における学習支援は，今後もさまざまな形で展開されていくのは間違いない。将来的には，Web 学習空間において収集可能な学習者の学習活動に関するログデータと認知診断テストを統合・併用することを視野に入れて，認知診断テストの活用方法を研究することも意味があるだろう。また，認知診断テストの実用性を高めるには，テスト項目のアトリビュート付与の自動化が不可欠である。現在，data-driven Q-matrix 自動作成の研究にも取り組んでいるところである。

謝辞

本研究は，情報・システム研究機構新領域融合研究プロジェクト「データ中心科学リサーチコモンズ」，および科学研究費補助金挑戦的萌芽研究26560134の助成を受けて実施されました。

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 149 2016/08/29 14:01:12

一般セッション７：ｅテスティング

９月９日（金）16:30 – 18:30　第２会場（B102）

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　151　2016/08/29　14:01:12

適応型テストの回答データを用いた新作項目の項目特性値推定

○杉山 剛　仁田光彦
株式会社リクルートキャリア測定技術研究所

1　はじめに

適応型テストの運用においては項目プールの拡充は重要なテーマのひとつである．項目プールの拡充の方法として，予備テストを行って新作項目の項目特性値をあらかじめ算出する方法のほかに，本番テストの中に新作項目を採点除外項目として紛れ込ませる形でデータ収集をおこなって項目特性値を算出する方法がある．本番テストの中に新作項目を紛れ込ませる方法には，本番テストで得られた受検者の特性値 θ を新作項目の項目特性値推定に使用することで本番テストの項目との等化済みの項目特性値が算出されるという利点や，本番テストの運用の延長上で項目プールの拡充が可能になるという利点がある．一方で，予備テストを行う方法の場合は実務上莫大なコストがかかるほか，運用上の負荷も大きい．

受検者の特性値 θ と項目への正誤を用いて項目特性値を推定する方法として，藤田ら(2009)が提案した回帰推定法がある[1]．θ をいくつかのレベルに分け，各レベルの正答率をもとに回帰計算により項目特性値を推定する方法であるが，安定した正答率を得るために一定数以上の回答情報が必要となる．藤田らによると 1000 名以上で安定するとされている．新作項目の場合は，十分吟味して作成したとしても項目特性値がよくないというケースも起こりうるが，回帰推定法の場合 1000 名分のデータを集めた上でないとその判断をおこなうことができないという意味においてやや非効率な面は否めない．

このほか，新作項目の項目特性値を随時推定し，その項目特性値をもとにどのような特性値 θ の受検者に出題するかをコントロールする方法については韓ら(2014)[2]の取り組みがある．

この発表では，受検者の特性値 θ と新作項目への正誤反応のみを用いて疑似最尤法による項目特性値の推定を行う方法と，この推定を毎日行うことで最新の推定状況をもとに新作項目の出題をコントロールする手法について報告する．提案手法により，適応型テストにおける新作項目のデータ収集および項目特性値推定の効率化に寄与することが期待できる．

2　擬似最尤法による項目特性値の推定

2.1　対数尤度関数の導出

適応型のテストに新作項目を紛れ込ませる形で N 人が受検した場合を考える．

受検の結果，新作項目以外の項目（項目特性値は既知）への正誤により推定された受検者の特性値 θ と，当該受検者の新作項目への正誤 u の N セットの組合せが得られる．

この組み合わせが得られる確率は，以下の式で表すことができる．

$$f(a,b|\mathrm{u},\theta)=\prod_{i=1}^{N}p(\theta_i)^{u_i}\,q(\theta_i)^{1-u_i}$$

また u と θ は既知なので定数とみなすと、尤度関数も同じ式となる。

$$L(a,b|\mathrm{u},\theta)=f(a,b|\mathrm{u},\theta)$$

一般セッション7

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　150　2016/08/29　14:01:13

ここで正答の確率 p(θ i)，誤答の確率 q(θ i)は 2 母数ロジスティックモデルでは以下のように表される。

$$
\begin{aligned}
p(\theta_i) &= \frac{1}{1+exp[-Da(\theta_i-b)]} \\
q(\theta_i) &= 1-p(\theta_i) \\
&= \frac{exp[-Da(\theta_i-b)]}{1+exp[-Da(\theta_i-b)]}
\end{aligned}
$$

尤度関数の両辺の対数をとって p(θ i)，q(θ i)を展開することで以下の対数尤度関数が得られる．

$$
\begin{aligned}
log(L(a,b|\mathrm{u},\theta)) &= \sum_{i=1}^{N}(u_i log(p(\theta_i)) + (1-u_i) log(q(\theta_i))) \\
&= \sum_{i=1}^{N}\left(u_i log(\frac{1}{1+exp[-Da(\theta_i-b)]}) + (1-u_i) log(\frac{exp[-Da(\theta_i-b)]}{1+exp[-Da(\theta_i-b)]})\right) \\
&= \sum_{i=1}^{N}((1-u_i)(-Da(\theta_i-b)) - log(1+exp[-Da(\theta_i-b)]))
\end{aligned}
$$

この対数尤度関数を最大化する a 値と b 値を，ニュートン・ラフソン法を用いて数値的に求める．実際には知ることのできない受検者の特性値 θ に本番テストの受検結果を使い，定数として扱っているため，擬似最尤法による項目特性値の推定となる．

2.2　初期値の算出

ニュートン・ラフソン法を使う場合，まず初期値を最適解の近くに設定する必要がある．

古典的テスト理論で項目の性能を示す点双列相関係数と正答率を使って，項目反応理論の a 値，b 値を求める以下の式が知られている[3][4]．

$$
\begin{aligned}
aa_0 &= \frac{Dr}{\sqrt{1-r^2}} \\
bb_0 &= \frac{z}{r}
\end{aligned}
$$

ただし，r は θ と u の点双列相関係数，D は定数(1.7)，z は正答率を標準正規累積確率に対応させたときの z スコア(正答率 p=P(Z>z)となる z である．

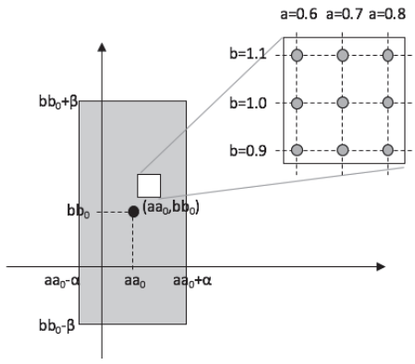
しかしニュートン・ラフソン法の初期値とするには，この aa_0，bb_0 は最適解の十分近くにあるとは限らない．そこで aa_0，bb_0 を中心に一定範囲を 0.1 刻みに格子状に調べる（グリッドサーチ）ことで，その中で最も対数尤度関数が大きくなる a 値と b 値の組み合わせを探し，その値をニュートン・ラフソン法の初期値とする(図 1)．

なおここで相関係数の絶対値が 0 や 1 だったり，全員が正答，全員が誤答の場合は aa_0，bb_0 は計算できない．計算できない場合は，基本的には人数が少なすぎるか項目の性能に問題がある（易しすぎ，難しすぎ，識別力が低すぎ)ことが原因なので，この時点においては推定不能とし，回答

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　151　2016/08/29　14:01:13

データが増えた後にあらためて推定することとする．

図１　グリッドサーチのイメージ図



※網掛け部の中のすべての格子点について対数尤度を計算する

2.3　ニュートン・ラフソン法による推定

ニュートン・ラフソン法は，求めたい値のベクトルβがあるとき，f(β)をβで偏微分した勾配ベクトルΔf(β)と２階偏微分したヘッセ行列Δ^2f(β)を使って，

$$\beta[k+1] = \beta[k] - \Delta^2 f(\beta)^{-1} \cdot \Delta f(\beta)$$

という式でβ[k]を順次更新していき，f(β)の変化量が閾値よりも小さくなったらそこで計算を終了，というアルゴリズムである．

今回求めたい値のベクトルは(a, b)なので，上記の更新式は以下となる．

$$\begin{pmatrix} a[k+1] \\ b[k+1] \end{pmatrix} = \begin{pmatrix} a[k] \\ b[k] \end{pmatrix} - \begin{bmatrix} \frac{\partial^2}{\partial a \partial a} LL(a,b) & \frac{\partial^2}{\partial a \partial b} LL(a,b) \\ \frac{\partial^2}{\partial b \partial a} LL(a,b) & \frac{\partial^2}{\partial b \partial b} LL(a,b) \end{bmatrix}^{-1} \begin{pmatrix} \frac{\partial}{\partial a} LL(a,b) \\ \frac{\partial}{\partial b} LL(a,b) \end{pmatrix}$$

※LL(a,b)は対数尤度関数

また，この推定結果の標準誤差はヘッセ行列の逆行列の対角成分の平方根をとって以下の式となる．

$$StdErr_a = \sqrt{\frac{-\frac{\partial^2}{\partial b \partial b} LL(a[k],b[k])}{\frac{\partial^2}{\partial a \partial a} LL(a[k],b[k]) \frac{\partial^2}{\partial b \partial b} LL(a[k],b[k]) - \frac{\partial^2}{\partial a \partial b} LL(a[k],b[k]) \frac{\partial^2}{\partial b \partial a} LL(a[k],b[k])}}$$

$$StdErr_b = \sqrt{\frac{-\frac{\partial^2}{\partial a \partial a} LL(a[k],b[k])}{\frac{\partial^2}{\partial a \partial a} LL(a[k],b[k]) \frac{\partial^2}{\partial b \partial b} LL(a[k],b[k]) - \frac{\partial^2}{\partial a \partial b} LL(a[k],b[k]) \frac{\partial^2}{\partial b \partial a} LL(a[k],b[k])}}$$

なお，この推定では当該受検で推定したθを定数としているが，実際にはこのθは本番テストの推定結果であり測定誤差を含んでいる．そのため算出された項目特性値の標準誤差はθの測定誤差を含まない分，過小評価されたものであることに留意する必要がある．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　152　2016/08/29　14:01:13

3　出題のコントロール

適応型テストを実施している大規模コンピュータシステム上に前述の算出ロジックを組み込むと，たとえば毎日の夜間処理ですべての新作項目の項目特性値および標準誤差を計算することができる．

その推定結果を用いることで以下のような出題コントロールが可能となり，効率よく新作項目のデータ収集をおこなうことができる．

① 項目特性値推定時の標準誤差が十分小さくなった項目はこれ以上の回答データは不要なので出題を打ち切る
② 一定以上の人数のデータが集まったにもかかわらず標準誤差が小さくならない項目は項目自体の性能に問題があると判断して出題を打ち切る
③ 前日の推定結果をもとに，b 値から大きくかけ離れた θ の受検者には，得られる情報量が少なく効率が悪いので出題しない（当該受検者には別の新作項目が割り当てられる）

4　まとめ

擬似最尤法による項目特性値の推定を行うことで，適応型テストにおいて本番テストに新作項目を紛れ込ませる形でデータ収集を行う場合に，データ収集の途中であっても最新の回答情報をもとに項目特性値と標準誤差を推定することが可能になる．

さらに得られた最新の推定結果をもとに適切な出題コントロールをおこなうことで，同程度の受検者数に対して，より多くの新作項目のデータ収集をおこなうことができるようになり項目プールの拡充が促進される．

今後は，標準誤差がどの程度であれば十分な精度といえるか，またどれくらいの人数をもってそれ以上推定精度があがらないと判断できるかなどについては，実際の出題を通じて慎重に吟味・調整していく必要がある．シミュレーションの結果とそれを踏まえた考察については当日発表予定である．

参考文献

[1] 藤田彩子，舛田博之(2009)：「適応型テストの回答データを用いた項目特性値の推定」日本テスト学会第 7 回大会発表資料集
[2] 韓 太哲，村木英治，吉川 厚，小林夏子，林 規生(2013)：「コンピュータ適応型テストの中で、項目パラメーターを適応的に推定（３）－２パラメーターモデルについて－」日本テスト学会第 11 回大会発表資料集
[3] Lord, F. M. (1980). Applications of item response theory to practical testing problems. Routledge.
[4] 加藤健太郎，山田剛史，川端一光 (2014)：「R による項目反応理論」オーム社

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　153　2016/08/29　14:01:14

大規模 CBT 実施における標準手法
ーAsia ATP のホワイトペーパーに基づく検討ー

齋藤 康寛
株式会社イー・コミュニケーションズ

１．はじめに

近年，高大接続システム改革で大学入試への CBT (Computer Based Testing)[1]の導入が検討されるなど，CBT に関する注目が高まっている一方で，日本では大規模な CBT を想定したテスト実施のガイドライン整備が十分な状態とは言えない．この度，グローバルなテスト業界団体である Association of Test Publishers (ATP) のアジア部門である Asia-ATP (A-ATP)が 2015 年に作成した，「大規模試験における標準手法に関するホワイトペーパー（White Paper on Standard Practices for Large-Scale Assessment Programs)」（以下，「本ホワイトペーパー」とする）を本学会に紹介する許可が得られたため，概要を紹介するとともに，既存のガイドラインと比較することにより，今後ガイドラインとして整備していくべき内容を検討した．

２．本ホワイトペーパーの概要

本ホワイトペーパーは，米国の Council of Chief State School Officers (CCSSO)と，The Association of Test Publishers (ATP) が，2010 年に最初に出版した Operational Best Practices for Statewide Large-Scale Assessment Programs の 2013 年版（以下，"Operational Best Practices" という）をベースに，Asia-ATP が Asia 向けにローカライズを進めているもので，その一部のチャプターを先行して Asia-ATP 会員向けに公表したものである．

Operational Best Practices は，ガイドラインとして守るべき基準を定めて制約を加えることを目的とするのではなく，テスト関連の業界に共通のベストプラクティスの情報を提供し，テスト実施の一助となることを目的としている．

また，Operational Best Practices は，2013 年度版を作成する過程において，"technology-based assessments" への対応のため大幅に改変されている．ここでいう"technology-based assessments" は，コンピューターを使用したテストを指し，CBT とほぼ同義である．本ホワイトペーパーは，その中でも特に"technology-based assessments"に関連した内容となっている，"Chapter 15. Technology-based Assessments and Technical Support" を含む２つの章を対象としており，今回の発表では，Chapter 15 の内容について紹介する．

３．" Technology-based Assessments and Technical Support" の内容

Chapter 15 は，本ホワイトペーパーでは 2 章に掲載され，以下の内容を含む．元の文書にはタイトルは無いが，本発表においては，便宜上タイトルを付与した．

No.	タイトル	内容
2.1	技術要件の定義	CBT を実施するにあたり，顧客とサービス提供者に必要とされる技術インフラの要件は，明確に定義され明文化されていなければならない．
2.2	デバイスやソフトウェアの仕様や設定の明確化	CBT による試験を安全に配信，管理するために必要とされるデバイスやソフトウェアの要件は明確に書面で説明されていなければならない．これらの要件は，マニュアルその他の書面で関係者に共有されるべきである．
2.3	アクセシビリティへの配慮	顧客とサービス提供者は協力して，CBT においてアクセシビリティを高めるためにどのような技術が利用可能かを検討するべきである．
2.4	セキュリティを考慮した技術や手順	顧客とサービス提供者に利用されるすべての技術や手順は，CBT 実施のすべてにおいて，セキュリティを維持すべきである．

[1] CBT は，狭義では Internet Base Testing(IBT)と区別するために，試験会場で試験監督での管理下で実施する形態のみを指す場合もあるが，ここではコンピュータを使用するテスト全般を指すこととする．

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 154 2016/08/29 14:01:15

2.5	技術サポートの提供	電話，電子メール，WEB 経由でのドキュメントなどのカスタマーサポートを通じて，CBT に関する技術サポートが，サービス提供者から顧客に提供されるべきである．
2.6	会場構成に関する推奨事項の記載	技術サポート資料にはテストセンター，コンピューター室などの設定・構成についての推奨条件を記述すべきである．
2.7	トレーニングの提供	サービス提供者から顧客に対して，CBT に関するトレーニングが提供されるべきである．トレーニングは，WEB，直接，動画，資料など複数の形態で提供されることが望ましい．
2.8	性能の測定・監視と拡張性	サービス提供者のソフトウェアならびにシステムの性能は CBT の運営の成功において重要であるため，測定され，常時監視されるべきである．また，試験の規模の増加，試験の種類の変化，技術の変化に合わせて対応できるように拡張性を持つべきである．
2.9	システム変更の管理	サービス提供者により，機能向上，不具合修正，技術の変化や基準の変更，顧客の要求などに対応するためのソフトウェアやシステムへの変更が行われるが，そのプロセスは適切に計画，管理され，顧客やその他の関係者に事前に説明するコミュニケーションが適切に行われているべきである．

４．既存のガイドライン／スタンダードとの比較

既存のガイドライン／スタンダートとして CBT に関する記述があるもののうち，以下の 2 つとの比較を行った．

- 「テスト・スタンダード」／日本テスト学会（編）(2007). 金子書房
- ITS Guidelines on Computer-Based and Internet Delivered Testing. / International Test Commission (ITC) (2005).

以下では，上記のうち前者を「テスト・スタンダード」，後者を「ITS ガイドライン」とする．

本ホワイトペーパーの 2.1～2.4 については，テスト・スタンダード，ITS ガイドラインのいずれにも同様の内容が含まれるが，2.5～2.9 については，記載が無いか，ある場合も簡易な記載にとどまっている．記載内容にこのような違いが出ている背景としては，以下のような点が考えられる．

- 本ホワイトペーパーが大規模な実施における「ベストプラクティス」の情報を提供するという位置づけであるため，より実際の運用上の課題に基づいた内容になっている．
- 作成された時期がテスト・スタンダードが 2007 年，ITS ガイドラインが 2005 年と古いのに対して，本ホワイトペーパーの元となっている文書が 2013 年と，比較的新しいため，CBT 利用の実績に基づいた内容が追加されている．

今後，日本において大規模な CBT に対応したガイドラインを整備するにあたり，上記 2.5～2.9 に含まれるような技術サポート，トレーニング，性能の監視，システム変更の管理などの項目も盛り込んでいくことが望ましいと思われる．

５．さいごに

本ホワイトペーパーを作成した Asia-ATP ならびに ATP は，日本での活動実績がまだ少なく，日本ではあまり認知されていないが，米国を中心に多くの団体が参加するテスト業界団体であり，テストに関する有用な情報を有している．今後，今回発表した資料の元となった Operational Best Practices の他の部分の内容や，ATP の保有する他の情報を精査し，有用な情報を発信していくことで，日本における CBT に関するガイドライン整備の一助になれれば幸いである．

＜参考文献＞

International Test Commission (ITC) (2005).
ITS Guidelines on Computer-Based and Internet Delivered Testing.
日本テスト学会（編）(2007). テスト・スタンダード．金子書房

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 155 2016/08/29 14:01:15

日本の初等中等教育における CBT 導入の意義についての考察

～ 理論、技術、運用の議論以外に何が必要なのか ～

○劉 東岳 (学研ホールディングス／学研教育総合研究所)

〈はじめに〉

ここ数年、日本では高大接続改革の枠組み内で、新しい試験制度の検討が議論されている。テストの専門家や教育の現場の人たちからは様々な慎重論・懸念が提示されているが、多くは理論、技術、運用のいずれかの観点から展開されている。

本発表では、CBT 導入の可能性に焦点を絞り、理論、技術、運用とは別の観点からアプローチする。高大接続改革の議論では新しい試験は高校生を対象にしたものに限定されるが、ここでの新しい試験が高校以前の教育にも良い影響を与えることが期待されていることは言うまでもない。そこで本発表では、風呂敷を広げて「初等中等教育に CBT 導入を検討する意義」ついて考察する。

〈CBT 概説〉

CBT とは、パソコンやモバイル機器などを使って試験を行う様々な実施形態の総称である。紙に印刷されていた問題を単純にディスプレイ上にデジタルに再現するだけではない。

・問題形式

CBT では、紙媒体の試験で扱えなかった形式の問題が出題可能である。図や写真を使うところを、CBT であればインタラクティブな 3D オブジェクトを使うことが可能である。解答の過程で解答者に能動的に 3D オブジェクトを操作させる、例えばオブジェクトを回転・拡大して情報を見つけ出す必要がある問題を出題するだけでも、紙に印刷される試験では測定できない能力を CBT では測定対象に含めることが可能になる。また、IT 技術者を対象にした資格試験で出題されるサーバー機器を設定する課題などのように、CBT では様々な実技試験をバーチャルに実施することが可能になる。

・出題ロジック

試験問題が紙に一旦印刷されてしまうと試験冊子のコンテンツを動的に操作することは難しい。しかし CBT では、出題の仕組みに様々なレベルでの柔軟性を織り込むことが可能になる。定められた条件(特に出題範囲、難易度、解答時間に関する条件)に従い、多数の試験問題を備えた試験問題プールより試験問題を受験者ごとに自動的に選定する LOFT(Linear-on-the-fly Testing)や、受験者の解答状況に応じて次に出題する試験問題を選択する CAT(Computer-Adaptive Testing)などの出題ロジックは、紙に印刷された試験冊子では実現できない CBT ならではの仕組みである。

受験者ごとにユニークな組み合わせの試験問題が出題される LOFT や CAT の仕組みは、「全ての受験者に内容が同じであることが公平性の大切な要素とする考え方」とは大きく異なる。LOFT や CAT の正当性を担保するためには、項目反応理論(Item Response Theory, IRT)の利用が不可欠となる。

・制度運用 ― 監督体制(配信モード)

インターネットに接続する環境さえあれば、自宅からでも、勤務先からでも、もしくはインターネットカフェからでも、オンラインで 24 時間受験が可能なタイプの試験から、専用の試験会場でしか受験できないタ

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 156 2016/08/29 14:01:16

イプの試験まで、実際のCBTの監督体制(配信モード)には様々なバリエーションが存在する。国際テスト委員会(International Test Commission)[1] はCBTの利用に関する国際ガイドラインを公表しているが、そこでは CBT における試験監督体制のカテゴリーとして以下の4種類(表1)を示している。

表1:国際テスト委員会によるCBTの配信モード

配信モード	説明
Open	何処でも誰でも。監督機能無し。(インターネット上で無制限に受験可)
Controlled	事前登録と ID・パスワード。監督機能無し。(インターネット上で受験を制限したテスト、実質的な本人確認無し)
Supervised	遠隔地の人による監督機能がある。(遠隔地よりインターネットを介して受験状況を監視されながら受験)
Managed	専用のテスト会場にて行われる。監督員が常駐し、受験環境が標準化され管理されている。

劉[2]は、CBT の監督体制によっては不正受験が起こりうる状況に対し、単に受験する側の意識の問題(モラルの問題)ではなく提供する側の選択の問題でもあると指摘した。試験制度の目的・状況に応じて適切な CBT の運用体制を選択することが重要で、不適切な運用体制を選択してしまうと、試験制度の意義そのものを揺るがしかねない事態となる恐れもあると注意を喚起している。

・制度運用 ― 実施日程

試験の提供頻度、すなわち受験機会の観点からも、実際の CBT の実施日程には様々なバリエーションが存在する。特定の日時にのみ提供されるイベント型と、特定期間中あるいは通年で継続的に提供されるウィンドウ型の2つのカテゴリーに分けられる。実際の CBT の運用では、受験需要や周辺業務の体制、あるいはセキュリティの観点などから実施日程を定めることになる。特に、実施日程の選択は出題ロジックの選択と大きく関わってくる。年に数回程度のイベント型であれば、丁寧に固定版を作りこむことで IRT を必要としないアプローチも検討可能だが、長期間継続的に提供する体制では LOFT などの出題ロジックを検討する必要性が出てくる。また、実際の運用では、ウィンドウ型であっても個人に対し無制限に受験機会を提供するわけではなく、多くの場合は再受験規定が必要となる。

〈高大接続改革における“新しい試験”〉

教育再生実行会議の「高等学校教育と大学教育との接続・大学入学者選抜の在り方について(第四次提言)」[3]では、高校教育の基礎的・共通的な学習の達成度を客観的に把握するためのテストと、大学教育を受けるために必要な能力判定のためのテストの 2 つの新しい試験を導入することについて言及している。いずれの試験についても、複数回受験・挑戦できるような仕組みで導入を検討するとし、後者については将来的に CBT 方式で実施することも検討するとしている。CBT の説明には、“数千～数万題の問題の中から、難易度が同じとなるよう問題を組み合わせて出題することにより、複数回受験しても安定した成績を示すことが可能となる”との記述があり、LOFT 等の CBT ならではの出題ロジックを前提にしていることがうかがえる。

文部学省中央教育審議会の答申「新しい時代にふさわしい高大接続の実現に向けた高等学校教育、大学教育、大学入学者選抜の一体改革について」[4]では、学力評価のための新たなテスト(仮称)の枠組みで、“高等学校基礎学力テスト”と“大学入試希望者学力評価テスト”の導入を宣言し、いずれの試験にも CBT の導入、そして年複数回実施が検討されるとしている。その後 1 年間に渡り高大接続システム改革会議が議論を重ね、その「最終報告」[5]で

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 157 2016/08/29 14:01:16

は、より具体的な内容が示された。改めてCBT導入の検討を進めると同時に、IRT 導入に関する記述も加えられた。

〈CBT 導入を検討する理由〉

一連の流れの中で、CBT 導入を高大接続の新しい試験に検討する理由は何だろうか。単純に、ICTの進歩・普及と教育の情報化に伴って必然的に現場から発生した需要、すなわち「これだけ学習環境のICT化が進む中で、いつまでも試験だけが紙というのは如何なものか」という話だとは思われない。実際、日本におけるCBTと利用事例は、ほとんどが高等教育以降の学生や社会人・プロフェッショナルを対象にした試験に限定されている。

・CBT ならではの問題形式

まず、英語に関しては、「CBT ならではの問題形式」を出題したいがためにCBT導入を検討していると言えるのかもしれない。言語の四技能、すなわち「聞く」、「話す」、「読む」、「書く」の能力を測定するため、今まで紙に印刷されているような試験問題では限界がある。たとえば「聞く」」と「話す」の能力測定に、イヤホンで課題を聞きマイクに向かって課題について口述するタイプの試験問題は CBT ならではであり、英語の民間資格・検定試験の多くが実際にCBTで提供されている。今のところ英語以外の教科では、「CBT ならではの問題形式」を出題したいがために CBT 導入を検討するという話には至ってないようだが、将来的に「思考力・判断力・表現力」の測定に新しいタイプの試験問題を開発される必要性を見越して、今から CBT 導入を検討するということだろうか。実際にCBT が導入され運用が始まってしまえば、教育測定の専門家たちが今までの印刷された試験問題のイメージから解き放たれ、「思考力・判断力・表現力」を測定対象に含む新しいタイプの試験問題を開発しようとする動きが後押しされる、という期待が持てるかもしれない。

・年複数回実施

高大接続システム改革会議の「最終報告」[5]では、少々トーンダウンしている印象も受けるが、複数回実施のベースには CBT を活用する発想がある。年複数回実施の構想の背景には「社会の多様性」があり、文部学省中央教育審議会の答申[4]では、繰り返して公平性(公正性)という言葉が使われている。文部科学省ではCBT導入の話とは明示的に関連した議論がされていないようだが、ここでは公平性の観点からCBT導入を検討する意義を考察する。

〈新しい公平性の視点〉

試験を測定ツールと考えれば、その妥当性と信頼性は重要な品質指標だが、試験を公益性の高い社会インフラと考えれば、そこには制度の公平性という視点からの品質も不可欠になる。

・「同じ内容」という公平性[6]

日本で長らく受け入れられている古典的テスト理論の枠組みでは、同じ問題でないと得点が比較できない。この制限があるために、「測定結果を比較される可能性のある全ての受験者には、同じ試験問題を出題する必要がある」ことになるが、これは同時に「全ての受験者にとって同じ内容という形の公平性」を提供する。そして実際の制度運用においては、「全ての受験者にとって同じ内容という公平性」の裏には「一斉受験」という受験機会を縛る制限が前提になる。

・「同等の機会」という公平性[6]

近年は、教育や職業の選択肢、キャリアのあり方や人生の優先順位なども含めて、社会の様々な側面で多様性が高まる傾向があり、その過程で「みんな同じで一緒」という類の公平性に対する風向きが変わりつつある。その公平性を担保するために今まで犠牲にしていたことが、社会の多様化が進むにつれて無視できなくなっているとも言える。初等中等教

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 158 2016/08/29 14:01:17

育における試験に限って考えても、一昔前であれば「児童・生徒は一律に同じようなタイミングで同じような試験を受ける」というのが社会常識であったのに対し、最近は、9 月入学、飛び級制度、教育機会のグローバル化などの議論が進むにつれ、また、様々な事情により登校すること・学習することが困難な児童・生徒が増加している現状を踏まえると、周囲との“シンクロ”を要求する「一斉受験」の試験体制が必ずしも望ましいと言い切れない状況がある。今まさに、他の選択肢が社会に求められている状況ではないだろうか。

どのような条件が満たされれば、我々は「一斉受験の公平性」の呪縛にとらわれないで済むのか。それは「異なるタイミング、異なる内容の試験であっても、公平性は損なわれていない」という認識を我々自身が持てる状況である。その答えの 1 つが CBT-IRT の導入であろう。IRT を適切に踏まえて CBT を活用することで、異なるタイミング、異なる試験問題で受験しても測定結果を比較できる。これは高等教育以降の試験では、すでに広く実用化されている技術である。

また、CBT-IRT の導入はリスク対策にも有効である。誤解を恐れずに言えば、試験実施を特定日に限定するのは「新しい良問を作成するのは困難である」「試験以外の業務も担当していて忙しい」といった類の試験を提供する側の論理であって、試験を利用する側が望んでいるわけではない。特定日でないと受験できないという制度では、その日に“たまたま” 体調不良や自然災害に見舞われた受験者は、“運が悪かった” とあきらめるしかなかったが、その状況は IRT や CBT のような試験技術の実用化に伴い、大幅な改善が可能になっている。

〈初等中等教育に CBT 導入するには何が必要か〉

大前提として、理論、技術、運用の議論は十分に尽くされなければならない。しかし、理論的に正しく、技術的に可能で、かつ運用面で現実的だとしても、それだけでは CBT 導入には十分ではない。教育の現場に「同じではなくても同等」だという公平感を受け入れる環境を醸成することが必要だ。「こんな仕組みで、本当に公平な試験制度なのか」と受験者・利用者に疑念を持たれては、試験としての表面的妥当性に大きな問題が残ることになる。

現実問題として、困難なのは CBT よりも IRT を受け入れる環境だろう。個人的には、初等中等教員免許更新の際に IRT について教員に学んでもらう機会を組み込むことから始められないかと思っている。

〈まとめ〉

初等中等教育における CBT 導入の意義は、測定対象の拡大の可能性と、多様性が高まる教育現場に求められる新しい公平性を支える仕組みの構築であると考えている。本発表では多様性の例として受験機会拡大の需要を取り上げた。

参考文献

[1] International Test Commission (2005). International Guidelines on Computer-based and Internet Delivered Testing. [www.intestcom.org]

[2] 劉 東岳 (2011).「試験における ICT の適切な活用方法について」, 教育システム情報学会研究報告, 26-1, pp25-28, 教育システム情報学会

[3] 教育再生実行会議 (2013).「高等学校教育と大学教育との接続・大学入学者選抜の在り方について(第四次提言)」. 2013 年 10 月 31 日

[4] 中央教育審議会 (2014).「新しい時代にふさわしい高大接続の実現に向けた高等学校教育、大学教育、大学入学者選抜の一体的改革について(答申)(中教審第 177 号)」. 2014 年 12 月 22 日

[5] 高大接続システム改革会議 (2016),「最終報告」. 2016 年 3 月 31 日

[6] 劉東岳(2014).「よく分かる CBT の仕組み」. プロメトリック株式会社テスト工学研究所. 2014 年 11 月 12 日

劉東岳 (LIU, Dongyue Louis) DL.Liu@mmf.gakken.co.jp

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 159 2016/08/29 14:01:18

上方カメラを用いたタブレット端末における
タッチジェスチャによる替え玉判別

○中村祐文*、渡邉悠二**、小方博之***
*成蹊大学大学院理工学研究科、**元成蹊大学大学院理工学研究科、***成蹊大学理工学部

1. 序論

1.1 研究背景

就職活動などでSPIやCAB、玉手箱といったWEBテストなどが自宅で受験できるようになっている。このように手軽に自宅受験ができる試験がある一方で、資格試験など重要性の高い試験には自宅受験が導入されていないのが現状である。その理由は試験時に監督者がいないため、不正行為を確認することができず、公平性・信頼性に問題点があることにある。

1.2 替え玉時の本人認証法の防止要件

本研究では、タブレット端末を利用した受験において、不正行為の一つである替え玉受験に着目する。そして替え玉の防止方法として試験中の本人認証法について検討する。本人認証を行い個人判別するためには4つの要件（継続性、透過性、同一性、耐攻撃性）を満たす必要がある[1]。

1.3 研究目的

タブレット端末上で行うフリック手形状動画データとそこから抽出した画像データから個人の特徴量を抽出する。そして得られた特徴量を分析し、本人認証を用いて個人を判別することを目的とする。

2. 先行研究

2.1 研究内容

小方らが行った実験[2]ではタッチジェスチャのうちタップを、渡邉らが行った実験[3]では横フリック、縦フリック、ズームアウトを扱った。しかし渡邉らの研究ではフリック開始点と終了点を指定していたため、被験者の動作が自然なものではなく、スライド動作に近くなってしまった。

2.2 先行研究との違い

横から撮影すると、右利きだけしか認証できないため、どちらの利き手でも認証ができる上からの撮影も新たに検討する。そのため横方向と上方向の2方向からのカメラ撮影を行った。

また、スライドに近い動きを改善するため、フリックの開始点のみを指定した。

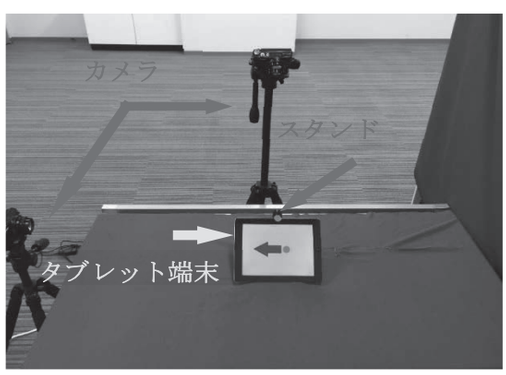
3. 実験

3.1 目的

タブレット端末操作時の横フリック、縦フリック動作を横方向、上方向の2方向からハイスピードカメラを用いて撮影を行う。ハイスピードカメラにすることで、普通のカメラよりディスプレイに触れた瞬間の画像が鮮明に取得できる。

3.2 実験環境

タブレット端末操作の横フリック、縦フリック時の手形状データを取得するために図1のような実験環境を構築した。カメラは上からと横から撮影できるように2台、机の真ん中には操作用のタブレット(iPad2:241mm×186mm×8.8mm)を設置している。



一般セッション7

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 160 2016/08/29 14:01:18

図 1　実験環境

3.3　実験方法

大学生と大学院生 34 人（男：24 人、女：10 人）を対象に実験を行った。

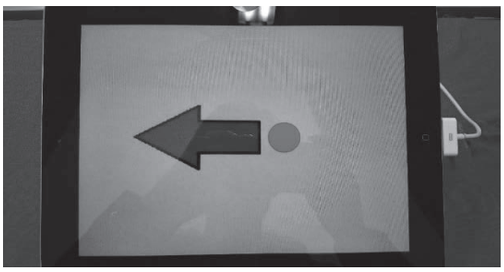


図 2　タブレット端末の実験画面

図 2 のディスプレイ上の丸いシールから矢印の方向へ横フリックをしてもらい、手を膝の上に乗せる。この操作を 10 回繰り返し行ってもらう。次に端末の向きを縦向きにし、縦フリックを先ほどと同様の操作を行ってもらう。

4.　判別方法

個人を判別するために撮影方向とフリックの種類、それぞれの特徴量を抽出した。実験で得られた動画からフリック動作の時間と、指先がディスプレイに触れた瞬間と離れた瞬間の 2 か所の静止画を抽出する。表 1 は今回抽出した個人の特徴量である。

表 1　抽出した特徴量

	横フリック	縦フリック
横撮影	・指の角度(1) ・人差し指の幅(1) ・指の開き具合(1) ・人差し指の突起部の座標(1)	・指先の距離、フリックする時間⇒速さ(2) ・人差し指の幅(1) ・指の開き具合(1) ・人差し指の角度差
上撮影	・指先の距離、フリックする時間⇒速さ(2) ・手の甲の幅(1) ・人差し指の長さ(1)	・手の面積(1) ・手の甲の幅(1) ・人差し指の長さ(1)

※(1)：触れた瞬間の画像を使用

(2)：触れた瞬間と離れた瞬間の画像を使用

得られた特徴量データを使い、k-最近傍法（k=1）、ロジスティック回帰分析、三層パーセプトロン（中間層のノード数：2）を本人認証の分析に用いた。

5.　実験結果

分析を行った結果、認証率は表 2 のようになった。①②③はそれぞれ k-最近傍法、ロジスティック回帰分析、三層パーセプトロンである。

表 2　各手法の認証率

	横フリック	縦フリック
横撮影	①　98.52% ②　97.62% ③　98.20%	①　96.41% ②　97.22% ③　97.43%
上撮影	①　97.34% ②　96.74% ③　97.01%	①　97.70% ②　97.32% ③　97.52%

6.　考察

実験結果より、それぞれ 98%前後の認証率を得ることができた。これは[3]で使っていた特徴量の他に時間や速さなど動画的特徴量を加えたからと考えられる。また横フリック、縦フリックともに撮影方向に関わらず大きな差がなかった。よって撮影方向は利き手の区別をしなくてよい上からのほうがよいと考えられる。

7.　結論

本研究はタブレット端末上で行うフリック手形状動画データとそこから抽出した画像データから個人の特徴量を抽出した。そして得られた特徴量を分析し、本人認証を用いて個人を判別することを目的とした。その結果、全体的に 98%前後の認証率が得られた。

また、カメラは横撮影と上撮影の認証率に大きな差はなかったため、上撮影用に 1 台だけカメラを用意すれば利き手にかかわらず本人認証ができることが確認できた。

8.　謝辞

本研究の一部は JSPS 科研費基盤(C)25350355 の助成を受けた。

9.　参考文献

[1] 橋本侑樹,村松大吾,小方博之「替え玉防止に向けたペンの持ち方認証法におけるなりすまし耐性の強化」電子情報通信学会論文誌, Vol. J96-A, No. 12, pp. 769-779, 2013.

[2]小方博之, 石川麻奈, 坂田智彦, 渡邉悠二, 竹囲年延, 鳥毛明「マルチタッチインタフェースにおけるタップ操作に着目した本人認証法による替え玉防止の提案(第 2 報)」日本テスト学会第 12 回大会, pp. 186-187, 2014.

[3] 渡邉悠二,小方博之,鳥海正則,中村祐文「タブレット端末上での複数種類のタッチジェスチャに着目した替え玉判別」日本テスト学会第 13 回大会, pp. 150-151, 2015

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　161　2016/08/29　14:01:19

オンライン受験における聴覚的不正行為防止を目的とした マスキング音が存在する時の集中力への影響

三浦 晋矢*、窪田 瑞穂*、○張 成龍**、小方 博之*
*成蹊大学理工学部、**成蹊大学大学院理工学研究科

1. 背景と目的

自宅受験等の監督者なしオンライン受験を公正に実施するために、筆者らはこれまで不正行為の自動検出法について検討してきた[1][2]。そのうち、聴覚的な手段を利用した不正行為に関しては、受験者の眼球運動から判別できるという結果が得られている[3]。しかし、人間の眼球運動の個人差は大きく、受験者の平常時の眼球運動特性が未知の場合、不正行為の判別精度は現状では十分ではない。そのため、筆者らは別の手段として聴覚利用型の不正行為を妨害することを考え、サウンドマスキングを利用することを提案した[4]。実際に計算課題と記憶課題に対して実験を行い、課題の実施に影響を与えないことを確認したので、報告する。

2. 聴覚利用型不正行為防止への サウンドマスキングの利用

聴覚利用型不正行為とはイヤホンなどの聴覚的な手段を悪用して解答に関連した情報を取得する行為である。これを防止するために、受験者には受験時に試験実施側が配布するカナル型イヤホンを装着してもらうこととする。また配布イヤホンのさらに内側に米粒大のイヤホンを装着するなどして、不正情報音声を取得しようという行為に対しては、配布イヤホンからマスキング音を流すことで対処することとする。

サウンドマスキングの効果については、これまで数々の研究で実証されているが[5][6]、人間がマスキング音に対して不快を感じる場合もあることが指摘されている[7]。本実験は、人間の音声を聴き取りにくくするマスキング音が、受験者に影響を与えるか否かについて確認することを目的とする。

3. 実験

3.1 マスキング音量の決定

以下の手順でマスキング音の音量決定を行った。

1) 不正情報音声になぞらえた会話音声が聞こえる最低音量、および日常会話と同等と感じる音量を測定した。会話音声としては映画「秋刀魚の味」から BGM 等の背景音のない日本語会話シーンを抽出して使用した。聴覚の健常な大学生、大学院生９名に被験者となってもらい、それらの音量の中央値を、最低音量（会話音量１）、日常会話と同等と感じる音量（会話音量３）とした。
2) 上の実験で得られた会話音量１、３、およびそれらの中間値の音量（会話音量２）の３条件で先の会話音声を流し、それらの会話内容がわからなくなるマスキング音の音量を測定した。８人の被験者に対して測定を行い、それらのマスキング音の音量の中央値を採用した。（それぞれマスキング音量１～３とする）

3.2 計算課題実験

まず、計算課題に関する実験を行った。

被験者には音楽プレーヤーに接続されたカナル型イヤホンを装着してもらい、マスキング音を聞きながら百ます計算を行ってもらう。マスキング音の音量は無音（マスキング音量０とする）およびマスキング音

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 162 2016/08/29 14:01:20

量１～３の４条件とした。健常な大学生、大学院生48名に被験者になってもらい、４条件すべてで課題を行ってもらった。４条件の順序の組合せは２４通りあるが、それぞれの組合せに２名ずつ割り当てた。

3.3 記憶課題実験

次に、記憶課題に関する実験を行った。

被験者には計算課題の場合と同様にマスキング音を聞きながら、パソコンのディスプレイに表示された15個の単語を記憶してもらう。各単語の表示時間は２秒であり、単語間に１秒間のインターバルが存在する。その後、被験者には暗記した単語を書き出してもらう。計算課題の場合と同様に、健常な大学生、大学院生48名に被験者になってもらい、マスキング音量０～３の４条件で実験を行った。

4. 実験結果

計算課題実験における平均解答時間と標準偏差をTable 1に示す。

Table 1　計算課題実験における解答時間の平均と標準偏差

マスキング音量番号	平均[sec]	標準偏差[sec]
0	95.5	20.1
1	97.2	21.7
2	95.9	17.7
3	96.8	21.5

マスキング音量に関して１元配置分散分析を行ったところ、統計的な有意差は見られなかった($F=0.61$, $p=0.61$)。

次に、記憶課題実験における平均解答数と標準偏差をTable 2に示す。

Table 2　記憶課題実験における解答数の平均と標準偏差

マスキング音量番号	平均[個]	標準偏差[個]
0	6.92	4.94
1	6.92	4.50
2	7.17	4.27
3	6.96	4.21

マスキング音量に関して１元配置分散分析を行ったところ、これについても統計的な有意差は見られなかった($F=2.65$, $p=0.93$)。

これらの結果より、いずれの課題についてもマスキング音の影響はないと考えられる。

5. まとめ

聴覚利用型不正行為を妨害するために、サウンドマスキングの利用を検討した。人間の音声を聴き取りにくくするマスキング音を計算課題、記憶課題を遂行している被験者に聞かせる実験を行い、影響がないことを確認した。

謝辞 成蹊大学理工学部 岩本宏之准教授にはマスキング音量決定に関する実験に協力していただくとともに、有用なアドバイスをいただきました。また、本研究の一部はJSPS科研費基盤(C) 25350355の助成を受けました。ここに感謝します。

参考文献

[1] 橋本、村松、小方「替え玉防止に向けたペン持ち方認証法におけるなりすまし耐性の強化」電子情報通信学会論文誌, vol. J96-A, no. 12, 2013.

[2] 小方、石川他「マルチタッチインタフェースにおけるタップ操作に着目した本人認証法による替え玉防止の提案(第２報)」テスト学会12回大会, 2014.

[3] 八木、小方、向山「眼球運動特徴を利用した試験時の二重課題遂行判別」テスト学会13回大会, 2015

[4] 小方、三浦「聴覚的不正行為防止および騒音低減へのサウンドマスキングの利用の検討」テスト学会13回大会, 2015.

[5] 宮島「位置情報を活用したサウンドマスキングシステムの開発」清水建設研究報告 90, 2013.

[6] 三戸、荒井、安「サウンドマスキングシステムにおけるデータベースを用いた音声マスカ作成法の提案」日本音響学会誌 vol. 71, no. 8, 2015.

[7] 茂出木、橋本他「スピーチプライバシー保護機能をもたせた電子ポスター」情報処理学会 インタラクション2011.

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　163　2016/08/29　14:01:21

能力検査受検にスマートフォンを使用することの影響とその要因

○園田 友樹，藤田 彩子
株式会社リクルートキャリア測定技術研究所

1. 研究の背景と目的

総務省の平成26年通信利用動向調査(2015)によれば，20代のインターネット利用の87.5%はスマートフォン(以下SMP)によるものが占めている．株式会社リクルートキャリアの調査(2016)によれば，大学生の94.8%がSMPを保有しており，就職活動をする学生にとってもSMPがメインデバイスとなっている．現在，Web上で受検するアセスメントやテストは採用場面において広く利用されている．SMPやタブレットなどの普及に伴い，受検にPCではなくSMPやタブレットが用いられる場合も増えると考える．

一方で，SMPでの受検，とりわけ能力検査の受検においていくつかの課題も指摘されている．Kingら（2014）の研究によると，Situational Judgmentなど，テストによっては，PC受検とモバイルツール受検の結果が同等とみなせるが，能力検査についてはモバイルツールが不利となる．LaPort（2016）の研究によれば，能力検査では問題の画面表示のされ方が得点に影響する事が確認されている．また，採用場面での利用を想定した時，受検者側の納得感も重要である．受検環境によっては受検者の納得感を損ね，不公平感を抱かせる可能性がある．

SMP などモバイルツールによるアセスメント受検については，海外を中心にここ数年実証的な研究が行われているが，国内での研究はほとんど見当たらない．本研究では，国内の能力検査について，PC で受検した場合と SMP で受検した場合でどのような違いが生じるか検証した．具体的には，デバイスの違いが問題の難易度にもたらす影響とその要因，および受検者からみた回答しやすさ・しにくさの程度とその要因について検証した．

2. 手続き

2-1. 調査項目

2-1-1.能力検査　本研究では，㈱リクルートキャリア(旧人事測定研究所)が開発した，基礎能力検査 GAT (General Ability Test)の Web テスト 1 版を用いた．これは， PC での受検を想定して開発された Web 配信のテストである．GAT は，言語尺度と非言語尺度の二つの下位尺度で構成されている．言語については，「語彙」「語句の並べ替え」「文章の論理展開や論旨の把握」などの下位領域で構成される．非言語については，「定価・速度の計算などの数量的情報の取り扱い」「表の計算」「確率・組合せ」「論理・推論」などの下位領域で構成される．

2-1-2.「回答のしやすさ」に関するアンケート　能力検査回答後，テストの回答のしやすさについて問うために「お使いの端末での回答のしやすさはいかがでしたか？」という質問に対し，「1:とてもやりやすかった」から「5:とてもやりにくかった」の 5 段階で回答してもらった．「4:ややりにくかった」「5:とてもやりにくかった」と回答した対象者からは，回答しにくさの具体的内容について自由記述で回答を得た．

2-2.調査の設計および対象者

一般セッション7

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　164　2016/08/29　14:01:22

今回は回答負荷を考慮し，言語検査（36 問，制限時間 30 分）を回答する群と非言語検査（22 問，制限時間 45 分）を回答する群の 2 つに対象者を分けてインターネット調査により回答を得た．実施形態は，①PC，②PC 画面をそのまま表示させた SMP(以下 SMP[PC])，③SMP 用に調整した画面を表示させた SMP(以下 SMP[調整])の 3 つで，対象者にはランダムに①～③の実施形態を割り振った．同一人物に複数の実施形態での回答を依頼しなかったのは，同じ内容の問題を複数回受検することによる練習効果を排除する為である．

2-3. データのスクリーニングとサンプリング

まず，得られた合計 2,799 名のデータから，ほとんど同じ回答を入力している等の明らかに回答傾向の偏った 196 名を除外した．実施形態をランダムに割り振ったが，PC 群と SMP[PC]・SMP[調整]群との間に，性別・年齢・学歴について割合が多少異なっていた．分散分析の結果，言語検査で学歴に，非言語検査では性別・年齢・学歴のすべてについて得点に有意な差が見られた為，PC 群についてランダムサンプリングを行い，PC 群の各属性の割合が SMP[PC]・SMP[調整]群の割合と同程度になるよう調整した．その結果，実際に分析に用いたデータの①～③の各群の属性ごと人数の内訳は【図表1】のようになった．

3. 結果と考察

3-1. 項目特性値に及ぼす影響

言語尺度，非言語尺度それぞれにつき，項目特性値(特に困難度パラメータ b)の比較を行った．項目特性値の算出には，R のパッケージ(difR)を使用した．

困難度パラメータ b の値を比較すると，言語の問題，非言語の問題ともに，平均的には PC で受検した場合が最も低く，次いで SMP[PC]，SMP[調整]となった(【図表 2】参照)． 問題が同じでも実施形態によって困難度が変化するといえる．その程度は平均すると困難度パラメータの 0.3 前後の違いで，ハイステークスなテストの場合は，無視できない違いと思われる．個々の問題を見ていくと，言語・非言語ともに，問題の下位領域による特徴的な傾向は見られず，各問題で異なる傾向が見られた．【図表 3】は．横軸が PC 受検での困難度，縦軸に SMP 受検での困難度(●は SMP[PC]，○は SMP[調整])をプロットしたものである． PC と SMP で困難度の大きな差があった問題数題(【図表 3】「非言語」のグラフ中で A，B，C の項目)について，困難度差が生じた要因について考察した．

《項目 A》 SMP[調整]は PC と同程度の困難度だが，SMP[PC]では困難度がかなり上がっている．問題画面を確認したところ，「－(マイナス)」の記号が SMP[PC]で見づらく，これを見落としたことによる不正解が多くなったことが考えられる(実際，マイナスがない場合に導かれる値を答えている受検者が他と比べて非常に多かった)．

《項目 B》 SMP[PC]は PC よりやや困難度が低いが，SMP[調整]

【図表 1】各テストの人数集計

●言語検査

【性別】

	男性	女性
PC	118	275
SMP[PC]	118	275
SMP[調整]	106	247

【年齢】

	22～24歳	25～29歳	30～34歳
PC	30	148	215
SMP[PC]	38	153	202
SMP[調整]	34	130	189

【学歴】

	大卒	大学院卒
PC	354	39
SMP[PC]	346	47
SMP[調整]	323	30

●非言語検査

【性別】

	男性	女性
PC	93	218
SMP[PC]	117	281
SMP[調整]	122	276

【年齢】

	22～24歳	25～29歳	30～34歳
PC	19	119	173
SMP[PC]	34	155	209
SMP[調整]	36	145	217

【学歴】

	大卒	大学院卒
PC	264	47
SMP[PC]	348	50
SMP[調整]	357	41

【図表 2】実施形態ごと項目特性値の平均値

		識別力	困難度
言語尺度	PC	1.23	−0.91
	SMP [PC]	1.12	−0.73
	SMP [調整]	1.24	−0.66
非言語尺度	PC	1.65	0.22
	SMP [PC]	1.60	0.35
	SMP [調整]	1.70	0.58

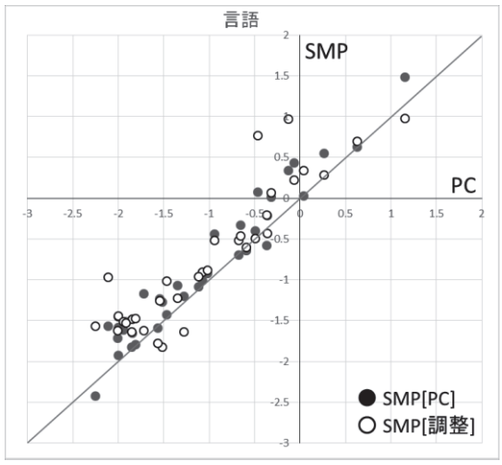
230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 165 2016/08/29 14:01:22

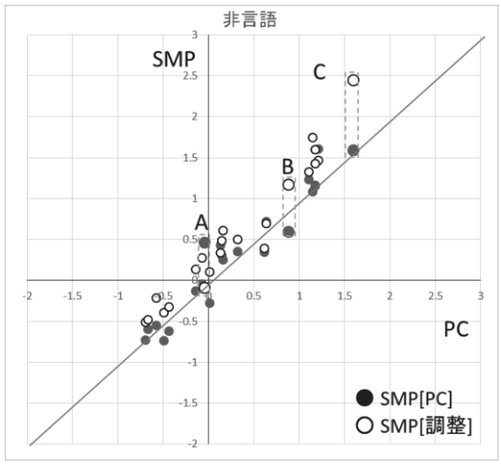
では困難度が高くなっている．この問題は，表を見て問いに答える 2 問組の問題の 2 問目で，SMP[調整]では表と問題を同時に見ることができず，スクロールしながら解かなければならない．

《項目 C》SMP[PC]は PC とほぼ等しい困難度だが，SMP[調整]では困難度が高くなっている．この問題は，分数を答えさせる問題で，PC と SMP[PC]は，回答欄の□/□が 1 行に収まっているが，SMP[調整]では「/」の後で改行され，一見分かりにくい回答欄となっていた．

単に SMP に適応した画面にすることが，必ずしも問題の難易度の低下にはつながっていないことが示唆される．A，B，C 以外の問題についても，SMP[調整]で SMP[PC]より困難度が高くなるものが多く見られ，改行位置やフォントなど，細かいレイアウトの影響は思った以上に大きいのかもしれない．

【図表 3】PC/SMP 実施の各項目の困難度の違い





3-2. 回答のしやすさに及ぼす影響

「回答のしやすさ」について実施形態による違いがあるか，言語・非言語それぞれについて分散分析を行った．その結果，言語(F(2,1136)=11.794, p<.001)，非言語(F(2,1104)=16.754, p<.001)共に実施形態による有意差があった．Tukey を用いた多重比較によれば，言語，非言語のいずれのテストについても回答のしやすさは①PC と③SMP[調整]は同程度で，②SMP[PC]が有意に回答しにくい結果となった．

回答しにくさの具体的内容を捉えるために，自由記述の回答の形態素解析を行った．形態素解析には RMeCab を用いた．全形態素の中から名詞・動詞・形容詞に該当するものについて出現頻度を計量し，上位 10 までの形態素を確認した．結果は【図表4】および【図表 5】の通りであった．言語問題では，PC での回答については問題そのものの文字量や読みにくさについて指摘するもの(「長い文章問題が読みにくい」等)が中心であった．一方で，SMP[PC]では文字の小ささに加え，画面サイズ(「画面が小さくイライラする」等)や選択肢ボタン(「ボタンが小さくて打ちづらい」等)についての言及が特徴的であった．【図表 4】からも②SMP[PC]に特徴的な言葉として，「拡大」「ボタン」が挙げられる．SMP[調整]についても同様に画面が小さいことによる文字の読みにくさについてのコメントが多く見られた．【図表 4】において，②SMP[PC]，③SMP[調整]にはあるが①PC にはない言葉として，「画面」「見にくい」「字」があり，また，トップに共通して「小さい」があるように，画面を SMP 用にしても，やはり見づらさを感じさせていることがうかがえる．

非言語問題では，PC での回答については，問題の難しさや計算の煩わしさについてのコメントが多い(【図表 5】においても，①PC 特有の言葉として「難しい」「わかる」「時間」などがある)．一方，SMP[PC]での回答では「拡大したりスクロールしたりという作業が手間」といった画面操作に関するものや文字入力についてのコメントが多く

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 166 2016/08/29 14:01:23

見られた. SMP[調整]についても文字の小ささ，文字入力に関するコメントが見られている.【図表 5】では，②SMP[PC]，③SMP[調整]だけに共通する言葉として，「画面」「小さい」「文字」「入力」「しまう」「なる」などがあり，やはり画面が小さい，入力がしにくい，という印象を持つ受検者が多いと思われる. また，SMP[PC]，SMP[調整]に共通して，「計算しているうちに画面が消えた」「すぐにロックがかかる」といった，問題を解く間の画面制御に関するコメントが見られた.

【図表 4】言語検査についての形態素解析結果

①PC		②SMP [PC]		③SMP [調整]	
形態素	順位	形態素	順位	形態素	順位
読む	1	小さい	1	小さい	1
文字	1	する	2	文字	2
問題	3	文字	3	する	3
文章	4	読む	4	問題	4
にくい	5	拡大	4	読む	5
小さい	6	にくい	6	画面	5
する	7	画面	6	にくい	7
づらい	8	見にくい	8	文章	8
いる	9	ところ	8	ある	9
多い	10	ボタン	8	見にくい	10
ある	10	字	8	字	10

アンケート回答の結果からは，PC と SMP[調整]において，回答しやすさに有意な差がなかったことから，SMP での受検が必ずしも回答しにくい印象を与えるわけではないが，SMP ならではの回答に与える影響が明らかになった. SMP の画面上での文字認識・回答操作の困難さ，問題文面の量に起因するスクロール等の操作を要すること回答への負荷を高めることが想定され，画面上での問題表示には十分な配慮が必要である. また，一定時間を経過すると画面のロックがかかるなどの問題もあり，開始前のインストラクションなど含めてきめ細やかな対応が必要である. 加えて，実施形態の違いは項目特性値ひいてはテスト得点に，ある程度の影響を与えると考えられ，実施形態として PC と SMP の両方を想定する場合は，実施形態の影響の大きい項目の発見・排除，また，実施形態の影響が小さい問題の作成などが求められる.

【図表 5】非言語検査についての形態素解析結果

①PC		②SMP [PC]		③SMP [調整]	
形態素	順位	形態素	順位	形態素	順位
問題	1	する	1	画面	1
計算	2	小さい	2	する	1
する	3	画面	3	入力	3
難しい	4	文字	4	小さい	4
時間	5	入力	5	文字	5
わかる	6	にくい	6	数字	6
ところ	7	なる	7	問題	6
にくい	8	問題	8	なる	6
ない	9	計算	9	にくい	9
づらい	10	しまう	9	ところ	10
				しまう	10

4. 今後に向けて

本研究では，SMP での能力検査受検における課題の一部が明らかになったが，どのような問題項目で実施形態による影響が大きくなるか等については，はっきりしないことも多かった. 今後に向けては，実施形態に影響を受けやすい問題の特徴を明らかにしていくとともに，SMP での受検に適した画面設計やインストラクションの在り方，回線や画面制御など想定される他の影響についても検証して行きたい.

5. 参考文献

Kate LaPort (2016) Mobile Assessment Comparing Traditional Cognitive, Cognitive Reasoning, and Non-Cognitive Performance. Mobile Equivalence: Expanding Research across Assessment Methods, Levels and Devices. Symposium presented at the 31th Annual Conference of the Society for Industrial and Organizational Psychology, Anaheim, CA.

株式会社リクルートキャリア (2016) 大学生の実態調査 2016

King, D. D., Ryan A. M., Kantrowitz, T., & Grelle, D. (2014). MIT Versus PCIT: Assessing equivalence, individual differences, and reactions. In T. Kantrowitz & C. M. Reddock (Chairs), Shaping the future of mobile assessment: Research and practice update. Symposium presented at the 29th Annual Conference of the Society for Industrial and Organizational Psychology, Honolulu, HI.

総務省 (2015) 平成 26 年通信利用動向調査

豊田秀樹(2013)『項目反応理論[中級編]』p78-p89 朝倉書店

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 167 2016/08/29 14:01:24

日本テスト学会第 14 回大会実行委員会

実行委員長	植野 真臣	（電気通信大学）
副実行委員長	宇都 雅輝	（電気通信大学）
大会実行委員（五十音順）	石井 隆稔	（東京理科大学）
	大森 拓哉	（多摩大学）
	加藤 嘉浩	（株式会社ベネッセホールディングス）
	橋本 貴充	（大学入試センター）
	宮澤 芳光	（東京学芸大学）
	山本 美紀	（電気通信大学）
	若山 昇	（帝京大学）

日本テスト学会第 14 回大会発表抄録集

2016 年 9 月 8 日発行

日本テスト学会第 14 回大会実行委員会

〒182-8585 東京都調布市調布ヶ丘 1-5-1
国立大学法人 電気通信大学 内

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd 168 2016/08/29 14:01:24

■理想の大学入試への改革と実現をめざして



新しい時代の大学入試
The University Entrance Examination in a New Era

繁桝算男 編著　　A5判・218頁　本体3,200円＋税

大学全入時代が到来した現代の日本において、各大学は
◆どのような受験生を合格させ、
◆受験生は自分に適した大学をどう選び、
◆また、大学は他大学との差異化をどう図るべきか。

大学入試の歴史と展望、論考、テスト理論や、各国の大学入試事情、世界的に著名な米国の大学入試制度の調査等を紹介。大学全入時代の現代において、多様な選抜システムによる理想の入試実現に焦点を当てる。

■テスト取り扱いの共通理解の指針と行動



テスト・スタンダード
日本のテストの将来に向けて

日本テスト学会 編　　A5判　240頁　本体4,000円＋税

テスト取り扱いの共通理解の指針と行動を示す

翻訳ではなく、日本独自のテスト規準である。テストの作成、実施、採点、結果の利用、管理など一連のテスト過程におけるあるべき姿と、日本の社会状況に沿った規準の提示である。

主な内容
- Ⅰ　基本条項
- Ⅱ　基本条項の解説
  - 0章　序
  - 1章　開発と領布
  - 2章　実施と採点
  - 3章　結果の利用
  - 4章　記録と保存
  - 5章　コンピュータを利用したテスト
  - 6章　テスト関係者の責任と倫理
- Ⅲ　テストの使われ方Q&A

■テストの点数っていったい何なのでしょう。

見直そう，テストを支える基本の技術と教育

日本テスト学会 編　　A5判　96頁　本体1,100円＋税

■適切な設計・実施・結果の有効運用のために

全国学力調査 日米比較研究

荒井克弘・倉元直樹 編著　　A5判　260頁　本体4,000円＋税

〒112-0012 東京都文京区大塚3-3-7　金子書房　☎03（3941）0111　FAX03（3941）0163
ホームページ　http://www.kanekoshobo.co.jp

北大路書房

〒603-8303
京都市北区紫野十二坊町12-8
☎075-431-0361　FAX 075-431-9393
http://www.kitaohji.com
振替 01050-4-2083

研究論文を読み解くための **多変量解析入門 基礎編**
ー重回帰分析からメタ分析までー　L．G．グリム・P．R．ヤーノルド編　小杉考司監訳　A5・356頁・本体3600円＋税　統計の初心者が高度な分析法を用いた研究論文を読み解くことができるように，概念的側面に特化して多変量解析を解説。できる限り数式を用いずに，パス解析・多次元尺度構成法・多変量分散分析・判別分析などを取り上げる。

研究論文を読み解くための **多変量解析入門 応用篇【仮題】**
ー2016年秋　発刊予定！ー　L．G．グリム・P．R．ヤーノルド編　小杉考司監訳　A5・約450頁・予価4400円＋税　統計の初心者が高度な分析法を用いた研究論文を読み解くことができるように，概念的側面に特化して多変量解析を解説。応用篇では，一般化可能性理論・項目反応理論・クラスター分析・正準相関分析・反復測定分析などを取り上げる。

Rによる心理学研究法入門
山田剛史編著　A5・272頁・本体2700円＋税　「心理学研究モデル論文集」「具体例に即した心理学研究入門書」「統計ソフトRの分析事例編」の3つの顔を持つテキスト。実際の研究例をもとに，研究法の基礎の紹介，研究計画立案のための背景や目的，具体的なデータ収集の手続き，Rでのデータ分析，研究のまとめやコメントなどで詳しく紹介。

M-plusとRによる **構造方程式モデリング入門**
小杉考司・清水裕士編著　A5・332頁・本体2500円＋税　M-plusは安価でありながらも，非常に高度な統計処理を簡単なコーディングで行うことができる。無償のフリーソフトRは，強力な統計のプラットフォームとして浸透している。本書は，両ソフトの操作およびコマンド記述の違いを比較しつつ，SEMの分析の実際を解説する。

改訂エンサイクロペディア　心理学研究方法論
W．J．レイ著／岡田圭二編訳　本体5000円＋税

心理学基礎実習マニュアル
宮谷真人・坂田省吾代表編集　本体2800円＋税

心理学教育のための傑作工夫集
L．T．ベンジャミン・ジュニア編／中澤　潤他監訳　本体2800円＋税

わかって楽しい心理統計法入門　Ver.2
松田文子・三宅幹子・橋本優花里著　本体2500円＋税

増補改訂　SPSSのススメ1
竹原卓真著　本体3200円＋税

改訂新版　初めての心理学英語論文
D．シュワーブ・B．シュワーブ・高橋雅治著　本体1900円＋税

現代の認知心理学1　知覚と感性
日本認知心理学会監修／三浦佳世編　本体3600円＋税

現代の認知心理学2　記憶と日常
日本認知心理学会監修／太田信夫・厳島行雄編　本体3600円＋税

現代の認知心理学3　思考と言語
日本認知心理学会監修／楠見　孝編　本体3600円＋税

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　169　2016/08/29　14:01:25

これからの教育に
フィットする。

日本データパシフィックの「e-Learning solution」

U-Assist

e-Learningコンテンツシリーズの総称。学校教育における、自学自習、授業補完、資格取得などの目的に合わせて利用できる教材を多数揃えました。

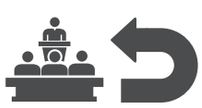


豊富なeラーニングコンテンツ



eポートフォリオ

プラグインを追加する事で、教職履修カルテや課外活動を記録できます。紙の成果物を取り込んで、学生生活全体をカバーするポートフォリオの構築も可能です。



学生生活を振り返る

学習記録ビューア

日々の授業で蓄積されたログや成果物をコースをまたいで見える化します。担当学生の活動の様子を確認しながら適切な指導を行う事が可能です。



学習記録の見える化

ラーニング・カンパニーでありたい。

日本データパシフィック株式会社

www.datapacific.co.jp

〒186-0002 東京都国立市東1-4-15 国立KTビル
TEL：042-573-6721　FAX：042-573-6728

230934_多摩大学_日本テスト学会発表論文抄録集_本文.indd　170　2016/08/29　14:01:26