外国語教育メディア学会 (LET) 関西支部 メソドロジー研究部会 第 10 号報告論集
住 政二郎・工藤 多恵・乗次 章子・山脇 野枝 (pp. 41–48)

アイテムバンクとAutomated Test Makerの開発と応用

住 政二郎 ・ 工藤 多恵 ・ 乗次 章子 ・ 山脇 野枝
関西学院大学

概要

関西学院大学では，スーパーグローバル大学創成支援事業の採択を受け，全学規模の英語教育に関する多くの施策が計画されている。2017 年度からは，全学規模の統一プレイスメントテスト，習熟度別クラス編成，および到達度テストが導入される。こうした変化は，英語教育の成果を学部横断的に検証可能にするのみならず，縦断的に英語力の推移を観察可能にする。しかし，こうした変化にはデメリットもある。大きな問題は学部独自の教育内容と各種統一テストの測定内容との齟齬である。各学部は専門性を活かした英語教育を行っている。著者らの所属する理工学部もその 1 つである。この問題を解決するために，理工学部では，到達度テストとしての利用が予定されている TOEIC® IP テストを念頭に，約 2,000 問のアイテムバンクを開発した。本稿では，開発したアイテムバンクの一部を古典的テスト理論と項目反応理論で分析し，アイテムバンクを授業で活用するために開発した *Automated Test Maker* について述べる。

Keywords: アイテムバンク，テスト項目分析，*Automated Test Maker*

1. はじめに

関西学院大学は，2014年9月，スーパーグローバル大学創成支援の採択を受けた。その後，学内では，英語教育に関するさまざまな施策が準備されている（関西学院，2016）。2017年度からは，全学規模のプレイスメントテスト，習熟度別クラス編成，および到達度テストが実施される。こうした変化には歓迎すべき点も多い。到達すべき目標を共有し，授業を行い，測定結果をもとに授業改善をする横断的な下地を全学規模で提供する。また，測定結果を経年的に蓄積し，英語教育の成果を縦断的に比較検討することも可能になる。しかし，各種の統一テストを導入することを不安視する声も多い。大きな問題は学部独自の英語教育の内容との齟齬である。また，測定結果をどのように評価するのか，という点について統一された見解がなく，測定結果のみが英語教育の直接的な結果として取り扱われる可能性もある。吉田（2009）は，多様な外部テストの活用が大学英語教育に混乱をもたらしていることを指摘し，テストが「何を」，「どのように」測定するものなのか，テスト利用者の理解が重要であることを指摘している。

著者らの所属する理工学部では，こうした問題にいち早く対応するために，到達度テストとして導入が予定されている TOEIC® IP テスト（以下，TOEIC）を念頭に，約 2,000 問からなる TOEIC 語彙・多肢選択問題のアイテムバンクを開発した。また，アイテムバンクを授業で活用していく *Automated Test Maker*（以下，ATM）も開発した。これは，データベースに保存されたアイテムバンクから自動的にテストを出力するシステムである。本稿では，開発したアイテムバンクの一部を古典的テスト理論と項目反応理論で分析し，ATM の活用について述べる。

2. アイテムバンクの開発

アイテムバンクの開発は，河上 源一 （2011）. 『TOEICテストに でる順英単語』に集録されているPart 2（501）からPart 5（2535）の単語を対象に，TOEIC Part 5の空欄補充・多肢選択問題のテスト形式に準拠して行った。選択肢は4つに統一した。理工学部では，すべての学生に，河上 （2011）の購入を義務づけ，1・2年生の英語科目の副教材として利用している。アイテムバンクの問題文および錯乱肢には，大学英語教育学会が定める基本語リスト *JACET 8000* の内，Level 4 までの語彙を使うことに配慮した。

3. 分析

3.1 材料

開発したアイテムバンクから，河上（2011）の Part 3（1001–1515）に対応する多肢選択問題から 100 問を抽出してテストを作成した。Part 3 に集録されている約 500 単語は，理工学部の学生が 1 年生後期に学習する必須単語になっている。理工学部の 1 年生の内，344 名が，後期開始時期，2016 年 9 月にテストを受験した。本稿では，その結果に分析を加えた。

3.2 方法

テスト結果には，古典的テスト理論と項目反応理論を使って分析を加えた。古典的テスト理論を使った分析については，島谷他（1999），前田（2003）を参考にした。問題項目の分析は，中村洋一（2002）.『テストで言語能力は測れるか—言語テストデータ分析入門』を参考にした。同書に同封されている分析ソフト *TDAP ver. 2.0* は，Windows XP 以降，すでに更新が終了していたため，同書の著者および関係者の許諾を得て，*R* で計算可能にして公開した（参照：https://rpubs.com/seisumi/ctt）。古典的テスト理論で分析を加えた各指標は，以下のとおりである。

DIFF: Item difficulty index（項目困難度）
DISC: Discrimination power index（項目弁別力指数）

AENO: Actual equivalent number of options（実質選択肢数）

ADIF: Appropriateness of difficulty（項目困難度適切度）
ADIS: Appropriateness of discrimination power index（項目弁別力適切度）
AAEN: Appropriateness of actual equivalent number of options（実質選択肢数適切度）

SADIF: Standard appropriateness of difficulty（標準項目困難度適切度）
SADIS: Standard appropriateness of discrimination power index（標準項目弁別力適切度）
SAAEN: Standard appropriateness of actual equivalent number of options（標準実質選択肢数適切度）

SATOT: Standard appropriateness total（標準適切度合計）

DIFF は，問題項目の正答率を表す。DISC は，受験者全体を対象に点双列相関係数を用いた項目弁別力を表す。AENO は，問題項目に対して実際にいくつの選択肢が機能したかを表す。ADIF，ADIS，AAEN は，前述した DIFF，DISC，AENO の適切性に関する指標である。ADIF，ADIS，AAEN を比較検討しやすくするために標準化したものが，SADIF，SADIS，SAAEN である。そして，その 3 つの指標を合計したものが SATOT である。各指標の詳細は，中村（2002）を参照にして頂きたい。

項目反応理論を使った問題項目の分析には *R* を使った。分析モデルには，パラメーターの取り扱いやすさと，モデルのデータへの適合性ではなく，反応データのモデルへの適合性を検討する本稿の趣旨を加味してラッシュモデルを選択した（靜, 2007, p. 301）。

2. 結果

2.1 基礎統計量

分析の結果，基本統計量は以下のとおりである（表 1）。正解は 1 問 1 点とし，100 点を満点とした。結果より，極端な片寄りのない得点分布であることが分かる。

表 1
テスト得点の基本統計量

平均	標準偏差	最小	最大	歪度	尖度	クロンバック α	標準誤差
52.44	14.95	22	88	-0.03	-0.85	0.91[0.90, 0.92]	0.82

2.2 古典的テスト理論と項目反応理論による分析

古典的テスト理論に基づく各指標は，テストの目的によって評価が異なり，各指標間のバランスも考慮する必要がある。例えば，DIFF に関しては，集団基準準拠テストの 4 択多肢選択問題の場合，最適困難度は 0.65 となる。DISC に関しては，一概に評価することはできないが，一般的には 0.300 以上が適正値と考えられている（中村, 2002, p. 88）。AENO に関しては，4 択の場合，2.500 以上が適正値と考えられている（中村，2002, p. 92）。

古典的テスト理論とラッシュモデルの結果の解釈の違いにも留意する必要がある。例

えば，古典的テスト理論では，DISC の値の高い項目は，弁別力の高い項目として評価されるが，ラッシュモデルの枠組みでは，モデルの期待よりも弁別力が高すぎるため不要な項目として判断される（靜, 2007, p. 361）。表 2 は，古典的テスト理論による分析結果から DISC を基準に降順に並べ替え，上位 5 項目を抜粋したものである。

表 2
古典的テスト理論による分析結果（抜粋）

ID	DIFF	DISC	AENO	ADIF	ADIS	AAEN	SADIF	SADIS	SAAEN	SATOT
70	0.668	0.533	2.718	0.915	0.397	0.998	0.597	0.742	0.584	1.924
95	0.629	0.531	2.905	0.993	0.393	0.998	0.629	0.738	0.584	1.952
56	0.620	0.521	2.928	0.990	0.373	0.989	0.628	0.719	0.565	1.913
86	0.557	0.521	3.210	0.864	0.373	0.990	0.576	0.719	0.567	1.863
62	0.680	0.512	2.584	0.891	0.355	0.953	0.587	0.703	0.496	1.787

表 3 は，ラッシュモデルによる適合度分析の結果に Wilson-Hilferty 変換を加え，アウトフィット *t* とインフィット *t* を付記し，インフィット *t* を基準に昇順にデータを並べ替え，下位５項目を抜粋したものである。

表 3
ラッシュモデルによる適合度分析結果（抜粋）

ID	Outfit MSQ	Infit MSQ	Outfit *t*	Infit *t*
86	0.829	0.859	-3.91	-3.90
95	0.821	0.843	-3.31	-3.78
56	0.813	0.855	-3.57	-3.55
70	0.789	0.838	-3.4	-3.5
62	0.805	0.852	-2.99	-3.06

興味深いことは，古典的テスト理論では高い弁別力として評価される項目 56, 62, 70, 86, 95 が（表 2），インフィット *t* の一般的な適正値（-2.00〜+2.00）の観点からすると，オーバーフィットの項目になる点である。尚，これらの項目は，アウトフィットとインフィットの平方平均の適正値（0.7〜1.3）（靜, 2007, p. 317）の観点からすれば不適合にはならない。また，一般的に，オーバーフィットの項目は，テストには余分な存在ではあるが，有害な存在ではないと考えられる（靜, 2007, p. 317）。

一方，テストにとって有害なのは，アンダーフィットの項目である。表 4 は，適合度分析の結果をインフィット *t* を基準に降順に並べ替え，上位５項目を抜粋したものである。

表 4

ラッシュモデルによる適合度分析結果（抜粋）

ID	Outfit MSQ	Infit MSQ	Outfit *t*	Infit *t*
16	1.222	1.199	4.30	4.82
7	1.261	1.197	4.74	4.62
94	1.288	1.192	3.95	3.68
35	1.288	1.180	4.04	3.52
100	1.166	1.146	3.09	3.47

これらの項目も，平方平均の解釈では，一概に不適合の項目とは言えないが，インフィット *t* の解釈ではアンダーフィットの項目になる。表 5 は，表 4 の結果を踏まえ，インフィット *t* で目立ってアンダーフィットとなった項目の古典的テスト理論による分析結果の一覧である。

表 5

古典的テスト理論による分析結果（抜粋）

ID	DIFF	DISC	AENO	ADIF	ADIS	AAEN	SADIF	SADIS	SAAEN	SATOT
7	0.404	0.072	3.520	0.558	0.005	0.907	0.450	0.370	0.408	1.228
16	0.575	0.075	2.815	0.900	0.006	0.842	0.591	0.371	0.282	1.244
35	0.332	0.067	3.689	0.415	0.005	0.917	0.390	0.369	0.427	1.186
94	0.326	0.048	3.924	0.403	0.002	0.994	0.386	0.367	0.576	1.328
100	0.401	0.146	3.633	0.552	0.022	0.945	0.447	0.386	0.481	1.314

表 5 から分かることは，DIFF や AENO の指標には目立った特徴はないものの，DISC の指標が各問題とも極端に低いことが分かる。このことは，問題項目への解答のランダム性が高すぎ，弁別力がモデルの期待値よりも極端に低いことを意味している。

図 1 は，各問題項目の困難度パラメーターとインフィット *t* との相対的な関係を示している。図 1 より，前述した項目以外にも改善を加える必要のある項目が存在することが分かる。ただし，適合度分析の結果の解釈は，サンプルサイズによっても変化する。また，平方平均の値だけではなく，ごく少数の受験者の得点の影響を検討するために残差を確認する必要もある。いずれにしろ適合度分析に絶対的なルールは存在せず，テストの目的を考慮しながら，古典的テスト理論による結果と項目反応理論の結果を対応させながら，問題項目の改善を図る必要がある。

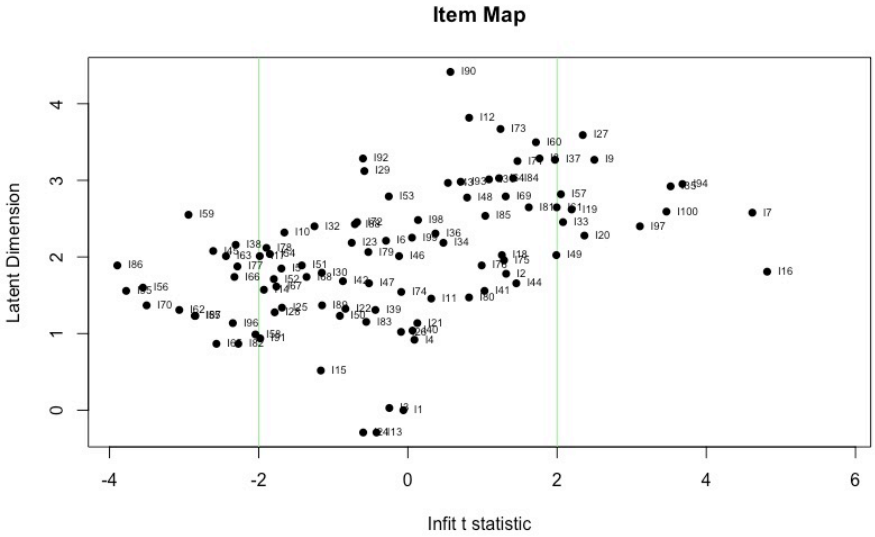


図 1. インフィット *t* の分布

4. Automated Test Maker

アイテムバンクを授業で活用するために，ATM（http://atm.lang-tech.net）を開発した。これは，サーバーに保存されたアイテムバンクから問題集，問題番号，出題 Part などを指定して（図 2），自動的に問題と解答の PDF を生成するシステムである（図 3）。アイテムバンクの編集機能も実装しており，改善内容をすぐに反映できるようになっている。尚，システム開発には，Version 2（http://ver2.jp）の大西昭夫氏の協力を頂いた。

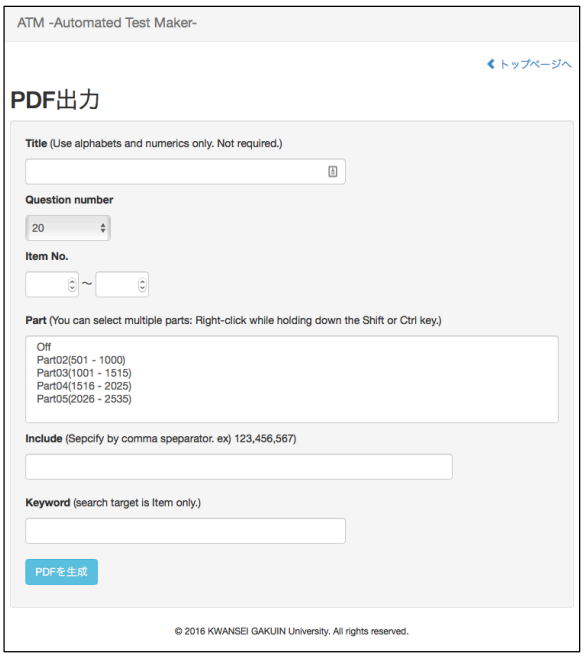
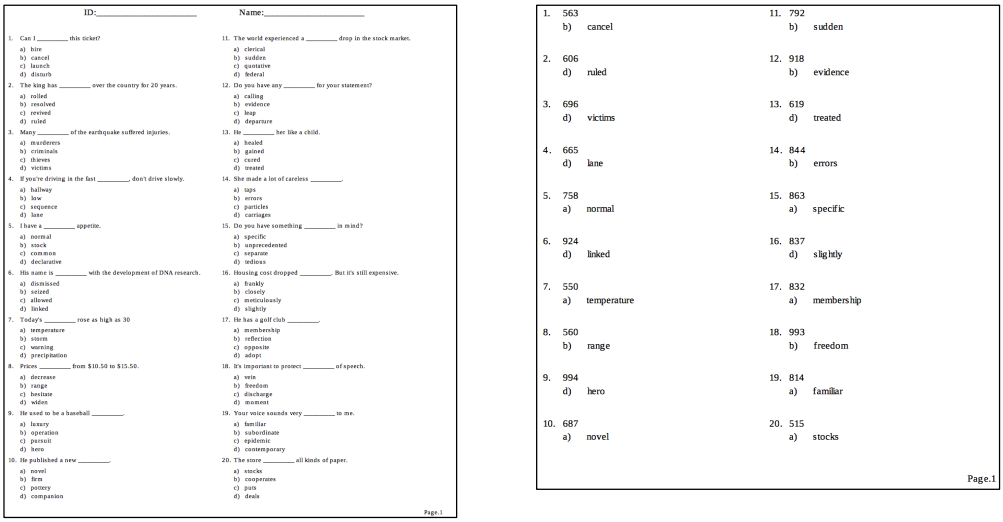


図 2. ATM の設定画面



ID:_____ Name:_____

1. Can I _____ this ticket?
a) hire
b) cancel
c) launch
d) disturb

2. The king has _____ over the country for 20 years.
a) rolled
b) resolved
c) revived
d) ruled

3. Many _____ of the earthquake suffered injuries.
a) murderers
b) criminals
c) thieves
d) victims

4. If you're driving in the fast _____, don't drive slowly.
a) hallway
b) low
c) sequence
d) lane

5. I have a _____ appetite.
a) normal
b) stock
c) common
d) declarative

6. His name is _____ with the development of DNA research.
a) dismissed
b) seized
c) allowed
d) linked

7. Today's _____ rose as high as 30
a) temperature
b) storm
c) warning
d) precipitation

8. Prices _____ from $10.50 to $15.50.
a) decrease
b) range
c) hesitate
d) widen

9. He used to be a baseball _____.
a) luxury
b) operation
c) pursuit
d) hero

10. He published a new _____.
a) novel
b) firm
c) pottery
d) companion

11. The world experienced a _____ drop in the stock market.
a) clerical
b) sudden
c) quotative
d) federal

12. Do you have any _____ for your statement?
a) calling
b) evidence
c) leap
d) departure

13. He _____ her like a child.
a) healed
b) gained
c) cured
d) treated

14. She made a lot of careless _____.
a) taps
b) errors
c) particles
d) carriages

15. Do you have something _____ in mind?
a) specific
b) unprecedented
c) separate
d) tedious

16. Housing cost dropped _____. But it's still expensive.
a) frankly
b) closely
c) meticulously
d) slightly

17. He has a golf club _____.
a) membership
b) reflection
c) opposite
d) adopt

18. It's important to protect _____ of speech.
a) vein
b) freedom
c) discharge
d) moment

19. Your voice sounds very _____ to me.
a) familiar
b) subordinate
c) epidemic
d) contemporary

20. The store _____ all kinds of paper.
a) stocks
b) cooperates
c) puts
d) deals

Page.1

1. 563 b) cancel
2. 606 d) ruled
3. 696 d) victims
4. 665 d) lane
5. 758 a) normal
6. 924 d) linked
7. 550 a) temperature
8. 560 b) range
9. 994 d) hero
10. 687 a) novel
11. 792 b) sudden
12. 918 b) evidence
13. 619 d) treated
14. 844 b) errors
15. 863 a) specific
16. 837 d) slightly
17. 832 a) membership
18. 993 b) freedom
19. 814 a) familiar
20. 515 a) stocks

Page.1

図 3. 出力された問題文と解答例

現在，ATM に集録されているのは TOEIC 形式の多肢選択問題のみであるが，今後は，指定教科書に準拠した読解問題，リスニング問題等を段階的に集録していく予定である。また，項目困難度を入力し，等化されたテストを生成できるようにする予定である。こうすることで同一科目で複数クラスを担当する場合にも，異なる定期テストが作成できるようになる。最終的には，改善が加えられた問題項目を整備し，到達度を揃え，客観的な指標で英語教育を検証できる体制を整備する予定である。

4. まとめ

本稿では，開発したアイテムバンクの一部に古典的テスト理論と項目反応理論を使い分析を加えた。分析の結果，未だ多くの改善の必要な項目があることが明らかになった。今後，それらの項目に対して，担当者間での協議を重ね改善を加えていく。また，問題項目の種類を増やし，ATM の内容を充実させ，客観的な指標で英語教育を検証できる体制を整備していく。

謝辞

中村洋一先生，大友賢二先生，秋山實様から，*TDAP* の *R* への書き直しとデータの 2 次利用の許諾を頂いたことに深く感謝いたします。

参考文献

関西学院 (2016).『中期計画の取組み 2016』Retrieved from http://www.kwansei.ac.jp/kikaku/kikaku_009760.html

前田啓朗 (2003).「到達目標型教育に向けた英語テストの改善―古典的反応理論と項目反応理論に基づいて」『広島外国語教育研究』, *6,* 131–140, Retrieved from

http://ir.lib.hiroshima-u.ac.jp/ja/list/HU_journals/AA11424332/--/6/item/15390

中村洋一 (著)・大友賢二 (監修) (2002).『テストで言語能力は測れるか—言語テストデータ分析入門』桐原書店.

島谷浩・木下正義・Terry Laskowski・髙梨芳郎・大津敦司・川尻徳 (1999).「日本と韓国の高校生に対する英語リスニング・テストに見られたテスト・バイアス：古典的テスト理論と項目応答理論に基づくデータ分析」『外国語教育評価学会研究紀要』*2,* 35–54, Retrieved from http://ci.nii.ac.jp/naid/110009610287

靜哲人 (2007).『基礎から深く理解するラッシュモデル−項目応答理論とは似て非なる測定のパラダイム』関西大学出版.

吉田弘子 (2009).「英語プレイスメントテスト分析−言語テストの観点から」『大阪経大論集』, *60*(2), 93–103, Retrieved from http://www.i-repository.net/il/user_contents/02/G0000031Repository/repository/keidaironshu_060_002_093-103.pdf