TOEIC 語彙学習のためのアイテムバンクの開発と実践への応用

住　　政二郎（理工学部）
工　藤　多　恵（理工学部）
山　田　一　美（理工学部）

要　旨

スーパーグローバル大学創成支援事業に採択されて以降、学内ではさまざまな施策が計画されている。英語教育に関しては、外部テストを使ったプレイスメントテストと到達度テストがある。こうした変化は、多様化する学生の英語力に対応するために必要なものであろう。しかし、全学規模の統一テストの導入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測定内容との齟齬である。到達度テストとして TOEIC® IP テスト（以下、TOEIC）の導入が予定されているが、各学部は英語教育の到達目標を独自に設定しており、TOEIC のスコア向上だけを掲げている訳ではない。理工学部では、こうした状況にいち早く対応するために、理系に特化した英語教育を行う一方で、TOEIC 語彙学習のためのアイテムバンクを開発した。アイテムバンクとは問題項目のデータベースのことを指す。また、アイテムバンクを活用した授業デザインを検討するために比較検証を行った。分析の結果、実験群に顕著な傾向が確認された。本稿では、特にアイテムバンクの開発および分析過程について論述する。また、本稿の最後には、教育目標・内容に沿ったアイテムバンクを学部独自で保有することの有意性と将来性についても論述する。

1. はじめに

スーパーグローバル大学創成支援事業に採択されて以降、学内の英語教育にも変化が見られる。主要なものとしては外部テストを活用した全学規模でのプレイスメントテストと到達度テストの実施計画がある（関西学院，2016，p. 6）。こうした変化は、多様化する学生の英語力に対応するために必要なものであろう。何よりも客観的な測定・評価と英語教育とが連環した仕組みと文化を全学に根づかせる重要な一歩となり得ることが期待される。

しかし、全学規模の統一テストの導入にはデメリットもある。１つは、学部独自の教育内容と測定内容との齟齬である。もう１つは、統一テストの結果のみが取り上げられ、学部独自の英語教育への理解を妨げる要因にもなり得ることである。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。

加えて、外部テストの問題項目は基本的に非公開である。このためテストの結果から受験者の

課題を具体的に把握することは困難である。本来であれば、プレイスメントテストで学生の弱点を把握し、習熟度別に英語力を伸ばす教育を行い、到達度テストで定点的に測定し、成果の確認と英語教育の反省材料にする一貫した仕組みが求められる。大河内・山中（2016）は、数学プレイスメントテストを独自に開発し、その結果によって初年次に成績不振となる学生の約50％を絞り込めることを明らかにし、困難を抱える学生の早期発見と教育サポートの連動性の大切さを指摘している。このことは、近藤ブラウン（2012）の教育と測定・評価は相互作用の関係にあり、測定・評価の役割を考慮せず、効果的な授業やカリキュラムは検討することができないという主張とも一致する。

プレイスメントテストや TOEIC を用いた到達度テストの導入など、社会的要請を含んだ全学レベルの英語教育と、専門性を踏まえた学部レベルの英語教育を両立させることは容易ではない。この両者を同時に実現するためには、まず両者の違いに目を向ける必要がある。言語テストには、大別して集団基準準拠テスト（norm–referenced test，以下，NRT）と目標基準準拠テスト（criterion–referenced test，以下，CRT）とがある。NRT は、一般的には熟達度テストとも呼ばれる各種の外部テストを指す（Brown, 2005）。本学の場合では、プレイスメントテストや到達度テストに利用する各種の外部テストが該当する。一方、CRT は、実際に科目を担当する担当者が、科目の到達目標に準じて作成するものである。本学の場合では、学部独自のテストがこれに該当する。一旦はテストの違いを把握し、次に重要なことは、各学部の特色ある英語教育の測定・評価指標となる CRT を客観的な指標で整備していくことであろう。こうすることで NRT との相対的な関連性を検討することが可能になる。また、NRT と CRT とが相互に関連することで、個人・学部・大学に統一且つ一貫した基準を定めながら、教育と測定・評価の好循環を生み出していくことができる。理工学部の英語担当者間では、上記の認識のもと、中・長期的な視点から英語教育を改善していくために、その第一歩として TOEIC 語彙学習のためのアイテムバンクの開発と実践への応用に着手した。

2. アイテムバンクの開発

アイテムバンクの開発は、担当者間でテストの仕様を協議することから開始した。その結果、１年生を対象に授業内外で TOEIC の語彙学習を支援する環境整備の必要性が明らかになり（工藤・住・山田，2017）、語彙学習に特化した問題項目を開発することになった。開発素材には、河上（2011）『TOEIC にでる順英単語』の Part 2 の英単語500個を選んだ。河上（2011）は、TOEIC に出題される単語を頻度順に並び替え、例文を添えた単語帳である。理工学部では、１・２年生の必修英語科目であるリーディングⅠ・Ⅱの副教材として利用しており、学期毎に単位修得の要件の一部として単語テストを課している。Part 2 の単語500個は１年生春学期に割り当てられている。

アイテムバンクの問題形式は、TOEIC Part 5 を参考に４択からなる多肢選択問題にした。正答以外の選択肢や、問題文を自作する際には、大学英語教育学会が定める基本語リスト JACET8000の内、Level 4（大学受験、大学一般教養初級）までの語彙を使うことにした。開発した多肢選択問題500問の問題文および選択肢のレベル構成は表１のとおりであり、Level 4 までの語彙が全体の77.40％を占めることが分かる。

表 1　JACET8000に準じたアイテムバンクの語彙レベル構成比

	level 1	level 2	level 3	level 4	level 5	level 6	level 7	level 8	over 8	others
%	30.49	21.28	13.74	11.89	6.04	3.52	2.16	1.72	6.04	3.13

CRT の場合、言語テストを開発する際には、構成概念的妥当性、内容的妥当性、結果的妥当性を検討する必要がある（近藤ブラウン，2012，p. 22）。今回の場合はテスト利用の目的がTOEIC の語彙学習に特化したものであり、内容的な偏りはあるが、今後、アイテムバンクを拡張していく段階で解消させることを予定している。具体的な問題項目については、工藤・住・山田（2017）を参考にして頂きたい。

3. 事前・事後テストと授業実践

500問からなるアイテムバンクを開発した後に、5 問刻みに問題項目を抽出し100問のテストを作成した。このテストを使って理工学部 1 年生509名を対象に 4 月に事前テストを、7 月には事後テストを行った。事前テストと事後テストは同じものである。事前テストの後に答え合わせは行わず、問題用紙は、事前および事後テストの後に回収した。

事前テストと事後テストの間には、アイテムバンクを活用した授業デザインを検討するために実験群（$n = 241$）と統制群（$n = 268$）を設け異なる実践を行った。実験群では授業と LUNA を活用して、授業内外での TOEIC 語彙学習環境を提供したが、統制群には提供せず、従来どおりの授業を行った。2 群の比較は事後テストの結果を用いて行った。授業実践と比較分析に関する詳細は、工藤・住・山田（2017）を参考にして頂きたい。本稿では、特にアイテムバンクの開発および分析過程について論述する。

4. 分析手法

問題項目の分析は、事前・事後テストで利用した100問のテストを対象に、古典的テスト理論の観点から正答率（item facility，以下，IF）、差異指数（difference index，以下，DI）、錯乱肢有効度（distractor efficiency，以下，DE）の検討を行った。その後、問題項目の一次元性と局所独立を確認し、項目反応理論を用いて問題項目の適合度分析を行った。最後に、参考としてベイズ推定を用いて統制群と実験群の事前・事後の差の比較を行った。

正答率（IF）は項目困難度（item difficulty）とも呼ばれ、文字通り各項目の正答率を表す。NRT に準拠した場合、適正値は .30〜.70（Brown, 2005, p. 76）であるが、到達度を測る CRT の場合はこの限りではない。差異指数（DI）は、特に CRT の項目分析や授業実践の評価の指標として重要な意味を持つ。これは、事前と事後テストの正答率の差で計算することができ、問題項目毎の学習者の熟達度を表す。錯乱肢有効度（DE）は、各選択肢が学習者によって選ばれた比率を表す。その値から各選択肢がどの程度有効に機能したかを検討することができる。

正答率（IF）、差異指数（DT）、錯乱肢有効度（DE）の指標は、特定の受験者集団のテスト得点に準じて算出される古典的テスト理論の一部である。古典的テスト理論は、受験者集団やテストの特徴を簡便な手法で把握するのに有用ではあるが、集団依存性と項目依存性の限界点も持つ。集団依存性とは、テストが英語力の高い学習者にとっては簡単になり、英語力の低い学習者

にとっては難しくなる特性である。項目依存性とは、簡単なテスト項目で構成されたテストでは平均点が上がり、逆に、難しいテスト項目で構成されたテストでは平均点が下がる特性である。古典的テスト理論では、テストの得点を他のテストと比較したり、素点に基づき経年的に学習者の英語力の変化を観察したりすることができない。

この問題を解決したのが現代テスト理論の項目反応理論（item response theory）である。項目反応理論の大きな特徴は、学習者の能力と項目の困難度を別々に且つ同一尺度上で推定できることである（加藤・山田・川端，2015，p. 4）。結果、項目反応理論は、受験者に依存しない測定、項目に依存しない測定を実現し、同一尺度を用いたテスト間の比較、問題項目の交換、学習者の経年的変化の観察などを可能にした。

項目反応理論には取り扱うパラメータの数によって複数の種類がある（住，2014）。本稿では、適合度分析には1パラメータ・ロジスティックモデルに分類されることもあるラッシュモデルを利用した（住，2013）。理由は、扱われるパラメータの数が少なく、結果の解釈が容易であることと、他のモデルよりも適用条件が厳しく、アイテムバンク開発の初期の段階では有益なフィードバックが得られると考えたからである。学習者 j が問題項目 i に正解する確率（P_{ji}）をラッシュモデルで表現すると、以下のとおりになる。θ は、学習者の能力パラメータ、δ は項目困難度パラメータを表す。

$$P_{ji}=\frac{\exp(\theta_j-\delta_i)}{1+\exp(\theta_j-\delta_i)}$$

観測応答データに項目反応理論を適応するためには、データがいくつかの前提条件を満たす必要がある。代表的なものに一次元性（unidimensionality）と局所独立性（local dependence）がある。一次元性とは、テストが単一の構成概念のみを測定することを指す。例えば、今回の場合、開発されたテストが単一の TOEIC 語彙力（θ）のみを測定していることを意味する。一次元性を確認するためにはいくつかの方法がある。代表的なものには、四分相関係数行列を求め、その固有値を計算し、固有値を降順に可視化した固有値プロットから判断する方法がある（加藤・山田・川端，2015，p. 140）。局所独立とは、各問題項目が互いに独立していることを指す。つまり、ある問題項目に正解したことが、他の問題項目の解答に影響を与えないことを意味する。局所独立の評価には、Yen（1984，1993）の Q_3 統計量が使われることが多い。Q_3 の値が .20を越える問題項目に関しては、検討を加える必要があると指摘されている。実際の観測応答データで厳密に一次元性や局所独立が成り立つことは考えにくいが、分析を進める事前に確認し、その結果を踏まえて問題項目の改善を図る必要はある（加藤・山田・川端，2015，p. 154）。

統制群と実験群の事前・事後テストの差の比較には、対応のある・ない2群の t 検定に代えてベイズ推定を行った。ベイズ推定は、近年、頻度主義に基づく検定に代わって注目を集めている。アメリカ統計学会は、p 値を用いた有意差検定に警鐘を鳴らすと共に、「ポスト $p<0.5$時代」を既に宣言している（ASA, 2016）。

5. 結果

5.1 正答率・差異指数・錯乱肢有効度

表２は、正答率、差異指数、錯乱肢有効度の一覧である。事後テストには、実験群と統制群の結果が混在するため、錯乱肢有効度は、事前テストの結果のみを利用した。選択肢の数に対応する錯乱肢有効度は、１～４の数字で表中に示した。表２では、便宜上、１列を正答に定めた。１列の値は正答率と一致する。pre-IF は事前テストの正答率、post-IF は事後テストの正答率、DI は post-IF から pre-IF を引いた差異指数を表す。分析の過程で、複数の選択肢が項目88で正答になり得ることが分かり、項目88の結果は除外した。

表２より、項目13の正答率は95.28％で、問題項目としては簡単すぎることが分かる。逆に項目18と77は正答率が15％以下であり、問題項目として何らかの歪みがあるか、難しすぎることが分かる。これらの問題項目は改善が必要であり、本来であれば、ラッシュモデルでの分析の前に省いてしまった方が良いものであるが、適合度との関連性を確認するために残すことにした。

錯乱肢有効度は、正答率の高い問題で偏りが生じ、正答以外の選択肢が有効に機能していないことが分かった。例えば、項目４，11，13，14，21，25，63，49には、選択肢の中に錯乱肢有効度が１ケタ以下のものがある。これらの項目と選択肢は早急な見直しが必要である。

興味深いのは差異指数である。項目77は、事前テストで最も低い正答率13.56％であった。以下のような問題である。以下の（　）の中の数字は，事前テスト時の錯乱肢有効度である。

77. Let me take a _____ of you.
　　a）shot　　（13.56％）
　　b）final　　（4.91％）
　　c）rest　　（68.57％）
　　d）jury　　（12.97％）

事前テスト後に項目77の正答率の低さを担当者間で協議した。その結果、多くの学生が c）rest（68.57％）を選択していることから、take a rest という表現に解答が影響を受けていることが分かった。そこで、正解は take a shot of～「～の写真を撮る、～を撮影する」という表現であるという指導を授業でするように実験群の担当教員に依頼をした。指導のなかった統制群（n＝268）の項目77の事後テストでの正答率は16.04％（43人）であった。統制群の事前テストの項目77の正答率は、12.31％（33人）であることから（DI＝3.73％）、統制群には事前・事後テストでほとんど変化がないことが分かる。一方、指導のあった実験群（n＝241）の項目77の事後テストの正答率は、47.30％（114人）であった。事前テストでは、14.94％（36人）であったことから（DI＝32.36％）、事前テストで明らかになった弱点を明示的に指導し、語彙学習環境を授業内外で提供することによって、到達度に有益な変化が見られることが明らかになった。

表２　正答率・差異指数・錯乱肢有効度一覧

Item	1	2	3	4	pre-IF	post-IF	DI	Item	1	2	3	4	pre-IF	post-IF	DI
1	65.23%	9.82%	2.75%	22.20%	65.23%	66.01%	0.79%	26	82.71%	3.34%	6.09%	7.86%	82.71%	88.41%	5.70%
2	71.51%	6.68%	7.27%	14.54%	71.51%	84.28%	12.77%	27	50.10%	20.43%	17.29%	12.18%	50.10%	60.90%	10.81%
3	77.41%	3.73%	12.57%	6.29%	77.41%	81.34%	3.93%	28	87.82%	5.11%	3.73%	3.34%	87.82%	91.55%	3.73%
4	83.10%	14.34%	0.79%	1.77%	83.10%	83.10%	0.00%	29	49.31%	15.72%	23.18%	11.79%	49.31%	62.28%	12.97%
5	78.59%	3.14%	8.84%	9.23%	78.59%	71.51%	-7.07%	30	66.40%	13.75%	14.93%	4.91%	66.40%	81.73%	15.32%
6	56.78%	19.84%	18.07%	5.30%	56.78%	58.35%	1.57%	31	44.79%	22.99%	10.61%	21.66%	44.79%	51.47%	6.68%
7	57.76%	8.64%	31.04%	2.55%	57.76%	67.58%	9.82%	32	44.40%	36.54%	14.34%	4.72%	44.40%	59.33%	14.93%
8	31.63%	14.15%	13.16%	40.86%	31.63%	38.51%	6.88%	33	41.06%	18.27%	41.06%	30.45%	41.06%	47.35%	6.29%
9	36.74%	5.89%	34.97%	22.40%	36.74%	36.94%	0.20%	34	80.55%	5.30%	8.25%	5.89%	80.55%	87.82%	7.27%
10	89.39%	3.14%	4.91%	2.55%	89.39%	89.59%	0.20%	35	34.18%	21.61%	20.63%	23.58%	34.18%	51.67%	17.49%
11	80.35%	11.39%	6.48%	1.77%	80.35%	86.25%	5.89%	36	85.27%	7.86%	4.52%	2.36%	85.27%	91.55%	6.29%
12	58.55%	11.00%	16.70%	13.36%	58.55%	67.39%	8.84%	37	35.56%	26.13%	20.83%	17.49%	35.56%	46.95%	11.39%
13	95.28%	3.54%	0.20%	0.98%	95.28%	96.46%	1.18%	38	57.56%	16.70%	17.29%	8.45%	57.56%	61.49%	3.93%
14	84.68%	6.88%	6.48%	1.96%	84.68%	84.87%	0.20%	39	61.69%	19.84%	11.20%	7.27%	61.69%	72.30%	10.61%
15	24.17%	14.73%	36.74%	24.36%	24.17%	25.54%	1.38%	40	71.32%	14.73%	6.68%	7.27%	71.32%	83.50%	12.18%
16	73.28%	7.27%	13.16%	6.29%	73.28%	78.59%	5.30%	41	57.96%	5.70%	14.73%	21.61%	57.96%	82.12%	24.17%
17	71.51%	7.07%	8.64%	12.77%	71.51%	78.00%	6.48%	42	78.39%	5.70%	9.04%	6.88%	78.39%	82.71%	4.32%
18	14.54%	24.75%	35.76%	24.95%	14.54%	18.47%	3.93%	43	73.67%	7.27%	8.25%	10.81%	73.67%	81.34%	7.66%
19	48.72%	25.34%	14.93%	11.00%	48.72%	52.85%	4.13%	44	35.95%	20.83%	26.52%	16.70%	35.95%	42.04%	6.09%
20	45.78%	37.72%	6.48%	9.82%	45.78%	53.05%	7.27%	45	54.03%	10.41%	20.04%	15.52%	54.03%	64.44%	10.41%
21	91.55%	0.79%	6.09%	1.57%	91.55%	93.71%	2.16%	46	59.92%	12.38%	16.90%	10.61%	59.92%	68.76%	8.84%
22	52.06%	31.63%	10.22%	6.09%	52.06%	57.96%	5.89%	47	37.72%	24.56%	17.88%	19.84%	37.72%	48.13%	10.41%
23	40.47%	8.06%	35.36%	16.11%	40.47%	40.28%	-0.20%	48	42.04%	23.97%	5.11%	28.88%	42.04%	50.49%	8.45%
24	84.87%	4.52%	5.50%	5.11%	84.87%	87.23%	2.36%	49	87.03%	1.96%	7.66%	3.34%	87.03%	91.36%	4.32%
25	89.39%	3.14%	2.55%	4.91%	89.39%	91.16%	1.77%	50	50.10%	20.43%	24.36%	5.11%	50.10%	60.90%	10.81%

Note. Item は問題項目番号を表す。１～４の数字は選択肢を表す。１列は正答の錯乱肢有効度で正答率と一致する。pre-IF は事前テストの正答率、post-IF は事後テストの正答率、DI は差異指数を表す。

表 2 （続） 正答率・差異指数・錯乱肢有効度一覧

Item	1	2	3	4	pre-IF	post-IF	DI	Item	1	2	3	4	pre-IF	post-IF	DI
51	82.12%	5.30%	7.07%	5.50%	82.12%	86.64%	4.52%	76	66.80%	7.47%	17.68%	8.06%	66.80%	73.28%	6.48%
52	58.35%	14.15%	10.61%	16.90%	58.35%	78.00%	19.65%	77	13.56%	4.91%	68.57%	12.97%	13.56%	30.84%	17.29%
53	81.93%	6.29%	2.95%	8.84%	81.93%	87.43%	5.50%	78	48.53%	12.77%	19.25%	19.45%	48.53%	64.83%	16.31%
54	42.44%	10.41%	19.45%	27.50%	42.44%	46.95%	4.52%	79	86.84%	2.75%	5.89%	4.52%	86.84%	89.98%	3.14%
55	66.60%	66.60%	15.13%	7.27%	66.60%	77.41%	10.81%	80	44.01%	14.93%	15.32%	25.74%	44.01%	51.47%	7.47%
56	88.21%	3.14%	4.52%	4.13%	88.21%	90.18%	1.96%	81	80.94%	4.13%	11.00%	3.93%	80.94%	82.91%	1.96%
57	71.91%	4.91%	8.25%	14.93%	71.91%	75.05%	3.14%	82	63.46%	17.88%	8.06%	10.61%	63.46%	70.73%	7.27%
58	50.49%	19.84%	9.82%	19.84%	50.49%	55.01%	4.52%	83	55.99%	18.27%	12.38%	13.36%	55.99%	70.33%	14.34%
59	83.30%	5.70%	6.68%	4.32%	83.30%	88.41%	5.11%	84	83.69%	5.30%	7.66%	3.34%	83.69%	88.41%	4.72%
60	74.07%	8.06%	11.98%	5.89%	74.07%	79.57%	5.50%	85	74.46%	12.57%	6.48%	6.48%	74.46%	80.55%	6.09%
61	26.13%	19.84%	17.68%	36.35%	26.13%	41.45%	15.32%	86	50.29%	16.70%	17.09%	15.72%	50.29%	70.92%	20.63%
62	71.12%	13.36%	10.61%	4.72%	71.12%	75.05%	3.93%	87	22.20%	11.00%	45.58%	21.22%	22.20%	49.51%	27.31%
63	91.16%	1.77%	3.54%	3.54%	91.16%	89.19%	-1.96%	89	63.06%	15.91%	12.18%	8.84%	63.06%	74.66%	11.59%
64	86.05%	6.48%	2.95%	4.52%	86.05%	86.25%	0.20%	90	84.87%	4.52%	6.09%	4.52%	84.87%	90.18%	5.30%
65	59.14%	17.29%	14.34%	9.23%	59.14%	67.19%	8.06%	91	55.80%	17.68%	13.75%	12.77%	55.80%	66.80%	11.00%
66	59.14%	13.36%	8.25%	19.25%	59.14%	72.30%	13.16%	92	59.92%	3.54%	21.22%	15.32%	59.92%	66.21%	6.29%
67	46.76%	19.45%	24.17%	9.23%	46.76%	50.29%	3.54%	93	36.15%	36.15%	15.52%	12.18%	36.15%	47.74%	11.59%
68	36.15%	27.31%	12.38%	24.17%	36.15%	41.65%	5.50%	94	62.48%	17.88%	11.79%	7.86%	62.48%	67.39%	4.91%
69	90.18%	3.54%	3.93%	2.36%	90.18%	91.16%	0.98%	95	32.81%	20.04%	23.77%	23.38%	32.81%	42.44%	9.63%
70	48.92%	12.18%	20.04%	18.86%	48.92%	63.26%	14.34%	96	23.18%	40.47%	24.95%	11.00%	23.18%	30.06%	6.88%
71	64.44%	10.81%	15.72%	8.84%	64.44%	79.57%	15.13%	97	38.11%	19.84%	14.54%	27.11%	38.11%	53.63%	15.52%
72	41.06%	20.83%	30.65%	7.07%	41.06%	55.60%	14.54%	98	82.71%	4.72%	4.72%	7.47%	82.71%	88.02%	5.30%
73	54.42%	12.77%	18.27%	14.54%	54.42%	69.74%	15.32%	99	66.80%	10.02%	6.88%	15.91%	66.80%	81.73%	14.93%
74	63.46%	6.29%	13.16%	17.09%	63.46%	73.48%	10.02%	100	62.08%	6.68%	8.25%	22.59%	62.08%	70.53%	8.45%
75	26.92%	15.52%	22.79%	34.77%	26.92%	48.53%	21.61%								

5.2　一次元性と局所独立

一次元性および局所独立の確認は、（加藤・山田・川端，2015，pp.138-154）を参照に行った。分析は、項目88を除いた全99問で行った。

事前テストの結果から四分相関係数を計算し、固有値を算出した *R*（ver. 3.3.1）の出力画面は、以下のとおりである。図１は、固有値をプロットしたものである。

```
eigen(polychoric.cor$rho)$values # 固有値の計算
 [1] 17.93  3.14  2.67  2.44  2.24  2.21  2.14  2.11  2.08  1.92  1.90  1.85  1.84
[14]  1.76  1.73  1.68  1.64  1.61  1.56  1.50  1.48  1.44  1.42  1.36  1.32  1.28
[27]  1.27  1.26  1.21  1.18  1.18  1.14  1.08  1.06  1.04  1.02  0.99  0.96  0.92
[40]  0.91  0.88  0.86  0.85  0.80  0.79  0.78  0.77  0.73  0.71  0.68  0.67  0.65
[53]  0.63  0.61  0.60  0.57  0.56  0.53  0.52  0.49  0.47  0.45  0.44  0.41  0.38
[66]  0.37  0.35  0.32  0.31  0.28  0.26  0.25  0.23  0.20  0.19  0.18  0.16  0.15
[79]  0.12  0.09  0.07  0.06  0.05  0.03  0.00  0.00  0.00  0.00  0.00  0.00  0.00
[92]  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
```

図１より、勾配がなだらかになるのは第２固有値以降で、ほぼ一次元性は確保されていることが分かった。しかし、古典的テスト理論での分析からも明らかになったように、改善が必要な問題項目が混在していることから、第２固有値以降を丁寧に観察すると完全な一次元性にはなっていないことも分かった。ラッシュモデルで適合度分析を行い、さらに改善が必要な問題項目を絞り込む必要性が明らかになった。

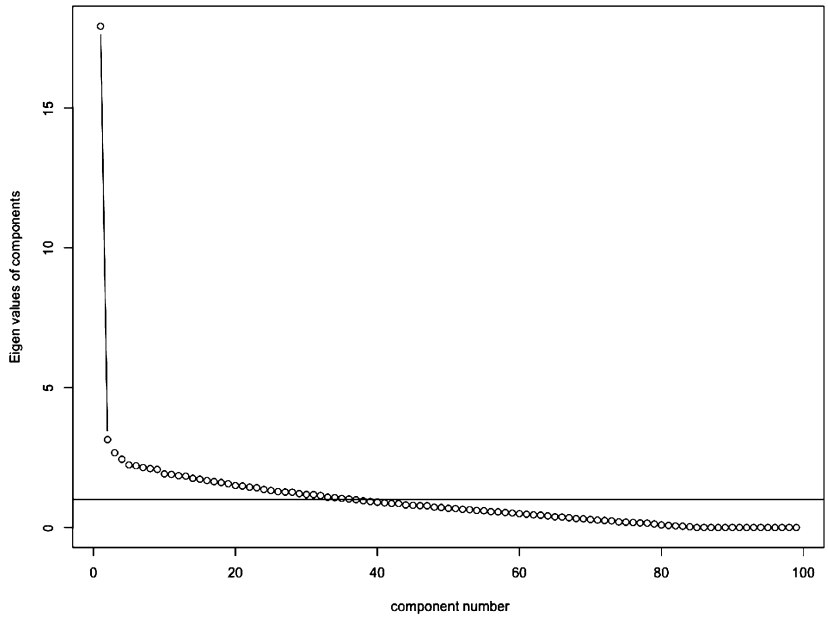

図１　固有値プロット

事前テストの結果を使って局所独立を確認した結果、.30を越える項目ペアが複数確認された。特に項目16（73.28）、26（82.71）、43（73.67）、90（84.87）には、目立って .30を越える項目

ペアが確認された。（　）の内の数字は、事前テストにおける正答率を表す。そして、これらの問題項目には興味深い共通点があることも分かった。それは、語彙の難度が低く、文法もシンプルで正答率が高い、ということである。例えば、項目16と項目42の Q_3 値は、.402であった。両項目を比較すると共に主語を補うシンプルな問題形式で、解答には共通する文法的知識が使われたことが考えられる。また、項目16の正答は equipment で、項目42の正答は impression で、共に JACET8000の語彙レベルは 2 であることも分かった。

5.3　項目反応理論

事前テストの結果をラッシュモデルで分析した。その結果、能力値パラメータは、$M = 0.58$、$SD = 0.73$、$\alpha = .89$、困難度パラメータは、$M = 0.00$、$SD = 1.09$、$\alpha = .99$であった。適合度分析は、許容範囲0.7〜1.3を基準に行った（靜，2007，p. 317）。その結果、インフィットでは許容範囲を超える項目は確認できなかったものの、アウトフィットで項目77が1.67、項目18が1.63のアンダーフィットであった。項目77の事前テストの正答率と困難度は13.56％と2.61で、項目18は14.54％と2.52であった。これらの項目は、選択肢を含めた問題文が適切ではなく、解答がランダムである可能性もあり、データ全体に影響を与えていることも考えられることから早急な改善が必要であることが分かった。一方、正答率が高く0.5以下のオーバーフィットの項目に関し

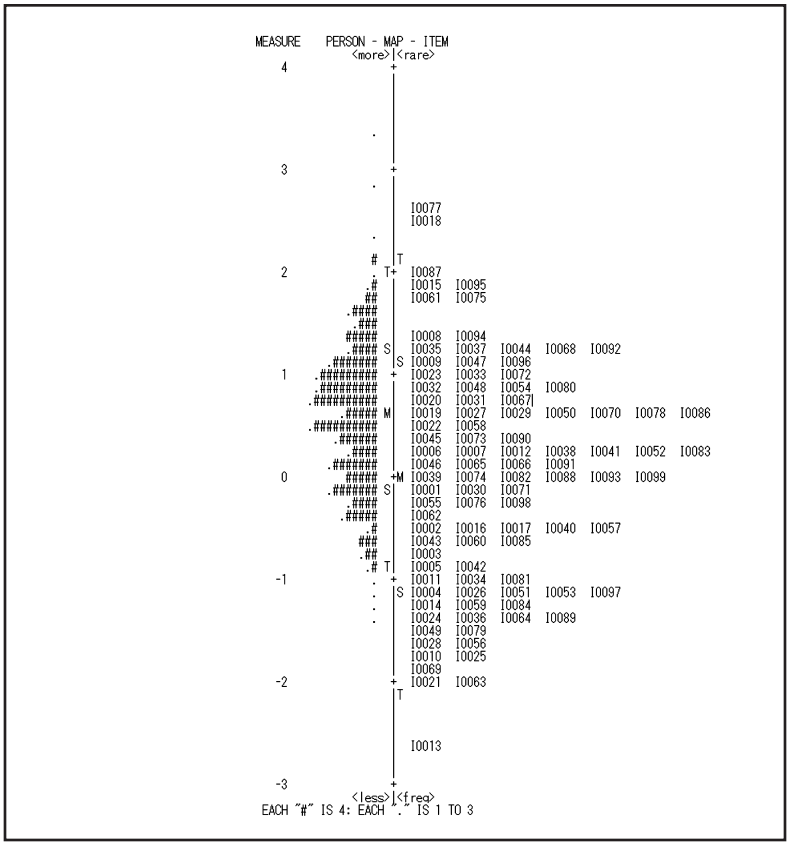

図 2　困難度による項目配置

ては、通常は問題視されないが、限られた問題数の中で適切に学生の英語力を測定し、今後の英語教育に生かしていかしていくためにも対応が必要と判断した。図２は、同一尺度上に学生の能力パラメータと問題項目の困難度パラメータを配置したものである。中央から左側に学生の能力パラメータが分布し、右側に困難度パラメータが分布する。図２より、双方の対応関係と、特に問題項目の分布を直感的に把握することができる。図２から、正答率が高く、あまり機能していないオーバーフィットの問題項目が、学生の能力パラ－メタ－の下部の方に多く点在していることが分かる。

5.4　ベイズ推定

アイテムバンクを活用した授業デザインの異なる統制群と実験群の事前・事後の差の比較のためにベイズ推定を行った。ベイズ推定を行うにあたっては、豊田（2016）を参考にした。具体的な授業実践や t 検定による分析結果については、工藤・住・山田（2017）を参考にして頂きたい。分析は事前テストの得点における統制群と実験群の差の比較を行い、次に、事後テストの得点における統制群と実験群の差の比較を行った。選択肢に問題のあった項目88は全員を正解とし、全100問（満点100点）での分析を行った。

事前テストの比較では、統制群の得点が正規分布 $x_1 \sim N(\mu_1, \sigma)$ にしたがい、実験群の得点も正規分布 $x_2 \sim N(\mu_2, \sigma)$ にしたがうと仮定した。標準偏差 σ は共通させた。テストは０～100点で採点されるので、母平均 μ_1 と μ_2 の事前分布は一様分布 $U(0, 100)$ とし、母標準偏差 σ の事前分布は一様分布 $U(0, 50)$ にしたがうと仮定した。

ベイズ分析では、2100のチェインを５つ発生させ、バーイン期間を1000とし、HMC 法によって得られた10万個の乱数で事後分布・予測分布を近似した。母数・生成量のすべてに関して有効標本数が多く、$\hat{R} < 1.1$ であり、事後分布・予測分布へ収束した。点推定には EAP を用いた。確信区間には95％確信区間を用いた。以下、（　）内の数字は誤差を、[　]内の数字は確信区間を表す。

表３より、事前テストでの母平均の差は1.76（1.11）[－0.43，3.94] であることから、平均的得点差は1.76点であった。効果量の推定値は、0.135（0.08）[－0.03，0.302] であり、偏差値換算では約1.3の差であった。非重複度は、0.553（0.033）[0.487，0.619] であった。非重複度は、0.5のときに２群が完全に重複していることを意味し、0.5から離れ、0.00や1.0に近づくほど非重複度が大きいと解釈できる。２群を比較して一方が他方の10％より大きい点にあるという命題の確からしさは、p（$U_3 > 0.6$）で評価できることから、事前テストにおいて、統制群と実験群は、ほぼ均質な集団と評価することができた。

表３　事前テストにおける統制群と実験群の比較

	EAP	post.sd	2.5%	5%	50%	95%	97.5%
平均の差	1.762	1.110	-0.433	-0.065	1.758	3.584	3.949
効果量	0.135	0.085	-0.033	-0.005	0.134	0.274	0.302
非重複度	0.553	0.033	0.487	0.498	0.553	0.608	0.619

事後テストの比較も、事前テストと同じ手順で行った。表 4 より、母平均の差は13.618（0.634）［12.376，14.866］であることから、平均的得点差は約13点であった。効果量の推定値は、1.038（0.062）［0.919，1.162］であり、偏差値換算では約10の差であった。非重複度は、0.850（0.014）［0.821，0.877］であり、p（U_3＞0.6）の基準に照らして、実験群に提供した TOEIC 語彙学習環境は一定の効果があったと考えられる。

表 4　事後テストにおける統制群と実験群の比較

	EAP	post.sd	2.5%	5%	50%	95%	97.5%
平均の差	13.618	0.634	12.376	12.579	13.618	14.657	14.866
効果量	1.038	0.062	0.919	0.937	1.037	1.141	1.162
非重複度	0.850	0.014	0.821	0.826	0.850	0.873	0.877

6. 考察

本研究では小規模ではあるが、アイテムバンクの開発を行い、古典的テスト理論と現代テスト理論を組み合わせ、アイテムバンクの分析に応用することで、双方の弱点を補いあい、客観的なフィードバックを参考にしながら、問題項目の改善点を明らかにすることができた。開発されたアイテムバンクには、未だ改善が必要な問題項目が多く存在することも分かり、今後へ課題を残す結果となった。

正答率が極端に高い問題項目は、ラッシュモデルでの適合度分析においてオーバーフィットとなる。逆に、正答率が極端に低い問題は、解答にランダム性がありアンダーフィットになる。こうした理論上の関連性を、実際の授業経験を踏まえ、具体的な問題項目と学生の反応データをもとに確認できたことは、今後の問題項目の開発および改善につながる重要な経験であった。また、項目反応理論では、テスト全体への影響が少ないことからあまり問題にされることのないオーバーフィットの問題項目でも、錯乱肢有効度を確認することによって、正答以外の選択肢が十分に機能しておらず、結果的にオーバーフィットになっている問題項目もあることが分かった。これは項目反応理論の適合度分析だけでは検出することができない改善点であった。

一次元性と局所独立は、実際のデータで厳密に成り立つことは考えにくいが、確認作業をとおして開発の段階では気がつけなかった改善点を客観的な指標で浮き彫りにすることができた。今回はアイテムバンクの開発素材が限られた範囲の語彙であり、また、問題形式も多肢選択問題のみであったことから、設問および選択肢に多様性を持たせることが困難であった。このことが、すべての問題に一貫して一次元性および局所独立を保証することをさらに難しくしていた。今後、アイテムバンクを拡張する際に、語彙の難度や問題の種類を増やしながら、この問題を改善していく必要がある。

授業実践に関する詳細は、工藤・住・山田（2017）に譲るものの、今回、アイテムバンクを開発し、事前テストを行い、結果の分析から指導のポイントを授業に活かし、その成果を事後テストで確認できたことは有益であった。教育、そして測定・評価の両輪を機能させるためにも、CRT の精度を向上させ、測る道具であるアイテムバンクを開発することは、大きな変化が待ち受けている今でこそ特に重要と考える。本来であれば、アイテムバンクの開発は、テスト仕様を

協議し、コーパスなどを活用して、場合によってはコーパスを開発してから語彙および表現を抽出し、問題項目の設計を行うものである。しかし、CRT のように、授業に即した到達度を測定するアイテムバンクを開発する場合には、既に教えている内容や使っている教材を活用しながらアイテムバンクを開発することが現実的である。本研究では、TOEIC 語彙のアイテムバンクを開発し、その結果を分析したが、今後、例えば、長年の研究を踏まえてすでに公開されている工学系 ESP の語彙リストを活用してアイテムバンクを開発することも考えられる。アイテムバンクを拡張し、問題項目の等化作業を行えば、授業と連動しながら授業内外で英語学習を支援できる適応型学習支援システムの開発に利用することもできる。

7. まとめ

学内外の英語教育を取り巻く変化に先んじて対応するために、理工学部では TOEIC 語彙学習に特化したアイテムバンクの開発に着手した。アイテムバンクは、まだ初期の段階で改善が必要な問題項目が多くあることが明らかになったが、本研究を通じて、開発、検証、分析までの知識と技術を英語科目を担当する専任教員間に蓄積することができたことは大きな資産となった。また、アイテムバンクを活用する授業実践にも並行して取り組み、その効果と今後の改善につながる示唆を得ることができた。本研究を通じて得られた示唆を活かし、理工学部の英語教育に関して、今後も教育と測定・評価の相互作用の高める活動に取り組んでいく。

謝辞

本研究のデータ収集には、理工学部英語非常勤講師の藤平先生、乗次先生、山脇先生からのご協力を頂きました。心より感謝申し上げます。

参考文献

American Statistical Association (2016, March 7). *American statistical association releases statement of statistical significance and p-values: Provides principles to improve the conduct and interpretation of quantitative science*. Retrieved from https://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf

Brown, J. D. (2005). *Testing in language programs: A comprehensive guide to English language assessment*. NY: McGraw-Hill.

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.

河上源一（2011）.『TOEIC テストに でる順英単語』, 中経出版.

近藤ブラウン妃美（2012）.『日本語教師のための評価入門』, くろしお出版.

工藤多恵・住政二郎・山田一美（2017）.「LUNA を活用した TOEIC 語彙定着のための実践」『関西学院大学高等教育研究』, *7*.

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

大河内佳浩・山中明夫（2016）.「プレースメントテストや高校の履修状況などのデータを用いた初年時成績不振者の早期発見」『日本教育工学会論文誌』, *40*, 45-55.

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.

住 政二郎（2013）.「ラッシュモデルの導出」『外国語教育メディア学会関西支部メソドロジー研究部会2012

年度報告論集』，83–101．Retrieved from http://www.mizumot.com/method/2012–07_Sumi.pdf

住 政二郎（2014）．「項目反応理論―1PLM，2PLM，3PLM、多段階反応モデル」『LET 関西支部メソドロジー研究部会2013年度報告論集』，*4*，34–62．Retrieved from http://www.mizumot.com/method/04–04_Sumi.pdf

豊田秀樹（2016）．『はじめての統計データ分析―ベイズ的＜ポスト *p* 値時代＞の統計学』，朝倉書店．

Yen, W. M.（1984）. Effects of local item dependence on the fit and equation performance of the three–parameter logistic model. *Applied Psychological Measurement, 2*, 125–145.

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

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．