論　文

古典的テスト理論を用いた2012年度新入生英語プレイスメントテストの分析と改善への提言

石　原　知　英

要　旨

本稿は，本学名古屋校舎で2012年度新入学生を対象に実施されたプレイスメントテストについて，その実施状況を報告し，今後の改善を視野に入れた分析結果をまとめる。具体的には，古典的テスト理論に基づく項目分析および信頼性の検討を中心に，(1) 得点の分布，(2) 信頼性，(3) 項目の難易度および弁別力，(4) 妥当性，の 4 つの観点から分析した。その結果，テスト全体としては概ね満足のいく分布であり，信頼性および妥当性も十分であると解釈された。ただし，問題や選択肢の改訂を行う（あるいは削除する）ほうがよいと考えられる項目がいくつかあり，今後の課題として指摘された。

キーワード：愛知大学名古屋校舎 1 年生，プレイスメントテスト，古典的テスト理論

1　はじめに

1.1　プレイスメントテスト実施の概況

愛知大学名古屋校舎では，2012年度新入生を対象に，英語プレイスメントテストを実施した。これは主に，入学生の英語熟達度を把握するとともに，1 年次必修科目の 1 つである Reading I/II のクラス編成のための資料とするためである。Reading I/II では，選抜方式による習熟度別クラス編成を実施しており，このプレイスメントテストの結果に基づき，発展ク

ラスと基礎クラスをそれぞれ 3 クラスずつ編成している。この編成方式により，発展クラスではさらなる英語力の伸長を目指し，また下位クラスでは少人数でのきめ細かな指導による対応が可能になっている[1)]。

1.2　テストの構成

2012年度のプレイスメントテストには，本学教員が作成した独自試験を用いた。4 月のガイダンス期間という短い時間内に，実施，採点およびクラス編成までを行うため，実用性に重点を置いたテストとなっている。

テストは全75問の構成で，解答時間は45分である。全ての設問は短文の空所補充問題で，すべて 4 択である。解答はすべてマークシートにマークする形式であった。問題は主に語彙や文法の知識を問うものである。これは，このテストスコアが Reading I/II のクラス編成のために用いられることと密接な関係がある。つまり，基礎的な語彙や文法の知識の習熟度をクラス内である程度一定の水準にすることで，授業内での解説の際に，どの程度の語彙や文法事項を，どの程度の時間をかけて説明する必要があるのか，判断することが可能となると考えられたためである[2)]。

このプレイスメントテストでは，学生の多様な英語習熟度を考慮し，設問の難易度が多岐にわたるように設計されている。具体的には，おおよそ英検 3 級程度の問題が10問，準 2 級程度の問題が20問，2 級程度の問題が30問，準 1 級程度の問題が15問の計75問の構成で，易しい問題から難しい問題へと配列されている[3)]。

1.3　本稿の目的

本稿の目的は，実施されたプレイスメントテストのスコアを，(1) 得点の分布，(2) 信頼性，(3) 項目の難易度および弁別力，(4) 妥当性，の 4 つの観点から分析し，今後の改善の提言を行うことである。

具体的には，(1) 得点分布から，全体的なスコアの散らばり具合を確認し，天井効果や床効果の有無を検討すること，(2) クロンバックの α および折半法による相関係数を計算し，テスト全体の信頼性を確認すること，(3) 正答率，項目弁別力および点双列相関係数から，不当に難しい，あるいは易しい問題をあぶりだし，今後の改善する必要がある項目を指摘すること，(4) 外部テストである TOEIC IP テストスコアとの相関係数を計算し，このプレイスメントテストの併存的妥当性を検討すること，の 4 点について言及する。

2　プレイスメントテストの分析

2.1　記述統計量と得点分布

テストの得点分布を確認するため，表 1 に学部別にみたプレイスメントテストの記述統計を，図 1 にヒストグラムを示す。

表 1　学部別にみたプレイスメントテストの記述統計

	n	*M*	*SD*	*95%CI*	*min*	*max*
全体	1684	40.27	10.02	[39.79, 40.75]	11	69
法	385	41.30	8.80	[40.42, 42.18]	12	66
経済	390	40.16	9.86	[39.18, 41.14]	13	63
経営	454	38.76	10.08	[37.83, 39.69]	11	67
国コミ	246	44.20	9.88	[42.96, 45.44]	19	69
現中	209	37.22	10.78	[35.75, 38.69]	15	67

注．プレイスメントテストは75点満点，*CI* ＝ Confidence Interval

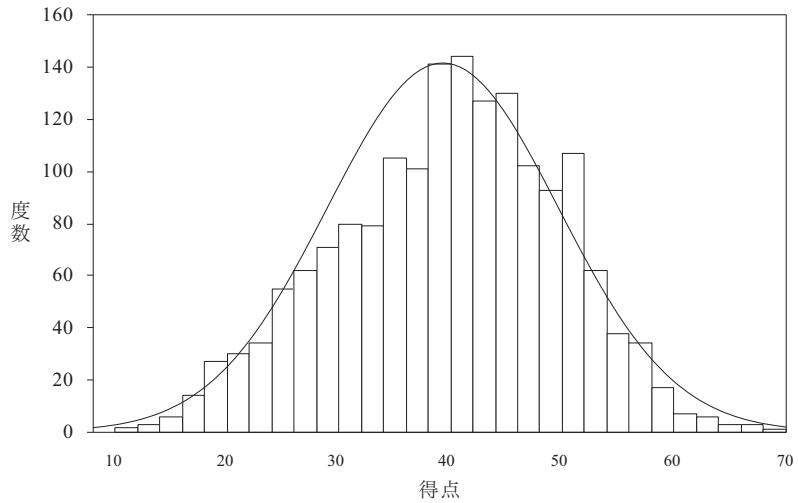

図 1　プレイスメントテストの得点分布

このテストの平均（40.27）は，得点可能範囲である 0 点から75点の中央に近く，標準偏差（10.02）および最小（11），最大（69）の値から，床効果も天井効果もないと解釈することができる。正規曲線からはやや右によっている（歪度−.197，尖度−.348）ものの，図 1 からも，極端に正規性を欠く分布ではないようである。

2.2　信頼性の検討

テストの信頼性とは，一つには，同じ受験者が異なる状況で受験した場合に，どの程度安定した得点が与えられるかということである。その意味では，安定性と言い換えることも可能である。もう一つは，テストの各項目間の得点にある程度の一貫性があるということである。こちらはテストの内的な整合性とみることがでる（斉田，2011）。

信頼性を検討するには，折半法によるものと，得点の分散を計算するクロンバックの α 係数を求めるものがある。前者は，個人の得点を，偶数問題と奇数問題の 2 つに分割し，相関係数を求める方法である。ただし，項目数が実際のテストの半分となるため，スピアマン・ブラウンの修正公式を用いて，信頼性係数＝（2 × 相関係数）／（1 ＋相関係数）で求める。後者のクロンバックの α 係数は，信頼性係数として広く知られる指標で，α＝[人数／(人数−1)]×[1−(項目分散の和／合計得点の分散)] で求める。

計算の結果，スピアマン・ブラウンの修正公式を用いた信頼性係数は .86，クロンバックの α 係数は .86であり，内的一貫性については十分な値であると解釈することが可能であろう。

2.3　古典的テスト理論に基づく項目分析

上述したように，テスト全体としての信頼性は，概ね満足いくものであった。この節では，全75問のテスト項目それぞれについて，古典的テスト理論に基づく項目分析を行い，今後の改善する必要がある項目を指摘する。

表 2 に，各設問の（1）項目難易度，（2）点双列相関係数，（3）項目弁別力指数，（4）実質選択肢数とその適切度を示す。

（1）の項目難易度は，正答した受験者の全体受験者に対する割合で，正答率と言い換えることもできる。0 から 1 までの値を取り，0 に近いほどその項目が難しいことを，1 に近いほど易しいことを意味する。当て推量による正解を考慮すると，適切な項目難易度は .625 である（大友，1996）。学生にとって不当に難しすぎる可能性があると考えられる .29以下の項目，逆に易しすぎると考えられる .81以上の項目について，下線を付した。

（2）の点双列相関係数は，その項目得点と合計点との間の積率相関係数であり，−1から 1 までを取る。渡部（2012）の判断基準に倣い，改良が必要とされる .29以下の項目に下線を，よくない項目であるとされる .19以下の項目に二重下線を付した。

（3）の項目弁別力指数は，当該項目によりどの程度成績上位者と下位者を区別できるかの指標である。具体的には，全体の正答数が多い成績上位27％と，正答数の少ない下位27％を設定し，それぞれの項目について，上位群の正答率から下位群の正答率を引いたものが，項目の弁別力指数である。この指標については，大友（1996）に倣い，改訂の要ありとされ

表2　プレイスメントテストの項目分析

項目	難易度	相関	弁別力	選択肢	適切度	項目	難易度	相関	弁別力	選択肢	適切度
1	.89	.27	.21	1.61	.98	39	.53	.41	.49	3.29	.98
2	.86	.20	.17	1.73	.96	40	.28	.33	.36	3.87	.96
3	.77	.40	.41	2.08	.89	41	.52	.24	.28	3.14	.90
4	.80	.26	.26	2.04	.98	42	.52	.21	.27	2.79	.75
5	.69	.15	.18	2.56	.98	43	.54	.27	.38	3.26	.99
6	.84	.40	.34	1.82	.95	44	.51	.28	.35	3.42	.99
7	.79	.40	.41	1.97	.87	45	.47	.26	.32	3.43	.94
8	.79	.31	.28	2.05	.94	46	.36	.29	.36	3.42	.84
9	.82	.37	.35	1.75	.78	47	.53	.27	.34	3.22	.95
10	.84	.41	.35	1.86	.99	48	.44	.29	.35	3.61	.97
11	.72	.22	.25	2.24	.85	49	.48	.34	.44	3.21	.87
12	.80	.43	.42	2.05	.97	50	.45	.25	.27	3.34	.89
13	.71	.32	.35	2.45	.95	51	.31	.13	.14	2.53	.52
14	.79	.41	.43	1.99	.88	52	.42	.31	.39	3.59	.95
15	.73	.37	.40	2.38	.99	53	.24	.24	.25	3.61	.87
16	.75	.38	.43	1.98	.76	54	.31	.29	.32	3.80	.94
17	.69	.33	.37	2.53	.96	55	.17	.03	.02	3.83	.97
18	.61	.48	.62	2.65	.82	56	.36	.32	.38	3.42	.84
19	.74	.30	.32	2.29	.94	57	.32	.08	.11	3.93	.99
20	.66	.32	.38	2.45	.83	58	.33	.30	.36	3.50	.85
21	.59	.37	.49	2.51	.73	59	.33	.17	.18	3.83	.96
22	.67	.39	.45	2.66	.97	60	.24	.20	.20	3.54	.85
23	.58	.38	.48	2.96	.91	61	.22	.25	.25	3.77	.93
24	.63	.37	.44	2.85	.99	62	.23	.18	.18	3.76	.92
25	.67	.39	.45	2.65	.96	63	.34	.35	.44	3.79	.95
26	.52	.17	.20	2.63	.69	64	.26	.05	.05	3.88	.96
27	.56	.28	.33	2.83	.83	65	.17	.16	.15	3.36	.81
28	.68	.27	.30	2.51	.92	66	.21	.18	.17	3.96	.99
29	.69	.41	.49	2.59	.98	67	.18	.26	.27	3.55	.87
30	.62	.32	.40	2.56	.80	68	.11	−.02	−.02	3.72	.99
31	.62	.42	.54	2.84	.94	69	.51	.39	.51	3.40	.99
32	.60	.20	.25	2.90	.93	70	.69	.39	.42	2.59	.99
33	.64	.31	.35	2.80	.96	71	.63	.35	.41	2.72	.91
34	.57	.34	.42	3.01	.91	72	.40	.28	.35	3.38	.85
35	.52	.33	.44	2.98	.84	73	.55	.36	.47	3.22	.98
36	.57	.34	.42	2.94	.89	74	.29	.16	.16	3.83	.95
37	.65	.34	.37	2.63	.90	75	.65	.40	.47	2.74	.97
38	.53	.19	.22	3.04	.87						

る .29以下の項目に下線を，除外もしくは改訂が必要とされる .19以下の項目に二重下線を付した。

（4）の実質選択肢数は，４つの選択肢のそれぞれに対する解答数をもとに，実質的に学生が選択した選択肢数を産出したものである。言い換えれば，いくつの選択肢に解答が分布しているかについての指標であり，数値の低い項目は，錯乱肢が実質的に機能していない可能性が高く，改訂の余地があると考えることができる。また，その適切度は，正答以外の錯乱肢が均等に選ばれた際に最適な分布になると仮定し，実質選択肢数が最適選択肢数にどれほど近接しているかを計算したものである。具体的な計算式は大友（1996）を参照されたい。2.00以下の実質選択肢数と，.80以下の適切度を持つ項目に下線を付した。

最初の数問（例えば項目１，２，６，９，10）は，多くの学生にとって易しすぎる項目である。しかし，入学直後の新入生を対象としたテストであることを鑑みると，こうした項目によって学生の不安を軽減するという意味において，必要な項目であると言えるだろう。

いくつか散見される弁別力の低い項目（例えば項目２，５，11，26，32，38，41，42，50，51）は，全体のスコアが高い学生が多く間違える（逆に全体のスコアが低い学生の正答率が高い）ことを示唆しており，難易度の高すぎる語彙が含まれていないか，またひっかけ問題になっていないかを確認する必要がある。

選択肢の適切度が低い項目（例えば項目９，16）は，４つの錯乱肢が有効に機能していない可能性がある。著しく選択されていない選択肢や，逆に集中している選択肢について，改訂する必要があるかもしれない。

また，項目50から68までは，正答率および弁別力の低い項目が並んでいる。問題の後半であるため，時間がなく無回答であったり，あるいは適当にマークしている可能性もあるが，本学学生にとって不適切な程に難しい項目である可能性もあるため，検討を要する。ただ，こうした問題群があることで，入学後に必要な英語の知識について示すことができるという意味では，単に易化すればよいというわけではなく，ある程度難易度の高い項目も残しておく必要がある。

2.4　妥当性の検討

プレイスメントテストの妥当性は，テスト得点でクラス分けをする相応性や適切性であると考えることができる。2012年度新入生のプレイスメントテストは，先述したように，主に Reading I/II のクラスにおける選抜クラス（発展クラスと基礎クラス）の編成に用いられている。Reading の授業運営や評価については，それぞれのクラスの担当者に任されているものの，こうした語句と文法の知識という基礎学力的なレベルという観点によってクラスを編成することで，授業に際してどのレベルの語彙や文法事項について詳細な解説が必要であ

るか，またどのレベルまではそれほど詳述する必要がないかについて，知見を得ることができる。その意味では，この形式のテストで相応の妥当性があると考えられる。

加えて，外部テストなどとの併存的妥当性について，1年次秋学期に実施されたTOEIC IPのスコアとの相関を確認することで検証する[4)]。その結果，プレイスメントテストとTOEIC IPテストのスコアの相関係数は，r =.59であった。また，プレイスメントテストで主に測定される語彙と文法の知識と特に関連すると思われるTOEIC IPテストのリーディングセクションに限ったスコアとの相関係数は，r =.64であった。これらの点から，外的な妥当性もある程度高いものであると結論付けることができる。

ただし，散布図から，TOEICスコア上位者において，ややばらつく様子がみられる。特に上位層におけるプレイスメントテストの妥当性については，今後もう少し詳細に検討する必要がある。

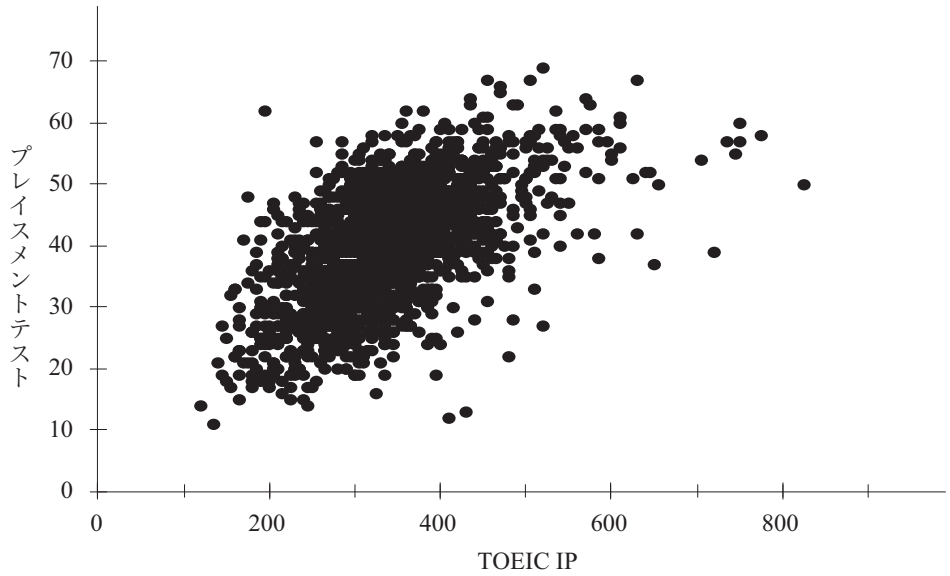

図2　プレイスメントテストとTOEIC IPテストスコアの分布

表3　プレイスメントテストとTOEIC IPテストスコアの相関係数（n=1650）

	TOEIC IP	Reading Section	Listening Section
プレイスメントテスト	.59	.64	.44

注．プレイスメントテスト（4月）とTOEIC IP（12月）の両方を受験している学生のみを分析対象とした。

3．まとめと改善についての提言

2012年度に実施した新入生プレイスメントテストについて，(1) 得点の分布，(2) 信頼性，(3) 各項目の難易度および弁別力，(4) 妥当性，の 4 つの観点から総合的に検討した結果，概ね狙い通りに機能していると結論づけることができる。

プレイスメントテストは本学の独自テストであるため，主に妥当性と信頼性についての懸念があったが，今回の分析から，概ね適切なテストであることが示された。当然，語彙や文法を多肢選択式でのテストであるため，総合的な英語学力，特にスピーキングやライティングなどの産出技能や，リスニングを中心とした音声言語の理解力，まとまった文章を読む力などについては，直接的に測定することができていない。ただ，今回のような簡便なプレイスメントテストでも，ある程度の妥当性と信頼性を保った評価が可能であったことは，実用性の面から，一定の評価が与えられるべきであろう。

最後に今後の課題と検討事項を 3 点挙げて，まとめとする。

第一に，選抜クラス数とそのカッティングポイントの設定についてである。現在は発展クラス，基礎クラスともに 3 クラスずつを編成しているが，今回の分析から明らかになったように，得点の分布がやや右寄りとなっている。これはつまり，上位レベルよりも下位レベルでのばらつきが大きいということである。時間割や教室の都合，また担当者の都合などにより，クラス数を増やすことが難しい場合には，例えば発展クラスを 2 クラス，基礎クラスを 4 クラスというようにすると，現状の分布をより適切に反映した編成が可能となるかもしれない。

表４　発展クラスと基礎クラスの編成人数

	発展クラス	基礎クラス
現状	70	71
M ± 2.0SD	19	48
M ± 1.5SD	87	109
M ± 1.0SD	272	293

第二に，項目の一部で改訂が必要であるという点である。全体として見た場合の妥当性，信頼性はある程度満足のいくものであった一方で，いくつかの項目では，難易度が不適切なもの，弁別力が低いもの，また選択肢が有効に機能していないものが散見された。こうした項目については，適宜修正を加え，より精度の高いテストとして改訂をすすめていくことが必要である。ただし，今回の詳細な結果は，テスト項目やテスト受験者に依存する（標本依存である）ため，違う受験者（例えば来年度の入学生）を対象とした場合，異なる項目統計

量が得られるだろうという点には注意を要する。そうした点を考慮した分析手法として，項目応答理論が近年注目を集めている（例えば斉田，2003; 前田，2003など）。必要に応じてこうした分析手法を取り入れ，更なる改善を目指すことも，今後の課題と言えるだろう。

第三に，プレイスメントテストとして外部テストを導入することの利点についてである。今回の結果から，本学独自のプレイスメントテストで十分な妥当性と信頼性が得られたことが明らかとなったが，その一方で，将来的には外部テストを導入する利点についても検討する必要があるだろう。外部テスト（例えばTOEIC Bridgeなど）は，社会的に認知度が高く，自分の手元にスコアが残ること，履歴書に記載することができることを考えると，学生本人にとっても受験する意味が大きい。独自テストは，あくまでクラス編成のための材料であるため，学生の受験に対する動機づけや，受験するメリットにやや欠けると言える。外部テストの導入には，金銭的なコストがかかる一方で，実施，採点などの運営面での手間を削減することができる。実際に2年次のPractical EnglishとTOEIC Iのクラス編成に際しては，全学的にTOEIC IPテストを導入して実施しており，2010年度の新入生クラス分けにはTOEIC Bridgeテストを利用したという実績もある。スコアの伸びを検討することなどを考慮しても，外部試験を利用する利点について，今後も継続的に検討していく必要があると考えられる。

註

1）選抜式クラス編成の成果と課題については石原（2012）を参照されたい。

2） 2年次のPractical EnglishおよびTOEIC Iのクラス編成には，1年次末に受験するTOEIC IPテストのスコアが用いられる。これは，TOEICのスコアベースでクラス内の習熟度をある程度一定にすることで，クラス内で扱う問題のレベルや目標のスコアを統一し，教育効果を上げようと考えるためである。

3）実際の採点では，設問の難易度を考慮した傾斜配点により100点満点で数値化されたが，本稿では，分析における簡便さのため，すべて各1点として再検討を行った。

4）このTOEIC IPのテストスコアは，プレイスメントテスト実施からおよそ10ヵ月経過した時点でのものであることには注意を要する。英語力の伸びは当然個人によって異なるため，例えばプレイスメントテストで低いスコアを取った学生が，基礎クラスでの指導によって，また当人の努力によって，学力が伸び，TOEICスコアが高いといった場合，ここでいう相関係数による併存的妥当性は，一見すると低いということになる。

参考文献

池田央.（1994）.『現代テスト理論』朝倉書店.

石原知英.（2012）.「愛知大学名古屋校舎2011年度Readingにおける選抜クラス編成の成果と課題

― TOEIC IP テストスコアおよびアンケートの分析―」『愛知大学語学教育研究室紀要　言語と文化』第27号，53–62.
大友賢二.（1996）.『項目応答理論入門』大修館書店.
斉田智里.（2003）.「高校入学時の英語能力知の年次推移―項目応答理論を用いた県規模英語学力テストの共通尺度化―」『Step Bulletin』vol. 15, 12–24.
斉田智里.（2011）.「第 2 章　英語学力測定論」石川祥一・西田正・斉田智里.（編著）『テスティングと評価― 4 技能の測定から大学入試まで』（pp. 30–58）大修館書店 所収
前田啓朗.（2003）.「到達目標型教育に向けた英語テストの改善：古典的テスト理論と項目応答理論に基づいて」『広島大学外国語教育研究』 6 ，131–140.
渡部倫子.（2012）.「日本語プレースメントテストにおける文法テスト項目の改訂」『広島大学大学院教育学研究科紀要　第二部』第61号，239–244.