異なるテスト分析からの結果に関する一考察

異なるテスト分析からの結果に関する一考察

津　村　修　志
盛　岡　貴　昭

1．はじめに

テストの得点は正答数の単純合計で算出されることが多い。しかし、この方法では正答数が同じであれば、同じ能力と判断されることになる。大友（2009）は、正答数に基づく得点について、「幾つの項目に正解したかという頻度の合計であって、それ以上の意味は持っていない（p.1008）」と言う。木村（2013）も「正答数による評価は、使用するテスト項目と受験者集団に依存した数値であるため、標準化された評価結果として扱うことはできない（p.63）」としている。例えば、入試などの合否判定では、仮に60点以上を合格と決めると、59点以下は不合格となってしまう。だが、60点を取った受験者と59点の受験者で、能力に差があると確信を持って言えるわけではない。他のテストを受けさせてみれば、順位が逆転する可能性も十分ある。不合格となった受験生は「運が悪かった」と言えなくもないが、公平な判定であったかどうか疑問に思う受験生がいるかもしれない。

能力の指標であるかのように扱われている従来の得点が、テストの難易度によって変化することも、単純に合計した得点があてにならない理由の１つである。易しい項目で構成されたテストでは得点が高くなるが、難しい項目が多いテストでは得点が低くなる。偏差値を用いて採点されることも多いが、そもそも平均値が違う２つのテストから別々に算出された得点を比較できるものではない。また、同一テスト内であっても、項目間で難易度に差があった場合、単純に合計することには大いに疑問が残る。このような理由から、単純合計によら

ない以下のようなテスト分析・採点が行われている。

項目応答理論（Item Response Theory：以降 IRT）は、公益財団法人日本英語検定協会が基礎開発した CASEC（Computerized Assessment System for English Communication）などで使用されている分析法で、斉田（2010）では、「個々のテスト項目に対する多くの受験者の応答パターンから、項目の特性値と受験者の能力値を推定していく（p.41）」理論であると説明されている。豊田（2002）は IRT のメリットとして、「異質な受験者が、異なる項目を、異なる日時に、異なる場所で受験したにもかかわらず、被験者は統一された処遇を受けることができる（p.24）」点を挙げている。これには、等化という手順を踏む必要があり、本研究の範疇にはないが、それによってテスト項目を一部変更して別の受験者に対してテストを実施した場合でも共通の尺度で受験者の能力を推定できると考えられている。理論の起源・展開・特徴、数学的な説明や分析方法などは豊田（2002）や山川（2008）に詳しいのでここでは割愛するが、単なる合計で得点を出す方法とは異なり、より厳密な測定方法であると言える。IRT を使ったプレースメントテストの分析については、山川（2008）、今井、伊東、中村ら（2009）があるが、これらには単純合計でレベル分けを行った結果との比較は見られない。

潜在ランク理論（Latent Rank Theory：以降 LRT）は Shojima（2007）で発表されたニューラルテスト理論（Neural Test Theory，NTT）のことで、順序尺度を仮定しており、学力を任意の段階で評価するのに適している。クラスター分析や潜在クラス分析が受験者のグループ分けに使用されるのに対し、LRT はグループに順序を与えることができる。数学的な説明は筆者らには難解であるため、Shojima（2007）を参照していただきたいが、この理論の意図するところはおおよそ以下のように読み取れる。

100点満点のテストを考えると、例えば69点と70点という1点の差は、受験者の能力の差を反映しているわけではない。テスト結果を考慮する際、1点刻みで評価できるほど、信頼性を上げることは相当困難である。また、1点刻みで示されるテスト得点の場合、1点や2点でも得点を上げようという意識が働くのが自然であろう。そうなると学習方法は、すぐに得点に繋がる、つまり即効性のあるテスト対策学習に傾いてしまうと考えられる。実際、多くの大学でも TOEIC などの資格試験対策の講座が開かれている。語彙力とテストスコアに高い正の相関があると分かれば、得点向上のために単語を集中して暗記することを学生に求め、テストの出題パターンを分析し、同様の問題を繰り返しやらせるなどは、従来の得点の出し方による典型的な負の波及効果と言える。

Shojima（2007）の理論は、従来の単純合計結果からでは実現できなかった、より公平で現実的なクラス分けを可能にしている。レベル分けテストに使用されるテストの場合、0点～100点という表し方は特に必要ではなく、任意のレベル数に分けるだけで十分であり、しかも、単純合計が寄せ集めの項目に対する正解数の合計でしかないのに対し、LRT は項目難易度と識別力を考慮したランク分けであるため、受験者にとってより公平なレベル分けが行える。木村（2013）は、LRT の潜在ランクは，順序尺度上の任意の段階に分けて表現されるため、クラス分けの判断が容易であり、そのため、プレースメントテストの分析に LRT が有用であると述べている。

LRT を用いた分析については、考案者の荘島（n.d.）を始め、小泉＆飯村（2010）や木村（2013）があり、荘島（n.d.）では、得点と LRT のランクとの Spearman 順位相関係数が .929

で、LRT によって得られる順位尺度は、正答数尺度と全く異なるものではなく、ある種の能力を反映していることを報告している。小泉＆飯村（2010）は、受験者を習熟度別に 3 クラスに分ける目的で語彙サイズテストを実施し、その結果を古典的テスト理論（Classical Test Theory：CTT）・ラッシュモデリング（テストを構成する項目の特徴を一つの母数で表現するモデル）・NTT（本稿では LRT を指す）で分析・比較を行った。ラッシュモデリングと LRT でのグループ分けの違いについては、LRT で Rank 2 となった受験者50名の内、4 名がラッシュモデリングで下位のグループに、2 名が上位のグループに振り分けられており、LRT で Rank 3 となった受験者52名の内、4 名が中位のグループに振り分けられていたが、それ以外はグループ分けが LRT とラッシュモデリングで一致していたことが報告されている。さらに小泉＆飯村（2010）は、CTT と LRT 間の項目難易度や識別力の対応を調査しているが、単純合計と LRT でのクラス分けの比較は見られない。一方、木村（2013）は、単純合計と LRT によるクラス分けを行った結果、両者の間で高い順位相関（125 人の実際のクラス分けで .95）を示したものの、2 つの分け方で異なるクラスになったケースが、125 人中42 人もいたことを報告している。

LRT が、有用な分析法であることは上で見た先行研究が示しているが、学力に偏りがある受験者集団でも同様の有用性を示すだろうか。筆者らの勤務先では大多数の学生が、中学校で習う文法事項を十分身に付けずに入学している。したがって、そうした学生はスローラーナーということになるだろう。英語学習に対して「嫌い」と回答する学生が 7 割を超えるような状況で、うまく受験者を識別することができるだろうか。信頼性が高いと言われる難しいテストを受けさせても、意欲を失って眠り込んでしまうのではないだろうか。あるいは、益々英語が嫌いになったりしないだろうか。そのように考えると、出来るだけ学生の自尊心を傷つけない、落ち込ませない、やる気を削がない工夫が必要だと思う。だから、筆者らは項目をより優しいものに差し替えることでうまくレベル分けができるようなテストを作りたいと考えている。そのような目的で、IRT や LRT 分析が行われている例は少ないと思われる。

加えて、項目分析に LRT が使用されている小泉＆飯村（2010）や木村（2013）のような例は未だに多くはなく、古典的と呼ばれる分析法が一般的である。LRT は、項目難易度や識別力を考慮してレベル分けを行うので、単純合計と比べて厳密な方法と言える。一方、後者による得点の算出は、特に同一テスト内の項目間で難易度が大きく異なるような場合には、信頼性に欠けると言わざるを得ない。受験者の能力を測定するテストで、難易度が考慮されずに採点が行われるのは公平とは言えない。幅広い受験者層を識別するためには様々な難易度の項目が必要であることは理解できるが、近年の大学生の学力低下を考えると、そもそもまったく身に付いていない文法事項がテストに紛れていることで、項目間の難易度が大きく変化することは明白である。したがって、プレースメントテストの項目も学生の実態に合わせて選択した方が、よりテストの目的が果たせると考えられる。また、項目の特性を調べることで、上位・中位・下位のどのレベルでどの文法事項を到達目標とすればよいかを判断する知見が得られるだろう。例えば、現在完了は中位のクラスでは扱えるが、下位のクラスでは触れるだけというような目安が得られることは大きなメリットである。上のような理由から、過年度のテストで扱った文法事項を少し整理して学生のレベルに比較的合った項目でテ

ストを作成した場合にどのような変化が見られるのかを確認するために、従来の方法とLRTによる項目分析を行うことにした。

そこで本研究は、より良いプレースメントテスト作成とそこから得られる知見の有益な活用を目的として、スローラーナーを対象とする擬似プレースメントテストを行った結果を基に、単純合計得点、IRT、LRTでレベル分けを行った場合にどれだけの差異が生じるかを確認し、さらに項目を受験者のレベルに近付けた場合により、適切なプレースメントテストとなり得ることを検証しようとするものである。

以下の3点が本研究の研究課題である。

1）単純合計､IRTの能力推定値、及びLRTのランク（10段階）の相関がどの程度か。
2）単純合計、IRT、LRTによる受験者の3分割を行った際に、どの程度差異が生じるか。
3）文法事項の中で特定の項目（be動詞、代名詞、疑問詞、3人称単数現在など）に焦点を絞って問題を作成した場合、信頼性、識別力に差が生じるか、また、分割方法による差が見られるか。

2．方法

2.1. 調査時期と対象

2014年、2015年、2017年4月、非外国語専攻の4年制大学1年生（経済学科、商学科、経営学科、公共経営学科）それぞれ458名（男子397名、女子61名)、379名（男子330名、女子49名)、138名（男子108名、女子30名）を対象として、文法テスト（2014年36問、2015年34問、2017年36問）を行った。筆者らの勤務先では、必修の英語クラスは実際のプレースメントテストにより3つのレベルに分けられるが、スローラーナーに焦点を絞るため、調査協力は大半下位クラスの学生に依頼した。2014年と2015年では、約6割が下位、3割が中位、1割が上位クラスの学生であった。2017年では、下位が約8割、中位が2割で上位クラスの学生は対象としていない。また、2017年は基本的な文法項目に焦点を絞って作成した。各年度で、最後まで回答していない者がそれぞれ10～20数名いたので、それらの回答は分析対象から除いている。上記の数値は除いた後の数値である。

2.2. 測定に使用したテスト

使用したテストは、いずれも筆者らが作成したもので、2011年から修正を加えながら、担当クラスで使用している。今回分析に使用したのは、実際にレベル分けに使用されたものではなく、2014年、2015年、2017年に英語担当教員数名の協力を得て授業中に行った文法力確認テストである。この目的は受講者のレベルと弱点を把握すると同時に、文法項目における到達目標を学習者に確認してもらおうというものである。それ以外の年度にも同様のテストを実施しているが受験者が40～60名程度と少なかったため今回の分析には使用しないことに

した。

2014年に使用したテストには、全36問中に中学生や高校生が学ぶ文法事項を網羅的に含めた。時制や仮定法に関する項目など、対象受験者には、少々難易度の高いテスト項目があった。2015年のテストは項目を34問に減らし、30項目は2014年のものをそのまま使用した。残り4項目は文法問題ではなく、適切な会話応答文を選択させる問題とした。これは、基本的な英語での応答を重視するべきだと考えたためであったが、会話文選択問題は項目としてあまり良い結果が出なかったことと、会話文選択問題を含めると、テスト紙面上、文法問題よりも大きなスペースを必要とし問題項目数が減ってしまうため、それ以降は使用していない。2017年は、学習者にとって難しいと考えられる項目（例えば仮定法の正しい形を選択する問題、正しい時制を選択する問題、不定詞・動名詞を使い分ける問題、など）を排除し、中学1年程度の項目に焦点を絞って作成した。2015年のテストと同じ項目は36問中、19問であった。項目にはbe動詞、代名詞、一般動詞に関する問題を多く使用した。これは、本学学生の過去の答案中にこれらに関する誤りが非常に多かったことと、これらの弱点を克服しないまま英語学習を続けても「分からない」「難しい」という意識が膨らみ、英語学習に対する嫌悪感を強めるだけだろうと考えたからである。

2.3. 手続き

各年度第2回目の授業で開始時の約20～25分を使い、4肢択一式の文法テストを実施した。単純合計、IRT、およびLRTで分析・採点を行い、3グループ（上位群、中位群、下位群）にレベル分けを行った。この時、できるだけ3グループのメンバー数に差が出ないように調整した。LRTではソフト上で調整ができるので問題なく行うことができる。IRTでも能力推定値が細かく算出されるので、3分割は容易に行える。一方、単純合計では同点の者は必ず同じグループに入るよう調整したため、一部でグループのメンバー数に差異が生じた。

なお、IRTは2値の正誤データに対して、2母数ロジスティック（2-parameter logistic, 2PL）モデルを採用したが、受験者数が少ないため本来であれば比較的少ない対象でも安定した結果が得られるラッシュモデリングを用いるところである。しかしながら、ラッシュモデリングは項目の識別力を考慮しないということなので、2PLを採用した。そのため、今回のIRTの結果は参考程度と考えている。また、2PLとラッシュモデリングの詳細についても、筆者らの説明が及ぶところではないので、山川（2008）を参照していただきたい。分析には、LRTと同様、荘島が開発したexametrika ver. 5.3を使用した。

3．結果と考察

3.1. 記述統計量から見た各テストの特性

Table 1は各年度のテスト記述統計量を表している。筆者らが意図した通り、2017年のテストは平均値が少し高くなっている。Figureはそれぞれのテストのヒストグラムである。容易に見て取れるように、どのテストにおいても天井効果や床効果はなかった。中央値と平均値の差が小さいことからも、分布の形状に大きな偏りがないことが分かる。ただし、2014

Table 1：各テストの記述統計量

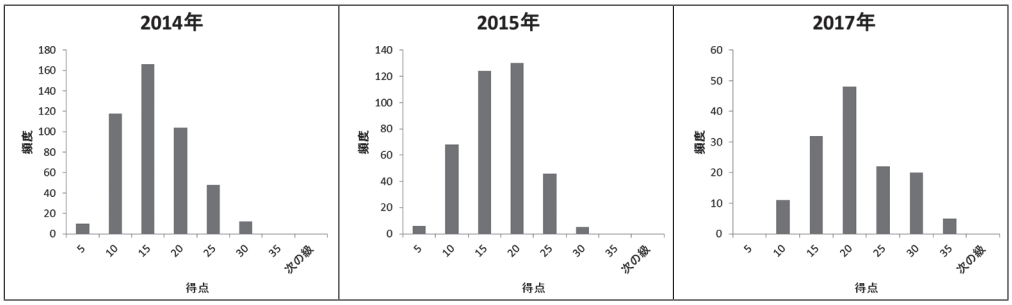
	2014年	2015年	2017年
受検者数	458	379	138
項目数	36	34	36
最小値	4	4	7
最大値	30	28	35
中央値	13	15	19
平均値	14.15	15.06	18.95
分散	27.87	22.55	36.27
標準偏差	5.28	4.75	6.02

年と2015年のテストには初級の学習者にとって難易度の高い項目が含まれているため、2017年のテストに比べて、少し左に偏っている。

3.2. 各テストの信頼性

信頼性は、難易度や内容が同じようなテストを同一人物が受験した場合に、ほぼ同じ結果を返すというようなテストの安定性を示す。物差しや体重計を例にとれば、測定するたびに違った値を示すような道具は信頼性に欠けると判断される。同様に、受けるたびに得点が大きく違っているようなテストは信頼性の低いテストということになる。しかし、普段現場で授業担当者が作成するテストは一般的に、パイロットを行うわけでもなく、項目分析を基に項目を厳選して作成しているわけでもないので、信頼性が高くなることは少ないと考えられる。

Table 2は各テストの2種類の信頼性係数を表にまとめたものである。古典的と言われる分析方法において、信頼性を報告する際によく用いられている α 係数では、2017年が.803となり、比較的高い値であると言える。しかし、α 係数は項目数が増えるほど高くなることが知られているので、項目数が36しかないことを考えると、高い信頼係数が得られることは期待していなかった。たった36問のテストでこのような比較的高い値が出たことに不安を感じたので、別の方法でも信頼性を見ることにした。この際に利用したのが、テスト全体を2分して2つのテストを仮定し、その得点間の相関係数を用いる折半法（Spearman-Brownの公式）である。その結果、2017年のテストは.780という値であった。2種類の信頼性係数から、



Figure：テスト結果のヒストグラム

Table 2：各テストの信頼性係数

	2014年	2015年	2017年
Cronbach's α	.757	.693	.803
Spearman-Brown	.757	.658	.780

Table 3：単純合計得点、IRT 能力推定値、LRT ランク（10段階）の相関係数

	2014年	2015年	2017年
単純合計 & IRT（Pearson 積率相関係数）	.982 ***	.971 ***	.992 ***
単純合計 & LRT（Spearman 順位相関係数）	.927 ***	.938 ***	.938 ***
IRT & LRT（Spearman 順位相関係数）	.963 ***	.976 ***	.961 ***

*** $p<.001$

36問のテストとしては、比較的信頼性が高かったと判断する。

他の年度の結果を見るとやはり、項目数の少ない2015年の値は少し低くなっているが、2014年の値と比べると大きく差が開いているわけでもなく、低すぎるとも考えられない。本来であれば、難易度が高すぎる、または低すぎる問題や識別力の低い項目を削除して、毎年テスト全体を更新していくのが理想であるが、実際のプレースメントテストとして使用しているわけでもなく、項目の良し悪しを見ることも今回の目的となっていたので、項目分析に基づく「更新」は行わなかった。

3つのテストは、受験したメンバーもテスト項目も異なるため、α係数を比較すること自体には大きな意味はない。しかし、項目が変わってもα係数の大きさがそれほど変わらない点には注目しても良いのかも知れない。結論はもっと大規模な調査に拠らねばならないが、スローラーナーに対しては、文法項目を網羅していなくても、それ自体がテストの信頼性を大きく損なう原因とはならないという可能性がうかがえる。

3.3. 単純合計、IRT、LRT によるレベル分け

レベル分けを行う前に、単純合計得点、IRT 能力推定値、LRT ランクの相関を確認した。ただし、LRT でのレベル分けでは 3 分割としたが、相関係数計算のために10分割を使用し、また、LRT のランクが順序変数であるため、単純合計と LRT、また IRT と LRT の相関は Spearman の順位相関係数を算出した。

Table 3 に示す通り、各年度どの組み合わせでも高い相関が見られた。これは、小泉＆飯村（2010）の結果と一致している。

Table 4 は、単純合計得点、IRT、LRT によるレベル分けと各グループのメンバー数をまとめたものである。例えば2014年のテストにおいては、単純合計で 4 ～ 11点の者が下位群、12 ～ 16点が中位群、17 ～ 30点が上位群とした。IRT によるレベル分けでは能力推定値が -3.239 ～ -1.619を下位群、-1.612 ～ -0.267を中位群、-0.264 ～ 2.523を上位群、さらに LRT では Rank 1を下位群、Rank 2 を中位群、Rank 3 を上位群と分類した。上位・中位・下位群のメンバー数の違いは単純合計得点で同点だった者がかなりいたためである。一方、分類方法が違ってもメンバー数は比較的均一にするようにしたので、2014年、2015年のテス

Table 4：単純合計、IRT、LRT によるレベル分けと各グループのメンバー数

		2014年			2015年			2017年		
		最小値	最大値	メンバー数	最小値	最大値	メンバー数	最小値	最大値	メンバー数
下位群	単純合計	4点	11点	163名	4点	12点	120名	7点	15点	43名
	IRT	-3.239	-1.619	163名	-3.242	-1.518	120名	-3.035	-1.246	43名
	LRT	Rank 1		163名	Rank 1		120名	Rank 1		47名
中位群	単純合計	12点	16点	153名	13点	17点	150名	16点	20点	48名
	IRT	-1.612	-0.267	153名	-1.499	-0.021	150名	-1.228	-0.057	48名
	LRT	Rank 2		153名	Rank 2		150名	Rank 2		48名
上位群	単純合計	17点	30点	142名	18点	28点	109名	21点	35点	47名
	IRT	-0.264	2.523	142名	-0.019	2.387	109名	0.024	3.402	47名
	LRT	Rank 3		142名	Rank 3		109名	Rank 3		43名

Table 5：分割方法の違いによるグループ間の移動と影響を受ける受験者の率

		2014年	2015年	2017年
IRT ↓ 単純合計	下位群 → 中位群	13	7	2
	中位群 → 下位群	13	7	2
	中位群 → 上位群	6	9	2
	上位群 → 中位群	6	9	2
	影響を受ける受験者の合計	38	32	8
	影響を受ける受験者の率	8.3%	8.4%	2.9%
LRT ↓ 単純合計	下位群 → 中位群	21	14	9
	中位群 → 下位群	35	10	5
	中位群 → 上位群	8	12	6
	上位群 → 中位群	10	24	2
	影響を受ける受験者の合計	74	60	22
	影響を受ける受験者の率	16.2%	15.8%	15.9%

トでは各グループのメンバー数が同じとなった。しかし、2017年のテストでは、IRT では単純合計のグループメンバー数に合わせることができたが、単純合計の同点の者を違うグループに振り分けることができないため、LRT のメンバー数では下位群と上位群において4名の差が生じた。

Table 5 は、各年度のテストを単純合計、IRT、LRT で3グループにレベル分けを行った際に、レベルの振り分けにどの程度の差異が出るかを表にまとめたものである。例えば2014年のテスト受験者を IRT で3分割した際に下位群に振り分けられていた受験者のうち13名は単純合計では中位群に振り分けられてしまう。同じテスト受験者を LRT で3分割した場合、下位群に振り分けられた受験者のうち21名は単純合計では中位群に入ることになる。単純合計との差異は IRT よりも LRT の方が大きいことが分かる。

LRT で分割した場合に影響を受ける受験者の率はどのテスト間でも大きな差はなく、15%～16%となっている。IRT での結果が約3%～8%程度であったため単純合計でレベル分けをする場合とそれほど変わらないとの見方もできる。一方、LRT での結果は、木村（2013）の結果同様、決して無視できる値ではないと言うべきであろう。仮に1000人受験者がいたとして、その内15%とすると150人が影響を受けることになる。これは入学試験のような合格・不合格という2分割の場合、合格するはずの75名が不合格となり、不合格となるはずの75名

Table 6：各テストの正答率平均値と点双列相関平均値

	2014年	2015年	2017年
正答率平均値	.393	.443	.526
点双列相関平均値	.382	.369	.458

Table 7：難易度、または識別力に問題がある項目数

	2014年	2015年	2017年
易しすぎる項目の数（正答率 .700以上）	0	2	9
難しすぎる項目の数（正答率 .300未満）	10	9	4
識別力が低い項目の数（点双列相関 .25未満）	8	4	1

が合格と判定されることを意味する。

3.4. 項目難易度と識別力

Table 6 は各テストにおける全項目の正答率平均値と点双列相関係数の平均値を示している。点双列相関係数は -1 ～ 1 の値で表され、各項目がどれだけ成績上位者と下位者を識別できるかを示す指標である。特定の項目においてテスト総得点が高い者ほど正解率が高く、低い者ほど正解率が低いようなとき点双列相関係数は高くなり、その値が高ければ識別力が高いと判断される。なお、この項目分析も「古典的分析法」と呼ばれるものの 1 つである。

基礎的な文法事項に絞って作成した2017年のテストは正答率平均値も点双列相関平均値も高くなっている。対象が変われば数値が変わるので、単純に比較することはできないが、正答率平均値が他と比べて高いのは、2017年の問題が、筆者らが意図した通り、比較的易しかったことが原因かもしれない。一方、点双列相関係数は項目の識別力を示すので、その平均値が高いということは、2017年のテストは他のテストと比べて全体的に識別力が高かったと推測できる。この数値も慎重に解釈するべきだが、このような比較を継続して行えば、スローラーナーに対しては、問題項目を基本的なものに絞った方がプレースメントテストとしての機能が高くなる可能性があることの根拠となり得る。

ただし、このような差が出るのは、もし受験者が難しい文法事項を学習していないとすると、当然の結果ということになる。それでも、2014年、2015年のテスト項目が高等学校の教科書の範囲を超えていないことを考えると、習ってはいるがまったく身に付いていない文法項目が少なくないことは容易に推測できる。まったく身に付いていない文法事項を項目に含めても、たいした情報が得られるわけでもないなら、テスト全体の識別力を上げるために、レベルに合った項目の選択はやはり重要である。

Table 7 は、各年度のテストを古典的分析法で見たとき、易しすぎる項目、難しすぎる項目、および識別力が低い項目の数をまとめた表である。項目難易度（正答率）や識別力の指標は対象が変われば違う値となるため、解釈には注意が必要であることはすでに述べた通りである。ただ、この表の示すところは、例えば2014年のテストでは、その年の受験者にとって易しすぎる項目は 1 つもないが、難しすぎる項目は10問あり、識別力が低い項目も 8 問あったというものである。2017年のテストは易しすぎる項目が 9 問と比較的多いが、識別力

Table 8：2014年のテストを LRT で分析したときの項目例

項目	正答率	点双列相関	項目参照プロファイル (IRP)			IRP 指標					
			Rank 1	Rank 2	Rank 3	Alpha	A	Beta	B	Gamma	C
6	0.146	0.133	0.141	0.121	0.171	2	0.050	3	0.171	0.500	-0.020
13	0.146	0.094	0.202	0.112	0.116	2	0.004	1	0.202	0.500	-0.091
21	0.212	0.272	0.183	0.189	0.259	2	0.070	3	0.259	0.000	0.000
36	0.212	0.100	0.231	0.180	0.217	2	0.037	1	0.231	0.500	-0.052
29	0.365	0.284	0.326	0.368	0.403	1	0.043	3	0.403	0.000	0.000
32	0.321	-0.100	0.383	0.334	0.246	1	0.000	1	0.383	1.000	-0.138
20	0.386	0.531	0.192	0.366	0.602	2	0.236	3	0.602	0.000	0.000
31	0.675	0.618	0.434	0.743	0.869	1	0.309	1	0.434	0.000	0.000

が低い項目は1問しかなく、テスト問題のレベルが合っていることが受験者を良く識別することに繋がる可能性がうかがえる。

Table 8は、古典的な分析による指標（正答率と点双列相関係数）に加えて、LRT で分析した際に算出される項目特性を要約する指標を、いくつかの項目を例にとってまとめたものである。以下、木村（2013，p.19）と小泉＆飯村（2010）がそれぞれの結果に基づいて数値の解釈について解説を行っているので、それらを参考にしながら、本研究の結果にあてはめて説明を試みる。

項目参照プロファイル（IRP）の Rank 1 ～ Rank 3 の数値は、その項目についての各 Rank に属する受験者が正解する確率を表している。例えば、項目6では、Rank 1に属する受験者が正解する確率は0.141であるのに対し、Rank 2の受験者が正解する確率は0.121しかなく、識別力に問題のある項目であることが分かる。

LRT 分析による正答率は、IRP 指標の Beta と B の値で見ることができる。基準となる値（木村、小泉＆飯村の研究ではどちらも0.5となっている）に最も近い潜在ランクを Beta、その時の正答率が B で表されている。Beta が高くて、B が低ければその項目は難しく、Beta が低くて、B が高ければ易しいと判断される。項目6は、Beta が3となっており、B が0.171と他と比べても低い。古典的分析法に基づく正答率も0.146と低いので、難しい項目であったと判断できる。項目番号の6、13、21、36は古典的分析法で特に正答率が低かったものである。その中で、項目13と36は、Beta はどちらも1で低いが、B の値も0.2程度と低いのでやはりこの年の受験者にとっては難しい項目であったということになる。

LRT 分析による識別力は、Alpha と A で示される。「隣り合う2つの IRP の値の差が最大となるペアの若い方の潜在ランク（木村，2013，p.19)」を Alpha，そのときの正答率の差が A で表されている。A が大きい項目は、Rank が Alpha 以上の学力の受験者と Alpha 以下の学力の受験者を見分ける力が大きい、つまり識別力の高い項目と判断される。項目6では、正答率の差が最大のペアは Rank 3と2であるため Alpha は2となり、その正答率の差を示す A、は Rank 3での正答率（0.171）と Rank 2の正答率（0.121）の差で0.050となる。項目31の A の値（0.309）と比べて項目6の識別力はかなり低い。表中の項目番号6、13、21、36、29、32は点双列相関係数でも A の値で見ても識別力が低く、特に項目32の A の値は限りなく0に近く、点双列相関係数は負の値となっている。

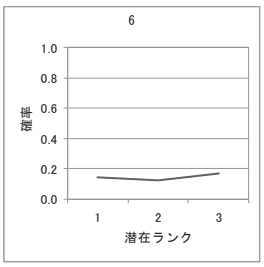
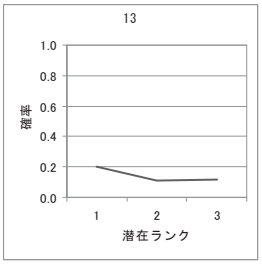
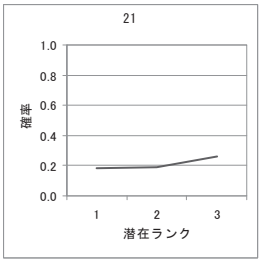
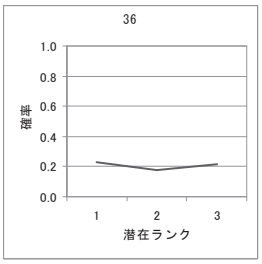
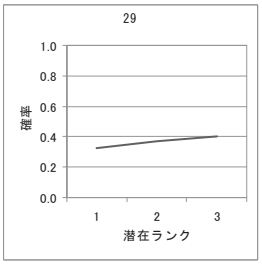
Gamma と C は、項目単調度（Item Monotonicity）を表す指標である。Rank が上昇すれば正答率も高くなる場合、そのような項目は適切であると判断できる。しかし、項目によってはどこかの Rank で下降してしまうものもある。項目単調度はその安定性を示すことになる。本研究では、Rank が 3 つあるので、Rank 1 と Rank 2、Rank 2 と Rank 3 という 2 つのペアの内、いくつのペアで下降しているかを示すのが Gamma で、項目 6 では Rank 1 と Rank 2、Rank 2 と Rank 3 という 2 つのペアの内、Rank 1 と Rank 2 においてのみ下降しているので、その割合（2 / 1 = 0.5）が Gamma、減少した値（0.121 − 0.141 = −0.020）が C の値となる。C の値が負となっているため、途中で下降したことが分り、項目としては好ましいものではないと判断できる。この項目単調度が一目で分るのが、項目参照プロファイルを基に描出された図（Table 9 の表中）である。

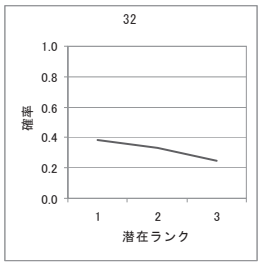
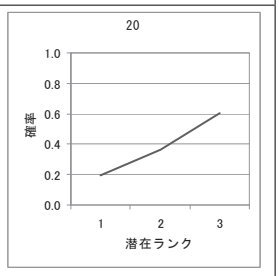
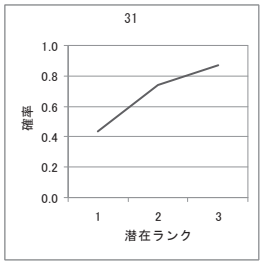
Table 9 は、Table 8 にある項目の問題本文と項目参照プロファイルに基づいて描出される図を表の形式にしたものである。図は、各 Rank（1 ～ 3）における正答率を直線で結んだ折れ線グラフになっている。一番上の項目 6 の図では、Table 8 でも見たように、Rank 1 に所属する受験者は 14.1 % の確率で正答しているが、Rank 2 の受験者はそれより少し低い確率（12.1 %）で正答していることが確認できる。識別力の高い項目であれば、項目 20 のように、Rank が高くなるにつれて正答率が上がるため、グラフは右上がりになる。Table 8 で A の値が限りなく 0 に近く、点双列相関係数が負の値となった項目はグラフ中のカーブが右下がりになっていて、下位の受験者の方が高い正答率を出し、上位の受験者の正答率が低い項目であることが一目瞭然である。このように、図を見ることでテストからの削除を検討するべき項目を見つけるのが容易になっている。

筆者らは、スローラーナーを識別するためには、学生のレベルに合った基本的な項目の方が、識別力も高くなり、適切な項目となると考えている。だから、代名詞、be 動詞と一般動詞の区別、疑問文や否定文といった、ごく初期の段階で習う文法事項が、筆者らの勤務先の学生には適していると考えていた。しかし 2014 年のテストでは、識別力の高い項目 20 と 31 は、どちらも現在完了形についての問題であった。現在完了は、スローラーナーにとってはそれほど身に付いている文法事項ではないと考えていたが、そうとも言えない結果となった。実際、項目 20 では Rank 1 に属する受験者の内 19.2 % しか正解していないものの、項目 31 では 43.4 % が正解していた。項目 31 のターゲットは現在完了形を形成する「have/has+ 過去分詞」という形を理解した上で正しい be 動詞を選択するというものである。ただし、この問題では、錯乱肢の is や was は、has との繋がり具合の不自然さから容易に除外できてしまう。さらに、being も後に –ing を伴う語が続いて不自然に感じる受験者も少なくないと考えられる。そのため、正解の been を選択した受験者が比較的多かったと推測している。実際、is を選択した者は全体の 9.6 %、was を選択した者は 12.8 %、being では 10% であった。つまり、識別力は高かったものの錯乱肢に魅力あるものがない易しい項目であったと言える。

一方項目 20 も、現在完了形（厳密には「現在完了進行形）」がターゲットとなっているが、since yesterday があるため現在完了との相性が良いことを覚えていればそれほど難しい項目であったとは言えない。実際、全体の正答率は 38.6 % であった。項目の良し悪しについて言えば、正解の has been raining を選択した受験者が 38.6 % であったのに対し、rains、

Table 9：2014年テストの項目例と項目参照プロファイル（図）

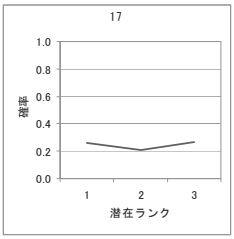
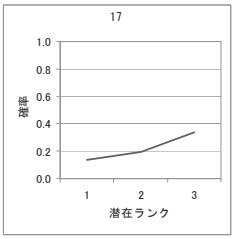
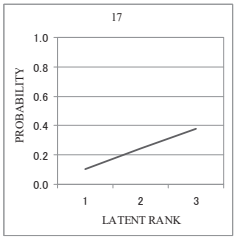
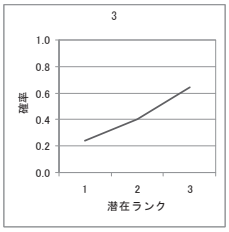
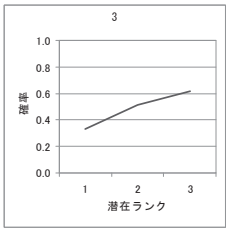
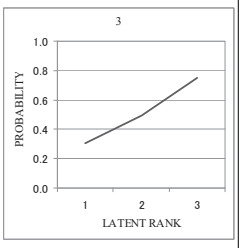
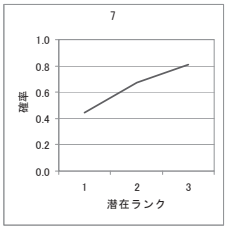
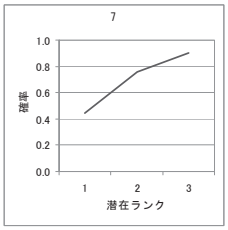
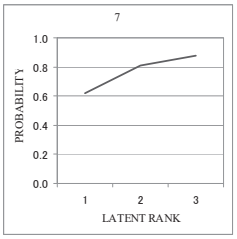
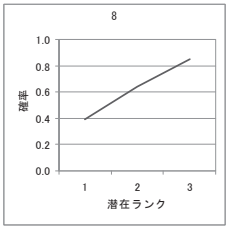
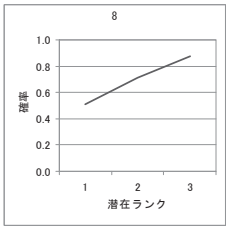
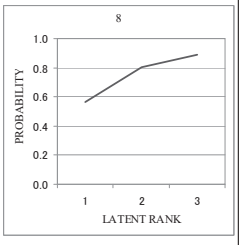
	項　　目	項目参照プロファイル
正答率・識別力ともに低い項目	6. I have two dogs. One is white and (　　) is black. (a) another　(b) other (c) the other　(d) it	
	13. My sister (　　) in 2008. (a) is married　(b) has married (c) got married　(d) has been married	
	21. I (　　) go out for a walk tomorrow. I'm not sure. (a) maybe　(b) will (c) can　(d) might	
	36. We remember (　　) this lady somewhere before. (a) seeing　(b) see (c) to see　(d) seen	
正答率はそれほど低くないが識別力が低い項目	29. There are three red balls and one yellow ball in the bag. Take (　　). (a) red　(b) red one (c) a red one　(d) the red one	

	項　　目	項目参照プロファイル
正答率はそれほど低くないが識別力が低い項目	32. (　　) do you study at home? ― From nine to eleven. (a) How many　(b) Where (c) When　(d) Why	
正答率はそれほど低くなく識別力も高い項目	20. It (　　) like this since yesterday. (a) rains　(b) is raining (c) will be raining　(d) has been raining	
正答率・識別力ともに高い項目	31. Tom has (　　) playing tennis for three hours. (a) is　(b) was (c) been　(d) being	

is raining、will be raining を選択した者がそれぞれ14％、29.5％、17.9％で、is raining を選択した者が少し多かったものの、錯乱肢もほぼ機能していたようである。Rank 1、2、3の正答率が、それぞれ19.2％、36.6％、60.2％と上がって下位・中位・上位の受験者をうまく識別できていることからも識別力が比較的高い良い項目であったことが分かる。現在完了の問題にも様々なパターンがあるため、たった2例では結論を出すことはできないが、少なくとも上の例から、現在完了形は筆者らが担当する学生のレベル分けに適した項目になり得ると予想できる。

筆者らが、基本的な文法事項に絞った、レベルの合った項目の方がスローラーナーには向いていると考える理由は以下のような結果による。Table 10は、2014年、2015年、2017年、すべてのテストで使われている項目例と項目参照プロファイル（図）である。項目17は、正答率も低く、2017年のグラフでは緩い右上がりの直線が見られるものの、全年度で識別力も低い項目である。扱っている文法事項は、受動態の正しい形を選択させるというもので筆者らは受験者にとって比較的難易度の高い項目であると考えている。各テスト中に受動態に関する問題が1問しかなかったので、受動態の問題が必ずしも難易度が高くなるとは言えないが、似通った結果が継続して出るようなら、テストからの削除を検討しても良い項目と考えられる。

Table 10：全年度のテストに共通の項目例と項目参照プロファイル（図）の変化

	2014年	2015年	2017年
17. SEIKO watches ( ) in Japan. (a) are made (b) made (c) make (d) are making			
3. Mike ( ) the piano. (a) play (b) plays (c) playing (d) have played			
7. Mary ( ) sleeping when I called her last night. (a) am (b) is (c) was (d) were			
8. ( ) does your father go to work? — By train. (a) How (b) How much (c) Which (d) Who			

これに対して項目 3、7、8 は、概ね正答率が比較的高く、グラフの傾きも右上がりになっているので、受験者の識別がほぼできていることが分る。これらの項目はそれぞれ、述語動詞の 3 人称単数現在形を選ばせるもの（項目 3）、過去進行形における正しい be 動詞を選ばせるもの（項目 7）、手段を問う際の疑問詞を選ばせるもの（項目 8）と、難易度の低いものである。それは、Rank 1 に属する受験者の正答率が20％を下回るものがないことでも確認できる。Table 8 で見た項目 6、13、21、36 と比べても、かなり基本的な文法事項を扱っ

た項目である。Table 9でも確認できるように、項目参照プロファイルは受験者集団によって変わってしまうので、長期に渡って同様の項目分析結果を見る必要があるが、少なくとも上の3、7、8の結果は、比較的容易な文法事項を扱った項目の方が、スローラーナーにとって概ね良いテスト項目になり得ることを示唆しているのではないだろうか。

4．結び

本研究は、プレースメントテストの採点方法を変えたり、項目を受験者のレベルに近付けることで違いが生じることを確認しようというものであった。単純合計によるレベル分けは、単に正解数を基に判断しているため、必ずしも受験者の能力差による分割ができていない。そこで、単純合計によるレベル分けと比較するためにIRTとLRTを使用して受験者の3分割を行った。3つの方法によるレベル分けは、相関は高いものの、メンバーの振り分けについては、少なからず差が生じていた。その差は単純合計による分割とLRTの間で特に顕著であった。IRTでの分析は、サンプルサイズが小さかったために今回の結果から導き出せるものはないが、LRTによるレベル分けが以下の点で優れており、プレースメントテストの分析に適していることが確認できた。1）難易度と項目識別力を考慮してレベル分けができること、2）信頼性が疑わしい1点刻みの得点ではなく任意のランクで分けられること、3）正答率、識別力、項目単調度という指標で詳しく項目の分析ができること。

さらに、単純合計による得点算出があてにならない理由の1つとして、項目間の難易度の差も挙げられる。難しすぎる項目が大半を占めるようなテストでは、習熟度の低い受験者をうまく識別できないということは容易に推測できる。したがって、本研究はLRTの項目分析を用いて、レベルに合った項目をテストに入れることで識別力が上がる可能性があるかどうかを確認しようとした。これは、身に付いていない学習内容をテストに含めても、できないのが当然であり、無駄な項目となってしまうことを改めて確認したに過ぎない。しかし実際、プレースメントテストとして使用されているテストのほとんどには網羅的に文法項目が含められており、受験者のレベルを予測して項目の選定が行われているというケースはあまりにも少ない。言い換えれば、テスト作成にあたって、どの程度までを基本的文法事項と捉えるかが考慮されることは滅多にないようである。むしろ、プレースメントテストの目的でTOEICやTOEFLなどがそのまま使用されている場合が多い。それが、受験者のレベルに合っていれば問題ないが、受験者の大半がスローラーナーであったなら、そうしたテストが「合わない」のは当然である。スローラーナーにとって難解極まりないそうしたテストが、学習者を語学学習から遠ざける結果にならないと言えるだろうか。

語学を学ぶ理由は様々である。しかし、知識を蓄積することに重点が置かれるなら、また、その知識もテストが終わった途端に忘れてもいいようなものなら、貴重な時間は他のことに使われるべきであろう。蓄積した知識を使う機会がないとすれば、学習に対する意欲が湧かないのも無理はない。学習することに意義を見出せない上に強制的に覚えさせられるとしたら、語学学習に対する嫌悪感を増幅させることになるだろう。実際、津村（2010）の質問に対する自由記述回答とテキストマイニングによる調査では、英語学習への意欲を失う原

因の頻度として、「覚えることが多い」は「分らない」に次いで2位となっていた。

語学学習がコミュニケーション重視にシフトしていることは明らかである。しかし、授業は必ずしもそうなってはいない。TOEIC を採用する企業が増えれば、TOEIC 対策講座が開設され、スコアを上げることがゴールであるかのような指導が行われる。受験する側も1点でもスコアを伸ばそうとする（ただし、TOEIC は5点刻み）。これが、1点刻みで能力を表そうとするテストの負の波及効果である。LRT のように段階的な評価が行われれば、語学学習も個々の知識の蓄積ではなく、段階的に実力を伸ばす指導に繋がると考えられる。そして、その最初の段階でクリアしなければならないのは、仮定法や間接話法などの難易度の高い文法項目ではないと筆者らは考えている。初期の段階では、基本的なコミュニケーションの素地が身に付いていなければならない。例えば、疑問文が作れなければ相手の情報を聞き出すことはできないし、相手を理解することはさらに難しい。文を組み立てるためには、文の要素がどのように並べられるかを知っていなければならない。代名詞が正しく使えなければならない。さらに、be 動詞と一般動詞が区別できていなければならない。それらを単に理解しているだけではなく、すぐ反応できるように訓練しなければならない。そのような文法事項を身体で覚えていないうちに、難解な長文や問題にチャレンジさせるなどは無謀である。そのような指導は、テストやその分析・採点法が見直されない限り無くならないのではないかと思う。

本稿は、小泉＆飯村（2010）や木村（2013）の研究に倣い、単純合計、IRT、LRT を用いて授験者のレベル分けを行い、その結果を比較し、さらに項目分析によってスローラーナーに適した項目を探ろうという試みであった。筆者ら自身にとってもたいへん難解な手法を使って分析を行ったため課題も少なくない。サンプルサイズが十分ではなかったため IRT の分析結果に信頼性を欠いたことに加えて、筆者らの勉強不足もあって説明が不十分であったかもしれない。さらに、テストで使った問題項目の良し悪し、選択についても議論するべき事柄は多いと考える。また、本研究では、文法テストを用いたが、文法力だけが語学力を反映しているわけではない。筆者らはコミュニケーション能力が重要だと考えているのでスピーキングのテストでも LRT を使った同様の分析を行う必要があると考える。

引用文献

今井新悟、伊東祐朗、中村洋一、菊池賢一、赤木彌生、中園博美、本田明子、平村健勝 .（2009）.「項目応答理論に基づくテストの得点―J-CAT の得点換算・解釈・利用法について」『大学教育』6 , 93－105.

木村哲夫 .（2013）.「潜在ランク理論を用いたコンピュータ適応型テストのためのアルゴリズムの提案と実装」.『早稲田大学審査学位論文』

小泉利恵、飯村英樹 .（2010）.「ニューラルテスト理論の特徴：古典的テスト理論・ラッシュモデリングとの比較から」.『日本言語テスト学会研究紀要』13（0）, 91－109. 日本言語テスト学会 . Retrieved from http://www7 b.biglobe.ne.jp/~koizumi/JLTA 2010 _Koizumi_Iimura_NTT.pdf

大友賢二 .（2009）.「項目応答理論― TOEFL・TOEIC 等の仕組み―」.『電子情報通信学会誌』92（12）, 1008－1012. Retrieved from https://www.ieice.org/jpn/books/kaishikiji/2009/2009121.pdf

斉田智里 .（2010）.「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』(30-58). 東京 : 大修館書店 .

Shojima, K. (2007). Neural Test Theory. *DNC Research Note*, 07-02. Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/Shojima2007RN07-02.pdf

荘島宏二郎 . (n.d.).「ニューラルテスト理論」Retrieved from http://www.rd.dnc.ac.jp/~shojima/ntt/ShojimaNKK08.pdf

津村修志 . (2010).「英語学習意欲喪失の要因と英語の好き・嫌いとの関係」『大阪商業大学論集　第5巻第5号（通算156号)』27-42.

豊田秀樹 . (2002).「項目反応理論　入門編」朝倉書店

山川修 . (2008).「項目応答理論を使った学生の能力推定とそれに対応した教材選択手法の開発」. サイエンティフィック・システム研究会教育環境分科会2008年度第1回会合資料 . Retrieved from https://www.ssken.gr.jp/MAINSITE/download/newsletter/2008/20080901-edu-1/lecture-3/paper.pdf