

【研究論文】

項目反応理論による
日本語リテラシーテストの検討（2）
—到達度テストの改善—

高 橋 あすみ
松 浦 年 男

研究論文

項目反応理論による日本語リテラシーテストの検討 (2) ——到達度テストの改善——

高橋 あすみ 松浦 年男
Asumi TAKAHASHI Toshio MATSURA

目次

I 問題

II 方法

1. 到達度テストの概要
2. 受検手続き
3. 対象者
4. 分析

III 結果

1. 項目分析の結果
2. 局所独立性と一次元性の確認
3. 項目母数の推定
4. テスト情報量
5. 学科・年度によるリテラシーの比較

IV 考察

1. 到達度テストの妥当性
2. 得点と潜在特性値 θ の比較
3. 改訂版の提案

V 総合考察

1. 初年次生の日本語リテラシー
2. 本研究の課題と今後の展望

[Abstract]

Japanese Language Literacy Test Using Item Response Theory (2): Improvement of the Achievement Test

Aim: First-year students at Hokusei Gakuen University took an achievement test to evaluate their Japanese learning proficiency. The study aimed to propose a revised version of the test by employing the Item Response Theory to confirm the validity of test items. **Methods:** The achievement test consisted of four sections and 60 multiple-choice questions. The data were from 5,372 first-year university students who answered the test between 2014 and 2020. We evaluated each item (correct response rate, non-response rate, and point-biserial correlation coefficient) for local independence and one-dimensionality. Subsequently, a 2-Parameter logistic model was applied to the remaining items. **Results:** The discriminative power of each item was estimated from 0.11 to 4.67. Seven items had discriminative power under 0.4. The difficulty parameter was estimated from -2.86 to 10.03 . When their latent traits were -1.8 , test information was 15.66. **Discussion:** We proposed a revised test with 47 items using the above analysis. Questions about several proverbs and honorifics were eliminated from the revised version. However, we need to elaborate on how to set questions which are either of educational or cultural significance in the Japanese literacy test. To compare the scores in the future, the achievement test and placement test must be equal.

I 問題

前稿¹では、北星学園大学（以下、本学）で大学初年次生の日本語リテラシーを測るテ

ストの妥当性を検証する必要性について論じ、4月に実施されている日本語の入学時基礎力調査（以下、プレイスメントテスト）を取り上げ、内容を検討した。続く本研究では、

キーワード：初年次教育，日本語リテラシー，学習成果，項目反応理論

Key words：initial education, Japanese language literacy, learning outcome, Item Response Theory

プレイスメントテストと対応して、初年次生の1年間の学修成果を見るために年度末に実施されている、日本語の「初年次到達度テスト」(以下:到達度テスト)を取り上げる。

到達度テストの妥当性を検討するために用いるのは、前稿と同様に項目反応理論 (Item Response Theory: 以下, IRT)である。IRTは、テスト項目に備わっている性質である項目母数(識別力 a と困難度 b)と、テストを受検する者の能力である潜在特性値 θ とを分けて推定し、検討することのできる理論である。IRTを適用するための項目分析と、推定される項目母数から、テストの信頼性であるテスト情報量を算出し、テスト項目それぞれの妥当性や、テストそのものがどのような対象者に適しているのかを検証することができる。日本語プレイスメントテストは、IRTに基づき元の70問から50問に構成しなおすことを提案した。60問から成る到達度テストについても、IRTによって妥当性の高い項目を検証し、改善を提案することが本研究の目的である。

II 方法

1. 到達度テストの概要

到達度テストは2013年度に100問を30分で解く形式で施行が始まった。2014年度からは60問を20分で解く形式に変更となっている。問題の構成は Table 1 のとおりであり、分野が(1)漢字、(2)語彙、(3)文法・敬語、(4)読解の4つに分かれている点はプレイスメントテストと同様である。相違点は、読解分野に「短文読解」以外の、「長文読解」と「図表読解」が含まれている点である。短文読解は段落のない文章であり、長文読解は4段落から成る文章である。図表読解は、クロス集計表を読み解く問いである。すべての問題は適切な解答を選択肢から選んでもらい、正答を1点、誤答を0点で採点する。得点は、

60問の合計点に10/6を掛けて、100点満点に換算した得点を算出する。本研究では素点のまま分析に用いた。

2. 受検手続き

受検対象である本学の初年次生全員に対して、2月初頭に大学から到達度テストの受検に関するメールが送られる。対象者は、決められた受検期間(約2か月間)に、本学の学習管理システム Moodle 上で、20分以内に解答を送信することが求められた。本研究へのデータ利用については、北星学園大学研究倫理審査委員会の承認を得た(21-第23号)。

3. 対象者

対象者は2014～2020年の間に到達度テストの解答を送信した4年制大学8学科の6,412名であった。そのうち、全問題に未受検の1,040名のデータを除外し、5,372名(有効数83.8%)の解答データを分析した。対象者の学科と受検年度の内訳は Table 2 のとおりであった。

4. 分析

前稿と同様の分析を行った。まずは、テストの信頼性に影響を与える項目を検討するために、各項目の正答率と未回答率を算出した。次に、対象者を60項目の合計得点の低い順から並べ、人数が最も均等な割合に分かれる5群に分けて各項目の正答率を算出し、得点の低い群から高い群にかけて項目の正答率が増加するかどうかを確認した。また、点双列相関係数を算出し、0.20以下の項目を確認した。

次に、IRTを適用するための前提として、局所独立性と一次元性が成り立つかどうかを確認した。局所独立性とは、ある問題の解答がほかの問題の解答に影響を与えないことであり、読解問題は局所独立性が確保されない可能性がある。そこで短文読解(Q46～

Table1 到達度テストの構成

分野	小項目	問題番号 Q
漢字	読みと書き	1～12
	四字熟語	13～15
語彙	語義	16～33
	ことわざ・成句	34～35
文法・敬語	受身・使役・可能, 敬語	36～45
読解	短文読解	46～50
	長文読解	51～55
	図表読解	56～60

50), 長文読解 (Q51～55), 図表読解 (Q56～60) の各項目の局所独立性を確かめるため, Yen's Q3統計量を算出した。以上の分析で, テストの信頼性に影響を与える項目を除外した後, 残った項目について次元性を確認するために, スクリーンプロット基準と因子寄与率, 信頼性 α 係数を確認した。

最終的に抽出された項目に対して, 2母数ロジスティックモデル (2-Parameter Logistic Model: 以下, 2PL) を適用し, 各項目の識別力 a と困難度 b を推定した。その後, 推定した項目母数を用いて到達度テストのテスト情報量, および対象者の潜在特性値 θ を算出した。

さらに, 到達度テストの合計点と潜在特性値 θ が, 対象者の所属学科と受検年度によって異なるかどうかを検討するために, 一元配

置分散分析を実施した。有意であった場合は, 多重比較を行った。

基本的な集計と分析は Excel で行った。IRT の適用には EasyEstimation(熊谷, 2013) を用いた。Yen's Q3統計量は R の `irt` パッケージを用いた。分散分析には HAD(清水, 2016) を用いた。

III 結果

1. 項目分析の結果

60項目の合計得点の平均は $M = 36.13$ ($SD = 8.68$) 点で, 最低0点, 最高57点であった。平均正答率は60.2% ($SD = 0.15$) であり, 0～95%の幅をとった。項目の正答率が85%以上あったのは6項目 (Q2, 8, 9, 13, 16, 43) で, 正答率が15%以下の項目はなかった。

テスト全体の平均未回答率は2.5%で, Q1では未回答率が0.3%であったが, 最終問題のQ60では15.2%であった。

60項目の合計点が小さい順から対象者を並べて, 人数が最も均等な割合になる5群に分けたところ, L群は0～29点 ($n = 1,119, 20.8\%$), LM群は30～35点 ($n = 1,157, 21.5\%$), M群は36～39点 ($n = 1,009, 18.8\%$), MH群は40～43点 ($n = 978, 18.2\%$), H群は44点以上 ($n = 1,109, 20.6\%$) にそれぞれ分かれた。各群の正答率を項目ごとに算出し,

Table2 到達度テスト受検者の人数の内訳

学科/受検年度	14	15	16	17	18	19	20	合計
英文	124	110	103	112	107	95	107	758
心理・応用コミュニケーション	90	92	96	70	87	77	90	602
経済	164	145	144	143	126	121	163	1,006
経営情報	90	114	95	117	81	78	86	661
経済法	112	120	111	107	110	106	97	763
福祉計画	84	84	95	71	80	68	90	572
福祉臨床	82	82	78	92	73	71	92	570
福祉心理	63	60	60	57	54	70	76	440
合計	809	807	782	769	718	686	801	5,372

L 群から H 群にかけての単調な増加が見られるか確認したところ、5 項目 (Q8, 35, 36, 39, 56) が該当しなかった。

次に、各項目の点双列相関係数を算出した。幅は $-0.07 \sim 0.48$ で、 0.2 以下の低い値を示したのは 10 項目 (Q8, 34, 35, 36, 39, 40, 42, 47, 52, 56) であった。ここまで算出した各項目の正答率と未回答率、群別の正答率、点双列相関係数を Table 3 に示した。

以上の基準の一つでも該当した項目は 15 項目あり、そのうち二つ以上の基準に該当したのは 5 項目 (Q8, 35, 36, 39, 56) であった。以降は、二つ以上の基準に該当した 5 項目を除外して、55 項目で分析することとした。

2. 局所独立性と一次元性の確認

3 種類の読解問題の局所独立性を確認した結果、短文読解 (Q46 ~ 50) の 5 項目について Yen's Q3 が 0.2 以上の項目はなかった。長文読解 (Q51 ~ 55) については、Q53 × Q55 (Yen's Q3 = $.26$) の間で局所独立性の成り立たない可能性が示された。また、図表読解 (Q57 ~ 60) については、Q58 × Q60 (Yen's Q3 = $.20$) の間に局所独立性の成り立たない可能性が示された。

長文読解の Q53 は、第三段落の「これ」という指示語の内容を答えさせる問題であり、Q55 は文章全体の趣旨を解答させる問題であった。二つの問いの正答率は長文読解の他の問題と比べて高く、Q53 が誤答でも、その他の文章を理解すれば Q55 の解答に必ずしも影響を与えないと考えられたことから、長文読解の項目はそのまま分析に残した。

それに対して図表読解の Q58 と Q60 は、クロス集計表から読み解かなければならない内容が部分的に同じであり、Q58 での読解や加算を誤った場合には Q60 でも誤答になる可能性があった。したがって、影響を受ける可能性のある Q60 を除外することとした。

以上の分析で残った 54 項目の一次元性を

確認した。スクリープロットを Figure 1 に図示した。第 1 因子の寄与率は 12.70 、第 2 因子は 2.58 、第 3 因子は 1.79 であり、第 1 因子で寄与率が十分高く、第 2 因子以降の寄与率の変化がゆるやかであることが確認できた。クロンバックの信頼性 α 係数は 0.86 であった。以上から、54 項目で一次元性が確保できると判断し、残った 54 項目に 2PL を適用することとした。

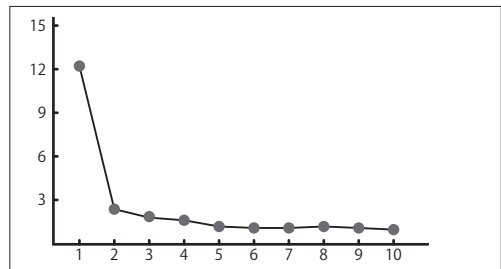


Figure1 54項目のスクリープロット

3. 項目母数の推定

54 項目に 2PL を適用し、各項目の識別力 a と困難度 b を推定した (Table 4)。識別力 a は $0.11 \sim 4.67$ の範囲をとり、平均 0.95 と全体的に十分な識別力を示した。識別力が 0.4 以下を示したのは 7 項目 (Q34, 40, 42, 45, 46, 47, 52) であった。困難度 b は $-2.86 \sim 10.03$ の範囲をとり、平均は -0.39 であった。困難度は Q42 ($b = 10.03$) と Q52 ($b = 5.42$) が外れ値と言えるほど高く、この 2 項目は点双列相関係数の値も識別力も低かった。この 2 項目を除いた困難度の平均値はマイナス 0.70 であった。

4. テスト情報量

54 項目から識別力の低かった 7 項目を除外し、47 項目の項目母数を推定しなおして、テスト情報量を算出した。テスト情報量は Figure 2 のとおり左側に尖った形となり、潜在特性値 $\theta = -1.8$ の場合に 15.66 で最大となったが、最大値からテスト情報量は顕著に減少した。

Table3-1 到達度テストの項目分析（Q1～34）

Q	全体の 正答率	群ごとの正答率					未回答率	点双列相関係数
		L	LM	M	MH	H		
1	81.8%	67.1%	78.7%	83.7%	87.9%	91.5%	0.3%	0.23
2	93.3%	80.8%	92.2%	95.4%	98.0%	99.0%	0.4%	0.29
3	63.4%	42.9%	54.0%	62.5%	74.6%	83.5%	0.4%	0.31
4	80.9%	61.5%	74.1%	83.2%	89.9%	95.8%	0.5%	0.33
5	36.5%	20.0%	23.3%	33.2%	41.6%	64.0%	0.7%	0.30
6	35.7%	22.9%	26.8%	33.3%	38.0%	56.7%	0.8%	0.23
7	66.7%	44.3%	61.1%	69.2%	74.4%	84.1%	0.6%	0.29
8	88.9%	76.0%	91.2%	91.3%	91.2%	92.4%	0.7%	0.18
9	95.2%	78.8%	97.6%	98.4%	99.6%	99.6%	0.6%	0.38
10	53.5%	31.0%	45.5%	52.1%	62.0%	76.6%	0.6%	0.32
11	83.5%	61.4%	80.3%	87.9%	91.7%	95.3%	0.7%	0.33
12	82.2%	59.0%	79.4%	87.1%	89.8%	95.1%	0.6%	0.35
13	93.5%	75.2%	95.1%	97.8%	98.4%	98.7%	0.7%	0.37
14	60.9%	33.8%	51.3%	61.3%	70.0%	87.0%	0.9%	0.38
15	62.7%	33.2%	53.4%	66.5%	74.2%	86.6%	0.7%	0.39
16	95.8%	78.0%	99.1%	99.6%	99.6%	99.9%	0.8%	0.44
17	54.8%	25.0%	42.9%	56.6%	69.1%	81.5%	0.5%	0.41
18	67.4%	39.8%	64.6%	70.7%	75.3%	85.7%	0.8%	0.35
19	70.4%	39.3%	60.2%	75.9%	84.7%	92.9%	0.6%	0.43
20	67.5%	41.2%	59.0%	71.4%	78.0%	87.6%	0.8%	0.36
21	40.8%	19.5%	30.9%	37.7%	52.0%	63.8%	0.9%	0.32
22	71.3%	44.2%	64.0%	72.6%	83.2%	92.0%	0.7%	0.37
23	81.6%	52.0%	76.5%	87.7%	94.0%	97.5%	0.7%	0.43
24	74.1%	46.3%	66.9%	77.1%	85.9%	93.6%	0.8%	0.39
25	80.9%	51.6%	75.5%	88.3%	92.5%	96.2%	0.8%	0.42
26	69.6%	34.0%	61.9%	73.9%	83.7%	94.4%	0.8%	0.46
27	56.0%	29.8%	41.1%	56.7%	68.0%	84.8%	0.7%	0.38
28	47.3%	21.8%	34.1%	42.7%	58.5%	78.4%	1.1%	0.38
29	78.5%	43.5%	72.7%	85.0%	92.3%	98.4%	0.8%	0.48
30	54.3%	23.9%	45.4%	57.7%	66.9%	77.1%	1.1%	0.38
31	79.8%	46.6%	77.4%	85.2%	91.5%	96.4%	1.1%	0.44
32	69.2%	31.5%	63.4%	74.3%	83.1%	92.6%	1.1%	0.45
33	53.8%	29.2%	50.1%	56.3%	62.4%	69.9%	1.1%	0.29
34	35.3%	23.5%	27.9%	31.5%	41.0%	51.6%	1.1%	0.20

注) 太字：項目分析の基準に該当した項目

Table3-2 到達度テストの項目分析 (Q35 ~ 60)

Q	全体の 正答率	群ごとの正答率					未回答率	点双列相関係数
		L	LM	M	MH	H		
35	38.5%	31.2%	32.3%	38.3%	38.2%	50.4%	1.2%	0.13
36	21.7%	21.0%	20.7%	20.8%	23.0%	21.6%	1.6%	0.00
37	66.5%	36.5%	58.9%	68.5%	77.9%	88.4%	1.5%	0.38
38	76.9%	41.9%	73.8%	83.9%	87.2%	93.2%	1.8%	0.41
39	20.0%	20.7%	16.9%	22.3%	17.2%	21.1%	1.9%	0.00
40	38.7%	24.8%	35.3%	35.9%	43.4%	51.1%	1.8%	0.18
41	62.6%	34.1%	53.6%	64.3%	73.3%	84.0%	1.9%	0.34
42	19.7%	13.5%	17.3%	17.9%	22.2%	25.8%	2.1%	0.09
43	88.2%	57.3%	88.7%	95.3%	96.0%	97.1%	2.0%	0.42
44	73.1%	43.2%	68.8%	75.0%	84.9%	87.3%	2.4%	0.34
45	43.9%	25.6%	40.1%	43.5%	46.4%	58.6%	2.8%	0.21
46	40.3%	26.2%	33.4%	36.3%	40.2%	60.2%	2.7%	0.20
47	30.3%	15.8%	26.8%	31.8%	33.9%	39.4%	3.1%	0.16
48	50.9%	25.5%	39.8%	52.9%	59.3%	70.3%	3.5%	0.30
49	58.0%	23.7%	45.7%	60.2%	71.4%	82.5%	3.3%	0.40
50	48.8%	26.6%	40.4%	48.4%	52.9%	67.3%	4.0%	0.26
51	78.3%	38.4%	67.6%	80.6%	92.8%	97.6%	4.6%	0.48
52	30.7%	24.4%	26.9%	29.0%	29.6%	36.2%	5.0%	0.07
53	64.9%	31.0%	47.6%	64.3%	74.5%	90.7%	5.9%	0.42
54	36.2%	17.4%	22.6%	29.7%	40.8%	59.8%	6.5%	0.28
55	61.4%	22.2%	44.0%	60.7%	72.1%	88.0%	7.6%	0.46
56	41.8%	42.1%	38.1%	36.5%	34.6%	39.1%	8.6%	-0.07
57	78.3%	45.5%	65.2%	74.0%	80.1%	91.7%	9.5%	0.35
58	69.0%	35.9%	54.0%	61.7%	71.5%	83.7%	11.6%	0.35
59	70.6%	45.7%	55.9%	60.6%	66.5%	79.2%	13.1%	0.23
60	66.1%	32.8%	48.3%	55.5%	65.6%	79.7%	15.2%	0.36

注) 太字：項目分析の基準に該当した項目

さらに、推定した項目母数を用いて対象者の潜在特性値 θ (事後期待値) を算出し、ヒストグラムを Figure 3 に示した。 θ の平均は 0.00 ($SD = 0.93$) で、 $-2.65 \sim 2.48$ の幅をとった。

5. 学科・年度によるリテラシーの比較

到達度テストの結果から算出した潜在特性値 θ と合計得点について、受検年度、及び所属学科による違いを一元配置分散分析によって比較した。結果を Table 5 に示した。

まず学科別では、潜在特性値 θ は有意な差が確認された ($F(7, 5366) = 58.83, p < .001, \eta_p^2 = .07$)。 θ が最も高い英文学科 ($\theta = 0.38$) は、福祉心理学科 ($\theta = 0.37$) 以外の学科より有意に高かった。福祉心理学科と心理・応用コミュニケーション学科 ($\theta = 0.23$) は、英文学科以外の学科より有意に高かった。福祉臨床学科 ($\theta = 0.02$) は、他の全 7 学科との間に有意差が認められた。 θ が負を示した残り 4 学科間では、有意な差は認められなかった。

Table4 到達度テスト54項目の項目母数

Q	a	SE	b	SE	Q	a	SE	b	SE
1	0.56	0.04	-2.86	0.18	29	1.47	0.05	-1.21	0.03
2	1.24	0.07	-2.61	0.10	30	0.84	0.03	-0.23	0.04
3	0.64	0.03	-0.95	0.06	31	1.35	0.05	-1.34	0.04
4	0.89	0.04	-1.87	0.08	32	1.21	0.04	-0.85	0.03
5	0.69	0.03	0.89	0.06	33	0.52	0.03	-0.31	0.06
6	0.50	0.03	1.25	0.09	34	0.39	0.03	1.63	0.14
7	0.63	0.03	-1.20	0.07	37	0.87	0.04	-0.92	0.05
9	2.30	0.11	-2.11	0.05	38	1.09	0.04	-1.35	0.05
10	0.68	0.03	-0.22	0.04	40	0.27	0.03	1.74	0.21
11	0.95	0.04	-2.00	0.08	41	0.74	0.03	-0.77	0.05
12	1.02	0.04	-1.79	0.07	42	0.11	0.03	10.03	10.03
13	1.82	0.08	-2.11	0.06	43	1.53	0.06	-1.80	0.05
14	0.86	0.03	-0.59	0.04	44	0.80	0.04	-1.42	0.07
15	0.95	0.04	-0.65	0.04	45	0.37	0.03	0.68	0.09
16	4.67	0.27	-1.87	0.03	46	0.36	0.03	1.12	0.12
17	0.99	0.04	-0.23	0.03	47	0.27	0.03	3.16	0.37
18	0.75	0.03	-1.09	0.06	48	0.59	0.03	-0.07	0.05
19	1.09	0.04	-0.98	0.04	49	0.91	0.04	-0.42	0.04
20	0.87	0.04	-0.97	0.05	50	0.50	0.03	0.10	0.06
21	0.71	0.03	0.59	0.05	51	1.48	0.05	-1.21	0.03
22	0.96	0.04	-1.13	0.05	52	0.15	9.97	5.42	1.11
23	1.39	0.05	-1.43	0.04	53	1.05	0.04	-0.73	0.04
24	1.01	0.04	-1.25	0.05	54	0.61	0.03	0.99	0.07
25	1.32	0.05	-1.43	0.04	55	1.22	0.04	-0.51	0.03
26	1.23	0.04	-0.87	0.03	57	0.87	0.04	-1.72	0.07
27	0.92	0.04	-0.31	0.03	58	0.73	0.04	-1.24	0.07
28	0.91	0.03	0.14	0.03	59	0.40	0.03	-2.28	0.19

注) a は識別力、b は困難度、SE は標準誤差を表す

学科別の得点についても有意差が認められ ($F(7, 5366) = 53.67, p < .001, \eta_p^2 = .07$), θ に見られたのと同様の箇所に、有意差が認められた。

受検年度別でも、潜在特性値 θ に有意差が見られた ($F(6, 5366) = 12.78, p < .001, \eta_p^2 = .01$)。2020年度受検者 ($\theta = 0.21$) が他のすべての年度と比較して有意に高かった。また、2016年度受検者 ($\theta = -0.16$) は、2014年度 ($\theta = 0.01$)、2018年度 ($\theta = 0.01$)、2019年度 ($\theta = 0.07$) の受検者より

も有意に低かった。得点についても有意な差が認められ ($F(6, 5366) = 12.36, p < .001, \eta_p^2 = .01$)、 θ と同様の箇所に有意差が確認された。潜在特性値 θ と合計得点の相関係数は $r = .95$ であった。

IV 考察

1. 到達度テストの妥当性

本研究は、本学で初年次生を対象に年度末に実施している日本語到達度テストについ

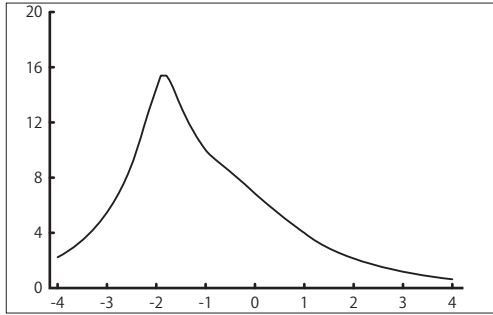


Figure2 47項目のテスト情報量

(注) 横軸が潜在特性値 θ 、縦軸がテスト情報量を表す。

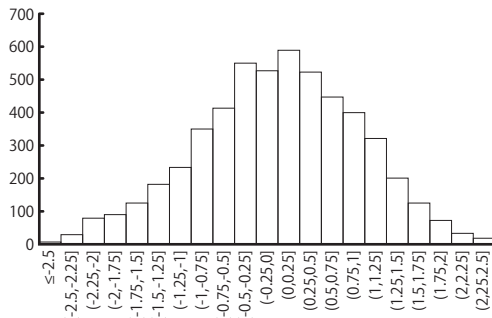


Figure3 対象者の θ のヒストグラム

(注)横軸が潜在特性値 θ 、縦軸が人数を表す。

Table5 学科・受験年度別の比較

	θ	SD	得点	SD
英文	0.38	0.86	34.42	6.82
心コミ	0.23	0.90	33.16	7.24
経済	-0.17	0.92	29.83	8.11
経営情報	-0.25	0.90	29.16	8.02
経済法	-0.17	0.88	29.72	7.97
福祉計画	-0.23	0.88	29.22	7.97
福祉臨床	0.02	0.86	31.41	7.36
福祉心理	0.37	0.93	33.92	7.66
14年度	0.01	0.95	31.14	8.17
15年度	-0.06	0.92	30.64	7.96
16年度	-0.16	0.91	29.96	8.10
17年度	-0.07	0.92	30.69	7.98
18年度	0.01	0.91	31.37	7.72
19年度	0.07	0.92	31.52	7.83
20年度	0.21	0.88	33.03	7.43

(注) 学科は略称で表記

て、IRTに基づいて妥当性を検討し、テストの改善を提案することが目的であった。

到達度テストはプレイズメントテストと比較して、問題数は10問少ないのに対し、読解問題が多いという相違点があった。そのため、終盤に行くにつれて未回答率はやはり増加し、解答時間20分で最後まで到達しなかった学生や、解答を諦める学生がいたことが推測される。さらに、テスト問題と各項目との整合性をみる点双列相関係数が、負や0の値をとるなど合計得点とは全く異なる傾向を示した難題といえる問も存在し、学生の日本語リテラシーを測るには明らかに不適切な項目も含まれていた。事前の項目分析とIRTの識別力から、IRTに適用できない項目や妥当性の低さが疑われる項目を13項目除外し、最終的に47項目が選定された。

小項目別にみると、「四字熟語」「語義」から除外された項目はなく、「読みと書き」からも除外されたのは1問のみであったことから、漢字分野や語彙分野は妥当性の高い項目で構成されていたことが分かった。一方で、語彙分野の「ことわざ・成句」は2問とも妥当性が低かった。「ことわざ・成句」の問題は、「嘴が黄色い」「鰻の寝床」ということわざの知識の有無が問われた。ことわざそのものを答える問題は、豊橋技術科学大学の日本語能力調査でも正答率が低く、その要因にはことわざの利用頻度の低さが挙げられていた(日比野, 2008)。本研究の結果からも、これらのことわざの知識の有無は、その他の項目が測定している日本語リテラシーと独立していることを示唆している。したがって、妥当性を確保しながらことわざの理解を問える

項目を検討する必要がある。例えば、あることわざを示してその意味を推測させるような形式の理解テストは、統合失調症にみられる思考障害のアセスメントへの活用も考えられており(松井・鳥居, 2007), 読解力を問う項目に類する可能性があることから, このような形式でことわざの理解を尋ねる方が識別力が高く, テスト問題として妥当かもしれない。

また, 「受け身・使役・可能, 敬語」においては10問中5問の妥当性が疑わしかった。この小項目はプレースメントテストにおいても問題数が半減したことから, 初年次生の日本語リテラシーを測定する問題として抜本的な吟味が必要である。特に本研究の結果は一貫して, 初年次生にとって会話における敬語表現のなじみが薄いことを示している。問題を再検討する際は, 初年次生の学修という観点に立ち, 教職員とメールでやりとりする上での敬語表現など, 使用頻度が高く学生に実

用的な内容を取り入れることが望まれる。

2. 得点と潜在特性値 θ の比較

選定された47項目の合計点と, IRTによって推定された潜在特性値 θ を学科と受検年度別に比較したところ, 合計点でも潜在特性値 θ でも, 同じ群間に有意差が確認された。以上から, 今回編成しなおした到達度テストは, 十分に対象者の能力を測定できる構成になったと考えられる。

3. 改訂版の提案

改訂版の到達度テスト47項目を困難度の降順になるように並べ替え, Table 6に示した。テスト情報量から, 改訂版は潜在特性値 θ が本研究の対象者の中でも特に低い値を示した者にとって信頼性の高いテストであることが分かった。前稿¹で述べたように, 潜在特性値 θ が低い学生を高い精度で識別できることは, リメディアル教育の必要性を吟味

Table6 日本語到達度テストの改訂版47項目の提案

Q	小項目	困難度 b	Q	小項目	困難度 b	Q	小項目	困難度 b
1	読みと書き	-2.86	25	語義	-1.43	37	受身・使役・可能, 敬語	-0.92
2	読みと書き	-2.61	31	語義	-1.34	43	受身・使役・可能, 敬語	-1.80
9	読みと書き	-2.11	24	語義	-1.25	44	受身・使役・可能, 敬語	-1.42
11	読みと書き	-2.00	29	語義	-1.21	38	受身・使役・可能, 敬語	-1.35
4	読みと書き	-1.87	22	語義	-1.13	41	受身・使役・可能, 敬語	-0.77
12	読みと書き	-1.79	18	語義	-1.09	48	短文読解	-0.07
7	読みと書き	-1.20	19	語義	-0.98	49	短文読解	-0.42
3	読みと書き	-0.95	20	語義	-0.97	50	短文読解	0.10
10	読みと書き	-0.22	26	語義	-0.87	51	長文読解	-1.21
5	読みと書き	0.89	32	語義	-0.85	53	長文読解	-0.73
6	読みと書き	1.25	27	語義	-0.31	54	長文読解	0.99
13	四字熟語	-2.11	33	語義	-0.31	55	長文読解	-0.51
15	四字熟語	-0.65	17	語義	-0.23	57	図表読解	-1.72
14	四字熟語	-0.59	30	語義	-0.23	58	図表読解	-1.24
16	語義	-1.87	28	語義	0.14	59	図表読解	-2.28
23	語義	-1.43	21	語義	0.59			

するのに有用である。その一方で、初年次の学修成果を確認するという到達度テストのもう一つの役割を考えれば、本学の初年次生の大多数にとって自らの学修の到達度を測るには更なる改善の余地があると言わざるを得ない。問題の追加や選択肢の精査を通して、到達度テストの難易度を再検討する必要がある。例えば、本研究で識別力の高かった項目(Q9や16など)を参考に、問題のレベルや選択肢の内容を検討するのがよい。また、学修の到達度を測る目的を重視するなら、正課との関連性を調べて内容を検討することも重要である。

V 総合考察

1. 初年次生の日本語リテラシー

前稿¹のプレイスメントテストの結果も併せて、本学の日本語リテラシーテストから明らかになった初年次生の日本語リテラシーを見ていこう。第一に、学生の所属学科や受験年度によって、日本語リテラシーに決まった傾向が見受けられた。所属学科に関しては英文学科、福祉心理学科、心理・応用コミュニケーション学科の3学科が、他の学科より相対的に高い日本語リテラシーを有していた。その理由として、学部・学科単位で求められる能力の影響が考えられる。英文学科と心理・応用コミュニケーション学科はともに文学部で、アドミッション・ポリシーにも書かれているとおり、言葉を扱うことが求められる。もちろん英文学科は主な対象は英語であるが、汎用的な言語運用能力が求められることは想像に難くない。また、心理系の学科の日本語リテラシーが他学科より高かったのは、一般的に心理学科の人气が高く、受験生の学力が相対的に高かったことが考えられる。

受験年度による傾向をみると、プレイスメントテストも到達度テストも2020年受験者の日本語リテラシーが顕著に高かった。これ

には、2020年初頭に始まった新型コロナウイルス感染拡大の影響を受け、受験生の層が例年から変化していたことが考えられる。この変化の要因を確かめるには、他のテストにも同様の傾向が現れているのか調べる必要がある。2018年度の日本語リテラシーテストの結果は、情報の知識を問うテストの結果と有意な正の相関があった(金子, 2021)ため、情報のテスト結果の推移を確認することは、考察に役立つであろう。また、この変化が一過性の影響かどうかを確かめるためには、本研究に含まれていない2021年以降の受験者まで追っていく必要がある。もし2021年以降の日本語リテラシーも2020年と同等の高さで維持されていることが示されれば、本研究で提案した改訂版のテスト情報量が表す特性値 θ の高さと釣り合わないため、テストの再構成が必要となるだろう。

2. 本研究の課題と今後の展望

本研究の課題を以下に3点挙げる。一点目に、対象者数の問題が挙げられる。到達度テストは、プレイスメントテストと母集団が同様であったにもかかわらず、受験者は合計して1,000名以上も少なかった。そのため、到達度テストの改訂の過程には、本来対象とする初年次生の6分の1が含まれていないことに留意する必要がある。到達度テストを受検しなかった層に性別など何らかの傾向が隠されている場合、今回の対象者と本来想定している対象集団の性質が実は異なり、改訂版が適さなくなることも考えられる。テストの妥当性の更なる改善を目指すには内容だけでなく受検手続きをも見直し、受検率を向上させる工夫が必要である。

二点目に、各項目の教育的意義の検討が不十分であることが挙げられる。到達度テストにおいて「ことわざや成句」の問いは妥当性が低く、小項目ごと削除した。しかしテストからの除外は、ことわざや成句に触れる機会

の更なる減少を招くことになる。ことわざや慣用句は、言語文化として向き合う必要があるものであり(宮城・石井, 2015), なじみが薄いという理由で扱うことを止めては, 初年次生の言語文化は乏しくなってしまうだろう。プレイスメントテストから削除された敬語表現¹も同様である。項目そのものの教育的意義とテスト全体としての釣り合いを考え, 初年次教育において内容的妥当性の高いテストを開発する努力は今後も継続する必要がある。

三点目に, プレイスメントテストと到達度テストの結果を個人に紐づけて分析しなかったことが挙げられる。到達度テストを受検している学生は全員プレイスメントテストも受検しているが, 二つのテスト結果から, 特定の個人がどのような日本語リテラシーを有していると判断されたのかは明らかではない。個人と紐づけてデータを整理し, 二つのテストから推定した特性値 θ の相関関係を確認すれば, 日本語リテラシーテストに関してより多くの情報を得ることができるとは思われる。

本研究で提案した改訂版をより精緻化する方法の一つに, 日本語リテラシーテストの外的妥当性の検証が挙げられる。例として, 本研究で推定した潜在特性値 θ が, 本学の日本語教育の成績や漢字能力検定といった外部テストの得点とどのように関連するかを調べることができる。精緻化の方法の二つ目に, プレイスメントテストと到達度テストの等化が挙げられる。具体的には, 今回の改訂版を参考に各テストから項目を抜粋して, プレイスメントテストと到達度テストで共通する項目と, 独自の項目を組み合わせた複数種類のテストセットを作成する。そして, 新しい対象者に回答してもらう共通項目法が現実的であろう。ただし, この計画には, すでに問題を解いたことのある本学の学生は参加できないため, 同等の学力水準を持つ他の大学の初年次生を対象者とするようになる。集団が異なる

と推定される項目母数も異なるため, 対象者の選定やテストセットの作成は注意深く行う必要がある。等化を通してプレイスメントテストと到達度テストを同一尺度上で比較できるようにになれば, 個人の学修成果に対する信頼性の高いフィードバックが可能になるであろう。

【謝辞】

本研究は, 2021年度北星学園大学特定研究費の助成を受けて実施したものである。

【注】

(1)高橋あすみ・松浦年男「項目反応理論による日本語リテラシーテストの検討(1)-プレイスメントテストの改善-」北星論集第60巻第1号, pp.25-30

【参考文献】

- 日比野浩信(2008) 大学生の日本語能力の現状・各論(ことわざ・慣用句・四字熟語・部首) -豊橋技術科学大学生の場合・雲雀野,(30), 121-135
- 松井三枝・鳥居幹樹(2007) ことわざ理解テストの開発. 富山大学杉谷キャンパス一般教育研究紀要, 35, 47-60
- 宮城信夫・石井勉(2015) 我が国のことわざ・慣用句・故事成語の指導について:平成27年度国語教科書5社の比較検討を通して. 教育学部紀要, 49, 103-110
- 金子大輔(2021) 大学の初年次生に対するプレイスメントテストと到達度テストの結果—初年次教育の学修成果の可視化の試み— 北星学園大学経済学部北星論集, 60(2), 34-42
- 熊谷龍一(2009) 初学者向けの項目反応理論分析プログラム EasyEstimation シリーズの開発. 日本テスト学会誌, 5, 107-118
- 清水裕士(2016) フリーの統計分析ソフト HAD-機能の紹介と統計学習・教育, 研究実践における利用方法の提案-. メディア・情報・コミュニケーション研究, 1, 59-73

