

【研究ノート】

〈報告〉北星学園大学平成11年度一般入学試験 英語多肢選択問題の分析

齋藤英敏

1 はじめに

本稿の目的は、平成11年度北星学園大学一般入学試験の英語の多肢選択問題 (multiple choice questions 以下 MCQ) の分析を行い、入学試験に示唆を与えることである。分析にはラッシュ測定モデルを用いた。まずこの分析方法について簡単に説明を行い、そして学部ごとの分析結果を示し、最後に考察を加える。

2 古典的テスト分析法とラッシュ測定法

ラッシュ測定法は、最近目覚ましい発展を遂げている項目応答理論 (item response theory) の中の最も基本的なモデルである。項目応答理論を理解するにはいわゆる古典的テスト理論 (classical test theory) の理解が不可欠である。ここでは両方の理論の概略を説明して、さらにラッシュ測定法の説明を簡単にやりたい。

MCQ の分析は、元来古典的テスト分析法を用いて行われてきた。この方法では主に次のような統計を算出することで、MCQ項目の良し悪しを判断してきた。

- (1)項目難易度(item difficulty/facility)
- (2)項目弁別力 (item discrimination)
- (3)点双列相関係数(point-biserial correlation)
- (4)信頼性係数 (reliability coefficient)

(1)は正答率と一般に呼ばれるもので、特定の項目でどれぐらいの被験者が正答したかと

いうことである。正答率の低い項目を集めたテストは当然難易度の高いテストとなる。

(2)は(1)を受験者の能力を考慮しながら考えたものである。例えば、受験者を最終的な得点ごとに4つのグループに分けるとする。つまり、一番高得点を得た上の25%、次の25%、その次の25%、下位の25%とわけたとする。弁別力は特定の項目において受験者の正答率が論理的であるかということである。例えばある項目で上位25%の正答率が最下位25%より悪かったとすると、この項目は論理的な正答率を得ていないということになる。つまり論理的な正答率を持つ項目とは、上位25%のグループが最も高く、下位25%が最も低く、その間のグループはその中間にくるといようなものである。

(3)はいわゆる相関係数であるが、これは項目ごとの被験者の出来、不出来とテスト全体での点数の相関関係を表したもので、(2)と同等にみなすことができる。この値が高い項目は良い項目である。

(4)の信頼性係数には幾つか種類があるが、いずれも被験者の解答の一貫性でテストの信頼度をみるものである。考え方としては、身長を測定するとき何度測定してもほぼ同じ結果がでるといような信頼性のあるテストを作りたいのである。基本的にはそれは、テストの真値 (true score) と実際の測定値 (observed score) の相関係数で得られる。最もよく用いられる α (アルファ) 係数や KR20 は実際にはテストを繰り返し行うことなく、そのテストに含まれる項目の標準偏差などを

用いて、テスト内の内的一貫性 (internal consistency) をもとに信頼性を算出する。

このような古典的テスト分析方法は大変有用で、これからも用いられ続けるだろうが、いくつかの問題点が指摘されている。まず第一に、信頼性係数が実際にはテストの特性として保持されるわけではなく、被験者集団によって変わることがあげられる。これは信頼性係数が被験者の項目における解答をもとに算出されるので避けることができない。またあるテストの信頼性が特定の被験者集団にだけ低い値が得られれば、そのテストはその被験者集団には向いていなかったということになる。

次に、二つの異なる、しかしながら同等とみなされる試験があって、それをあまり期間をおかずに受けたとする。最初と次の試験の点数が大きく異なった場合その人の能力はどう判定すべきであろうか？これが意味することは、結局のところ、人の能力はそのテストに含まれる項目によって異なって判定されるということである。これはテスト項目依存性と呼ばれる特性で、測定される能力はテストの項目から独立していない。

ラッシュ測定法では、古典的テスト分析法の持つこれらの問題点を解決し、さらに以下に述べる特性を持つ。第一に項目依存性であるが、ラッシュ測定法では項目の難易度などをあらかじめ決定して、能力の算定時に固定するのである。すると誰がそのテストを受けても項目の難易度と信頼性などは変化がしえない。また異なった項目を持つ同等なテストが既知の難易度をもとに容易に作れるし、異なった項目の試験でも、同一人物が受ければ同じ能力が算定される。

またラッシュ測定法では、項目弁別力是一定とする前提をもっている。つまりどの項目においても能力の高い人は能力の低い人よりできるということである。この前提が成り立たない項目は不適合 (misfit) 項目として取

り除くことが可能である。

さらにラッシュ測定法では、被験者の能力と項目難易度をロジットという同一単位で表示するので被験者と項目が同一スケール上で比較が可能という特質を持っている。ロジットはある人がある項目で正答得る確率を、同じ項目で正答を得られない確率で割ったものを自然対数で表したものである。

3 分析方法

前述のラッシュ測定法を用いて、それぞれ学部ごとに MCQ の分析を行った。ラッシュ測定法では、主に次のような統計を用いて MCQ 項目およびテストの良し悪しを判断する。

- (1)項目難易度 (item difficulty/facility)
- (2)内的適合性 (infit meansquare)
- (3)外的適合性 (outfit meansquare)
- (4)信頼性係数 (reliability)

(1)は古典的テスト分析法と共通であるが、ここでは前述のようにロジットという単位を用いて表示し、被験者の能力も同様の単位で表す。

(2)はラッシュモデルと実際のデータとの差からできる残差値 (residuals) を用いた統計指標で、項目難易度の値の近辺の解答のモデルへの適合性をみる。これは次の図1を見ると理解しやすい(図1は理想的モデルでこのようなことは現実にはありえない)。この図は縦軸に最も簡単な項目から最も難しい項目を並べ、横軸に最も高得点の被験者から低得点の被験者を並べたものである。1は正答、0は誤答を示している。内的適合性は、項目難易度を決定する場所あたりの解答が適切かどうかを示す。例えば図1の第二項目では20人の被験者のうち16人が正答をしているが、図2の現実的なモデルの第二項目では、この左から数えて16人目あたりで少し変動がある。このような変動を統計的に示しているのがこの指標である。この値が大きい場合、項目に

なんらかの問題があり項目難易度を決定するのが困難になる。尚, Wright & Linacre (1994)により, 入学試験などの場合はこの値は0.8から1.2の間にくるべきであると提案されている。

(3)も(2)と同様に残差値を用いた指標であるが, これは名前通り外側の解答パターンの適切さを示すものである。例えば図2の最も簡単な項目では, 両端の解答にはほとんど変動がない。しかし図2の4番目の項目では能力の高いものが間違えていたり, 能力の低いものが正答していたりする。このような場合, 外的適合性指標の値が高くなる。これも内的適合性指標同様, 0.8から1.2が適切な値である。尚, 両指標とも, 平均0の標準化z点値を伴い, これが2.0以内であれば有意差はないので適切となる。

最後に(4)であるが, これは古典的テスト法同様真値と測定値の相関値である。ラッシュでは項目と被験者の能力を別々に算出するためそれぞれ信頼性係数がある。

4 被験者

文学部英文科受験者は533名, 経済学部2学科は各816名, 495名, 社会福祉学部3学科は各329名, 302名, 397名であった。

項目	被験者	
	EASY	HIGH LOW
		11111111111111111100
		1111111111111111110000
		111111111111111110000000
		111111111111111000000000
		111111111111000000000000
		111111110000000000000000
		111110000000000000000000
		110000000000000000000000
	HARD	110000000000000000000000

図1: 項目と被験者の関係を示した理想的なモデル

5 試験問題

問題は学部ごとに異なる。文学部は17問がMCQで残り14問が記述解答である。MCQの第1問から第5問までは短文を聞いて英語の質問に答える問題(以下LSG1-5と表示)であった。第6問から第10問は少し長い対話を聞いてやはり英語の質問に答える問題(以下DLG1-5)であった。第11問から16問は6パラグラフ程の文を読んで, 英語の質問に答える問題(以下RDG1-6)であった。第17問は1パラグラフの文の空所補充で, これだけ10の選択肢があった(以下FLL1)。他は全て4つの選択肢があった。

経済学部は9問がMCQで, 7問が記述解答であった。MCQの最初の2問は, 4パラグラフ程の文を読んで空所を補充する問題(以下FLL1)と, 本文にでてきた単語のアクセントの問題(以下ACT1)であった。第3問から第7問は, やはり4パラグラフ程の文を読んで英問英答する問題(以下RDG1-5)。最後の2問は, 対話の中の文の語順正序問題である(以下ODR1-2)。選択肢数はFLL1が5択, ACT1は6択, RDG1からRDG5は4択, ODR1とODR2は7択であった。

社会福祉学部はやはり9問がMCQで, 9

項目	被験者	
	EASY	HIGH LOW
		111111111111101111100
		11111111011111110110100
		1111111110111111000000
		1011111111100101000010
		1111111111110000000000
		1111111101000000000000
		1111101100010000000000
		1111100001000000000000
		1011000000101000000000
	HARD	1000100010000000000000

図2: 項目と被験者の関係を示した現実的なモデル

問が記述解答であった。MCQの最初の2問は、3パラグラフ程の文のなかの語句と近い意味の語を選ぶ問題(以下MNG1)と、同じ文に出てくる語のアクセントの問題(以下ACT1)。第3問から第7問は、やはり3パラグラフ程の文を読んで英問英答する問題(以下RDG1-5)。最後の2問は、対話の中の文の空所補充である(以下FLL1-2)。選択肢数はRDG1からRDG5が4択で、他は5択である。尚、具体的な内容は「北星学園大学入試ガイド2001年版」を参照されたい。

6 結果

信頼性

表1のようにいずれも項目信頼性は.99で高い信頼性が確認された。しかし被験者信頼性は.64 .49 .45といずれも低い。この主な理由は、いずれの学部でも問題に使用された項目では被験者の能力をカバーしきれていないということである。図3, 4, 5はこの傾向を示している。この三つの図は左側に被験者、右側に項目をそれぞれ能力値、項目難易度、によって配置したものである(単位はロジット)。被験者は最も能力の高いものが図の上で、下に行くほど低くなる。項目は最も難しいものが図の上で、下に行くほど易くなる。文学部では項目が比較的被験者全体に対応して、広く分布している。しかし経済学部、社会福祉学部では全体の受験者のおよそ半分の者の能力を測る項目しかない。つまり残り半分のものにとっては、このMCQ項目は簡単過ぎたということである。これは両学部ともおよそ50名ずつのMCQ満点の者がいることから明らかである。

図3, 4, 5でcutoff pointとreal cut-

offとあるのは、それぞれ定員数の位置と実際の合格者数の位置である。例えば文学部では、定員が71名で実際の合格者は178名であった。

図3でわかるように、実際の合格者数の位置にはテスト項目がない。しかし文学部はその前後、また全体に項目が広く分散している。問題なのは図5の社会福祉学部である。図5では実際の合格者数の位置に全く項目がなく、また全ての項目がその下のほうに集まっている。つまりMCQはほとんど合格者弁別の役にたっていない。図4の経済学部では実際の合格者数の位置に項目が集まっており、これらの項目が合格者弁別に役に立っていることがわかる。しかし社会福祉学部同様、項目は広く分散していない。

不適合項目

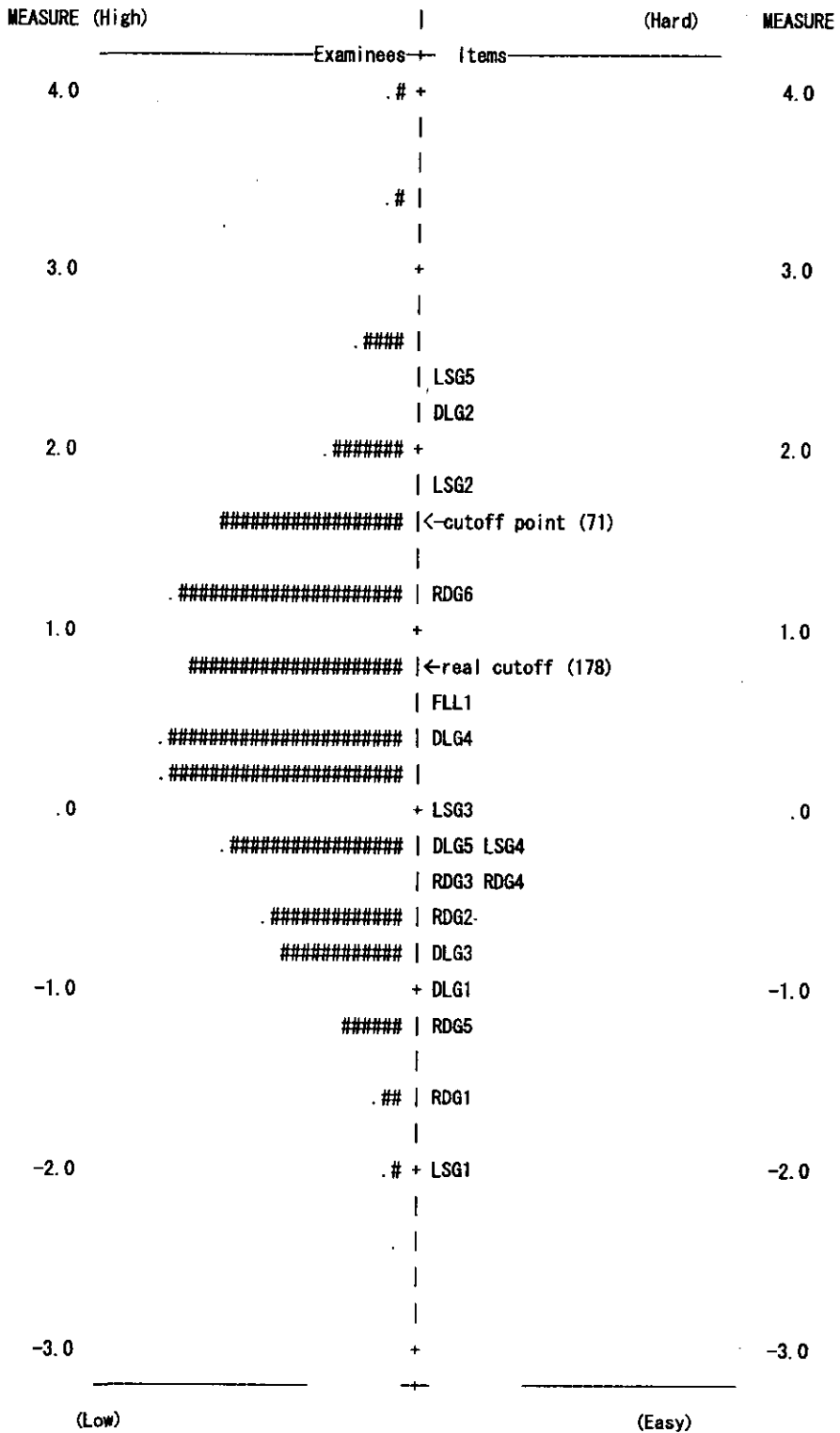
表2, 3, 4はそれぞれ最も難易度の高い項目から低いものへ並んでいる。Wright & Linacre (1994)の提案する不適合項目の基準に、4つの適合性統計指標(内的、外的適合性のMSQとZ点値)全てがあてはまる項目は見つけられなかった。ただし文学部のRDG5に、強い不適合の傾向が見られた。RDG5の外的適合性MSQ, Z点値ともWright & Linacre (1994)の基準値を上回っている。表2のように、RDG5は項目難易度-1.19を持つ、文学部MCQ全17問のなかで3番目に易い問題である。

図6は文学部の被験者の不適合解答を示したものである。前出の図1, 2のように、縦軸に最も難易度の低い項目から高い項目、横軸に最も能力の高いものから低いものへと配置してある(横軸の数字は被験者のID番号である)。ここでは特に不適合度の高い解答のみを示している。ここで0とあるのは、正答できると予測されるのにとれなかったものが示され、逆に1は、間違えると予測されるところで、正答されたという意味である。例

表1: 信頼性係数

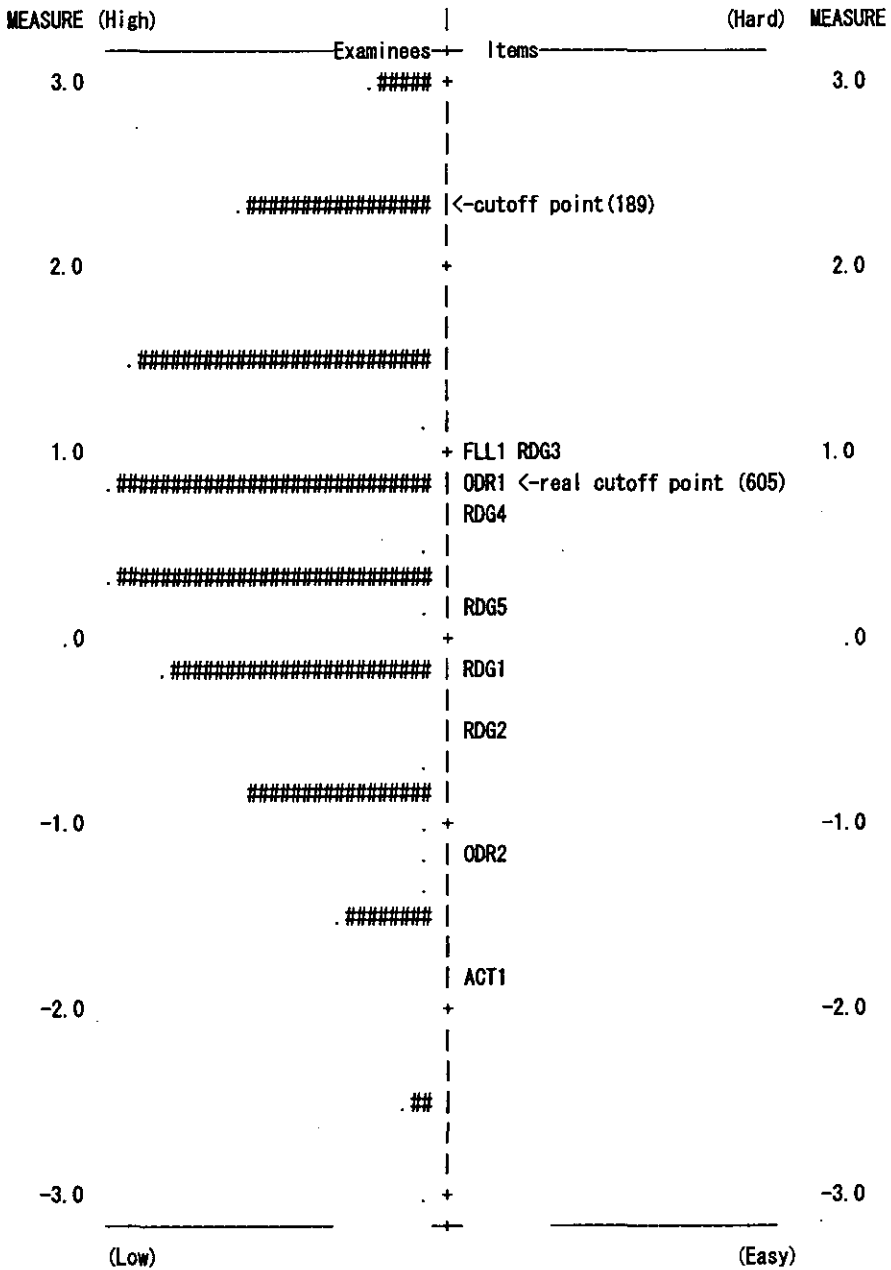
	文学	経済	福祉
項目信頼性	.99	.99	.99
被験者信頼性	.64	.49	.45

＜報告＞北星学園大学平成11年度一般入学試験英語多肢選択問題の分析



注：'#' = 3名；'.' = 1から2名

図3：文学部の項目と被験者の位置関係



注: '#' = 8名: '.' = 1から7名

図4: 経済学部的项目と被験者の位置関係

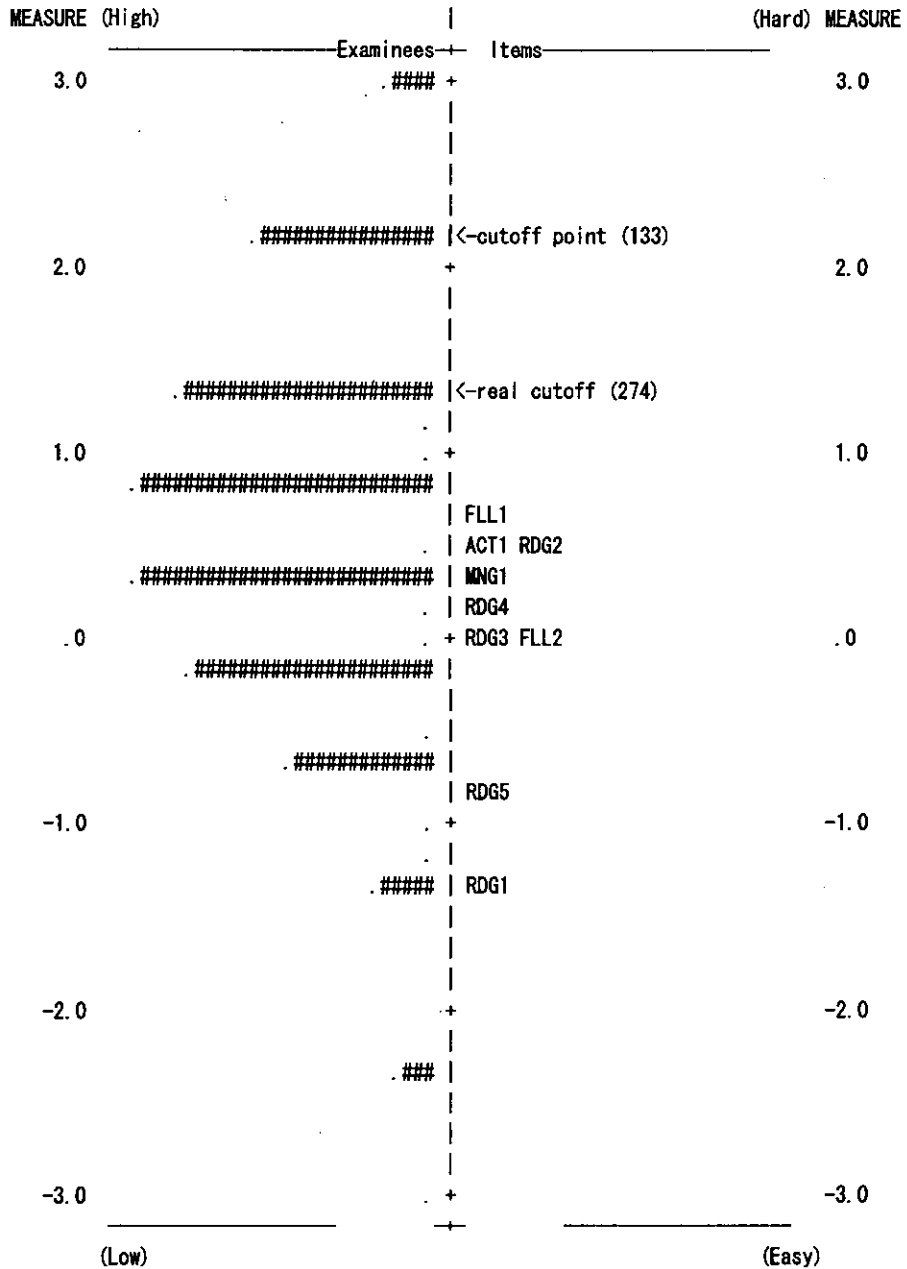


図5: 社会福祉学部 of 項目と被験者の位置関係

表2：文学部のテスト項目分析

難易度	内的適合性		外的適合性		点双列 相 関	項目
	MSQ	Z点	MSQ	Z点		
2.30	.96	-.5	.96	-.3	.19	LSG 5
2.27	1.01	.2	1.02	.2	.17	DLG 2
1.79	1.00	.0	1.01	.1	.20	LSG 2
1.10	1.12	2.8	1.15	2.3	.11	RDG 6
.62	.93	-2.0	.91	-1.8	.31	FLL 1
.47	.81	-6.1	.76	-5.5	.46	DLG 4
.00	1.12	3.1	1.18	3.1	.10	LSG 3
-.16	.90	-2.8	.86	-2.4	.35	LSG 4
-.18	1.09	2.2	1.07	1.1	.14	DLG 5
-.42	.95	-1.3	.99	-.1	.28	RDG 4
-.48	.99	-.3	.99	-.1	.24	RDG 3
-.62	1.06	1.2	1.09	1.2	.16	RDG 2
-.79	1.07	1.4	1.12	1.3	.13	DLG 3
-.98	.97	-.4	.92	-.8	.25	DLG 1
-1.19	1.06	.9	1.45	3.5	.10	RDG 5
-1.65	.92	-1.0	.76	-1.8	.29	RDG 1
-2.08	.97	-.3	.78	-1.3	.22	LSG 1

表3：経済学部のテスト項目分析

難易度	内的適合性		外的適合性		点双列 相 関	項目
	MSQ	Z点	MSQ	Z点		
1.08	.99	-.2	1.00	-.1	.18	RDG 3
.92	1.06	2.1	1.07	1.8	.12	FLL 1
.85	1.03	1.1	1.03	.7	.15	ODR 1
.66	.97	-1.3	.96	-1.0	.21	RDG 4
.19	.95	-2.2	.94	-1.7	.24	RDG 5
-.16	.99	-.2	.99	-.3	.19	RDG 1
-.52	.97	-.9	.96	-.8	.21	RDG 2
-1.20	.98	-.4	1.01	.1	.19	ODR 2
-1.82	1.60	1.0	1.09	.8	.09	ACT 1

えば右下に1が集まっているが、これは難易度の高い項目を能力の低い被験者が正答したので、不適合度が高い。問題となるRDG5は、比較的能力の高いものがかかなり間違えているのがわかる。

実際に問題を見ると、筆者の意図を確認する問題で、文章全体を把握する能力が求められる。ここにその質問と選択肢を挙げる。また括弧内の数字はそれぞれの選択肢を選んだ人数である。正解は4である。

would the author agree with?

1 People need not protect animals because they don't have rights. (28)

2 People are not responsible for animals because they don't have rights. (60)

3 Supporters of animal rights should stop protecting animals. (18)

4 People should think of better arguments for protecting animals. (426)

(RDG5) Which of the following statements

選択肢4のみが断定を避けた表現で、他の選択肢は断定的で、かなり正解が明白のよう

		EXAMINEES	
		45544332 5331 42211 43321 432154434332144332221	
		12033605623093 08882844975329798935279875998659814	
		93986753993067563292871666249839429302245831482673	
ITEMS	easy high	-----low	
LSG1	-2.08 M0.....000000.0.00.....		
RDG1	-1.65 N0000.....0.....		
RDG5	-1.19 A 00.....0.0.000000000.....0.....		
DLG1	-.98 L0.....0.....		
DLG3	-.79 D0.0.....0.....		
RDG2	-.62 E ...00.....0.....		
RDG3	-.48 I0.....0.....		
RDG4	-.42 J ..0...0.....0.....		
LSG4	-.18 F0.0.....		
DLG5	-.16 P0.....		
LSG3	.00 B0.00.0.....0.....		
DLG4	.47 Q0.....		
FLL1	.62 O 0.....		
RDG6	1.10 C1...1...1...		
LSG2	1.79 H1111		
DLG2	2.27 G11.1...11111111.....		
LSG5	2.30 K11111...1...1...		
		Hard	low

		45544332653313542211844433219432154434332144332221	
		12033605923097 08882871975329798935279875998659814	
		93986753 9306 63292 66624 839429302245831482673	

図6：文学部被験者の不適合解答

な気がする。しかし本文では，“Animals are not able to understand what rights are and are certainly unable to respect yours.”のように強い主張を含んだ文が含まれている。つまり4以外を選択した被験者はこのような文によって誤った推論を導き2などを選択してしましたことになる。しかしこの問題では、質問、選択肢を見ても特に問題はみられなく、これはおそらくいわゆるケアレスミスをしたものが高得点者に多かったと考えるのが妥当であろう。

7 考察

本稿の分析では特に問題となる項目は見つからなかった。しかし社会福祉学部のMCQの問題が実際の合格決定、つまり合格者数点での被験者の弁別に役にたっていないことがわかった。今回の分析ではMCQの問題のみを扱っており、実際には記述式の問題が被験者の弁別に役立っていると考えられる。

この問題に関しては、二つの提案が考えられる。まず第一に社会福祉学部のMCQの問題に難易度の高いものを加えることである。そうすることで、より広範囲の能力の弁別が

表4：社会福祉学部のテスト項目分析

難易度	内的適合性		外的適合性		点双列 相 関	項目
	MSQ	Z点	MSQ	Z点		
.69	1.06	2.3	1.09	2.3	.08	FLL 1
.58	1.04	1.5	1.10	2.7	.10	ACT 1
.57	.96	-1.7	.94	-1.7	.19	RDG 2
.36	1.02	.7	1.03	.9	.13	MNG 1
.10	1.04	1.4	1.07	1.6	.11	RDG 4
.01	1.03	1.0	1.03	.6	.12	RDG 3
-.08	.96	-1.4	.95	-1.3	.20	FLL 2
-.88	.91	-2.3	.83	-2.6	.25	RDG 5
-1.36	.95	-1.0	.96	-.4	.18	RDG 1

可能になり、また実際の合格者数の位置で弁別できる項目が含まれる可能性が高くなる。

第二は項目プールである。これは日本ではなかなか馴染みにくいだが、世界最大規模のテストTOEFL やTOEICでは10年以上前から導入されているし、アメリカの大学入試であるSAT, GRE, GMATなどでも同様である。項目プールとは沢山の項目をまずパイロット調査をして項目難易度を決定してプールを作る。そのプールの中から、項目難易度の組み合わせで同等の試験を作ることができる。項目は何度か再利用されたのち、プールから捨てられる。現在センター試験もアメリカのSATのように一年に何度も受けられるようにするため、項目プールを導入するようである。これはそう簡単に実現可能ではないが、いくつかの大学が共同すれば可能である。またこのシステムが導入できれば、作られた項目が一回で捨てられる無駄がないのに加え、毎年多くの問題を作らなければならない手間が省ける。しかもほぼ同等の難易度を持った試験を作ることが簡単にできるのである。

今回の分析では記述問題が含まれていないため、一般入学試験英語の全容がわからなかった。それを含めた項目分析を行い、その反省を次回にいかすことが必要である。

[参考文献]

- Wright, B., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement: Transactions of the Rasch Measurement SIG*, 8, 370.
- 北星学園大学 (2000). 「北星学園大学入試ガイド2001年版」。(無料配布資料)