原　著

明倫歯誌 1(1) 39－43 (1998)

英語歯科用語に関するテスト項目の研究(1)
－項目困難度，項目弁別力指数，モデルとデータの適合度の検討－

広瀬浩二
明倫短期大学歯科衛生士学科

A Study of Test Items on English for Dentistry (1) :
Consideration on Item Difficulty, Item Discrimination Power Index, and Model-Data Fit

Koji Hirose
Department of Dental Hygiene and Welfare, Meirin College

本研究の目的は，英語歯科用語のテスト項目を作成し，それぞれのテスト項目の項目困難度，項目弁別力指数とモデル適合度を検討し，より適切なテスト項目の作成に寄与することである。18のテスト項目を作成した。各項目は，英語での歯科用語と4つの選択肢からなる。4つの選択肢のうちの一つは，あて推量を排除するための選択肢である。被験者は，歯科衛生士学科に所属する短大1年生75名である。18項目について古典的テスト理論による分析と項目応答理論による分析の両方を行った。各項目を項目困難度の観点から分析した場合，最も適切なテスト項目と考えられるのは18項目中，項目(14)だけであった。易しいと考えられる項目は14項目あった。項目弁別力指数の観点から分析した場合，項目(1)が不良項目で除外若くは改善の余地があると考えられ，項目(9)は改善の余地がある。モデルとの適合度の観点から分析した場合，項目(1)(9)(10)の3項目がモデルと適合せず，テスト項目から除外した方が良いとみられる。

以上の分析を基にしたうえで，モデルとの適合度を重視し項目(1)(9)(10)を作成し直すことにした。今後，更に，実験を重ね適切なテスト項目を多く作成集積し，item bankの完成につなげたい。

キーワード：英語歯科用語，英語テスト，項目困難度，項目弁別力指数，モデルとデータの適合度

The purpose of this research was to make test problems of English for Dentistry, and to examine each item from the viewpoints of item difficulty, item discrimination power index, and model-data fitness. Eighteen multiple choice test items related to English for Dentistry were created and examined. Subjects were 75 college students who belong to the Department of Dental Hygiene. Each test item was analyzed by both classical test theory and item response theory. When analyzed for the item difficulty, only No. 14 was considered a very appropriate test item, and fourteen items were found to be too easy. When analyzed for item discrimination power index, No. 1 was considered an inappropriate item. No. 9 was also considered to be weak. When analyzed for model-data fitness, Nos. 1, 9, and 10 were not suitable and should be excluded. Consequently, Nos. 1, 9, and 10 will be rewritten. More appropriate test items will be created and compiled to complete the item bank.

Key words : English for dentistry, Language test, Item difficulty, Item discrimination power index, Model-data fit

緒　言

「英語II」では歯科衛生士学科の学生に対し毎回18の英語歯科用語を指導している。その指導内容は，主として単語の発音・日本語訳・簡潔な意味内容である。学生にとって馴染みの薄い英語歯科用語が多いため，指導及び評価方法に工夫が必要である。評価方法の一つとして指導済みの英語歯科用語の定着を図るため毎時間，前回に学習した用語の小テストを実施している。この評価方法の妥当性を検討する必要性が生じた。

原稿受付　1997年12月20日，受理　1998年2月18日
連絡先：〒950-2086　新潟市真砂3-16-10　明倫短期大学　TEL. 025-232-6351（内線163）

そこで，本研究では，英語歯科用語の習得の度合を評価するためのテスト項目を作成し，古典的テスト理論（Classical Test Theory：以下CTTと省略）で得られた項目困難度と項目弁別力指数，項目応答理論（Item Response Theory：以下IRTと省略）で得られたモデルとの適合度を検討することにより，より適切なテスト項目の作成を目的とした。

IRTの研究は，Bachman[1]，Henning[2]や言語テスト専門学術誌Language Testing[3]等にみられるが，日本では，まだOhtomo（大友賢二）et al.[4]など数例がみられるだけである。IRTの研究成果はすでに米国のETS（Educational Testing Service）によってTOEFL(Test of English as a Foreign Language)の分析に利用されるなど，その重要性が認識されてきた。

理論的背景

CTT

言語テストにおいて，CTTによる項目分析は「正答数に基づく得点」を基盤としていることから，次のような不都合な点があった（Alderson et al.[5]）。(1)受験したテスト項目によって得点が変わる。(2)得点は，そのテストを受験した受験者によって変わる。(3)線形性（linearity）がない。つまり，「間隔尺度」であり「比率尺度」でない。

項目困難度

受験者にとって，適切な項目困難度とは何か。一般に，difficulty rangeの中間地点に到達するあたりが，最も適切であるといわれる。また，難しすぎたり易しすぎる項目をどのように除外したらいいのか。以下の点に配慮すべきである。(1)指導目標との関連。充分吟味した項目内容の選択。難しい項目と易しい項目の適度な配置。(2)受験者の心理的側面を考慮した易しい項目からの配列。特に，テスト不安を経験した受験者にとって有効である。(3)少ない項目数では，たとえ困難度の不適切な項目があっても棄却できない。そのため，多数の項目の用意と予備実験の必要性。

項目弁別力

ある項目を最終的に受容するか棄却するかの決定基盤として項目困難度だけでは不十分で，項目弁別力を苦慮する必要性がある。項目弁別力指数が0.25あるいはそれ以上あれば良い，とされる（Henning[2]）。

実質選択肢数

錯乱肢を考慮に入れた実質的な選択肢を求めるのは重要だが，意味のない錯乱肢は以下の理由から避けるべきである。(1)錯乱肢としてあまり役立たない。(2)指導に負の「波及効果」を及ぼすことがある。

IRT

CTTでは，テストによって算出された数値がテスト受験者の特性によって変化するといった不都合な点があった。一方，IRTの利点は，次の3点である。(1)どんな異なったテストを用いても共通の尺度上で能力測定が可能である（Test-free person measurement）。(2)どんな受験者集団に実施しても共通の項目特性に関する値を求めることができる（Sample-free item calibration）。(3)能力ごとにわかる測定の精度（Multiple reliability estimation）。

IRTで用いられるモデルは3つあるが，本研究では，比較的小規模の受験者に適用可能であり，計算・分析・解釈の点で比較的簡便である「1パラメータ・ロジスティック・モデル」別名「ラッシュ・モデル」を使用する。本モデルでは，その測定が受験者の能力と項目困難度に限定される（Alderson et al.[5]）。ここでは，特に，モデルとの適合度を検討する。

対象および方法

対象

本実験の対象は，本短大歯科衛生士学科1年生75名である。テストは平成9年12月に通常の授業時間内で行った。

方法

テスト項目の作成

平成9年11月までに指導した165の英語歯科用語の中から18のテスト項目を作成した。各項目は以下に示すように，最初に英単語を挙げ，次に該当する日本語の選択肢を3つと当て推量の要素を排除するための選択肢dを配列した。受験者は，各項目において英単語に相当する日本語訳をa～dの中から選び，解答欄に記入した。尚，a～cのいずれが正解であるかわからない場合には，d.？を選択するように指示した。

(1) carving knife（ ）a. 切断刀，b. 彫刻刀，c. 両刃刀，d.？
(2) mallet（ ）a. 触診，b. 悪性腫瘍，c. 槌，d.？
(3) central incisor（ ）a. 中切歯，b. 側切歯，c. 乳切歯，d.？
(4) lateral incisor（ ）a. 側切歯，b. 乳切歯，c. 中切歯，d.？
(5) recall（ ）a. 顎関節検査，b. 定期検査，c. 知覚検査，d.？
(6) block anaesthesia（ ）a. 伝達麻酔，b. 表面麻酔，c. 浸潤麻酔，d.？

(7) objective symptom（　）a. 自覚症状，b. 間接症状，c. 他覚症状，d.？
(8) inverted cone bur（　）a. 倒立円錐バー，b. 仕上げバー，c. 円形バー，d.？
(9) silver impregnation method（　）a. 内観法，b. 銀含浸，c. 鍍銀法，d.？
(10) dry socket（　）a. 口腔乾燥，b. ドライソケット，c. 歯周ポケット，d.？
(11) floss silk（　）a. 塗蠟絹糸，b. 外科用絹糸，c. 処女絹糸，d.？
(12) milk teeth（　）a. 誕生歯，b. 新生児歯，c. 乳歯，d.？
(13) cyst（　）a. 嚢胞，b. 歯根嚢胞，c. 歯肉嚢胞，d.？
(14) incision of abscess（　）a. 歯槽膿瘍，b. 口蓋膿瘍，c. 膿瘍切開，d.？
(15) saliva ejector（　）a. 安静時唾液，b. 排唾器，c. 神経節唾液，d.？
(16) dressing forceps（　）a. 麦粒鉗子，b. 抜歯鉗子，c. 骨鉗子，d.？
(17) extraction of the tooth（　）a. 脱落歯，b. 人工歯配列，c. 抜歯，d.？
(18) dental extracting forceps（　）a. クランプ鉗子，b. 抜歯鉗子，c. 麦粒鉗子，d.？

分析方法

テスト結果のデータ分析に際し，大友賢二と中村洋一作成（Copyright : Kenji Ohtomo & Youichi Nakamura）の Test Data Analysis Program : TDAP Ver. 1.0 (1996) を使用した。統計的に処理されたデータを項目困難度・項目弁別力・モデルとの適合度の3点から解釈した。

結　果

基礎統計

(1) raw scores の総計：1235 (2)最小得点：3 (3)最大得点：18 (4) Median：17 (5) Range：15 (6)平均：16.467 (7)分散：5.876 (8)標準偏差：2.424 (9) Skewness：−2.934 (10) Kurtosis：11.438

かなり高い得点域に平均がきており，Skewness も負の数値を得たことから全体に易しい問題であったことがわかる。しかし，そのこと自身問題とはならない。

表1． CTT による項目分析結果

RANK	NO.	DIFF	DISC	AENO	ADIF	ADIS	AAEN	SATOT
1	14	0.667	0.580	2.589	0.917	0.506	0.921	1.852
2	4	0.947	0.731	1.302	0.357	1.000	0.987	1.782
3	17	0.880	0.596	1.626	0.490	0.550	0.969	1.652
4	16	0.907	0.553	1.508	0.437	0.441	0.995	1.602
5	8	0.893	0.548	1.569	0.463	0.428	0.983	1.599
6	7	0.787	0.637	1.861	0.677	0.684	0.767	1.575
7	18	0.960	0.573	1.236	0.330	0.488	1.000	1.559
8	11	0.920	0.543	1.433	0.410	0.419	0.977	1.553
9	2	0.973	0.646	1.152	0.303	0.718	0.924	1.544
10	3	0.947	0.535	1.278	0.357	0.402	0.909	1.422
11	15	0.960	0.545	1.213	0.330	0.422	0.904	1.407
12	5	0.987	0.646	1.073	0.277	0.716	0.822	1.395
13	12	0.987	0.646	1.073	0.277	0.716	0.822	1.395
14	1	0.920	0.077	1.433	0.410	0.006	0.977	1.373
15	6	0.920	0.422	1.397	0.410	0.216	0.896	1.360
16	9	0.920	0.381	1.397	0.410	0.170	0.896	1.340
17	10	0.973	0.442	1.152	0.303	0.242	0.924	1.337
18	13	0.920	0.564	1.322	0.410	0.466	0.726	1.251

<NOTES>
RANK＝ 標準適切度合計による順位（Rank in SATOT order）
DIFF＝ 項目困難度指数（Item difficulty index）
DISC＝ 項目弁別力指数（Discrimination power index）
AENO＝ 実質選択肢数（Actual equivalent number of options）
ADIF＝ 項目困難度適切度（Appropriateness of difficulty）
ADIS＝ 項目弁別力適切度（Appropriateness of discrimination power index）
AAEN＝ 実質選択肢数適切度（Appropriateness of actual equivalent number of options）
SATOT＝標準適切度合計（Standard appropriateness total）

CTT による項目分析結果

表1は，CTT により，困難度・点双列相関係数による項目弁別力・実質選択肢数等の数値を求めた結果である。

信頼性

(1)信頼性係数：0.825 (2)標準偏差：2.424 (3)測定の標準誤差：1.015

信頼性係数が 0.825 と高い数値を示していることから，本実験で得たデータはかなり信頼できる。また，測定の標準誤差が 1.015 であった。これは，15 点を取った受験者の中で，68%の人の真の得点は 13.985 から 16.015 に入るし，95%の人の真の得点が 12.97 から 17.03 に入ることを示す。

IRT による項目分析結果

表2は，IRT の1パラメーター・ロジスティック・モデルにより，項目困難度パラメーター・モデルとの適合度を算出した結果である。

表2．IRT による項目分析結果

Item No.	Final Calib.	Standard Error(d)	Fit(t)
1	0.288	0.492	6.331
2	−1.206	0.807	−5.470
3	−0.283	0.586	−3.105
4	−0.283	0.586	−4.735
5	−2.096	1.127	−9.139
6	0.288	0.492	−0.107
7	1.919	0.356	−0.747
8	0.717	0.439	1.171
9	0.288	0.492	4.664
10	−1.206	0.807	6.724
11	0.288	0.492	−1.388
12	−2.096	1.127	−9.139
13	0.288	0.492	−1.296
14	3.024	0.356	0.222
15	−0.672	0.667	−2.913
16	0.515	0.462	−1.767
17	0.901	0.420	0.549
18	−0.672	0.667	−3.466

考　察

CTT による項目分析では，特に，項目困難度と項目弁別力指数に注目した。項目困難度は 0.500 が最も適切である，といわれる。本研究で作成した 18 項目の中では，項目番号(14)が最も適切なテスト項目であると考えられる。(2)(5)(10)(12)(15)(18)はかなり易しい項目であった，と判断されるし，項目(1)(3)(4)(6)(9)(11)(13)(16)も易しい部類に属すると考えられる。しかしながら，2.1.1 で指摘したように易しい項目も配列の仕方によっては有効に活用できることになる。また，項目弁別力指数では，項目番号(1)が不良項目で除外若くは改訂の必要があり，(9)は改善の余地のあることがわかった。

IRT による項目分析では，特に，モデルとの適合度に注目した。モデルとの適合度は，2.00 が危険域である。2.00 あるいはそれ以上の場合は，応答妥当性を欠きモデルとは適合しないと考えられる（Henning[2]）。モデルとの適合度では，項目番号(1)(9)(10)の3項目がモデルと適合せず，テスト項目から除外したほうが良いといえる。一方，負の適合度を持つ項目が多かった。−2.00 あるいはそれ以下の場合，モデルと過剰適合していると考えられるが，テスト項目から除外する理由にはならない。

項目(1)(9)(10)は，CTT における項目困難度分析では，どれも易しい項目に属している。また，項目弁別力指数でも(1)は不適切であり，(9)は改善の余地のあることが判明した。少なくともこの3項目については作成し直した方が良いと思われる。また，テスト形式は CTT や IRT による分析を意図したため，多肢選択の形式とした。英語の歯科用語に対応する日本語の選択肢を設定したが，逆にすれば，難易度は増すかもしれない。これについて検討の余地はある。

本研究の限界は，次の2点である。(1)項目数を少なく制限したうえでの分析であったこと。(2)受験者数が 75 名であったこと。項目数に関しては，今後更に増強することが可能である。また，受験者数に関しては，One-Parameter（Rasch）Model で 100 名，Two-Parameter Model で 200 名，Three-Parameter Model で 1,000 名必要である（Alderson et al.[5]）という指摘を踏まえると，やや受験者数が不足であったと言わざるをえない。

結　論

本研究では，英語歯科用語の評価のため，18 のテスト項目を作成し学生に実施した。テスト結果を統計的に処理した。CTT と IRT により分析した。分析では，項目困難度・項目弁別力・モデルとの適合度に注目した。その結果，18 項目のうち項目(1)(9)(10)の3項目を作成し直すことにした。

今後の課題は，英語歯科用語に関するテスト項目を増やすことと項目作成の際，言語テストとして最適な形式を模索すること，である。形式を整え，受験者数を確保し，より多くのテスト項目のデータを蓄積することが，不変性を維持した項目銀行の形成につながる。

文　献

1）Bachman, L. F. : Fundamental Considerations in Language Testing. Oxford University Press., New York, 1990.
2）Henning, G. : A Guide to Language Testing. Heinle & Heinle Publishers., Boston, MA, 1987.
3）Alderson, J. C. and Bachman, L. F. (eds.) : Language Testing. Edward Arnold., London, 1985.
4）Ohtomo, K., Asano, H., Hattori, T. and Yoshie, M. : Item Difficulty of English Language Tests for Japanese Students : The Rasch Model Calibration. *JACET Bulletin*, **18**, 109-125, 1987.
5）Alderson, J. C., Clapham, C., and Wall, A. D. : Language Test Construction and Evaluation. Cambridge University Press., Cambridge, 1995.