

原 著

## 英語歯科用語に関するテスト項目の研究(5) —テスト妥当性の検証—

廣 瀬 浩 二

明倫短期大学 歯科衛生士学科

### A Study of Test Items on English Terminologies for Dentistry (5) —Examining Test Validity—

Koji Hirose

*Department of Dental Hygiene and Welfare, Meirin College*

本研究では、平成12年度に実施した英語歯科用語のテスト項目の妥当性を検証することを目的とした。妥当性は内部構成概念証明法によって検証した。被験者は、歯科衛生士学科に所属する短大1年生86名である。テストは、50の歯科用語に関するものと7つの対話文から構成された。テストは、ピアソンの積率相関係数(妥当性)とクロンバックの $\alpha$ 係数(信頼性)によって分析した。信頼性係数は、0.886であった。テストの信頼性は証明された。2項目を除いて(20と43)、歯科用語に関するテスト得点の相関は対話文に関するテスト得点の相関より高かった。したがって、2項目以外は意図した構成概念を測定していることが証明された。これら2項目については、測定しようとするものを測定しておらず、書き直す必要があるであろう。

キーワード：英語歯科用語、構成概念妥当性、内部構成概念証明法、信頼性

The purpose of this research was to examine the test validity of the "English for Dentistry" examination which was made and given as one of the formative tests in the academic year of 2000. The internal construct validation was determined to employ among some validation methods in order to investigate the test validity. Subjects were 86 college students who belonged to the Department of Dental Hygiene. The test was composed of 50 dental terms and 7 dialogues. Both categories of the test were analysed by the Pearson's product moment correlation coefficient (validity) and the Cronbach's coefficient alpha (reliability). The reliability coefficient for the test was 0.886. Therefore the test was highly reliable. The correlation coefficients between an individual score and total scores on dental terms were consistently higher than those between an individual score and total scores on dialogue, except for two test items (Nos. 20 and 43). Consequently, we have evidence that we are measuring separate and identifiable constructs, excluding the two test items. On these two items we could not measure what we intended to do and will need to rewrite each item's options.

**Key words :** English for dentistry, Construct validity, Internal construct validation, Reliability

#### 緒 言

筆者は本学歯科衛生士学科「英語Ⅱ」における歯科用語に関するテスト項目を作成し実施してきた。そのテスト形式は、伝統的な筆記テスト(paper-and-pencil language test)であり、多肢選択式(multiple choice

format)である。テストの目的は指導内容の評価のためであり、到達度テスト(achievement test)である。作成した各項目は古典的テスト理論と項目応答理論の両者により細部にわたり分析を行い改良を重ねてきた。項目数に関しては、当初18項目であった項目銀行を、50にまで拡大してきた。これまで歯科用語に関するテ

スト項目を何度か作成し、実施してきたが、同テストの妥当性については言及されないままであった。そこで、本研究では、筆者<sup>1)</sup>が作成し実施したテスト項目の妥当性を中心に検証することとした。

言語テストの妥当性検証 (validation in language testing) は、その目的と適合する必要がある<sup>2)</sup>、本研究では、授業で指導したことを学生がどの程度習得したのかを確認することを目的とした。

妥当性は、内的妥当性 (internal validity) ・外的妥当性 (external validity) ・構成概念妥当性 (construct validity) の3つに大きく分類される<sup>3)</sup>が、ここでは構成概念妥当性について検証を試みた。構成概念妥当性は「測定したい能力、つまり構成要素の指標としてテスト得点をどの程度まで解釈できるかの限度を示す」特質であり、その検証の手段には様々な方法がある<sup>4)</sup>。本研究では、内部構成概念証明法 (internal construct validation)<sup>2)</sup>に従うこととした。この方法に従えば、「英語Ⅱ」の試験における歯科用語に関するテスト部門と同一試験の他のテスト部門の関係を検討すればよいことになる。例えば、歯科用語の項目1の得点をV1、歯科用語の総得点をVTとし、対話文問題の総得点をDTすると、「V1 VT > V1 DT」といった関係が成立すれば、歯科用語に関するテスト部門の構成概念妥当性があることを証明できる。本研究では、筆者のテスト項目について主としてこの内部構成概念証明法によって構成概念妥当性を証明した。

## 対象および方法

### 1. 対象

本研究の対象者は、本短大歯科衛生士学科1年生86名である。テストは平成13年1月の後期試験期間内に実施した。

### 2. 方法

#### 1) テストの目的

本テストは、本学歯科衛生士学科「英語Ⅱ」の指導内容に関して、学生の学習状況を客観的に評価することである。

#### 2) テストの構成

本研究で使用したテストは、2種類のテストからなる。50のテスト項目から成る英語歯科用語に関するものと対話文の穴埋め問題である。

英語歯科用語に関するテスト項目の特徴は次の通りである。①テストタイプは多肢選択式で、各テスト項

目とも基幹部分 (stem) は日本語の歯科用語名である。②選択肢 (options) は英語の歯科用語名である。③選択肢の数は4である。

対話文の穴埋め問題では、授業中、学習した歯科衛生士と患者の会話の一部を取り上げた。7つの短い対話例を提示し、合計20か所に適する語句を挿入する形式である。挿入すべき語句は選択肢として示してある。また、空欄を含む英文には日本語訳も付与した。

### 3) 分析方法

テスト結果のデータ分析に際し、Item Response Theory (IRT) のラッシュモデル (Rasch model) を使用する。このモデルは次の式によって記述できる。

$$P\{x_{vi}=1 \mid \beta_v, \delta_i\} = \exp(\beta_v - \delta_i) / [1 + \exp(\beta_v - \delta_i)]$$

ここで、Pは正答確率である。 $\beta_v$ は受験者の能力、 $\delta_i$ はテスト項目の困難度をさす。また、正解の場合は、 $x_{vi}=1$ となり、不正解の場合は、 $x_{vi}=0$ となる。 $\exp$ は、指数 (exponent) を指す。

信頼性の推定に関しては、クロンバックの $\alpha$ 係数 (Cronbach's coefficient alpha) を使用する。数式は次の通りである。

$$\alpha = 2 \left( 1 - \frac{S_{\text{odd}}^2 + S_{\text{even}}^2}{S_t^2} \right)$$

ここで、 $S_{\text{odd}}$ は奇数番号のテスト項目の標準偏差 (SD)、 $S_{\text{even}}$ は偶数番号のテスト項目の標準偏差、 $S_t$ はそのテスト総得点の標準偏差である。

内部構成概念証明法では、ピアソンの積率相関係数 (Pearson's product moment correlation coefficient) を使用する。次の式によって求められる。

$$r_{xy} = \frac{\sum (Y - \bar{Y})(X - \bar{X})}{NS_y S_x}$$

分析に際し、大友賢二と中村洋一作成のTest Data Analysis Program: TDAP Ver.1.0 (Copyright: Kenji Ohtomo & Youichi Nakamura, 1996)を使用した。基礎統計量の算出には、FUNDAMENTAL STATISTICSのプログラムを、Classical Test Theory (CTT) による分析にはITEM ANALYSISのプログラムを、IRTによる分析にはITEM RESPONSE THEORYのプログラムをそれぞれ使用した。また、ピアソンの積率相関係数の算出には、Stat View 4.0を使用した。

## 結果と考察

表1は、英語歯科用語のテスト項目に関する基礎統計量を示したものである。

表1. 基礎統計量

受験者数	(Number of examinees) : 69
得点の総計	(Sum of the raw scores) : 2274
最小得点	(Minimum score) : 4
最大得点	(Maximum score) : 48
中央値	(Median) : 34
範囲	(Range) : 44
平均	(Mean) : 32.957
分散	(Variance) : 67.694
標準偏差	(Standard deviation) : 8.228
歪み	(Skewness) : -0.823
尖度	(Kurtosis) : 1.249

平均値は32.957と比較的高い位置にあるため、Skewnessは負の数値を示している（-0.823）。最高得点と最低得点の差（Range）はかなりの幅があり（44）、分散（67.694）の数値も大きい。Kurtosisの数値がかなり小さい（1.249）ことから、得点の分布曲線は平坦なカーブを描き、得点分布の集中が回避される結果となっている。

次に、信頼性（reliability）・構成概念妥当性（construct validity）について順を追って検討した。

表2. 信頼性係数

アルファ係数	(Coefficient alpha)	0.886
標準偏差	(Standard deviation of the test)	8.228
測定の標準誤差	(Standard error of measurement)	2.781
得点の総和	(Sum of the raw scores)	2274
平均値	(Mean of the raw scores)	32.957
分散	(Variance of the test)	67.694
項目分散の和	(Sum of the item variance)	8.934
平均通過率	(Average proportion of passing)	0.659
平均不通過率	(Average proportion of failing)	0.341

信頼性：信頼性係数は0.886であった。観察された得点の分散の約89%は真の得点の分散に依存し、説明のつかない誤差分散は11%に過ぎなかった。このことから、信頼性係数は十分に高い数値であり、信頼性のあるテストであったといえる。

測定の標準誤差は2.781であった。もし、40点を取った受験者がいたと仮定すると、その中で、68%の人の真の得点は37.219から42.781に入り、95%の人の真の得点が34.438から45.462に入ることとなる。許容範囲の誤

差であるといえる。

妥当性：ここでは、以下のように内部構成概念証明法（internal construct validation）によって、構成概念妥当性を検証した。

表3. 用語各テスト項目と下位テスト総得点との相関係数一覧

°V1VT=.095	>	°V1DT=.023
°V2VT=.228*	>	°V2DT=.177
°V3VT=.200+	>	°V3DT=.000
°V4VT=.276*	>	°V4DT=.174
°V5VT=.245*	>	°V5DT=.223*
°V6VT=.417**	>	°V6DT=.366**
°V7VT=.544**	>	°V7DT=.359**
°V8VT=.171	>	°V8DT=.059
°V9VT=.232*	>	°V9DT=.196+
°V10VT=.259*	>	°V10DT=.044
°V11VT=.348**	>	°V11DT=.215*
°V12VT=.566**	>	°V12DT=.276**
°V13VT=.492**	>	°V13DT=.277**
°V14VT=.338**	>	°V14DT=.234*
°V15VT=.252*	>	°V15DT=.162
°V16VT=.627**	>	°V16DT=.555**
°V17VT=.572**	>	°V17DT=.457**
°V18VT=.526**	>	°V18DT=.273*
°V19VT=.506**	>	°V19DT=.423**
°V20VT=.325**	<	°V20DT=.369**
°V21VT=.504**	>	°V21DT=.315**
°V22VT=.378**	>	°V22DT=.309**
°V23VT=.294**	>	r V23DT=.112
°V24VT=.545**	>	°V24DT=.381**
°V25VT=.410**	>	°V25DT=.270*
°V26VT=.542**	>	°V26DT=.328**
°V27VT=.490**	>	°V27DT=.298**
°V28VT=.250*	>	°V28DT=.187+
°V29VT=.466**	>	°V29DT=.298**
°V30VT=.225*	>	°V30DT=.168
°V31VT=.342**	>	°V31DT=.215*
°V32VT=.107	>	°V32DT=.041
°V33VT=.309**	>	°V33DT=.232*
°V34VT=.735**	>	°V34DT=.527**
°V35VT=.281**	>	°V35DT=.130
°V36VT=.410**	>	°V36DT=.240*
°V37VT=.322**	>	°V37DT=.179
°V38VT=.339**	>	°V38DT=.247*
°V39VT=.493**	>	°V39DT=.419**
°V40VT=.396**	>	°V40DT=.286**
°V41VT=.536**	>	°V41DT=.311**
°V42VT=.496**	>	°V42DT=.427**
°V43VT=.335**	<	°V43DT=.396**
°V44VT=.342**	>	°V44DT=.214*
°V45VT=.652**	>	°V45DT=.348**
°V46VT=.486**	>	°V46DT=.383**

「V47VT=.318**	>	「V47DT=.268*
「V48VT=.345**	>	「V48DT=.181+
「V49VT=.303**	>	「V49DT=.189+
「V50VT=.270*	>	「V50DT=.174
+p<.10 *p<.05 **p<.01		

表3に示されたように、歯科用語テスト項目のNo.20とNo.43以外の項目では、「V1 VT」>「V1 DT」の関係が成立し、構成概念妥当性が証明された。ここまでは、量的分析によって明らかになったことである。しかし、これらNo.20とNo.43の2つの項目でなぜ妥当性が証明されなかったかについては、質的分析に委ねざるをえない。

No.20（現病歴）に関しては、history of the present illnessが正解であり、誤答が38あった。その中、history of the presentive illnessと回答した学生が22名あり、history of the present illnatureと回答した学生が10名あり、history of the preset illnessと回答した学生が6名あった。illnessの誤認識が10に対し、presentの誤認識が28と約3倍であった。ここから、No.20のテスト項目では、単語presentの認識の誤りが多数を占めたことがわかる。

また、No.43（小臼歯）では、正解がpremolarで、誤答が19あった。誤答数は、No.20に比較して少なかった。誤答の中、premetroと回答した学生が8名、premonitorと回答した学生が6名、premiereと回答した学生が5名と誤答は分散した。そもそもpremolarは、pre+molarという構成になっていることを指導してあった。単語molarの認識が不十分であったと考えられる。

以上の分析をふまえて、選択肢を再度作成すること

が必要となろう。また、本テストは目標基準準拠テスト（criterion-referenced test）であるという性格上、指導の際にもこれらの項目に配慮すべきであろう。

## 結 語

本研究では、英語歯科用語に関するテストの有用性を検証するため、本学歯科衛生士学科で実施した「英語Ⅱ」の試験結果を分析した。信頼性（reliability）・構成概念妥当性（construct validity）について考察を加えた。信頼性は、高い数値のもと証明できた。構成概念妥当性については、用語に関する50のテスト項目の中、2項目を除いて証明できた。

テストの妥当性を証明することはたいへん困難であるが、妥当性を高めることはさらに困難なことである。実際、妥当性は低くても信頼性は高いというテストは存在する。しかし、逆に、信頼性が低くて妥当性が高いテストは存在しない。今後も、信頼性と妥当性の検証を継続し、より良いテスト作成に取り組みたい。

## 文 献

- 1) 廣瀬浩二：英語歯科用語に関するテスト項目の研究(4) — 項目銀行の拡張 —。明倫歯誌4(1)：54-62, 2001.
- 2) Henning G : A Guide to Language Testing. pp97-100, Heinle & Heinle Publishers., Boston, MA, 1987.
- 3) Alderson J C, Clapham C and Wall A D : Language Test Construction and Evaluation. pp183-185, Cambridge University Press., Cambridge, 1995.
- 4) Brown J D : Testing in Language Programs. pp231-262, Prentice Hall Regents., Upper Saddle River, New Jersey, USA, 1996.