# 到達目標型教育に向けた英語テストの改善：
## 古典的テスト理論と項目応答理論に基づいて

前田啓朗
広島大学情報メディア教育研究センター
外国語教育研究系

### 1．研究の背景

国立教育政策研究所（2002）によって，平成14年度から小学校・中学校における到達目標に準拠した評価（いわゆる絶対評価）が進められることとなり，到達目標の設定と並んで適正な評価を行うことの必要性が改めて認識されている。高等学校においても同様で，平成15年度新入生から年次進行で，この評価方法が導入される。

初等・中等教育におけるこの動きは高等教育段階にあっても同様で，広島大学（2002）では大学計画委員会による『到達目標型教育に向けて』という答申が評議会で説明され「学長から今後これを具体化するための方策について検討したいとの報告があった」とされている。これは，到達目標を設定し，その目標に沿った教育活動を行い，教育効果を適切に測定し，測定結果を教育活動改善のための資料とするという，教育の質を高めるための方針である。そこで測定結果をもとに行われる評価は，初等・中等教育において導入される，いわゆる絶対評価（到達目標に準拠した評価）である。

この理念においては，横断的に英語学力を妥当性と信頼性のある方法で測定・評価するだけでなく，縦断的な英語学力の推移をも分析結果として提示する必要がある。しかもその推移とは，集団に準拠したようなテスト受験者集団内の相対的・順位的なもの（いわゆる偏差値など）ではなく，あくまで目標に準拠した，それぞれの受験者による目標への到達度でなければならない。

前述の答申（広島大学，2002）においては英語学力測定にはTOEIC等を用いることが提案されているが，このように項目応答理論に基づいて分析し，評価としてスコアを算出する（The Chauncey Group International, 2002）タイプの分析が，受験者集団に準拠せずに到達目標に準拠した評価を可能にするものとして注目されている。

完全に受験者集団の影響を受けることなくスコアを算出するには，多様な受験者による解答データや大量な問題データによる，いわゆるアイテム・バンクが必要となる。国内では，大学による組織的な先行研究としては筑波大学外国語センター（1994）がもっとも包括的な成果であり，島谷他（1999）では日本と韓国の高校生英語学習者を対象としたテスト・データの分析が行われ，基礎データが示されている。しかし，安定して学習者の「真の能力」を推定するために使用可能な基礎データを公開した研究は，その数がアイテム・バンクとして十分であるとはいい難い。

そこで，本研究は前田（2002）で示された問題に対する解答データを分析し，英語学力測定のための基礎的研究として資料を提示する。この問題と解答のデータを利用する理由は，データが利用可能であったことと，問題データは実用英語技能検定（財団法人日本英語検定協会）において使用された問題から選ばれた50問であるためにかなりの妥当性が予想されること，そして受験者集団は多段抽出によって行われた依頼に応じた38高等学校からの1,554人であるためにかなりの程度で日本の高校生英語学習者を母集団として想定しうること，である。また，分析においては，古典的テスト理論と項目応答理論の双方を用いる。古典的テスト理論は受験者集団に準拠し

た，素点に基づく分析方法であるが，その範囲内において解釈を行うには簡便かつ十分である場合も存在する。より洗練された分析理論である項目応答理論については，基礎的な研究としてデータを提示し，古典的テスト理論による結果とあわせて分析を試みる。

## 2．分析

### 2.1．材料

分析には前田（2002）で得られたデータが用いられた。このテスト問題は実用英語技能検定の過去問題から選ばれた全50問（4択）であり，4級（8問）・3級（12問）・準2級（14問）・2級（12問）・準1級（4問）から構成される（資料参照）。

受験者は38高等学校の1,584人である。多段抽出によって学校単位で依頼を行い，これに応じた各学校において1クラスの生徒が対象となった。

これらのデータは解答用紙をもとにデータ入力の専門業者によってそれぞれ2度入力されて照合する方法でベリファイされ，電子化されている。

### 2.2．手続き

分析には TDAP Ver. 2.0（Ohtomo *et al*, 2002）を用いた。このソフトウェアは古典的テスト理論と項目応答理論の双方を用いてテスト・データを分析できるものである。使用法や結果の解釈法を示す日本語書籍（中村，2002）があることからこのソフトウェアを選択した。また，具体的な手順や着眼点については島谷他（1999）を参考にした。

### 2.3．古典的テスト理論に基づく分析

表1にテスト得点の代表値などを挙げる。得点は1問ごとに正解であれば1点とし，満点50となっている。平均（24.380）は得点可能範囲の中央に近く，標準偏差が8.816であることと最小（4）と最大（45）の値から，床効果も天井効果もないと解釈できる。分布の形状は正規分布と比較して左右方向への歪みはほとんどなく（歪度0.068），尖度（-0.850）から尖りが幾分少ないとはいえるものの，極端に正規性を欠く分布ではないと解釈できる。

**表1　テスト得点の基本統計量**

| 平均 | 標準偏差 | 最小 | 最大 | 歪度 | 尖度 | $\alpha$ | 標準誤差 |
|---|---|---|---|---|---|---|---|
| 24.380 | 8.816 | 4 | 45 | 0.068 | -0.850 | 0.882 | 3.025 |

古典的テスト理論に基づく分析結果は，表2に示す。問題番号欄の記号は，FOは4級，THは3級，PTは2級，TWは2級，POは1級の過去問題であることを示す。

表中の変数は，DIFF（Item Difficulty Index：項目困難度）が正答率を示し，DISC（Discrimination Power Index：項目弁別力）が点双列相関係数による受験者全体を対象とした項目弁別力を示す。たとえばFO01についてはDIFFが高く，ほとんどの受験者が正答していることが示される一方で，DISCは低く，正答率が高すぎるという天井効果のために弁別力があまりないということが解釈できる。逆にDIFFが低すぎる場合にも床効果のために弁別力に問題が出る例もある（TW10やPO02等）。AENO（Actual Equivalent Number of Options：実質選択肢数）は，この場合4つの選択肢のそれぞれに対する解答数をもとに，実質的に受験者が選択した選択肢数

を算出したものである。他の数値とあわせて考慮し，これが低すぎる場合には正解のほかの錯乱肢を改善すること等が考えられる。ADIF（Appropriateness of Difficulty:項目困難度適切度）およびAAEN（Appropriateness of Actual Equivalent Number of Options:実質選択肢数適切度）については，前者はDIFF，後者はAENOの適切性に関する指標である。

表2　古典的テスト理論に基づく分析

| 問題 | DIFF | DISC | AENO | ADIF | ADIS | AAEN | SADIF | SADIS | SAAEN | SATOT |
|---|---|---|---|---|---|---|---|---|---|---|
| FO01 | 0.823 | 0.393 | 1.911 | 0.604 | 0.182 | 0.973 | 0.504 | 0.474 | 0.593 | 1.571 |
| FO02 | 0.755 | 0.479 | 2.197 | 0.740 | 0.297 | 0.933 | 0.556 | 0.555 | 0.475 | 1.586 |
| FO03 | 0.867 | 0.407 | 1.689 | 0.516 | 0.198 | 0.965 | 0.470 | 0.485 | 0.571 | 1.526 |
| FO04 | 0.525 | 0.486 | 3.103 | 0.799 | 0.310 | 0.888 | 0.578 | 0.564 | 0.343 | 1.486 |
| FO05 | 0.672 | 0.498 | 2.555 | 0.907 | 0.329 | 0.914 | 0.620 | 0.577 | 0.420 | 1.616 |
| FO06 | 0.786 | 0.355 | 2.028 | 0.678 | 0.144 | 0.912 | 0.532 | 0.448 | 0.415 | 1.394 |
| FO07 | 0.676 | 0.550 | 2.417 | 0.898 | 0.433 | 0.844 | 0.616 | 0.650 | 0.212 | 1.478 |
| FO08 | 0.824 | 0.476 | 1.918 | 0.601 | 0.293 | 0.988 | 0.503 | 0.552 | 0.637 | 1.692 |
| TH01 | 0.604 | 0.594 | 2.989 | 0.958 | 0.546 | 0.983 | 0.639 | 0.729 | 0.623 | 1.992 |
| TH01 | 0.649 | 0.549 | 2.748 | 0.952 | 0.433 | 0.965 | 0.637 | 0.650 | 0.569 | 1.856 |
| TH03 | 0.590 | 0.487 | 3.056 | 0.931 | 0.311 | 0.985 | 0.629 | 0.565 | 0.630 | 1.824 |
| TH04 | 0.718 | 0.527 | 2.425 | 0.813 | 0.384 | 0.970 | 0.584 | 0.616 | 0.585 | 1.785 |
| TH05 | 0.552 | 0.490 | 3.144 | 0.854 | 0.317 | 0.951 | 0.599 | 0.569 | 0.528 | 1.695 |
| TH06 | 0.604 | 0.453 | 2.910 | 0.958 | 0.259 | 0.944 | 0.639 | 0.528 | 0.509 | 1.676 |
| TH07 | 0.699 | 0.497 | 2.477 | 0.851 | 0.327 | 0.944 | 0.598 | 0.576 | 0.509 | 1.683 |
| TH08 | 0.777 | 0.573 | 2.138 | 0.697 | 0.489 | 0.969 | 0.539 | 0.690 | 0.581 | 1.810 |
| TH09 | 0.746 | 0.444 | 2.307 | 0.759 | 0.246 | 0.981 | 0.563 | 0.519 | 0.617 | 1.699 |
| TH10 | 0.631 | 0.563 | 2.745 | 0.987 | 0.465 | 0.921 | 0.650 | 0.673 | 0.439 | 1.762 |
| TH11 | 0.460 | 0.391 | 3.392 | 0.669 | 0.180 | 0.917 | 0.529 | 0.473 | 0.427 | 1.428 |
| TH12 | 0.659 | 0.465 | 2.560 | 0.932 | 0.276 | 0.885 | 0.629 | 0.540 | 0.333 | 1.502 |
| PT01 | 0.440 | 0.521 | 3.582 | 0.630 | 0.372 | 0.966 | 0.514 | 0.607 | 0.573 | 1.694 |
| PT02 | 0.732 | 0.449 | 2.225 | 0.787 | 0.252 | 0.874 | 0.574 | 0.523 | 0.300 | 1.397 |
| PT03 | 0.361 | 0.381 | 3.739 | 0.472 | 0.169 | 0.951 | 0.453 | 0.465 | 0.528 | 1.447 |
| PT04 | 0.332 | 0.350 | 3.676 | 0.414 | 0.140 | 0.912 | 0.431 | 0.444 | 0.414 | 1.290 |
| PT05 | 0.516 | 0.450 | 3.257 | 0.783 | 0.254 | 0.940 | 0.572 | 0.525 | 0.497 | 1.594 |
| PT06 | 0.322 | 0.360 | 3.857 | 0.394 | 0.149 | 0.969 | 0.423 | 0.451 | 0.582 | 1.456 |
| PT07 | 0.446 | 0.377 | 3.545 | 0.643 | 0.165 | 0.959 | 0.519 | 0.463 | 0.552 | 1.534 |
| PT08 | 0.399 | 0.556 | 3.658 | 0.548 | 0.446 | 0.952 | 0.482 | 0.659 | 0.531 | 1.673 |
| PT09 | 0.588 | 0.485 | 3.037 | 0.927 | 0.308 | 0.972 | 0.627 | 0.562 | 0.591 | 1.781 |
| PT10 | 0.530 | 0.469 | 3.278 | 0.809 | 0.283 | 0.971 | 0.582 | 0.545 | 0.586 | 1.713 |
| PT11 | 0.480 | 0.384 | 3.297 | 0.711 | 0.173 | 0.905 | 0.545 | 0.468 | 0.394 | 1.406 |
| PT12 | 0.486 | 0.485 | 3.418 | 0.722 | 0.307 | 0.961 | 0.549 | 0.562 | 0.559 | 1.670 |
| PT13 | 0.358 | 0.501 | 3.816 | 0.466 | 0.336 | 0.975 | 0.451 | 0.582 | 0.601 | 1.633 |
| PT14 | 0.323 | 0.308 | 3.612 | 0.395 | 0.105 | 0.886 | 0.424 | 0.420 | 0.337 | 1.182 |
| TW01 | 0.242 | 0.187 | 3.882 | 0.234 | 0.036 | 0.961 | 0.362 | 0.372 | 0.558 | 1.292 |
| TW02 | 0.273 | -0.032 | 3.630 | 0.297 | 0.001 | 0.878 | 0.386 | 0.347 | 0.314 | 1.047 |
| TW03 | 0.265 | 0.266 | 3.912 | 0.279 | 0.076 | 0.971 | 0.380 | 0.400 | 0.588 | 1.368 |
| TW04 | 0.383 | 0.392 | 3.728 | 0.515 | 0.181 | 0.963 | 0.470 | 0.474 | 0.564 | 1.508 |
| TW05 | 0.247 | 0.309 | 3.627 | 0.245 | 0.106 | 0.876 | 0.366 | 0.421 | 0.306 | 1.093 |
| TW06 | 0.484 | 0.378 | 3.389 | 0.718 | 0.167 | 0.947 | 0.548 | 0.463 | 0.516 | 1.527 |
| TW07 | 0.286 | 0.265 | 3.788 | 0.322 | 0.076 | 0.933 | 0.396 | 0.400 | 0.476 | 1.272 |
| TW08 | 0.372 | 0.288 | 3.616 | 0.495 | 0.090 | 0.916 | 0.462 | 0.410 | 0.425 | 1.297 |
| TW09 | 0.308 | 0.230 | 3.862 | 0.366 | 0.056 | 0.965 | 0.413 | 0.386 | 0.569 | 1.368 |
| TW10 | 0.162 | 0.111 | 3.760 | 0.073 | 0.012 | 0.949 | 0.301 | 0.355 | 0.522 | 1.178 |
| TW11 | 0.260 | 0.125 | 3.794 | 0.270 | 0.016 | 0.932 | 0.376 | 0.358 | 0.472 | 1.206 |
| TW12 | 0.312 | 0.409 | 3.821 | 0.374 | 0.201 | 0.953 | 0.416 | 0.488 | 0.534 | 1.437 |
| PO01 | 0.232 | 0.004 | 3.805 | 0.213 | 0.000 | 0.936 | 0.354 | 0.347 | 0.484 | 1.185 |
| PO02 | 0.165 | 0.009 | 3.830 | 0.080 | 0.000 | 0.970 | 0.303 | 0.347 | 0.586 | 1.236 |
| PO03 | 0.187 | 0.091 | 3.756 | 0.124 | 0.008 | 0.933 | 0.320 | 0.353 | 0.475 | 1.147 |
| PO04 | 0.271 | 0.185 | 3.869 | 0.293 | 0.036 | 0.958 | 0.385 | 0.372 | 0.549 | 1.305 |

これらを比較しやすくするために線形変換を施し，平均0.500・標準偏差0.100に標準化した値がSADIF（Standard Appropriateness of Difficulty: 標準項目困難度）・SADIS（Standard Appropriateness of Discrimination Power Index:標準項目弁別力）・SAAEN（Standard Appropriateness of Actual Equivalent Number of Options:標準実質選択肢数適切度）であり，これら3指標を加算したSATOT（Standard Appropriateness Total:標準適切度合計）は平均1.500として，数値が大きいものほど適切度が高い項目として判断できるものである。

### 2.4. 項目応答理論に基づく分析

古典的テスト理論の枠組みにおいては，たとえばDIFF（項目困難度）の最大値が1.000で最小値が0.000であるように，あくまで正答（誤答）数に基づく分析だけが可能であり，受験者集団に依存した分析結果しか得ることができない。これを克服するために，項目応答理論を用いて項目ごとのテスト情報関数を示す。そして基礎資料として言語テスト研究に資するとともに，テスト改善のための指標とする。

分析モデルには項目困難度パラメータ（$b$）を推定する1母数ロジスティック・モデル（Raschモデル）を用いた。項目応答理論の枠組みにおいては項目弁別力パラメータも推定する2母数ロジスティック・モデル，さらに当て推量パラメータも含む3母数ロジスティック・モデルが存在するが，テストを運用する上での取り扱いが容易であり，母数の推定も容易であるRaschモデル（豊田，2002a）を用いた。

**表3　項目応答理論に基づく分析**

| 問題 | $b$ | SE ($b$) | $t$ | 適合性判断 | 問題 | $b$ | SE ($b$) | $t$ | 適合性判断 |
|---|---|---|---|---|---|---|---|---|---|
| FO01 | -1.808 | 0.070 | -4.313 | 過剰適合 | PT06 | 0.791 | 0.057 | 1.067 | 適合 |
| FO02 | -1.338 | 0.062 | -4.325 | 過剰適合 | PT07 | 0.188 | 0.054 | 0.063 | 適合 |
| FO03 | -2.189 | 0.079 | -9.933 | 過剰適合 | PT08 | 0.410 | 0.055 | -5.297 | 過剰適合 |
| FO04 | -0.169 | 0.054 | -2.479 | 過剰適合 | PT09 | -0.463 | 0.054 | -3.213 | 過剰適合 |
| FO05 | -0.872 | 0.057 | -5.583 | 過剰適合 | PT10 | -0.192 | 0.054 | -2.425 | 過剰適合 |
| FO06 | -1.537 | 0.065 | 0.450 | 適合 | PT11 | 0.032 | 0.054 | 0.796 | 適合 |
| FO07 | -0.895 | 0.057 | -7.071 | 過剰適合 | PT12 | 0.006 | 0.054 | -3.041 | 過剰適合 |
| FO08 | -1.818 | 0.070 | -11.037 | 過剰適合 | PT13 | 0.608 | 0.056 | -3.403 | 過剰適合 |
| TH01 | -0.538 | 0.055 | -7.497 | 過剰適合 | PT14 | 0.788 | 0.057 | 3.889 | 不適合 |
| TH01 | -0.756 | 0.056 | -6.986 | 過剰適合 | TW01 | 1.244 | 0.063 | 7.863 | 不適合 |
| TH03 | -0.472 | 0.055 | -3.499 | 過剰適合 | TW02 | 1.056 | 0.060 | 15.381 | 不適合 |
| TH04 | -1.123 | 0.060 | -8.286 | 過剰適合 | TW03 | 1.107 | 0.061 | 5.842 | 不適合 |
| TH05 | -0.293 | 0.054 | -3.565 | 過剰適合 | TW04 | 0.488 | 0.055 | 0.621 | 適合 |
| TH06 | -0.538 | 0.055 | -2.821 | 過剰適合 | TW05 | 1.209 | 0.062 | 3.733 | 不適合 |
| TH07 | -1.018 | 0.058 | -5.889 | 過剰適合 | TW06 | 0.015 | 0.054 | 0.316 | 適合 |
| TH08 | -1.474 | 0.064 | -12.186 | 過剰適合 | TW07 | 0.985 | 0.059 | 5.589 | 不適合 |
| TH09 | -1.281 | 0.062 | -4.367 | 過剰適合 | TW08 | 0.537 | 0.055 | 4.427 | 不適合 |
| TH10 | -0.669 | 0.056 | -6.362 | 過剰適合 | TW09 | 0.864 | 0.058 | 5.440 | 不適合 |
| TH11 | 0.128 | 0.054 | 0.704 | 適合 | TW10 | 1.817 | 0.073 | 12.373 | 不適合 |
| TH12 | -0.807 | 0.057 | -3.438 | 過剰適合 | TW11 | 1.133 | 0.061 | 11.288 | 不適合 |
| PT01 | 0.218 | 0.054 | -3.264 | 過剰適合 | TW12 | 0.844 | 0.058 | -0.301 | 適合 |
| PT02 | -1.199 | 0.060 | -3.770 | 過剰適合 | PO01 | 1.308 | 0.064 | 15.397 | 不適合 |
| PT03 | 0.593 | 0.056 | 0.331 | 適合 | PO02 | 1.791 | 0.072 | 17.751 | 不適合 |
| PT04 | 0.739 | 0.057 | 2.048 | 不適合 | PO03 | 1.617 | 0.069 | 11.173 | 不適合 |
| PT05 | -0.132 | 0.054 | -1.199 | 過剰適合 | PO04 | 1.067 | 0.060 | 7.726 | 不適合 |

分析結果は，各項目の項目困難度パラメータである $b$，測定の標準誤差である SE ($b$)，分析モデルとの適合度 $t$ とそれに基づく適合・過剰適合・不適合の判断について，それぞれ表3に示される。今回の分析では受験者数が多いことから，各受験者に関するパラメータは検討の対象とせず，各項目に対するパラメータのみを扱う。項目困難度パラメータは平均0.000・分散1.000として算出されており，理論的に上限値も下限値も存在しない。多くの受験者データを得たために，測定誤差である SE ($b$) は非常に小さい値になっている。

適合度 $t$ は，分析に利用したロジスティック・モデルに対する適合（逸脱）の指標であり，+2.000を超えるものは逸脱が大きいものとして不適合，-2.000を下回るものは過剰適合，として判断される。分析モデルから逸脱していて不適合の値を示す項目は削除したが，過剰適合を示す場合には理論と観測変数の合致について問題はないと判断できるため，特に措置しないものとした。

適合度を基準に考慮した結果，15問については不適当な項目であると判断した。特に日本英語検定協会によって「大学初級程度」の難易度とされる準1級（PO）は4問中全問，「高校卒業程度」の難易度とされる2級（TW）は12問中9問が，不適合すなわちこの受験者集団（高校2年生の1月から3月期）にとっては，項目応答理論を用いて分析をするにあたって不適切である，と判断できる。

## 3．まとめ

項目応答理論に基づく分析結果において不適合と判断された項目は，古典的テスト理論の枠組みにおいても同様に標準適切度合計が比較的低く，実施に適さない問題であると判断された。これらの項目を改善することによって，より適切に英語学力を測定することが可能になると考えられる。

テスト全体の改善に際しては，古典的テスト理論や項目応答理論に基づく分析結果に従って適さない項目を削除するだけでなく，各項目を改善することも重要である。古典的テスト理論に基づくならば実質選択肢数をもとに検討し，実質選択肢数が高ければ錯乱肢が正答に紛らわしすぎる可能性や，低ければ錯乱肢が正常に機能していない可能性も考えられる。また項目困難度（正答率）が高ければその項目は簡単すぎるのであるから問いを難しくすること，低ければ問い自体を再検討する方法も考えられる。項目応答理論に基づく場合には，適合度に基づいて問題のある項目を洗い出すほかに，困難度パラメータを用いてテスト全体が十分に弁別力を持つように項目を取捨選択することも，冗長性をなくす観点からは重要である。

問題は，これらの指標は共変するために，ある指標だけに着目することはできないことと，言語テスト自体の構成概念的妥当性や表面的妥当性も考慮する必要があることである。ある指標を解釈する際には他の指標も検討するなど慎重になると同時に，項目自体の妥当性についても十分に留意しなければならない。

ただし，言語テスト作成段階における妥当化については従来多くの開発や改善が試みられている一方で，受験後の段階において受験者集団に依存しないパラメータの推定を可能にする項目応答理論を用いて分析を行い，その結果を考慮したうえで妥当化を行ってテストを改善するという面においては，その方法論も研究成果もいまだ発展初期にあるといえよう。縦断的に学習成果を測定して到達目標型教育に資することができるような言語テストを作成するために，これまでの言語テスト理論に基づくテスト開発に併せ，受験データに基づく観点からの改善も不可欠である。

本研究は，そのための基礎的研究として，分析結果を示した。

**付記**

本研究の一部は，財団法人日本英語検定協会による「第14回『英検』研究助成（研究題目：高校生英語学習者の学習方略使用と学習達成）」によって行われた。

**参考文献**

The Chauncey Group International. 2002. *TOEIC Technical Manual.* Available: [http://www.toeic.com/pdfs/TOEIC_Tech_Man.pdf] [December. 2002]

国立教育政策研究所. 2002. 『評価規準の作成，評価方法の工夫改善のための参考資料－評価規準，評価方法等の研究開発（報告）－』. Available: [http://www.nier.go.jp/kaihatsu/index.htm] [December. 2002]

広島大学. 2002. 「評議会だより（平成十四年七月九日）」. 『広大フォーラム34期3号』. Available: [http://www.hiroshima-u.ac.jp/Committee/forum/34-3/hyougikai.html] [December. 2002]

前田啓朗. 2002. 「高校生英語学習者の学習方略使用と学習達成」. *STEP Bulletin,* 14, 26-38.

中村洋一(著). 大友賢二(監修). 2002. 『テストで言語能力は測れるか』. 桐原書店.

大友賢二. 1996. 『項目応答理論入門』. 大修館書店.

Ohtomo, K., Y. Nakamura, and M. Akiyama. 2002. TDAP (Test Data Analysis Program) Ver. 2.0.

島谷浩・木下正義・T. Laskowski・高梨芳郎・大津淳史・石井和仁・川尻徳. 1999. 「日本と韓国の高校生に対する英語リスニング・テストに見られたテスト・バイアス：古典的テスト理論と項目応答理論に基づくデータ分析」. *Japan Language Testing Association (JLTA) Journal. 2,* 35-54.

豊田秀樹. 2002a. 『項目反応理論［入門編］』. 朝倉書店.

豊田秀樹. 2002b. 『項目反応理論［事例編］』. 朝倉書店.

筑波大学外国語センター. 1994. 『外国語検定制度開発ワーキンググループ研究報告書』

資料4　調査に使用された英検の過去問題(正解は下線の選択肢)

FO01. A: You look really nice in that dress, Sally.　Shall I (　　) your picture?
B: Sure.　Thank you, Jane.
1 write　　2 lose　　3 take　　4 find

FO02. Mike (　　) fifteen minutes for John at the station.
1 waited　　2 wanted　　3 silent　　4 took

FO03. Emily's mother made a very (　　) dress for her.
1 glad　　2 main　　3 silent　　4 pretty

FO04. A: Who (　　) the game first?
B: I do.
1 plays　　2 played　　3 playing　　4 to play

FO05. A: How did you spend last weekend?
B: I went (　　) with my friend Mary.
1 swim　　2 swimming　　3 swam　　4 swum

FO06. A: Where were you this afternoon, Sally?
B: I (　　) in the computer shop with my friend.
1 were　　2 am　　3 was　　4 are

FO07. The train will (　　) for Kyoto at two o'clock this afternoon.
1 leave　　2 take　　3 keep　　4 like

FO08. Rob and his mother are (　　) cookies now.
1 make　　2 made　　3 making　　4 makes

TH01. A: Mary, the TV is too loud.　Will you turn it (　　) a little?
B: OK, Dad.
1 down　　2 from　　3 at　　4 above

TH02. A: When should I throw (　　) these old newspapers?
B: Just leave them.　I'll take them to the recycling center tomorrow.
1 open　　2 away　　3 about　　4 through

TH03. A: How long has your grandfather (　　) in Tokushima?
B: All his life.
1 lived　　2 lives　　3 living　　4 to live

TH04. A: Bob, can I borrow your camera?
B: Sure, but it doesn't (　　) very well.　It's really old.
1 return　　2 work　　3 live　　4 stay

TH05. A: Don't forget to tell Jenny about the concert tomorrow.
B: Oh, yes.　I'll call her (　　) now.
1 ever　　2 still　　3 quite　　4 right

TH06. A: The girl (　　) the music room is a new student.
B: I know.　Her name is Nancy Brown.
1 clean　　2 cleans　　3 cleaning　　4 cleaned

TH07. A: Emily, do you know (　　) the Spanish test is?
B: I don't really know.　Maybe it's next Thursday.
1 when　　2 where　　3 what　　4 why

TH08. A: Mom, I've (　　) writing the letter in Japanese.
B: That's great, Fred.　Your Japanese pen pal will be surprised to get it.
1 finish　　2 finishes　　3 finishing　　4 finished

TH09. My father has a great (　　).　He can remember all his friends' telephone numbers.
1 memory　　2 mistake　　3 space　　4 health

TH10. A: How long has your family lived in Tokyo?
B: Since I was a boy. I moved here at the ( ) of six.
1 age 2 event 3 season 4 time

TH11. I have two cousins. One works as a police officer and the ( ) is a businessman.
1 one 2 another 3 others 4 other

TH12. First, I'd like to introduce ( ). My name is Akiko Sato.
1 mine 2 my 3 me 4 myself

PT01. A: Could you ( ) why you didn't come to soccer practice yesterday?
B: I had to study for a test.
1 argue 2 perform 3 offer 4 explain

PT02. You're driving too fast. Slow ( ), or you might hit another car.
1 under 2 down 3 away 4 into

PT03. Children came into the toy shop one ( ) another to buy the new game.
1 after 2 or 3 with 4 before

PT04. I went to my sister's wedding today. She looked so beautiful ( ) her white wedding dress.
1 in 2 on 3 at 4 upon

PT05. A: I hope to see the North Star tonight.
B: Oh, that shouldn't be too hard since it's one of the ( ) stars in the sky.
1 furthest 2 brightest 3 smartest 4 lightest

PT06. In many villages in southern France, people still use a ( ) of making wine that hasn't changed in hundreds of years.
1 law 2 sort 3 method 4 material

PT07. A: I'm calling to ( ) you that Sarah's birthday is on the 15th.
B: Don't worry. I'll be there.
1 allow 2 support 3 treat 4 remind

PT08. When Ted and Nick began to fight in the classroom, Chris ( ) them and stopped the fight.
1 practiced 2 received 3 continued 4 separated

PT09. ( ) case of fire, do not use this elevator. Please use the stairs.
1 For 2 On 3 To 4 In

PT10. A: Hello. This is Carol Adams. Can I speak to Ms. White?
B: I'm sorry, she's away on ( ). She'll be back next Monday.
1 job 2 trade 3 business 4 company

PT11. My brother can speak two foreign languages. One is Chinese and the ( ) is Spanish.
1 someone 2 another 3 other 4 others

PT12. When I was a boy, the neighbor's dog ( ) me. I still have the marks from the dog's teeth on my leg.
1 bit 2 hit 3 ate 4 met

PT13. We took the train to Tokyo Disneyland in order to ( ) the heavy traffic.
1 avoid 2 prevent 3 keep 4 accept

PT14. A: How are you ( ) along? I haven't seen you much recently.
B: Pretty good. I just got back from vacation.
1 doing 2 making 3 getting 4 living

TW01. Every year many new brands of beer go on the ( ). Sometimes they're only sold during one particular season.
1 market 2 shop 3 stock 4 store

TW02. A: Was your vacation expensive?
B: Yes, it was. But we felt that didn't matter as ( ) as we really enjoyed ourselves.
1 long 2 well 3 high 4 much

TW03. Lucy and her sister, Mary, do not get along very well and (　　) with each other nearly every day.
1 discuss　　2 argue　　3 stick　　4 consider

TW04. A: Why does Robert have such a sad (　　) on his face?
B: His cat died last night.
1 expression　　2 statement　　3 pressure　　4 delivery

TW05. Jeremy has put (　　) $200 each month. He is planning to buy a new car with the money.
1 over　　2 aside　　3 through　　4 along

TW06. I was surprised when I felt a tap (　　) my shoulder. I turned around and saw that it was Adam.
1 in　　2 at　　3 behind　　4 on

TW07. A: I sometimes (　　) money to a group that helps homeless people.
B: That's great. I should do that, too.
1 reform　　2 adopt　　3 require　　4 contribute

TW08. When David told his joke to the audience, everyone burst (　　) laughing.
1 up　　2 out　　3 around　　4 away

TW09. When I told my soccer coach that my knee hurt, he recommended I (　　) a doctor at once.
1 see　　2 to see　　3 seeing　　4 seen

TW10. To (　　) for making me wait over 30 minutes for delivery, the pizza shop let me have the pizza for free.
1 substitute　　2 apply　　3 exchange　　4 compensate

TW11. The food in Italy was so good that I (　　) a lot of weight while I was there on vacation.
1 brought up　　2 put on　　3 got on　　4 took in

TW12. During Tracy's trip around the world, her money (　　) short, so she had to call home and ask her parents to send her some more.
1 ran　　2 left　　3 sold　　4 turned

PO01. A fire on the top floor of the college dormitory forced the (　　) of dozens of students.
1 exposition　　2 evacuation　　3 declination　　4 desertion

PO02. Hospital staff are required to (　　) all instruments prior to their use in surgery in order to avoid the spread of disease.
1 exterminate　　2 extinguish　　3 neutralize　　4 sterilize

PO03. Based on present data, researchers predict that sea temperatures will ( ) rise in the next decade or two.
1 inevitably　　2 intentionally　　3 respectively　　4 voluntarily

PO04.Robert needs clothes three sizes larger than he did last year. He's really (　　) weight.
1 gotten on　　2 put on　　3 set up　　4 added up

注：FO=4級，TH=3級，PT=準2級，TW=2級，PO=準1級

ABSTRACT

# Test Improvement toward Criterion-referenced Evaluation and Education: Based on Classical Test Theory and Item Response Theory

Hiroaki MAEDA
Department of Foreign Language Research and Education
Information Media Center, Hiroshima University

This study reports test data analyzed by both Classical Test Theory and Item Response Theory as fundamental language testing research. Since Criterion-referenced Evaluation has been chosen for introduction into Japanese elementary and junior high schools in 2002, Japanese high schools in 2003, and Hiroshima University in 2004, measurements such as paper tests should be consolidated for the purposes.

Since norm-referenced evaluation has assumed a main role in evaluation, language testing research has focused on pre-test stages of test administration. This has occurred because a test is inferred to be sufficient when it can show ranking of test takers with validity and reliability at the time the test is conducted. However, in criterion-referenced evaluation, tests are reasonably inferred to measure longitudinal transitions of learners' proficiency.

Item Response Theory is recommended to analyze test data because it can estimate item characteristics regardless of test takers' proficiency. Classical Test Theory, which is comparatively simple, but limited by test takers, is used to show fundamental data. Results obtained by these analyses represent highly reliable open data for Japanese EFL learners. Those data can be used for further test development or test improvement.