5 月 30 日

# 第 2 章　英語学力測定論 (p.38~47)

## 3. 項目分析　(p.38 ~)

■ 項目分析 (item analysis): テスト項目の統計的な特徴 (適切な得点分布・信頼性) を明らかにすること。

→ 受験者: 個々のテスト項目の回答状況について診断的な情報の提供

→ 教員・プログラム開発者: 指導改善に役立つ情報の提供

→ テスト開発者・作成者: 事前のテストの得点分布やテストの難易度レベルの調整・内的一貫性の信頼性の向上・有効に機能していない項目の事前の発見や修正に役立つ情報提供

### 3.1 古典的項目分析

各項目の得点の積み上げが合計点となる。

① 正答・誤答: [ 2 値型採点 ] 正答 = 1 点, 誤答 = 0 点

② 部分点を与える方法

→ 受験者の各項目への応答状況が数値化されるため、項目得点の分布を統計的に記述することが可能。
項目分析の結果、各種の項目統計量 (難易度・弁別力) が算出される。

■ 項目難易度: 各項目がどのくらい難しいかを示す指標。正答や誤答の潜在的な問題を診断できる。

□ 項目難易度指標 (item difficulty index, $p$): $0 \leqq p \leqq 1$
2 値型採点項目:　正答者数／全受験者数　の割合
部分点採点項目: 項目得点の平均値／(部分点の最大可能値－最小可能値)　の割合
錯乱肢を選んだ受験者の割合・各部分点を得た受験者の割合を求めることも可能。

■ 項目弁別力: 各項目がどれだけ成績上位者と下位者を分別できるかの指標。
得点分布の形状や信頼性に影響を与える。

□ 項目弁別力指標 (item discrimination index, $D$): $-1 \leqq D \leqq 1$, プラスの値が大きいほど弁別力が高い。
(その項目に回答した成績上位群／全体) – (成績下位群／全体)
成績上位群・下位群の人数は、合計点でそれぞれ 3 分の 1 あるいは 27%

□ 点双列相関係数 (point-biserial correlation coefficient): 項目得点と合計点の相関係数を計算する手法
2 値尺度の項目得点⇔間隔尺度としての合計点　の積率相関係数
※ある項目の弁別力が高ければ、その項目の正答者の合計点はより高く、誤答者はより低い。
→ 項目得点とテストの合計点との間には強い正の相関関係がある、という予想を前提とするもの。

□ 弁別力指標は教室内のテストで用いられる一方で、点双列相関係数は大規模テストで用いられる。

■ 項目バンク: テストの品質管理として、項目の内容と項目特性値の情報を記録する。

■ 項目選択の基準:

【 集団基準準拠テスト 】

□ 項目難易度が 0.5 前後の項目を中心とする。
→ 難易度が極端な場合は、弁別力指標が低い傾向にある。

□ できるだけ弁別力の高い項目を選ぶ ($D \geqq 0.3$, $D < 0.2$ の場合は削除か修正が望ましい)
→ 標準偏差を大きくすることができ、内的一貫性の信頼性を高めることができる。

【 目標規準準拠テスト 】　合否の分岐点の割合に近い項目難易度指数を持つ項目を選択する。

3.2 古典的項目分析の限界

(1) 標本に依存した記述統計量しか得られない。【標本依存】

項目統計量→ テストを受験した特定の集団や受験者に依存する。

テスト得点→ テストを構成する特定の項目群に依存する。

→ 受験者集団が異なる場合は項目統計量の比較・テストが異なる場合には、受験者の得点の比較が困難

→ 学力の伸びを測るために平行テストが必要だが、古典的項目分析に基づく項目統計量から作成することは困難。

(2) ある項目の項目特性値とある受験者の能力水準値とを結びつける情報が得られない。

(3) 項目という測定の 1 面 (相: facet) のみしか扱っていない。

・ スピーキングやライティングのテストでは、複数のタスクがありそれを複数の評定者が評価する。

→ 評定者の相対的な厳しさの情報を得る必要がある。

4. 項目応答理論

4.1 考え方

■ 項目応答理論 (item response theory: IRT):

・古典的テスト理論の限界を克服するために開発された測定モデル

・テスト項目に対する多くの受験者の応答パターンから、項目の特性値と受験者の能力値を推定する。

→ 能力値は仮定された潜在特性尺度上に位置づけて表される。

→ 項目の困難度と受験者の能力値とが独立の特性値として推定される。

■ 受験者の学力・能力 ($\theta$) は項目に対する正誤の応答パターンから推定される。

→ 推定の前提 = 「能力の高い受験者はよりその項目に正答する確率が高い」

■ 局所独立の仮定 (local independence assumption): ある特定の能力値 $\theta$ を持つ人の項目への応答はそれぞれの項目で互いに独立である。

= 1 次元性の仮定 (unidimensionality assumption): すべての項目が 1 つの特性のみを共通して測定している。

4.2 項目特性曲線

■ 項目特性曲線 (item characteristic curves): テスト項目の特性を表す曲線で、ある能力値を持つ受験者がその項目に正答する確率を表す。受験者の能力値が高くなれば、正答率も高くなるという前提。

(a) 項目困難度パラメタ (difficulty parameter): 項目特性曲線が右にあるほど、項目を正答するために要求される能力値レベルが高いので、難しい項目である。

(b) 項目識別力パラメタ (discrimination parameter): 項目特性曲線の傾きが急なほど、能力値の高低をより明確に識別する (項目 1, 3)。

(c) 当て推量パラメタ (guessing parameter): 能力値の低い受験者の正答確率が 0 でない場合は、当て推量で正答する可能性を示す。多肢選択式項目でよく見られる。

※パラメタ: 項目特性曲線の形状を決定するのに必要な定数。

■ 項目応答理論の基本モデル式: 能力値 $\theta$ の受験者が項目 $j$ に正答する確率。

→ ロジスティック関数を用いると数学的に扱いやすい。

□ 1 パラメタ・ロジスティック・モデル(1PLM = ラッシュモデル): (b) のみを含める (a = c = 0)

□ 2 パラメタ・ロジスティック・モデル(2PLM): (a), (b) を含める ( c = 0)

□ 3 パラメタ・ロジスティック・モデル(3PLM): (a), (b), (c) を含める

→ 含めるパラメタが増えるほど、安定した推定値を得るために多くの受験者が必要になる。

※2 値型 IRT モデル (dichotomous IRT model): 項目得点が 1 か 0 で採点される 2 値型項目を対象とする。
⇔多値型 IRT モデル: 部分点や段階点で採点される。

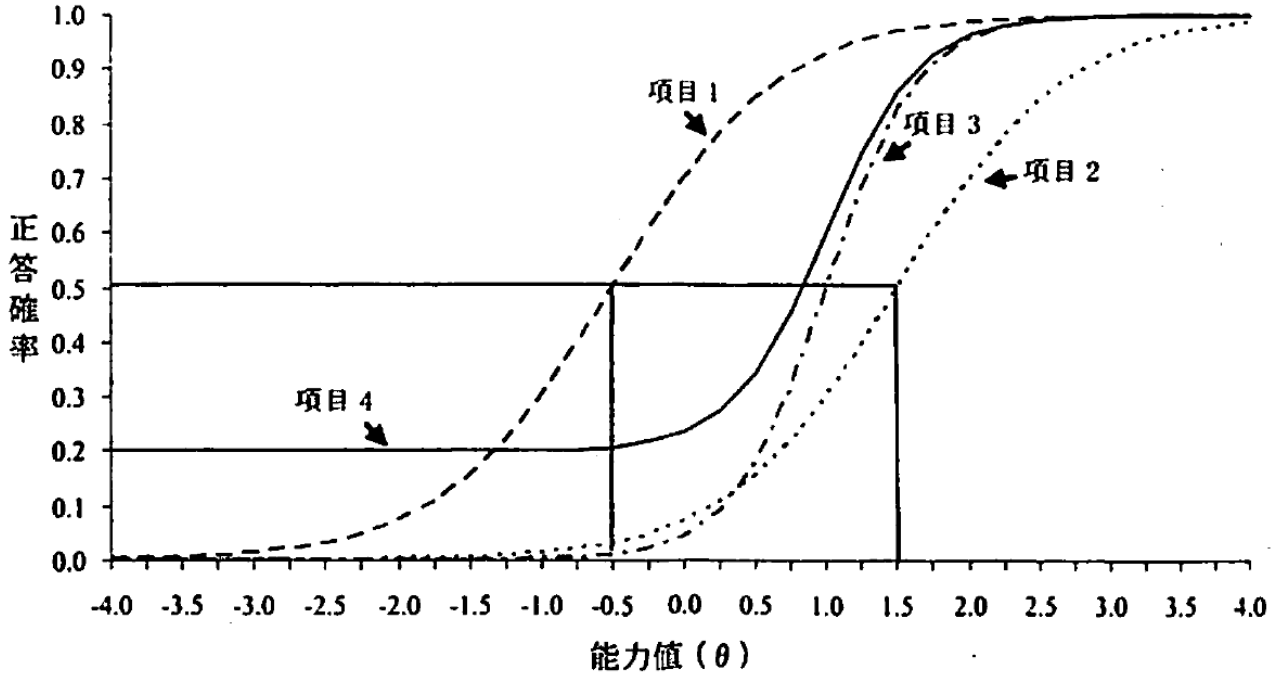


図 1　項目特性曲線の例 (p. 42 より抜粋)

4.3 情報量と情報関数

- 項目情報量 (amount of item information): 受験者の能力値水準を推定する際、ある項目が表す情報量。項目情報関数で表される。
- 能力値水準に応じて情報量が異なる。→ 情報量が大きいと精度の高い推定値 (estimates) が得られる。
- テスト情報量 (amount of test information $I(\theta)$): テストを構成する各項目の項目情報量を合計したもの。
- テスト情報関数 (test information function): ある能力値で与えられるテスト情報量を表した曲線。テストが提供する各能力値水準における情報量の大きさを推定する。
- 推定の標準誤差 (standard error of estimation): 各能力値水準における測定精度。　1／　√テスト情報
- 異なる能力値水準において最大の項目情報量を提供する項目を選別することで、測定精度の高いテストを作成できる。

【集団基準準拠テスト】幅広い能力値水準で最大の情報量が得られる項目を選択する。

【目標規準準拠テスト】合否の分岐点に近い能力値で最大の情報量が得られる項目を選択する。

4.4　テスト開発への応用

■ 項目応答理論の特徴

・受験者集団に存在しない項目特性値が推定できる。

・特定のテスト項目群に依存しない受験者の能力値が推定できる。

・受験者ごとにそのテストによる測定精度を評価することができる。

→ テスト (尺度) の等化を容易にするため、テスト開発の測定モデルとしてよく用いられる。

□ テストの等化 (equating): テストの異なる版から得られた測定結果を相互に比較可能にするため共通尺度上で表すための手続き。Ex. TOEFL PBT のテストはどの回のテストも比較可能とされている。

■ 英語力の経年変化の研究: 項目応答理論の等化により、複数年度の学力テストが比較可能となる。

・高校生の英語力やセンター試験で測定される英語力が年々低下している。

・英語力の学校間格差が広がっている。

■ その他の役割

・DIF (特異項目機能) の項目の検出

・共通尺度上で困難度や識別力が推定されている項目バンクの作成

・学力の国際比較研究

4.5　多相ラッシュモデル

■ ラッシュモデル: 1 つの相 (fact) のみを分析のプロセスで扱う。

■ 測定の相: 測定のプロセスでテスト得点に影響を与えると考えられる側面。
ex. 項目・タスク・評定者・回数・版・テスト方法　→ 各相は複数の条件 (condition) を含む。

■ 多相ラッシュモデル: ラッシュモデルの拡張モデルの 1 つ。複数の相の分析を行うことが可能。
ex. 受験者が複数のタスクを与えられてタスクごとに異なる評定者によって評定される場合
→ 項目の相対的な困難度に加えて評定者の相対的な厳しさの程度の情報が得られる。
→ モデルに適合しない項目・評価者を特定 → 項目の修正・削除や評定者の採否・再訓練

■ 多層ラッシュモデルには FACETS program (Linacre & Wright, 1993) が用いられる。

■ FACETS: ロジット尺度 (平均 0, 標準偏差 1 に標準化された尺度) 上で、項目の困難度や評定者の意厳しさの程度を推定値として示す。

| ロジット値 | 項目 | 評定者 |
|---|---|---|
| プラス (大きい) | 難しい項目 | 厳しい評定者 |
| マイナス (小さい) | 易しい項目 | 甘い評定者 |

→ Infit: データとモデルの適合度を表す指標。
±2 以上だと適合度が悪い (misfit)　→ 項目・評定者の変更・修正に生かせる

■ FACETS では相の組み合わせから一貫性が見られない測定に関する情報も提供される。【交互作用】
1) 評定者と受験者, 2) 項目と受験者, 3) 項目と評定者 の 3 種類の組み合わせの場合
1) ある特定の集団には甘い採点をする評定者
2) ある項目群について一貫性のない、またはバイアスのある評定をする評定者
3) 特定の項目群に特別な応答をする受験者