精度の高いテストの研究*

樋口三郎 (理工学部)†

概要

精度の高いテストの設計を目的とする, 古典テスト理論, 項目応答理論によるテスト結果の分析とテストの設計の考え方, テスト結果の分析のためのソフトウェアを紹介する.

1 はじめに

本研究は, テスト理論の知識と実践を本学の教員集団に普及し, それによってよりテストの精度の向上, さらには教育の改善に資することを目的とする.

テストとは, 受験者の内部にあるひとつの能力の (1 個の数値で表現できる) 値を, 外からの働きかけと測定によって推定するものである.

テストは様々な目的で行われる. 授業科目において学習者の評価を行うためのテスト (授業内テスト)[1], 入学者などを選抜するためのテスト (入学試験), 一般に何らかの観点から幅広い受験者の能力を測定するためのテスト (TOEFL, TOEIC など) などがある. 本研究では, 実技ではなくペーパーテストを中心に考察する.

テストの設計, 実施, 結果の解析に有用な情報として, 計量心理学に起源を持つテスト理論の立場からの解説として文献 [2, 3] がある.

学習理論およびインストラクショナルデザインの立場からは, (ペーパーテストと限らない) テストは, 学習の入口と出口を規定するという機能を持つ. また, テストは学習者の状態を通して教材・教育を評価するためのツールと考えられる. 学習の効果測定に関するカークパトリックの 4 段階評価モデルでは, レベル 2 のラーニングにおいてテストを使用することが想定される [4].

また, これらと重なるが, 第二言語習得, 日本においては英語教育の分野で, 言語テストが活発に研究されていることは特筆に値する [5, 6]. この分野の知見は他の分野にも有用と考えられる.

2 テストの品質

テスト理論では, 妥当性 (validity), 信頼性 (reliability) という観点からテストの品質を評価する. また, 特に言語テストの分野においては, これらに加えて, 波及効果 (washback effect) という機能が強調される.

■妥当性 テストが, 測りたい能力を測定するものになっているか, という観点である. 例えば, 英語の能力を測定するためのテストでも, 適切に構成されないと, 英語の能力がなくても, 常識やテストに対処する力や (丸) 暗記力があれば正解できてしまうことがある. このようなテストは, 妥当性に欠けると表現する.

* 龍谷大学大学教育開発センター 2010 年度自己応募研究プロジェクト

† http://hig3.net

■信頼性　テストを繰り返し実施したときに, 偶然の要素があっても, 毎回ほぼ同じ結果を与えるかどうか, という観点である. 例えば, 多数の九九の問題からなるテストは (学習が停止しているなら) 毎回ほぼ同じ結果を与えるだろうが, トリッキーな補助線が必要な幾何の問題 1 個からなるテストは, たまたま補助線を思いつくかどうかで, 毎回の結果が大きく異なるかもしれない. 前者は信頼性が高く, 後者は信頼性が低いと言える.

妥当性があるという前提のもとで, 信頼性の高いテストが, 精度の高いテストであると考えることができる.

一般的には, 数多くの独立な客観評価方式の小問からテストを構成することにより, テストの信頼性を高めることができると考えられている. また, このように構成することにより, 古典テスト理論, 項目反応理論による分析が容易になる.

■波及効果　受験者はテストでよい結果を出すために熱心に学習することがある. 逆に言うと, 適切なテストを行えば, 受験者をよい学習に導くことができる. このような機能を波及効果という. 波及効果を持つためには, 妥当性は必要であり, また信頼性とも関係する.

3　古典テスト理論

古典テスト理論 [7, 8] では, 特定のテストにおいて, 各受験者の点数 X は, 受験者の能力により定まる真の得点 T と, 確率変数である誤差 e の和として得られると考える:

$$X = T + e.$$

信頼性は, 受験者の集合を考えたときの X の標本分散 $V(X)$ と T の標本分散の比 $V(T)/V(X)$ として定義される. 古典テスト理論の中心的な課題は, テスト実施結果からテストの信頼性を評価することである.

3.1　信頼性の評価

受験者とテストを固定したとき, T は一定だが, e は平均 0 で毎回変化する. そこで, 信頼性を評価するには, 実際に同じ受験者に同じテストを繰り返して受験してもらって, 点数の変化を観察すればよい (テスト-再テスト法). しかし実際には, 問題や解答の過程を記憶したりしていれば, 同じ状況での受験は困難である.

一方, テストの構成から点数が同じになることが確実な, 別のテストを受けてもらう方法もある (平行テスト). これをより簡易化した方法として, 1 個のテストを前半後半 2 つのテストとみなし, 点数を比較する方法がある (折半法). さらに突き詰めると, かならずしもちょうど 2 個に分割する必要はなく, すべての小問について点数がどう分布するかから信頼性の情報が得られることがわかる. これは, 内的一貫性から信頼性を求める考え方である.

3.2　Cronbach の α 係数とテストの点数の誤差

誤差の分散に対して適当な仮定をおき, 各小問の点数の分散の関数としてテストの信頼性を評価する標準的な量が, Cronbach の α 係数である. この α から, テスト全体の得点に想定される誤差の大きさを推定することができる.

3.3　信頼性を高めるための一般的な工夫

テストを構成する各小問がテスト全体の信頼性に寄与しているかどうかを判定する指標として, 弁別力, 弁別係数, 項目特性曲線などがある. テスト実施後に, これらを検討して, 必要なら小問を入れ替え, テストを改善

図 1　Moodle 1.9 の小テストモジュールの分析機能．正解ファシリティとは項目容易度, 識別指数とは弁別力, 判別係数とは弁別係数のこと．

することができる．ただし, このような改善作業は, テスト問題を非公開にして繰り返し利用するときに効果的なものになる．

3.4　古典テスト理論による分析のためのソフトウェア

無料で使用できるものとして, 統計解析用プログラミング言語 R[9] の `CTT`,`psy`,`psych` などのパッケージに, Cronbach の α を計算する関数が含まれる．これらは, アーカイブ C-RAN[10] に収録されている．R 本体, およびパッケージの多くはオープンソースで開発が行われており, 新たなパッケージ, ソフトウェアを開発する上でも有効な情報源となる．

また, 次節で述べる, 項目反応理論による分析のためのソフトウェアの多くにも, 古典テスト理論による分析を行う機能が一部含まれている．

本学の e ラーニングサイト ReLS で使用している LMS, Moodle の ‘小テスト (Quiz)’ モジュールには, Moodle 上で行ったテストについて, 古典テスト理論におけるいくつかの指数を計算して表示する機能がある (表 1)

本研究プロジェクトでは, `http://www.a.math.ryukoku.ac.jp/~hig/eproj/testing/` において, データファイルをアップロードすると古典テスト理論に基づいて分析し表示するサービスを試作し, 試験的に提供している (図 2,3)

4　項目反応理論

古典テスト理論では, 特定の受験者集団を固定して, 小問の難易度やテストの信頼性を考えた．これに対して, 項目反応理論 (項目応答理論) は, 小問の難易度と受験者の能力ごとに分離して考えることができる．その結果として, この程度の能力を測るときには精度がどのくらいであるかを予言したりすることができる．

TOEFL, GRE, GMAT などの高い精度が求められるテストで使用されている．教科書として [11] がある．また, Rasch モデリング [6] は項目反応理論とは異なる思想に基づくものだが, 項目反応理論と関係づけて考えることができる．

テストスコア解析サービス

古典テスト理論を用いて, 各小問別, 各受験者別のスコアのデータを解析します.

基本的な統計量に加えて次のものを計算します.

テスト全体についての情報

- 信頼性係数(弱同族測定=Cronbachのα)(説明)
- テストの合計点に想定される誤差(説明)

小問=項目についての情報

- 項目容易度 FI=facility index(説明)
- 項目弁別力 DI=discrimination index(説明)
- 項目弁別係数 DC=discrimination coefficient(説明)
- 項目特性曲線 ICC=Item Characteristic Curve(説明)

下のフォームから, CSVファイルをアップロードしてください.

下のフォーマットに従った, CR/LF改行の(mac形式ではない)CSVファイルである必要があります.

	小問1	小問2	…	小問N
配点	小問1の満点	小問の満点	…	小問Nの満点
学生1	スコア	スコア	…	スコア
学生2	スコア	スコア	…	スコア
…	…	…	…	…
学生M	スコア	スコア	…	スコア

例: CSV

タイトル, N,M,テスト全体の合計点などは書かないでください.

スコアには数値を, 小問1には問や学生を区別するための任意のラベル(文字列)を書いてください. ただし, 学籍番号や学生名は使わないでください.

個人が特定できる情報が送信されても, サーバ側では無視し, 処理も保存もしません. しかし通信路は保護されていませんので, あらかじめ個人が特定できる情報は除いてくださるようにお願いします.

選択... Upload

このサービスは, 龍谷大学大学教育開発センター 2010年度自己応募研究プロジェクト 精度の高いテストの研究 の成果物です.

Saburo Higuchi 樋口三郎 hig@math.ryukoku.ac.jp http://www.math.ryukoku.ac.jp/~hig/,

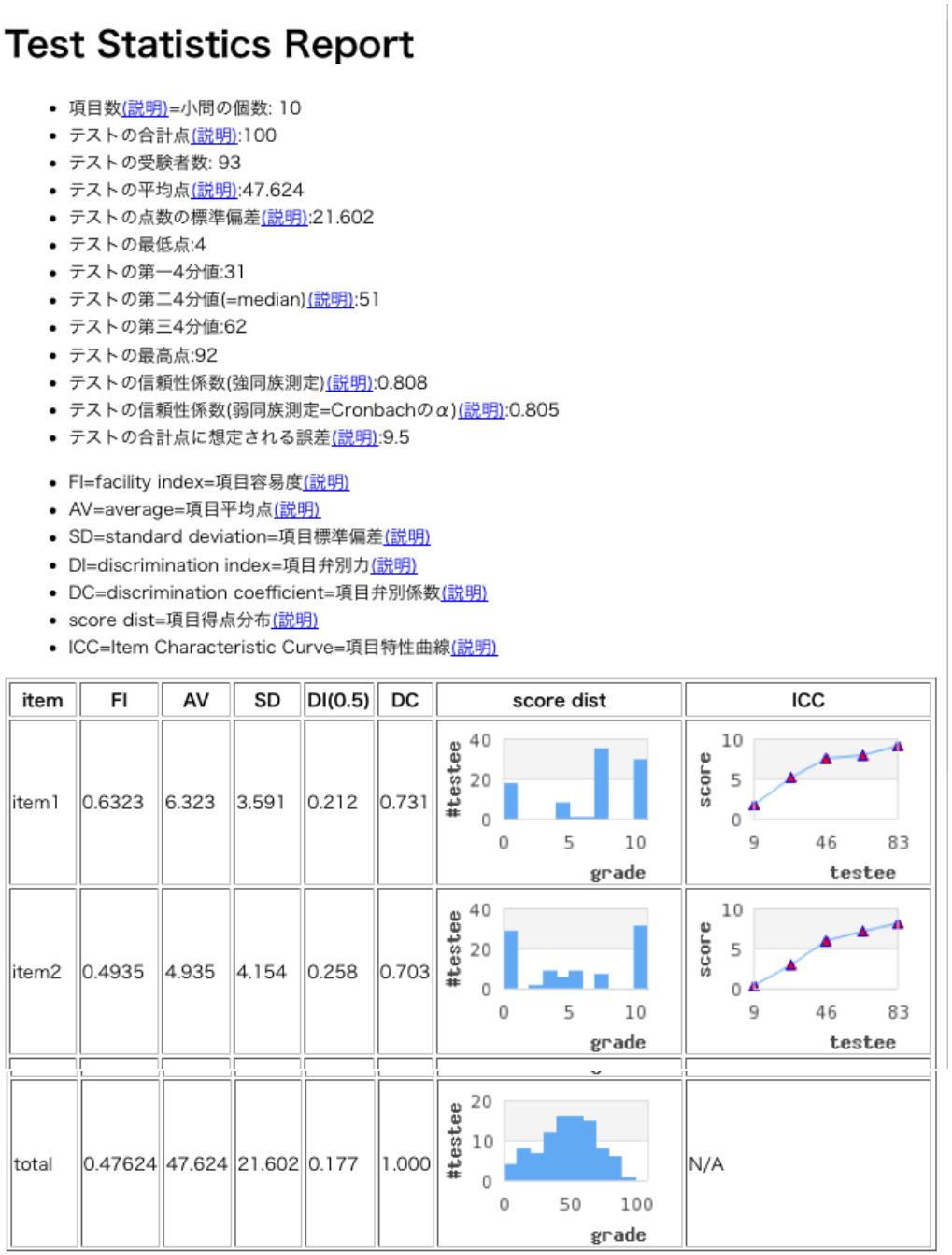
図 2　テストスコア解析サービスのトップおよびアップロードページ

4.1　項目反応理論の考え方

項目反応理論では, 各受験者が能力値 θ , 各小問が, 弁別力, 難易度などのパラメタ群 $\boldsymbol{a}_k$ をもち, 正答の確率が非線形な関数 $p(\theta, \boldsymbol{a}_k)$ により定まると考える. 関数形とパラメタの個数の異なるいくつかのモデルが使用されている.

受験者の解答パターンが得られると, 能力値と小問のパラメタを推定することができる. これにより, 受験者グループとは独立な問題の難易度を知ることができる. 原理的には, (受験者または問題に重なりがあれば) 入試における A 日程と B 日程の点数を換算することができる.

また, 小問の特定の集合であるテストが, どの能力値の受験者に対してはどの程度の精度を持つかを考えることができる. 原理的には, 合否ボーダー (と予想される能力値) のあたりの精度が特に高いテストや, 広い範囲



Test Statistics Report

- 項目数(説明)=小問の個数: 10
- テストの合計点(説明):100
- テストの受験者数: 93
- テストの平均点(説明):47.624
- テストの点数の標準偏差(説明):21.602
- テストの最低点:4
- テストの第一4分値:31
- テストの第二4分値(=median)(説明):51
- テストの第三4分値:62
- テストの最高点:92
- テストの信頼性係数(強同族測定)(説明):0.808
- テストの信頼性係数(弱同族測定=Cronbachのα)(説明):0.805
- テストの合計点に想定される誤差(説明):9.5

- FI=facility index=項目容易度(説明)
- AV=average=項目平均点(説明)
- SD=standard deviation=項目標準偏差(説明)
- DI=discrimination index=項目弁別力(説明)
- DC=discrimination coefficient=項目弁別係数(説明)
- score dist=項目得点分布(説明)
- ICC=Item Characteristic Curve=項目特性曲線(説明)

item	FI	AV	SD	DI(0.5)	DC	score dist	ICC
item1	0.6323	6.323	3.591	0.212	0.731	#testee 40 20 0 / 0 5 10 grade	score 10 5 0 / 9 46 83 testee
item2	0.4935	4.935	4.154	0.258	0.703	#testee 40 20 0 / 0 5 10 grade	score 10 5 0 / 9 46 83 testee
total	0.47624	47.624	21.602	0.177	1.000	#testee 20 10 0 / 0 50 100 grade	N/A

図 3　テストスコア解析サービスの解析例

で一定の精度を保つテスト, などを用途に応じて構成することができる.

4.2　項目反応理論による分析のためのソフトウェア

項目反応理論による分析を行う商用のソフトウェアは多く存在する. BILOG-MG, MULTILOG などである. また, Rasch モデルに特化したものとして WINSTEPS,Facets などがある.

日本発の無料のソフトウェアとして, EasyEstimation, Exametrika[13] などがある.

また, R のパッケージとして, catR,plink などがある. 古典テスト理論および項目反応理論に関する R パッケージのまとめとして文献 [14] がある.

項目反応理論によるパラメタ推定には多量の計算が伴うためか, サーバ側で計算を行うような Web サービスは発見できなかった.

5　考察

項目応答理論による解析とアイテムバンク (よい小問の集合) の構築には大きなコストがかかる. 必ずしもすべてのテストでこのような手法をとる必要はないと考えられる.

継続的に, 幅広い受験者に対して, 求められる精度の測定を行いたい場合, および, 非常に高い精度の測定を求められる場合 (入学試験など) には, 項目反応理論を用いてテストを設計することが適切とされている.

一方, 授業内テスト, 期末試験においては, 波及効果を重視するとともに, 授業の目標を達成できたかに限って測定するテストで十分と考える. しかしその場合も, 各小問が適切かどうかを, 古典テスト理論, 項目応答理論を用いて検証することが有効である.

参考文献

[1] 京都 FD 開発推進センター（編）. マンガ FD ハンドブックおしえて!FD マン [成績評価編]. 京都精華大学事業推進室, October 2010.

[2] 池田央. テストの科学 — 試験にかかわるすべての人に. 日本文化科学社, 1992.

[3] 日本テスト学会（編）. 見直そう, テストを支える基本の技術と教育. 金子書房, 2010.

[4] 鈴木克明. 教材作成マニュアル—独学を支援するために. 北大路書房, 2002.

[5] 靜哲人. 英語テスト作成の達人マニュアル. 英語教育 21 世紀叢書. 大修館書店, 2002.

[6] 靜哲人. 基礎から深く理解するラッシュモデリング—項目応答理論とは似て非なる測定のパラダイム. 関西大学出版部, 2007.

[7] 池田央. 現代テスト理論. 行動計量学シリーズ. 朝倉書店, 1994.

[8] 植野真臣, 荘島宏二郎. 学習評価の新潮流. 行動計量の科学 4. 朝倉書店, 2010.

[9] The R Project. The R project for statistical computing. `http://www.r-project.org/`.

[10] The R Project. The comprehensive R archive network. `http://cran.r-project.org/`.

[11] 豊田秀樹. 項目反応理論 [入門編]—テストと測定の科学. 朝倉書店, 2002.

[12] 熊谷龍一. 項目反応理論と EasyEstimation のページ. `http://irtanalysis.main.jp/`, 2011.

[13] 荘島宏二郎. Exametrika. `http://antlers.rd.dnc.ac.jp/~shojima/exmk/`, 2011.

[14] Patrick Mair. CRAN task view: psychometric models and methods. `http://cran.r-project.org/web/views/Psychometrics.html`, 2011.

付録 A　研究実施記録

- 2010 年 4 月–2011 年 2 月 資料図書・雑誌の購入, 調査
- 2011 年 2 月 28 日 (月) 講演会 (深草) テスト理論を生かした問題作成と結果の分析, 靜哲人氏 (埼玉大学教育学部)
- 2011 年 3 月 4 日 (金) 講演会 (瀬田) テスト理論を生かした問題作成と結果の分析, 橋本貴充氏 (大学入試センター研究開発部)

- 2011 年 3 月 4 日 (金) 講演会 (瀬田) 項目間の依存関係の分析, 橋本貴充氏 (大学入試センター研究開発部)
- 2011 年 3 月 テスト評価報告書 (例) 公開
- 2011 年 3 月 テスト結果分析 Web サービス公開
- 2011 年 3 月 4 日 (金) 大学教育開発センター指定・自己応募研究プロジェクト発表会

付録 B　成果物の公開

本研究プロジェクトの成果物である,

- 研究報告書 (本文書)
- 期末試験のスコアの解析例を説明した文書
- テストスコア解析 Web サービス (採点結果をデータファイルとしてアップロードすることにより解析が行える)

を, `http://www.a.math.ryukoku.ac.jp/~hig/eproj/testing/` で公開する.