テスト理論に基づいた項目分析支援システムの開発と評価

林貴史[†1]　高木正則[†1]　山田敬三[†1]　佐々木淳[†1]

概要：多くの教育現場では，学習者の能力や学習効果を測定するためにテストが実施されている．しかし，テスト作成者が出題した問題の良し悪しを理論的に評価することは少ない．その要因として，テスト項目の分析に必要となるテスト理論などの専門知識の不足や，項目分析の結果から次回作問時の改善点を導き出すことが難しいことが考えられる．本研究では，テスト受験者の解答結果に基づいた次回作問時の改善点の把握を目的とし，テスト理論に基づいた項目分析の結果や次回作問時のアドバイスを提示するシステムを開発した．開発したシステムを本学のリメディアル科目の担当教員に利用してもらった結果，システム利用前よりも項目特性や改善点をより具体的に把握できるようになったことが示された．

キーワード：項目分析，テスト理論，作問支援システム

Development and Evaluation of a Test Item Analysis Support System based on Test Theory

TAKAFUMI HAYASHI[†1]　MASANORI TAKAGI[†1]
KEIZO YAMADA[†1]　JUN SASAKI[†1]

Abstract: In many education fields, various tests have been performed to measure the intelligence and knowledge of learners. It is difficult for test creators to evaluate that each test item was whether the good or bad theoretically. As a factor, usual test creators are not experts on test theory and they have less ability to analysis the test items theoretically. So, it is difficult to improve the test quality in the time of test creating. In this study, in order to improve the test quality, we have developed a Test Item Analysis Support System based on Test Theory. In the system, the previous test results are analyzed, reported and appropriate advice is provided for next test creation. After a remedial course instructor in our university had used our system, it was shown that they could grasp test items characteristics and concrete issues toward the next test improvement points better than before using the system.

Keywords: Test Item Analysis，Test Theory，Support System for Creating Test Questions

1. はじめに

多くの教育現場では，学習者の能力や学習効果を測定するためにテストを実施し，テストの得点から各学習者を評価する．しかし，テストに出題された問題（以下，項目）の良し悪しを評価することは少ない．その要因として，項目の分析に必要となるテスト理論や統計学などの専門知識が不足していることが考えられる．また，項目の分析結果から，作問時の改善点を導き出すのも難しい．そこで，我々は作問者がテスト受験者の解答結果に基づいて次回作問時の改善点を把握することを目的とし，テスト理論に基づいた項目分析支援システムを提案・開発した．本システムでは，項目の分析結果や次回作問時のアドバイスをシステムから提示する．これにより，作問経験の少ないテスト作成者でも項目を評価でき，次回作問時に作成される問題の質向上や作問者の作問スキルの向上が期待できる．

本稿では，テスト受験者の解答データに基づく項目分析結果の提示方法や，作問アドバイスの生成ルールならびに提示方法について述べる．また，岩手県立大学ソフトウェア情報学部の初年次に開講されている「情報基礎数学 A」（以下，基礎数 A）の授業時に実施している確認テストの解答結果を活用し，科目担当教員に本システムを利用してもらった結果から本システムの有効性を評価する．

†1 岩手県立大学大学院ソフトウェア情報学研究科
Graduate School of Software and Information Science, Iwate Prefectural University

2. 関連研究

樋口[1]は，テスト理論の知見を有さない教授者が容易にテスト理論を用いて各小問別，各受験者別のスコアデータを解析できる Web アプリケーションを開発している．基本的な統計量に加え，テスト全体についての情報である信頼性係数やテストの合計点に想定される誤差の情報を計算できる．また，熊谷[2]は，項目反応理論によるテスト分析を行うソフトウェアの Easy Estimation を開発している．Easy Estimation は研究目的に限り無料で利用できる国産のフリーソフトウェアであり，GUI（Graphical User Interface）によりマウス操作のみで分析できるため，テスト分析の入門者においても容易に分析できる．多母集団分析や一部の項目母数固定による分析など，実用上必要な分析オプション

も用意されている．これらのシステムでは，テスト結果の分析といった点で本研究と類似するが，次回作問時の作問アドバイスまでを対象としていない点で異なる．

3. 研究課題と課題解決へのアプローチ

本研究では，①テスト理論の知見を有さないテスト作成者でも理解できるようなテスト分析結果の表示方法の解明と，②次回テスト作成時の参考になる作問アドバイスの生成ルールの構築が研究課題となる．課題①については，項目の良し悪しを視覚的に判断できるように，良い問題と悪い問題を一覧で表示する．また，設問解答率分析図[3]（4.2節で後述）やヒストグラムなどのグラフを活用した表示方法を検討する．課題②では古典的テスト理論や項目反応理論によって算出される各項目のパラメータ（項目難易度，項目識別度，S-P 表分析から得られる注意係数の値など）から次回作問時における改善点（アドバイス）を生成するルールを構築する．

4. 項目分析支援システム

4.1 システムの概要

提案する項目分析支援システムの概要を図 1 に示す．項目の良し悪しは出題意図によって判断基準が異なるため，システム利用者はテスト受験者の解答データだけでなく各項目の出題意図も本システムに入力する（図 1 ①）．項目分析モジュールでは古典的テスト理論や項目反応理論などを駆使して，各項目を分析する（図 1 ②）．分析の際には，出題意図を考慮して適切な分析方法や評価基準を採用する．例えば，授業で教えた内容のうち最も基本的で全員が理解していることを確かめる問題であれば，全員が正解しても不適切な難易度の問題とは判断しないようにする．項目説明 DB は，表 1 に示す基準（文献[4]～[8]を参考に項目の評価指標・評価手法・評価基準を設定）を元に信頼性，難易度，識別度，注意係数の各数値に対応させた項目に関する説明が登録されている．項目分析モジュールでは，解答データを分析した結果と項目説明 DB で参照し

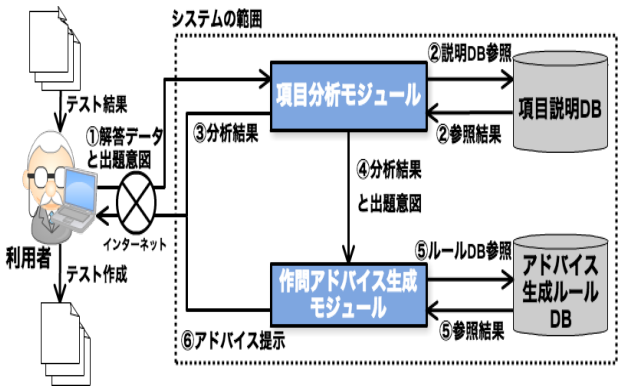

図 1　システム概要

Figure 1　System summary.

表 1　項目の評価指標・評価手法・評価基準

Table 1　Evaluation indicators and evaluation methods and evaluation criteria of the test items.

指標	手法	基準
信頼性	再テスト法 平行テスト法 折半法	“信頼性 0.80 以上”であれば，信頼性が高い[5]
	内部一貫法(クロンバックのα)	“信頼性 0.71 以上”であれば，テスト全体の信頼性が高い[6] “信頼性 0.71 未満”であれば，項目数や受験者人数が少ないかテスト全体の信頼性が低い
難易度	古典的テスト理論 項目反応理論	“難易度 0.4 未満”または ”難易度 0.80 以上”の項目は難易度が不適切な項目， ”難易度 0.3 未満” または ”難易度 0.90 以上“は修正及び検討が必要な項目[4]
識別度	古典的テスト理論 項目反応理論	“識別度 0.3 未満”は識別度が低く不十分な項目，”識別度 0.2 未満”は識別度が極端に低く合否判定には直結しない項目[4]
注意係数	S-P 表分析	“注意係数 0.5 以上 0.75 未満”は注意すべき項目 “注意係数 0.75 以上”は特に注意が必要な項目[7] [8]

た各数値の補足説明を利用者にフィードバックする（図 1 ③）．また，この分析結果と出題意図は作問アドバイス生成モジュールに渡される（図 1 ④）．作問アドバイス生成モジュールでは，作問アドバイス生成ルール DB を参照（図 1 ⑤）して生成した作問アドバイスを利用者に提示する（図 1 ⑥）．

4.2 項目分析モジュール

一般に，テストは信頼性と妥当性によって評価されるが，妥当性は定量的評価が難しいことから，本研究ではまず信頼性を評価する機能について検討した．信頼性を評価する方法には内部一貫法，折半法，再テスト法，平行テスト法など数多くの手法がある[4]．今回はテスト全体の信頼性を評価するために一般的に広く用いられている内部一貫法のクロンバックのα係数を利用して信頼性を評価することとした．

個々の項目の評価では，現状，利用者が入力した出題意図や各項目の予想正答率に基づく項目分析方法が未実装であるため，本稿では出題意図を考慮しない分析方法について述べる．個々の項目を評価するパラメータには，項目反

応理論や古典的テスト理論で算出した難易度と識別度，S-P 表分析を元に算出した項目注意係数を利用する．現状の分析では，単一のテストを想定しており，複数のテストの分析結果を比較するような機能に関しては検討段階であるため，項目反応理論よりも分析時間が短時間でおこなえる古典的テスト理論を用いている．分析結果は，項目説明 DB を参照し，能力を判定する上で適切と思われる問題グループ（以下，良問グループ）と不適切と思われる問題グループ（以下，悪問グループ）に分けて具体的な数値や説明とともに提示する．表 1 の基準値に基づき難易度が 0.4 以上 0.8 未満で識別度 0.4 以上かつ注意係数 0.5 未満の項目を良問グループとした．また，難易度が 0.4 未満または 0.8 以上で識別度 0.3 未満かつ注意係数 0.5 以上の項目を悪問グループとした．良問にも悪問にも該当しない項目に関しては，標準的な問題群としてアドバイスを省略し，良問グループ・悪問グループを含んだ全項目を最後に設問解答率分析図と難易度・識別度・注意係数で表示する．

項目のパラメータだけではなく視覚的にも分析結果を表示するため，テストを構成している項目の良し悪しを視覚的に確認できる設問解答率分析図を用いる．特性図の作図方法は，まず受験者のテストの合計得点を昇順に並べ替え，受験者を 5 群に等分割する．5 群において最も受験者のテスト合計が高い群をレベル 5（以下，Lv5）とし，最も受験者のテスト合計が低い群をレベル 1（以下，Lv1）とする．割り切れない場合には， Lv5 から降順に割り振っていく．次に各群の正答率を計算する．縦軸に正答率，横軸に 5 群をとるグラフをプロットする．各プロットを直線で結ぶことにより作図する．作図方法に基づき，実際のデータから作図した設問解答率分析図を図 2 に示す．図 2 に示した項目は左側解答者群（テスト全体の得点が低い解答者群）の正答率が低く，右側解答者群（テスト全体の得点が高い解答者群）の正答率が高くなっているため識別力が高く，測りたい特性がよく識別されている項目であることを示している．

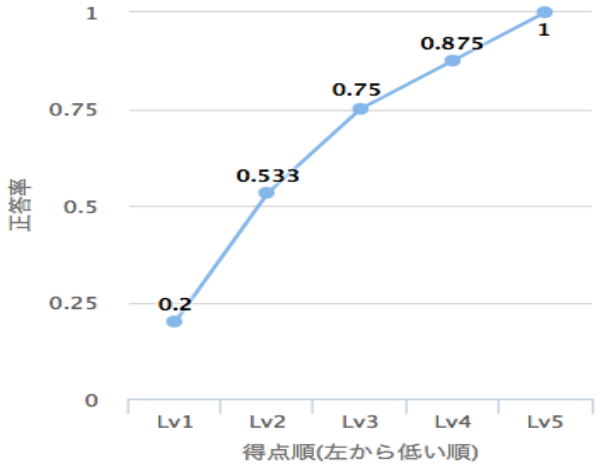

図 2　設問解答率分析図

Figure 2　Quintile Item Response Chart.

4.3 作問アドバイス生成モジュール

図 3 に作問アドバイス生成の流れを示す．項目番号 1 の難易度 0.26，識別度 0.16，注意係数 0.76 の問題の場合，①表 1 で示した基準値に当てはめ，難易度/識別度/注意係数の特徴の判断を行う．次に，②その問題に関する特徴をまとめた文章を作問アドバイスとして生成する．生成される作問アドバイスの一部文例を表 2 に示す．現在，作問アドバイスは全部で 60 種類登録されている．

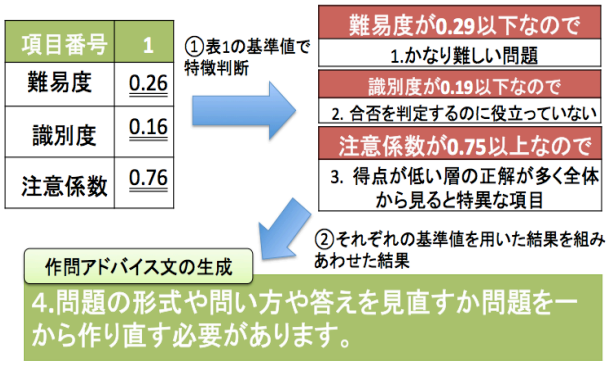

図 3　作問アドバイス生成の流れ

Figure 3　Problem creating advice generation of flow.

表 2　作問アドバイス生成例

Table 2　Problem creating advice generation example.

条件	作問アドバイス例
難易度 0.3 未満 識別度 0.2 未満 注意係数 0.75 以上	・問題の形式や問い方や答えを見直すか問題を一から作り直す必要が有ります。
難易度 0.3 未満 識別度 0.2 未満 注意係数 0.5 以上 0.75 未満	・合否判定には機能していない問題であり、問題の問い方や答えがあっているか注意してみる必要が有ります。
難易度 0.3 未満 識別度 0.4 以上 注意係数 0. 5 未満	・難しい問題を作る際はこの問題をベースに問題を作るか、修正すると次回のテストでも使えるでしょう。
難易度 0.4 以上及び 0.80 未満 識別度 0.4 以上 注意係数 0. 5 未満	・最も理想的な問題であり、次回以降のテストにおいても使えるでしょう。この問題をベースにすると理想的な問題ができそうです。

5. システムの開発

上記の考え方に基づき項目分析支援システムを開発した．なお，本研究では出題意図を考慮した分析結果やアドバイスの提示は未実装である．開発言語は PHP，JavaScript，HTML，データベースには MySQL を用いた．本システム

で項目を分析する際に，テスト結果を正誤の 2 値データに置き換える．置き換え方法としては，各項目の正誤を 0 か 1(誤答：0，正答:1)で表し，項目反応データとし扱う．項目反応データを本システムに入力し，出力された分析結果の一例を図 4，図 5 に示す．本システムでは，まずテスト全体の結果が表示される（図 4）．そして，テスト全体の結果の下に，良問と悪問の一覧と作問アドバイスが表示される（図 5）．図 5 中の各項目番号のリンクをクリックすると同画面上にポップアップ表示され，項目の詳細な特徴説明が表示される（図 6）．同画面で各項目の特徴説明を表示することにより，利用者のページ遷移の回数を少なくし，ページ読み込みの回数を最小限にしている．

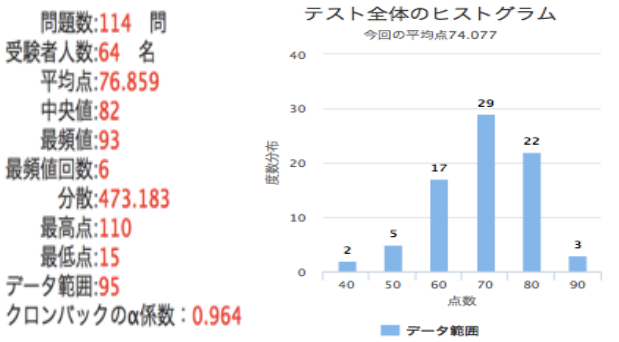

図 4　テスト全体の結果

Figure 4　Overall results test.

図 5　良問と悪問の一覧および作問アドバイス表示画面

Figure 5　List of good problem and a bad problem, And display advice for creating a question of the item.

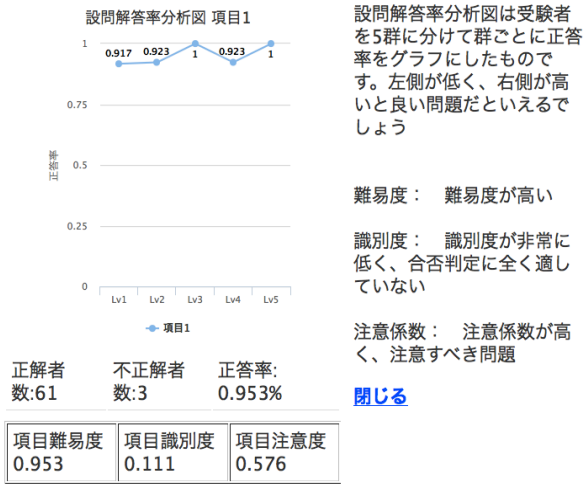

図 6　各項目の特徴説明

Figure 6　Feature description of each item.

6. システムの評価実験

6.1 実験方法

本システムの有効性を検証するために，岩手県立大学ソフトウェア情報学部で平成 27 年度前期に開講された基礎数 A（ソフトウェア情報学部 1 年生 64 名が履修）の授業中に実施した理解度を確認するための小テスト 4 回分計 39 問の解答結果を利用して実験を行った．本実験では，科目担当教員 1 名（高校の数学教諭を定年退職された非常勤講師）にシステムを利用せずに項目分析や次回以降のテスト作成における改善点の考察を行ってもらった後，本システムを利用して再度同じ項目の分析や改善点の考察を行ってもらった．また，実験の最後に項目分析の負担に関するアンケートを実施した．

本システムの利用前は，テスト全体の平均点や各項目の正答率，小テストごとに学生の得点をヒストグラムにしたものを紙に印刷して渡し，項目の分析をしてもらった．表 3 に項目分析の際に提示した分析に関する指示内容を示す．分析の際にはこれらの分析項目を記載した紙を渡し，分析内容を記述してもらった．システム利用の際には，本システムの説明を 5 分程度で行ったあとにシステムから提示される情報に基づいて分析してもらった．システムの分析処理時間は，64 名のデータ分析で平均してわずか 5 秒程である．科目担当教員による分析時間はシステム利用前と利用後それぞれで約 30 分ずつ行った．なお，システムに入力する解答結果のデータは，著者が問題毎の正解・不正解の 2 値反応データを CSV ファイルとして用意した．この際，数学における配点の都合上，部分点があった場合には 0 とすることで，2 値反応データに置き換えた．

6.2 実験結果

システム利用前の表 3（1）に対する分析では，正規分布を目指している問題，必ず解いてほしい問題，ヒストグラムの人数に偏りがあるかなどを把握したことが記述されていた．一方，本システム利用時には，学習した者と学習していない者をはっきり区別できる問題があったことを把握したことが記述されていた．

表 3（2）に対する分析では，システム利用前は 4 つの改善点，システム利用後は 7 つの改善点が記述されていた．利用前の改善点では，テスト全体に関する改善点が記述されており，特定の項目に関しての改善点の記述はなかった．利用後の改善点では，テスト全体に関する改善点の他に特定の項目に関する改善点が 7 つ記述されていた．具体的には，全受験者が解けると想定していた基本的な項目を Lv1 や Lv2 の群が解けていなかったことから，項目の問い方が曖昧であった可能性や，一部の受験者の理解が困難な単元である可能性が考えられたため，項目の見直しや一部の学生への重点的な指導の必要性を感じたなどの記述がみられた．

表 3　分析時に提示した分析項目

Table 3　Analysis items presented at the time of analysis.

(1).提示された内容から把握できた確認テストの結果について自由に記述してください
(2).提示された内容から把握できた各テスト問題の改善点を記述してください

6.3 考察

本実験の結果から，項目の分析をする際にはヒストグラムや平均点，正答率だけでは，項目の詳細な分析が難しいことが推察された．本システムの利用時には，ある項目について具体的な改善点を記述していた．これは，設問解答率分析図や良問グループ・悪問グループなど，視覚的かつ直感的にどの問題を改善しなければいけないのかが分かる工夫が有効だったと考えられる．一方で，本システムで悪問グループとされた項目について，テスト作成者が悪い項目だとは判断しておらず，良グループの項目について改善点を記述しているケースもみられた．これは，テスト作成者の出題意図によって良問や悪問になる評価基準が変化することが影響したと考えられ，テスト作成者の出題する意図を考慮した分析結果を表示する必要性が確認できた．

7. まとめと今後の課題

7.1 まとめ

本稿では，作問者がテスト受験者の解答結果に基づいて次回作問時の改善点を理解することを目的とし，項目の分析支援システムの開発を行った．項目分析結果の，難易度と識別度と注意係数の結果から，それぞれの特徴を説明した DB と次回作問時のアドバイスルール DB の生成を行い，作問アドバイス支援システムを開発した．開発した本システムの評価実験として，リメディアル科目の担当教員に実際にシステム利用前と利用後でどの程度改善点を見つけることができるかの実験を行い，システムが有効であることが示すことができた．

7.2 今後の課題

今後は，出題意図を考慮した項目分析や作問アドバイスについて検討する．この部分はこれからのシステムで重要な機能であり，出題意図を考慮した分析結果の表示によりテスト作成者側が出題意図にそぐわない問題などを除外や修正する際に役立つ機能になる．また，今回行った実験結果を踏まえてより利用者が使いやすい UI，UX を考慮したシステムに改良する．

参考文献

[1] 樋口三郎.『テストおよびアイテム分析 Web サービスの開発』. 教育システム情報学会第 39 回全国大会講演論文集，2014, p.377-378.

[2] 熊谷龍一．初学者向けの項目反応理論分析プログラム EasyEstimation シリーズの開発．日本テスト学会誌，2009，5, p.107-118.

[3] 吉村宰：大学入試センターにおけるテストデータベースによる項目分析，植野真臣，永岡慶三（共編），e テスティング，培風館，2009，pp.167-190.

[4] 大友賢二.『言語テスト・データの新しい分析法 項目応答理論入門』．大修館書店，1996.

[5] 山森光陽．前田啓朗(編)『英語教師のための教育データ分析入門』．東京:大修館，2004，pp.4－12.

[6] Nunnally,JumC.PsychometricTheory 2ndEdition．New York:McGraw-Hill Book Company, 1978.

[7] 佐藤隆博：S-P 表の入門(教育実践文庫 3），明治図書出版社, 1985.

[8] 藤垣雅司，藤垣康子，中島光洋.「注意係数の規格化： S-P 表における反応パターンの指数について」．日本科学教育学会，1985，Vol.9，pp.260-261.