項目反応理論に基づく中学校理科の定期テストの分析

An Analysis of Regular Science Tests in Junior High School Based on Item Response Theory

○菊池蒼雅*[1]，久坂哲也*[1]
Soma KIKUCHI*[1], Tetsuya HISASAKA*[1]
*[1]岩手大学
*[1]Iwate University

［要約］本研究の目的は，中学校理科の定期テストを対象に項目反応理論（IRT）に基づいて分析を行い，項目困難度や項目識別力から改善が必要な問題の特徴を見いだすことである．中学校第一学年で実際に実施された 2 つの期末試験（計 85 問）を対象に，項目反応理論の前提となる一次元性と局所独立性について検討を行った結果，2 つの仮定を満たすことができたため項目困難度や項目識別力について出題の観点や解答形式ごとに算出した．その結果，選択式問題の項目識別力がやや低いことなどが示された．それらについて選択肢の内容に着目すると，実際は 4 択問題であっても文脈からほぼ自動的に 2 択まで絞られる問題が散見された．したがって，選択式で思考の結果のみを問うのではなく，思考の過程を問う問題に改善することで項目識別力が上がると推察された．
［キーワード］項目反応理論，中学校理科，定期テスト，項目困難度，項目識別力

Ⅰ．問題の所在

学校教育において「テスト」は重要な役割を担っている．例えば，各校の入学者を選抜するためには入学試験が実施され，在学者の学力を評価するためには OECD が進める PISA や IEA（国際教育到達度評価学会）が進める TIMSS，文部科学省国立教育政策研究所が進める全国学力・学習状況調査などが実施される．また，中学校や高等学校では教科担任が作成した学期ごとの中間テストや期末テスト（i.e., 定期テスト）なども実施され，学期末の評定や調査書作成の材料となることも多い．

そもそもテストとは，「能力，学力，性格，行動などの個人や集団の特性を測定するための用具であり，実施方法，採点手続き，結果の利用法などが明確に定められているべきもの」と定義され，項目の性能を評価する指標として困難度や識別力が用いられる（日本テスト学会，2007）．また，テストを科学的に運用するための統計学的なテスト理論として，古典的テスト理論（classical test theory; CTT）や項目反応理論（item response theory; IRT）が知られている（光永，2017）．両理論において困難度や識別力が用いられるが，理論間で定義や算出方法が異なるため注意が必要である．ただ，古典的テスト理論には，いくつか問題があることが知られている．例えば，困難度や識別力が受験者の特性の分布に完全に依存してしまうことや，テスト得点という受験者の能力に関する情報が項目の集まりであるテストそのものに依存してしまうことである（豊田，2012）．そのため，TOEFL や PISA などはこれらの問題を克服できる項目反応理論に基づいて運用されている．また，全国学力・学習状況調査においても項目反応理論を導入する動きが見受けられる（文部科学省，2020）．

これまで国際的または全国的な調査においてはテスト理論に基づいた項目設計や評価が既に行われていたり，近い将来導入されたりするが，定期テストについては目が向けられてこなかった．当然，教員個人が定期テストをテスト理論に基づいて運用することは現実的ではない．しかし，子供の人生や進路に大きな影響を与える定期テストに対して統計学的な分析を行い，子供の学力を反映できていない問題の特徴を見いだすことは，「指導と評価の一体化」を実現する上でも非常に意義のあることである．

そこで本研究では，中学校理科の定期テストを対象として項目反応理論に基づいて分析を行い，改善が必要な問題の特徴を見いだすとともに，その改善案を提供することを目的とする．

Ⅱ．研究の方法

1．手続きと対象

岩手県内の中心部に位置する中学校 1 校に理科教員が作成した試験問題及びその解答データの提供を依頼し，2021 年度に第 1 学年で実施された 2 つの期末試験

データ（以下，T_1，T_2 とする）を得た．問題数は計 85 問で，これを分析の対象とした．

2. 分析方法

得られた解答データについて，正答を 1，誤答及び無答を 0 に変換し，問題の観点（1：知識，2：思考），解答形式（1：選択式，2：短答式，3：記述式），領域（1：エネルギー，2：粒子，3：生命，4：地球）でコーディングを行った．

また，項目反応理論が前提とする一次元性や局所独立性といった 2 つの仮定を満たしているか検討を行った上で 2 パラメタ・ロジスティックモデルに基づき，項目困難度と項目識別力を周辺最尤推定法により推定し，観点，解答形式，領域ごとに基礎集計を行う．さらに，本研究では項目困難度が中程度で項目識別力が低い問題を「改善が必要な問題」と操作的に定義し，その共通項を原因として見いだし，改善案を提供する．なお，分析には R (ver. 4. 1. 2) および R Studio (ver. 2022. 02. 0) を使用した．

Ⅲ. 結果と考察

1. 一次元性と局所独立性の確認

項目反応理論の仮定を満たしているか確認するため T_1 と T_2 について，それぞれテトラコリック相関行列を用いて因子分析を行った結果，固有値の減衰状況からどちらも 1 因子構造が妥当であり，一次元性の仮定を満たしていると判断した．なお，T_1 と T_2 それぞれの固有値の減衰（スクリープロット）については図 1，2 に示す．

また，局所独立性の指標として，Yen の Q3 統計量（Yen's Q3 statistic）を算出した結果，ほとんどの項目間において絶対値が 0.2 を下回ったことから，局所独立性の仮定を満たしていると判断した．

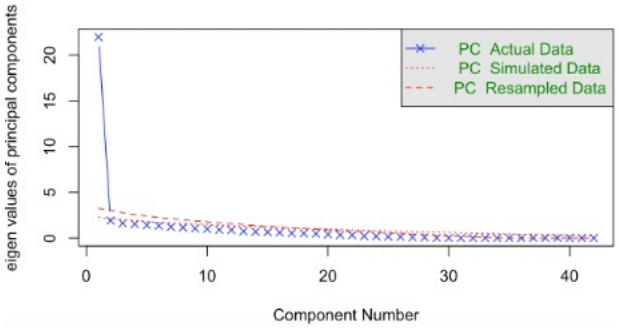

図 1　T_1 のスクリープロット

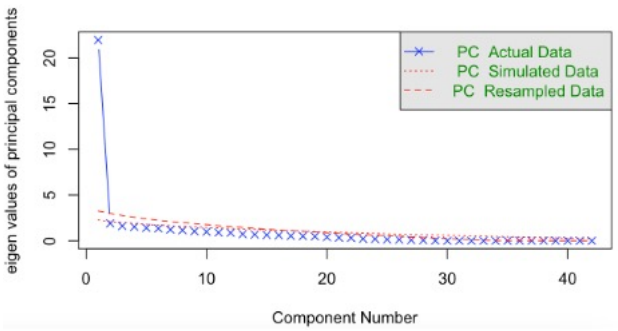

図 2　T_2 のスクリープロット

2. テスト情報曲線

項目反応理論の仮定を満たしていることが確認されたことから，2 パラメタ・ロジスティックモデルに基づき分析を行った．

T_1 と T_2 のテスト情報曲線を図 3，4 に示す．図 3 より T_1 は $\theta = -2.0 \sim 0.5$ において情報量が高く，中でも $\theta = -0.8$ において情報量が最大になっている．そのため，T_1 は $\theta = -2.0 \sim 0.5$ の集団の識別において有用なテストであったと考えられる．また，図 4 より T_2 は $\theta = -1.7$ において情報量が最大であるが，他の能力値との情報量の差が大きい．そのため，T_2 は $\theta = -1.7$ の集団の識別においてのみ有用なテストであったと考えられる．様々な能力値の生徒がいる中で，特定の能力値のみを識別できるテストではなく，幅広い能力値を識別できるテストが望ましいと考える．

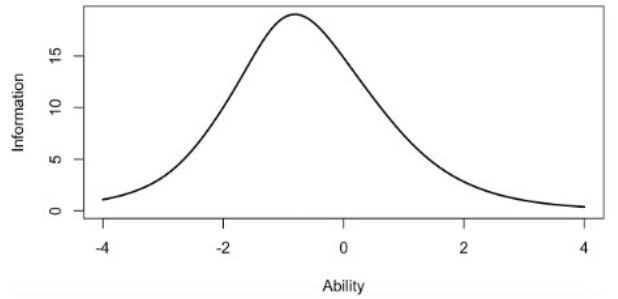

図 3　T_1 のテスト情報曲線

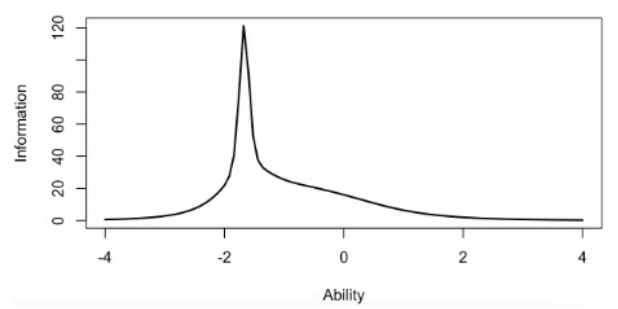

図 4　T_2 のテスト情報曲線

3. 項目困難度と項目識別力

T_1 と T_2 それぞれの項目特性曲線を図 5，6 に示す．図 5 より，項目 2 と項目 24 の項目特性曲線が水平であることから，項目識別力が低いことが示された．項目 2 は項目困難度 0.37，項目識別力 0.22，項目 24 は項目困難度 0.14，項目識別力 0.09 であった．図 6 より，項目 13 と項目 22 の項目特性曲線が水平であることから，項目識別力が低いことが分かった．項目 13 は項目困難度 0.53,項目識別力 0.24,項目 22 は項目困難度 0.19，項目識別力 0.16 であった．T_1 の項目 24 や T_2 の項目 22 は項目困難度が低いために，項目識別力が低くなったと考えられる．T_1 の項目 2 や T_2 の項目 13 は項目困難度が中程度であるにも関わらず，項目識別力が低い項目であるため特に改善が必要な問題と解釈できる．

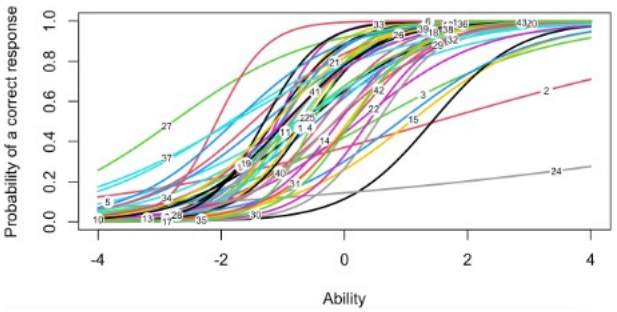

図 5　T_1 の項目特性曲線

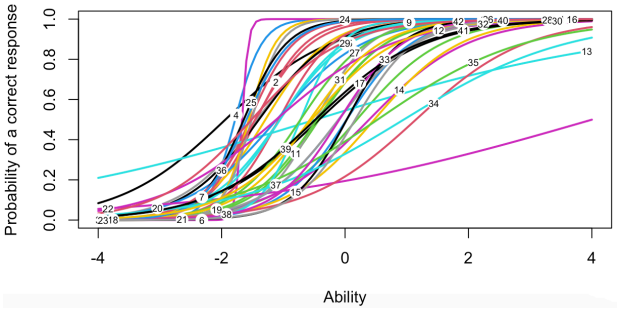

図 6　T_2 の項目特性曲線

次に，T_1 と T_2 における項目困難度と項目識別力の散布図を図 7 と図 8 に示す．項目困難度が低くなるほど，項目識別力が低くなる傾向が見られた．項目困難度が低い場合，正答する生徒が減少して項目得点とテスト得点の相関が弱くなるため，項目識別力が低下すると考える．また，T_1 と T_2 の全項目について観点，解答形式，領域ごとに分け，それぞれについて項目困難度と項目識別力の基礎統計量を算出した結果を表 1 に示す．

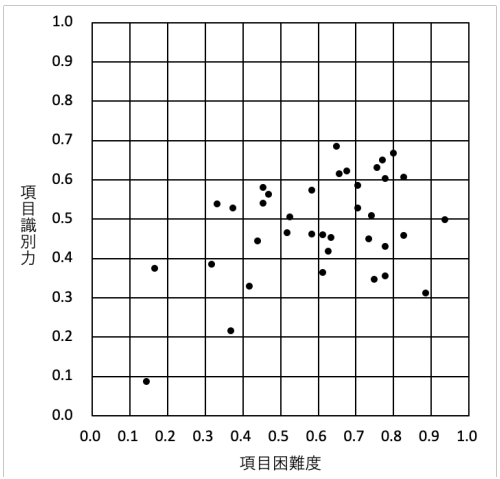

図 7　T_1 の散布図

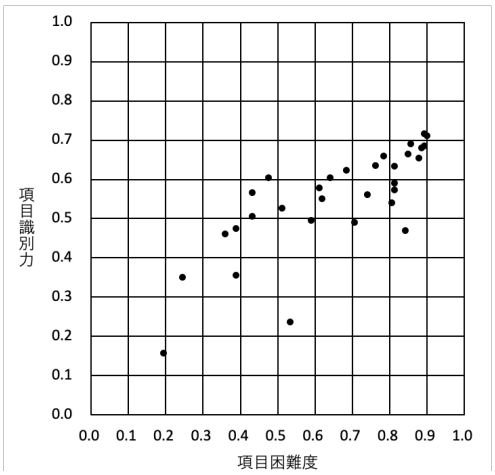

図 8　T_2 の散布図

まず，観点別に見ると項目困難度については，思考問題の方が知識問題より高くなった．思考問題の項目困難度が高かった理由として，思考問題では必要な知識を想起し，その知識を活用して表現しなければならないことが挙げられる．項目識別力については，知識問題の方が思考問題よりも高くなった．知識問題の項目識別力が高くなった理由として，知識問題では知識習得の有無を直接的に問うことが原因であると推測される．

次に，解答形式別に見ると項目困難度については高い順に，選択式，記述式，短答式となった．短答式の項目困難度が最も低くかった理由として，短答式には生徒全員が正答するような確認問題を多く含んでいたことが考えられる．項目識別力は高い順に，短答式，記述式，選択式となった．選択式の項目識別力が最も低かった理由として，選択式は一定の確率で正解すること（当て推量）が考えられる．

さらに，領域別に見ると項目困難度については高い順に，地球領域，粒子領域，生命領域，エネルギー領域となった．項目識別力については高い順に，エネルギー領域，地球領域，生命領域，粒子領域となっていた．ただ，領域別については問題数や単元に偏りがあるため，現時点においてはこの結果に対する積極的な解釈は避けたい．今後，定期テストの学年や単元などレパートリーを増やしていく中で，結果の一般化可能性について検討を重ねていく．

Ⅳ．総合考察

本研究では，中学校理科の定期テストを対象に，項目反応理論に基づいて分析を行った．その結果，項目困難度が中程度であるにも関わらず，項目識別力が低く改善が必要な問題が含まれていた．熊谷・荘島 (2015) は，識別力母数の基準はないものの，0.4 を下回るような項目は小さいといってよいと指摘している．本来，定期テストは，生徒の学力を適切に評価するためのものであるが，今回分析の対象となった定期テストでは，一部の項目において識別力が低い問題が散見され，学力を適切に反映できていない可能性が示された．学校現場では，定期テストの問題について正答率にのみ着目して評価を行っているが，項目識別力を参考に，正答している生徒の学力にも着目しながら問題の改善を図ることが必要であると考える．

また，定期テストを受ける生徒の能力は様々であり，生徒の能力に幅広く対応して識別できるテストである必要があると考える．図 4 に示したように，ごく特定の能力集団において有用なテストを作成するのではなく，様々な項目困難度において項目識別力が高い問題を集め，様々な能力の生徒を識別できるテストを作成することが必要になる．

文献

熊谷龍一・荘島宏二郎 (2015) : 心理学のための統計学 4 : 教育心理学のための統計学 : テストでココロをはかる，誠信書房.

光永悠彦 (2017) : テストは何を測るのか : 項目反応理論の考え方，ナカニシヤ出版.

文部科学省 (2020) : 全国的な学力調査の CBT 化検討ワーキンググループ中間まとめ「論点整理」(案), Retrieved from https://www.mext.go.jp/content/20200929-mxt_chousa02-000010171-3.pdf (accessed 2022.11.10).

日本テスト学会 (編) (2007) : テスト・スタンダード : 日本のテストの将来に向けて，金子書房.

豊田秀樹 (2012) : 項目反応理論 [入門編] 第 2 版，朝倉書店

表 1　観点，解答形式，領域ごとの項目困難度と項目識別力の基礎統計量

		項目困難度		項目識別力	
		M	*SD*	*M*	*SD*
観点	知識 (36)	.687	.182	.550	.115
	思考 (32)	.562	.199	.473	.141
解答形式	選択式 (28)	.590	.208	.476	.132
	短答式 (24)	.663	.176	.544	.119
	記述式 (16)	.642	.208	.534	.141
領域	エネルギー領域 (19)	.699	.210	.561	.148
	粒子領域 (27)	.598	.206	.482	.121
	生命領域 (10)	.621	.146	.485	.144
	地球領域 (12)	.588	.179	.534	.097

注) カッコ内の数値は問題数を示す