中学校理科の定期テストは良い問題なのか
― 項目反応理論による検討 ―

久坂　哲也*，菊池　蒼雅*，佐々木　聡也**，平澤　傑**，小原　翔太**

（令和５年２月１日受理）

要　約

本研究は，中学校理科の定期テストを対象に項目反応理論（IRT）を用いて識別力や困難度などを算出して評価するとともに，出題趣旨の観点や解答形式ごとの項目パラメタの特徴とその要因について検討することを目的とした。2021年度に中学校で実際に行われた第３学年理科の定期テスト４つ（計162問）について分析した結果，(1) テスト情報量のピーク幅が極端に狭いテストが含まれていること，(2) 識別力が低く理科の学力を反映できていない項目が各テストに散見されること，(3) 短答式の問題は識別力が高くなること，(4) 記述式の問題は困難度が高くなること，などが示された。

問題と目的

多くの中学校では各学期の中間や期末の時期に，普段指導にあたっている教師が作問した定期テストが行われている。そのテスト結果は，学期末の観点別学習状況評価（A, B, C）や評定（中学校では５段階）の主な材料となっている。また，各学年における評定は，高等学校の推薦入学者選抜や一般入学者選抜における調査書の学習の記録（内申点）の算出に利用され，自治体や高等学校によっても異なるが学年進行に伴って換算の重みが増すことが多い。当然ながら，評定の材料となるのは，ノートやプリントの記録，レポートや作品，授業内での発言や行動など多岐にわたるが，定期テストの得点の影響は少なくない。それゆえ，生徒個人の人生にも影響を及ぼしていると考えられるが，定期テストの信頼性や妥当性が評価されることは皆無に等しく，設問項目の作成や評価基準の設定といったテスト運用が教員個人の力量や裁量に委ねられているのが現状である。そこで本研究では，中学校において過去に実施された理科の定期テストを対象としてテスト理論に基づいた分析を通して信頼性や妥当性を検討する。

テストを科学的に運用するための代表的な理論として，古典的テスト理論（Classical Test Theory; CTT）と項目反応理論（Item Response Theory; IRT）が知られている（光永，2017）。古典的テスト理論では，各項目の正誤データとテストの合計得点との相関係数を識別力，各項目の正答率を困難度として用いるが，これらは受験者集団の特性に依存したり，合計得点という受験者の能力を示す情報は項目の集まりであるテストそのものに依存したりする（豊田，2012）。一方，項目反応理論では受験者の能力を間隔尺度（潜在特性尺度）上に位置付けて表現し，識別力や困難度といった指標を能力（特性値θ）と分離して表現することで，古典的テスト理論における標本依存性や項目依存性といった問題を克服することができる。

したがって，本研究では項目反応理論を援用し，定期テストを構成する設問項目の識別力や困難度などを算出して評価する。また，一般的に定期テストは観点別学習状況評価に照らし合わせ，知識・技能を測る問題（以下，知識問題）と，思考・判断・表現を測る問題（以下，思考問題）で出題するため，出題趣旨の観点や設定された解答形式ごとの特徴について検討を行う。

＊岩手大学教育学部，＊＊岩手大学教育学部附属中学校

方　法

分析の対象

本研究では，岩手県内の中学校１校にデータ提供を依頼し，2021年度に第三学年で４回実施された理科の定期テストデータを得た。データには，出題の観点が付されたテスト問題と小問及び個人ごとの採点結果が含まれていた。

なお，データの提供を受ける際は，事前に生徒の氏名を削除し，出席番号とは関係なくランダムにIDを割り振って並び替えを施した状態で提供を受けた。各テストの受験者は140名であった。

変数のコーディング

４つのテストをそれぞれT_1からT_4とし，各テストの問題に通し番号を割り当てた。１つの小問内に設問が複数ある場合は，それぞれに通し番号を振った。その結果，設問項目は計162問であった。解答データ（採点結果）については，配点に関係なく誤答及び無答を０，正答を１とした２値データに変換した。また，出題趣旨の観点（１：知識，２：思考）と解答形式（１：選択式，２：短答式，３：記述式，4:その他）をコーディングした。コーディングの作業は第二著者が行い，その結果を第一著者が確認した。

分析の手続き

本研究では，項目反応理論に基づき分析するため，事前にその前提となるT_1からT_4の一次元性と局所独立性についてそれぞれ確認した。次に，２パラメタ・ロジスティックモデル（Two-Parameter Logistic Model: 以降では2PLMと略記）に基づき，項目パラメタの推定を行った。2PLMは，

$$P_j(\theta) = \frac{1}{1 + \exp\left(-Da_j\left(\theta - b_j\right)\right)} \qquad (1)$$

で表される。(1)式中の添え字jは項目の識別子を表す。また，θは潜在特性値，a_jは項目jの識別力，b_jは項目jの困難度を表す。Dは尺度因子定数と呼ばれ，正規累積分布を用いた項目特性曲線に近似させるための定数で，本研究では$D = 1.0$（ロジスティック計量）を用いた。

結　果

以下の分析では，R（ver. 4.2.2）及びRStudio（ver. 2022.12.0）のpsychパッケージとirtoysパッケージに含まれるest関数とltmパッケージを用いた。

基礎集計

各テストの基礎集計として設問の項目数，信頼性係数としてクロンバックのα係数，正答率の平均値と標準偏差を算出した（Table１）。α係数は.91から.96と高い値であった。また，正答率は.62から.70であった。一般的に，定期テストは平均正答率60％から70％前後を目指して作成することが多く，今回分析対象となったテストはその範囲内であった。

Table１　基礎集計（N=140）

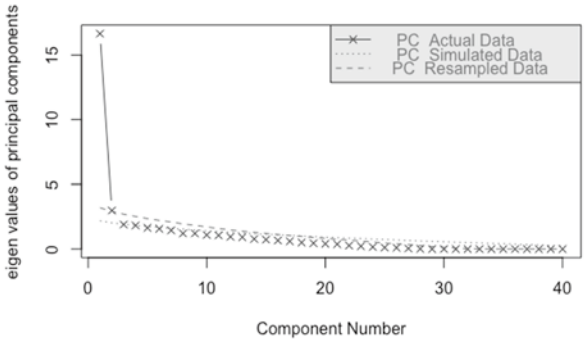
	J	α	M	SD
T_1	40	.91	.66	.21
T_2	41	.93	.68	.21
T_3	43	.92	.70	.20
T_4	38	.96	.62	.30

一次元性と局所独立性の確認

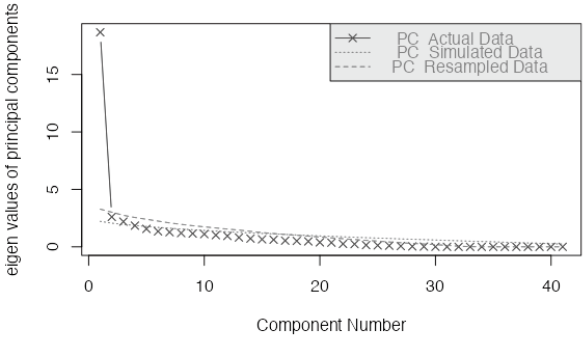
項目反応理論の前提となる２つの仮定について確認した。まず，一次元性（unidimensionality）について確認するため，T_1からT_4についてテトラコリック相関行列を用いたカテゴリカル因子分析と平行分析を行った。固有値の減衰を表すスクリープロットと平行分析の結果をFigure 1に示す。各テストにおいて固有値が１を超える因子が複数見られたが，因子の減衰状況や平行分析の結果から１因子構造が妥当であると判断した。ゆえに，一次元性の仮定を満たしていると判断した。

次に，局所独立性（local independence）について確認するため，YenのQ3統計量（Yen’s Q3 statistic）を算出した。各テストのごく僅かな項目間において絶対値が0.2を超える箇所が散見されたが，大問や領域を跨っている箇所でも見受けられ，出題内容的に関連は薄いと判断した。ゆえに局所独立性の仮定を満たしていると判断した。

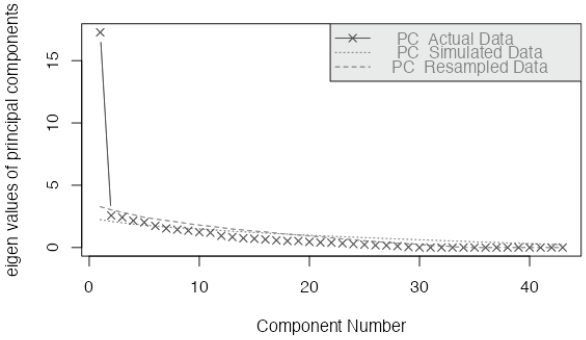
a：T_1（項目数40）



b：T_2（項目数41）



c：T_3（項目数43）



d：T_4（項目数38）

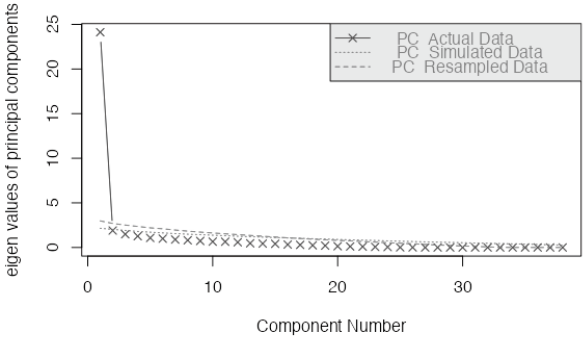


Figure 1　スクリープロットと平行分析

テスト情報量

テスト全体の測定精度を検討するため，テスト情報関数（Test Information Function; TIF）を用いてテスト情報曲線（Test Information Curve; TIC）を描いた（Figure 2）。なお，クラメール・ラオの不等式によりテスト情報量の平方根の逆数によって能力値θの測定の標準誤差（Standard Error of Measurement; SEM）や能力パラメタの95％信頼区間も算出できる。

T_1，T_2，T_4のテスト情報曲線を見ると，情報量のピーク幅が極端に狭いことが読み取れる。また，T_2においては情報量のピークが$\theta = -2$付近であり，偏差値換算で約30程度の低い学習者のみに対して高い弁別性を有していることが示された。

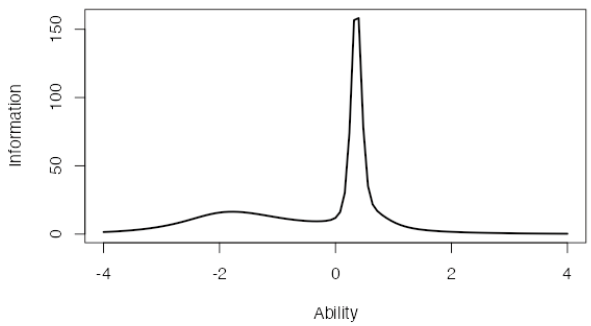
一方，T_3は他のテストに比べて情報量のピーク幅が広く，$\theta = -2$から$\theta = 0$付近の学習者集団に対して高い測定精度をもつことが示された。

項目特性曲線

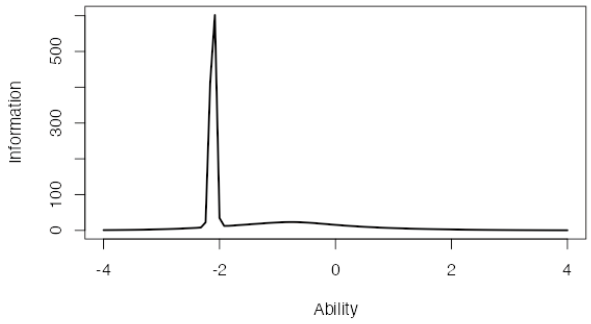
テストを構成する個々の項目の性能を検討するため，項目特性曲線（Item Characteristic Curve; ICC）を描いた（Figure 3）。識別力（a_j）が高ければ曲線の傾きは大きくなる。また，2PLMでは正答率0.5のときの能力値θが困難度（b_j）となる。なお，明確な基準はないが，一般的な目安として識別力が0.4を下回るような項目は，識別力が小さいと判断される（熊谷・荘島，2015）。また，全項目のICCを合計したものがテスト特性曲線（Test Characteristic Curve; TCC）となる。

T_1からT_4までの項目特性曲線を概観すると，どのテスト内においても識別力が低い項目が散見される。例えば，T_1の項目u_{30}は$a_{30} = 0.354$, $b_{30} = 0.809$であり，T_3の項目u_{34}は$a_{34} = 0.187$，$b_{34} = 1.608$である。これらの項目は，テスト全体の正答率（合計得点）が理科の学力を反映していると仮定するならば，理科の学力を識別できていない項目となる。また，項目特性曲線が交差している項目も見受けられる。これは，能力値θによって正答率が逆転することを意味する。ゆえに，困難度の数値のみで項目の難易度を単純比較はできない。

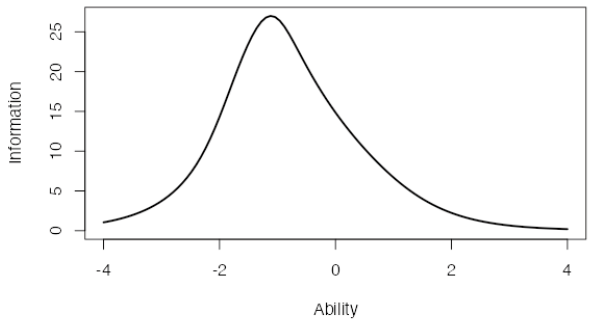
a：T_1（項目数40）



b：T_2（項目数41）



c：T_3（項目数43）



d：T_4（項目数38）

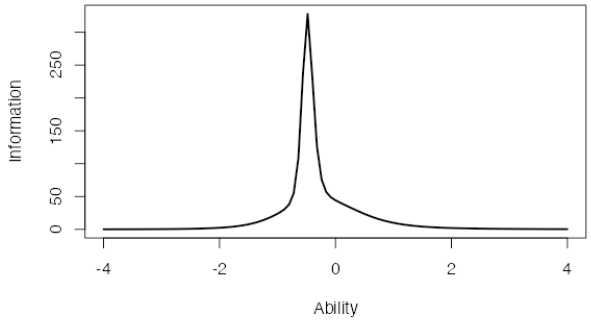
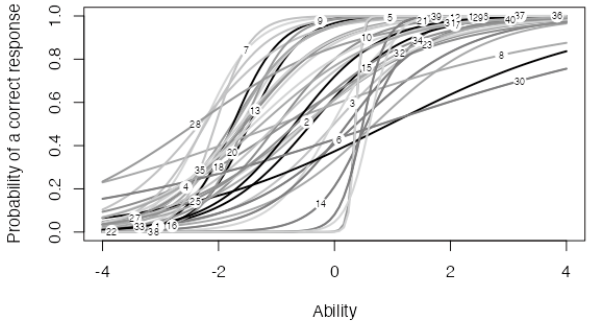
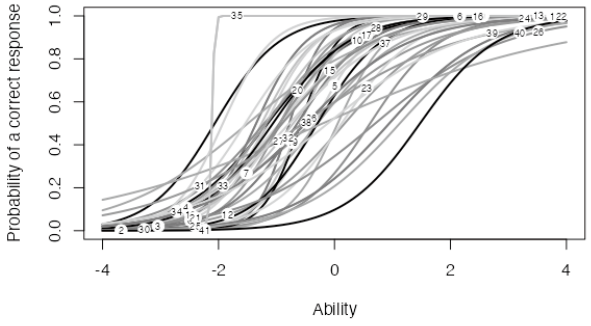


Figure 2　テスト情報曲線

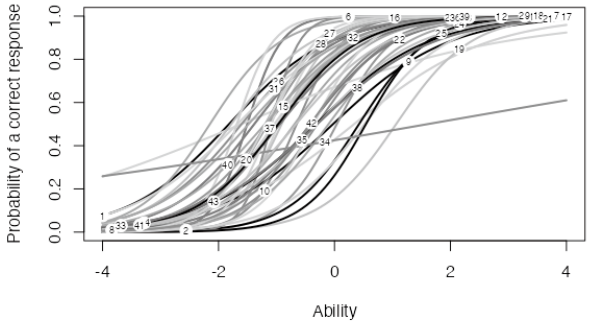
a：T_1（項目数40）



b：T_2（項目数41）



c：T_3（項目数43）



d：T_4（項目数38）

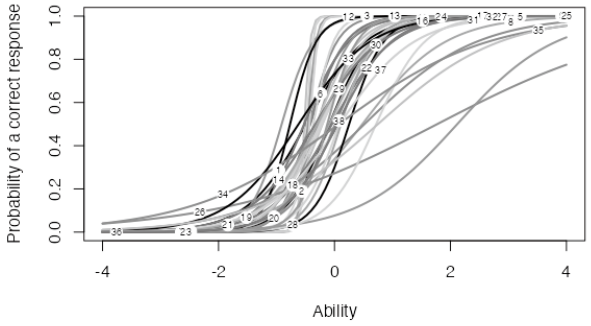


Figure 3　項目特性曲線

項目パラメタの基礎統計量

テストごとの項目パラメタの基礎統計量をTable 2に示す。テスト間の差を検討した結果，識別力は$F(3, 158) = 1.64$, $n.s.$,$\eta^2 = 0.03$となり有意差は認められなかったが，困難度は$F(3, 158) = 6.96$, $p < .001$, $\eta^2 = 0.12$となり有意差が認められたため，多重比較（Bonferroni法）を行った結果，T_4はT_1, T_2, T_3より高いことが示された（$p < .01$）。また，各項目パラメタについて出題趣旨の観点（知識，思考）と解答形式（選択式，短答式，記述式，その他）でクロス集計を行った（Table 3）。識別力と困難度を従属変数，観点と解答形式を独立変数とする多変量分散分析を行った結果，観点と解答形式の主効果は有意であり（順に$F(2, 153) = 10.73$, $p < 001$, $\eta_p^2 = .123$; $F(6, 306) = 3.41$, $p < .01$, $\eta_p^2 = .063$），交互作用は有意ではなかった（$F(6, 306) = 0.67$, $n.s.$, $\eta_p^2 = .013$）。下位検定の結果，観点においては思考の困難度は知識よりも高いこと（$p < .001$），解答形式においては短答式の識別力は選択式より高く（$p < .05$），記述式の困難度は他の解答形式よりも高いことが示された（$all\ ps < .01$）。

Table 2　テストごとの基礎統計量

	J	識別力		困難度	
		M	SD	M	SD
T_1	40	2.411	3.732	−0.828	0.994
T_2	41	3.746	9.495	−0.657	0.846
T_3	43	1.646	0.736	−0.733	0.813
T_4	38	4.037	4.751	−0.051	0.605

考　察

本研究の目的は，項目反応理論に基づいて定期テストを構成する設問項目の識別力や困難度などを算出して評価することであった。

Figure 2よりT_1, T_2, T_4のテスト情報量は，ごく限られた能力の学習者集団に対して高い測定精度をもつことが示された。定期テストの受験者は，その学校に在籍する生徒全員であるため能力値の幅は広い。したがって，局所的な弁別性をもつ定期テストでは，その限られた能力の受験者集団以外の受験者では標準誤差が大きくなるため，学力を測定するという視点では問題があることがわかる。一方，T_3のテスト情報量は他に比べてピーク幅が広く，$\theta = -2$から$\theta = 0$程度の学習者に対して高い測定精度をもつことが示され，他の3つと比較して適当なものであったと考察できる。ただ，ピークが$\theta = -1$付近であるため，グラフ全体が右側にシフトするとより良いテストに近づくと推察される。

Figure 3より各テストを構成する項目パラメタの特徴が示された。このグラフでは，識別力（a_j）が高ければ曲線の傾きがどこかの特性値θの箇所で急激に大きくなるが，すべてのテストにおいて傾きが小さく水平方向に伸びている項目が散見された。特に顕著であるのがT_1の項目u_{30}（$a_{30} = 0.354$, $b_{30} = 0.809$）とT_3の項目u_{34}（$a_{34} = 0.187$, $b_{34} = 1.608$）である。T_1の項目u_{30}は，実験で残ったクエン酸150cm^{3}を中和するために必要な水酸化ナトリウム水溶液の量について，選択肢ア〜オ

Table 3　観点及び解答形式ごとの基礎統計量

	識別力				困難度			
	知識		思考		知識		思考	
	M	SD	M	SD	M	SD	M	SD
選択式	1.821	2.546	1.529	0.885	−0.844	0.736	−0.089	1.016
短答式	3.363	4.572	4.950	6.206	−0.869	0.743	−0.179	0.724
記述式	1.929	0.815	4.292	10.884	−1.423	0.625	−0.813	0.783
その他	1.786	0.868	4.500	9.624	−0.350	0.716	−0.143	0.855

注）解答形式のその他は，作図，計算，並び替えである。

の中から１つ選択して解答する問題であったが，選択肢の内容から３つは明らかに間違いであることが容易に想像できるため，実質，二択問題になっていたと推測される。したがって，正解となる選択肢を判断できなくとも残りのどちらか一方を解答すれば50％の確率で正解するため識別力が低くなったと解釈できる。T_3の項目 u_{34}は，タマネギの根の細胞分裂が観察できるプレパラートについて，観察された細胞の数が示された表を見て３つの中から１つ選択して解答する問題であった。この問題は選択肢が初めから３つと少ないことに加え，表中に示された細胞数が６，15，150となっており１つだけ突出して多かった。正解は１つであるため数値が近い２つは誤答であること，また細胞分裂という言葉から細胞の数が多いものが正解であることは，細胞分裂について理解していなくても容易に判断できるため識別力が低くなったと解釈できる。一方，識別力が高い項目も見受けられた。例えば，T_2の項目 u_6（$a_6 = 3.511$，$b_6 = -0.777$）やT_3の項目 u_{18}（$a_{18} = 4.125$, $b_{18} = -1.091$），T_4の項目 u_{15}（$a_{15} = 5.793$，$b_{15} = -0.364$）である。T_2の項目 u_6は，硫酸（H_2SO_4）と水酸化バリウム（$Ba(OH)_2$）の化学式を呈示した上で，この２つを混ぜ合わせたときの化学反応式を書く問題であり，T_3の項目 u_{18}は，木片を2N の力で20cm 動かしたときの仕事（J：ジュール）を計算させる問題であった。また，T_4の項目 u_{15}は，秋分の日の太陽の動いた道筋を透明半球に作図させる問題であった。このように実験の化学反応式を問うたり計算や作図を求めたりする問題は，知識を概念的に理解していたり，技能を活用できるレベルで身に付けたりしていなければ解けない問題であるため識別力が高くなったと解釈できる。

Table２のテストごとの項目パラメタの比較から，識別力に有意差は認められなかったが困難度に有意差が認められた。今回，分析対象としたのは１年間の中で実施時期が異なる計４つの定期テストのため出題範囲がエネルギー，粒子，生命，地球の全領域であり，テストによって領域が大きく異なる。したがって，テストによって困難度が異なる原因として領域による違いも想定される。また，T_4は公立高校入試前に実施された最後のテストであったため，作題した教師には高校入試を見据えて難易度の高い設問項目セットを用意した意図があった可能性もある。ただし，これらは本データから結論付けることはできず，推測の域を出ないため積極的な解釈は控えたい。

出題趣旨の観点と解答形式ごとの項目パラメタの比較からともに主効果が有意であることが示され，思考問題の困難度は知識問題よりも高いことや短答式の識別力は選択式よりも高いことなどが示唆された。また，クロス集計の結果（Table３），統計的な有意差は認められなかったが，知識問題を短答式で問う場合の識別力は高い値を示し，思考問題を選択式で問うた場合の識別力は低い値を示した。知識と思考のどちらを問うにしても，選択式には偶然正答する確率（当て推量パラメタ）が混入するため識別力が下がると考えられるが，思考問題の困難度は知識問題よりも有意に高かったため，その確率は増すと推察できる。したがって，思考問題を選択式で出題することには注意が必要である。

本研究の示唆と限界

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を用いて分析を行ってきた。その結果，テストによってテスト情報曲線に差異があることに加え，ピーク幅が局所的なテストが多いこと，各テストに許容範囲を超えるほど識別力が低い問題が含まれることなどが示された。また，知識問題を短答式で出題すると識別力が高くなる傾向があることや，思考問題を選択式で出題すると識別力が低くなる傾向があることなども示唆された。

近年，全国の一部中学校で定期テストを廃止したり見直したりする動きも見られるが（読売オンライン，2022），多くの中学校で定期テストが重要な役割を果たしていることは確かである。それにも関わらず，現在の教員養成課程カリキュラム

の中でテストを含む教育測定や教育統計について専門的に学ぶ制度はなく（木村，2010），教員個人の力量に任せられているのが現状である。テスト理論の知識を教授することは，信頼性や妥当性といったオーセンティックな学習評価や指導と評価の一体化の実現に欠かせない重要な概念について理解したり，テストの効用とその限界について実感したりできるといった教育効果がある（木村・西郡，2017）。折しも，現在の教育政策ではEBPMやEBEといったエビデンスベースの教育の推進が求められているが，エビデンスをつくったり利用したりするために必要な基礎的・専門的知識が学校教員に不足していると言わざるを得ない。その意味で，本研究が定期テストの現状に警鐘を鳴らすとともに，これからの社会を賢明に生き抜くために必要となる資質・能力を，エビデンスベースで子供たちに確実に育成することができる教育の実現に僅かでも寄与することを筆者らは願っている。

最後に，本研究の限界を２点挙げる。１点目は，サンプルサイズに不足があることである。2PLMを用いて分析するには，サンプルサイズをより大きくする必要がある。しかし，１つの定期テストを受験するのはその学校に在籍する該当学年の生徒のみであるため，定期テストを分析対象とする場合にはサンプルの確保が困難となるため，より確かな知見を得る方法を検討する必要がある。

２点目は，定期テストの分析に項目反応理論を援用することが適切かについて，データに基づく検討が行われていないことである。ただしこれは，そもそも定期テストが指導改善を目的としているのか，それとも評価や評定を目的としているのかの議論に立ち戻ることになる。両者で求められるテストの望ましい在り方や運用の仕方は異なると考えられる。また，理科のテストでは問題解決や探究の過程に沿って出題されることもあるため，項目反応理論の前提となる局所独立性を仮定し難い実態もある。今後は，定期テストの分析により適したモデルや方法を検討する必要がある。

引用文献

菊池蒼雅・久坂哲也（2022）項目反応理論に基づく中学校理科の定期テストの分析，日本科学教育学会研究会報告，37, 2, 21–24. https://doi.org/10.14935/jsser.37.2_21

木村拓也（2010）日本における「テストの専門家」を巡る人材養成状況の量的把握，日本テスト学会誌，6（1），29–49. https://doi.org/10.24690/jart.6.1_29

木村拓也・西郡大（2017）教養教育段階におけるテストに関する授業開発と実践：「テスト学教育」の効果測定，日本テスト学会誌，13（1），49–68. https://doi.org/10.24690/jart.13.1_49

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

豊田秀樹（2012）『項目反応理論［入門編］第２版』朝倉書店

読売オンライン（2022）脱「一夜漬け」，中学校で定期テスト廃止広がる：小テスト・論文で日頃の学び重視へ（2022年11月11日記事）. Retrieved from https://www.yomiuri.co.jp/kyoiku/kyoiku/news/20221111-OYT1T50185/

付　記

本研究は，第二著者が2022年度に岩手大学教育学部へ提出した卒業論文の一部データを再分析及び再構成したものである。また，菊池・久坂（2022）でも項目反応理論に基づく中学校理科の定期テストの分析に関する研究発表を行っているが，本論文とは分析対象のデータが異なる。なお，本論文に関して，開示すべき利益相反関連事項はない。

Appendix　全項目の2PLMにおける項目パラメタの推定値

	T_1（$J=40$）		T_2（$J=41$）		T_3（$J=43$）		T_4（$J=38$）	
項目	識別力	困難度	識別力	困難度	識別力	困難度	識別力	困難度
1	2.297	−1.466	1.797	−2.064	1.147	−1.819	1.960	−0.494
2	0.864	−0.525	1.638	0.615	2.492	−0.652	2.540	0.013
3	1.187	−0.007	2.138	−1.255	1.552	−0.984	19.623	−0.518
4	1.684	−1.763	0.987	−0.404	1.507	−0.967	3.662	−0.910
5	1.939	−1.607	2.090	−0.350	1.513	−1.036	2.477	−0.057
6	0.989	0.375	3.511	−0.777	3.408	−1.505	3.235	−0.419
7	2.302	−2.250	1.642	−0.909	2.532	−0.964	19.781	−0.480
8	0.395	−0.940	1.360	−0.292	1.465	−0.690	1.399	0.481
9	2.480	−1.747	3.121	−0.584	1.003	−0.025	2.649	−0.086
10	1.427	−0.970	1.282	−1.210	1.593	−0.296	4.240	−0.431
11	1.285	−2.324	1.841	−1.135	2.281	−0.911	2.819	−0.207
12	1.372	−1.571	1.243	0.177	2.056	0.040	14.662	−0.457
13	1.328	−1.553	45.070	−2.114	0.862	0.339	14.433	−0.429
14	2.389	0.552	2.333	−1.350	1.329	−0.944	2.606	−0.524
15	1.960	−0.036	1.259	−0.944	0.445	−1.606	5.793	−0.364
16	1.010	0.693	2.356	0.057	1.650	−2.148	2.390	−0.087
17	1.260	−0.400	1.372	−1.152	1.661	0.424	4.459	−0.766
18	0.719	−0.842	1.631	−0.948	4.125	−1.091	1.036	0.517
19	1.885	−2.224	2.620	−0.582	1.560	1.063	2.097	−0.269
20	2.640	−1.554	2.119	−0.938	1.878	−1.154	3.545	−0.295
21	1.308	−1.472	2.253	−1.178	1.369	0.520	2.110	−0.241
22	1.422	0.222	1.684	−0.640	1.383	−0.416	1.918	−0.044
23	0.894	−0.511	1.585	0.126	2.062	−0.361	3.080	0.081
24	1.802	−1.711	0.899	−1.492	1.456	−0.103	3.444	0.112
25	0.538	0.961	1.804	−0.293	1.907	0.542	3.367	0.253
26	1.350	−1.538	0.900	0.706	1.305	−1.603	0.555	1.772
27	1.082	−0.990	1.375	−0.712	1.438	−1.780	2.739	−0.312
28	0.742	−2.400	1.481	−1.172	2.003	−1.192	1.177	2.118
29	1.292	−1.627	2.366	−1.924	3.229	−1.683	2.317	−0.212
30	0.354	0.809	1.201	0.947	1.399	−1.384	1.775	−0.330
31	1.002	−1.394	1.695	−1.536	1.072	−1.676	2.395	0.719
32	0.992	−0.473	0.469	−0.200	1.212	−1.480	2.080	−0.298
33	1.372	−0.686	1.543	−1.060	1.038	−0.321	1.632	−0.619
34	0.987	−0.661	1.274	−0.899	0.187	1.608	0.780	0.076
35	3.332	−2.049	45.070	−2.114	1.139	−0.294	0.955	0.750
36	16.618	0.358	0.741	−0.541	1.445	−1.145	3.942	−0.124
37	19.041	0.361	1.166	−0.777	2.596	−1.093	1.561	0.089
38	4.996	0.487	1.105	−0.494	0.995	−0.300	2.160	0.050
39	4.968	0.605	0.871	−0.112	2.034	−1.183	—	—
40	0.955	−1.256	1.190	1.128	1.227	−1.199	—	—
41	—	—	1.509	1.476	1.422	−0.970	—	—
42	—	—	—	—	1.526	−0.414	—	—
43	—	—	—	—	1.278	−0.662	—	—