IRT尺度値を利用した中学校理科のパフォーマンスの解釈について

―電力の課題を例に―

柴　山　　直*
千　葉　陽　子**

思考力・判断力・表現力等の高度な複合的学力を育てるとされるパフォーマンスアセスメントは，その一方で，評価者の主観的な判断に頼る部分が大きいため評価結果が安定しない欠点をもつ。本研究では，中学校理科を例に，パフォーマンスの解釈にIRT尺度値を用いることで客観性を担保することを試みた。具体的には，中学2年生に理科の学力を測る客観式調査とパフォーマンス課題の2つのテストを課した。そのデータを用いて，客観式調査については，同時尺度調整法によって第1学年と第2学年の項目を同一尺度上に位置づける等化を行った。その上で，パフォーマンス課題の結果を分析すると，予備調査，本調査ともに学力のレベルが上がるにつれて着目する観点が増える傾向があることが見いだされた。

キーワード：パフォーマンスアセスメント，中学校理科，IRT，等化，ルーブリック

1　問題と目的

知識基盤社会，グローバル社会を生きる現代の子どもたちには，基礎的・基本的な知識・技能の習得や思考力・判断力・表現力，共存・協力が必要とされている。平成25年度までに全面実施がなされた現行学習指導要領のもとでは，基礎的・基本的な知識・技能の習得，知識・技能を活用して課題を解決するために必要な思考力・判断力・表現力等がすべての教科において重視され，これらをいわば車の両輪として相互に関連させながら伸ばしていく教育活動が行われることとなった（中央教育審議会，2008）。特に算数・数学，理科では観察・実験，課題学習を充実させるように求められており，理系教科の重視，思考力・判断力・表現力の重視が近年の動向である。

このような学力像に基づいて，2010年に改訂された指導要録では「関心・意欲・態度」，「思考・判断・表現」，「技能」，「知識・理解」の4つの観点が設定された。これらの観点のうち，思考力・判断力・表現力の評価にはパフォーマンスアセスメントの方法が用いられているとしている。パフォーマンスアセスメントとは，アメリカにおいて1980年代後半に登場した「真正の評価」論に伴って開発されたものである。この評価においては，どれほど学習目標を達成できたかを採点者が質的に判断を

*教育学研究科　教授
**教育学研究科　博士課程前期

する。この際，主観的な判断に陥らないように，評価者を複数にする，評価基準表（ルーブリック）を用いる等の対策がなされることが多い。しかしながら，ルーブリックには元より主観が含まれている。石井（2010）によると，一般的なルーブリックの開発のための手順の一例は，①試行としての課題を実行し多数の児童生徒の作品を集める，②あらかじめ数個の観点を用いて作品を採点することを同意しておく，③それぞれの観点について一つの作品を少なくとも3人が読み，6点満点で採点する，④次の採点者にわからぬよう付箋に点数を記して作品の裏に貼り付ける，⑤全部の作品を検討し終わった後で全員が同じ点数をつけたものを選び出す，⑥その作品を吟味しそれぞれの点数に見られる特徴を記述する，というものである。

この例からわかるように，ルーブリック開発の過程において，解答の採点や観点の設定は評価者の経験に裏付けられて行われていることがわかる。ルーブリックは本来，主観を可能な限り排除する目的で使用されるものであるにもかかわらず，主観的な指標に基づいて作成されている。

先に述べた日本の現状から，近年注目されており，様々な場面で採用されているパフォーマンスアセスメントにはルーブリックの作成がほとんど経験に基づいて行われているという問題があると言える。パフォーマンスアセスメントは信頼性が低く，主観に基づく判断から逃れられないとしばしば言われるが，主観を取り除くための機能を持つルーブリック自体も主観的な判断によって作成されている。つまり，このルーブリック作成法が続く限り，主観からは根本的に逃れられないのである。

そこで，パフォーマンスアセスメントの信頼性を高めるために，IRT モデルに基づく尺度値 θ をパフォーマンスの解釈の参考とする。この手法においては，まず，妥当性，信頼性ともに高い客観式テストから推定される尺度値 θ によって学力が保証される。そして，その保証された学力に基づいてどのようなことができるかといった観点でパフォーマンスを記述することによって，そのパフォーマンスの段階を位置付けることができる。

本研究では，佐藤・柴山（2013）で提案されたルーブリック作成の手法をもとに，中学校理科を題材にし，パフォーマンスの解釈に客観式テストから得た尺度値 θ を用い，尺度値 θ とパフォーマンスの関係を明らかにすることを目的とする。この手法によって得られたパフォーマンスの特徴や観点をルーブリック開発時に用いることで，信頼性を担保するに十分な仕様のものが作成できると期待される。

2　予備調査

本研究で用いる課題の選定のために，宮城県内 A 市立 a 中学校の協力を得て，平成26年1月に予備調査を行った。調査対象者は第1学年1学級33名，第2学年1学級34名であった。この調査においては，客観式調査（理科）とパフォーマンス課題の2つの問題冊子を配布し，解答してもらった。客観式調査（理科）は，両学年ともに，新潟県における平成18年度「全県学力調査」から未履修項目等を削除した15項目からなる。この調査結果から，尺度値 θ を推定した。項目パラメタ（識別力，困難度）の推定は新潟県における平成18年度「全県学力調査」のデータ（受検者：中学校第1学年22035

名，第2学年21520名，実施：平成18年1月）を用い，それぞれの学年についてあらかじめ行った。受検者パラメタ（尺度値θ）の推定は最尤法による。なお，これらのパラメタ推定にはEasyEstimation（熊谷，2009）を使用した。パフォーマンス課題は，両学年ともに第1分野と第2分野から1題ずつ，計2題を出題した。なお，これらの課題は全国の公立高等学校入試問題を参考にしてa中学校教員と作成した。

第1学年の客観式調査において全問正答者は1名，全問誤答者は0名であったため，尺度値θの推定が不可能な受検者は1名であった。この1名を除いた際の尺度値θの平均値は-0.388，標準偏差0.895，最大値1.463，最小値-1.781であった。尺度値θが0.307から1.463までの受検者と全問正答者をH群（N=11）とし，-0.704から-0.001までの受検者をM群（N=10）とし，-1.781から-0.899までの受検者をL群（N=12）として3群に分割した。

第1学年のパフォーマンス課題は，第1分野の項目が圧力に関する実験手順を問うものであった。この課題は，スポンジの上にレンガを置き，レンガの面によってどのようにスポンジが沈むのかを調べ，面積と圧力について考察することを想定したものであった。この項目について学力群ごとにパフォーマンスの傾向を見ると，H群については，実験の操作と判断基準について想定した範囲内で記述をしている受検者が多く見られた。また，スポンジの上にレンガを置くといった操作のみを記述する受検者が最も少ないことから，問題文の説明と同様の状況を再現する実験を行うということが理解できていると考えられる。M群については，手順について正しく記述できる受検者は同時に面積と沈み方の関係性についても正しく記述できる傾向が見られた。その一方で，受検者の思考を理解することが難しいパフォーマンスが最も多く見られた。L群については，実験の操作について正しく記述ができた1名と，操作や判断記述については記述がないものの，面積と沈み方について正しく記述ができた1名以外のほとんどはスポンジの絵を描くにとどまるか，「スポンジの上にレンガを置く」という記述をしていた。この結果から，学力群のレベルが高くなるにつれて実験のイメージを伝えることができると言える。しかしながら，この結果は，理科に関わる力というより問題文の読み取りといった国語に関わる力を測定している部分が多く占めている可能性がある。したがって，この項目は調査項目としては適切でないと判断した。

一方，第2分野の項目は，植物の光合成と呼吸についての実験結果を記述するものであった。この項目は，4つの試験管の変化を予想し，試験管内で起きている現象について説明をすることを想定したものである。H群に属するほとんどの受検者は望ましいパフォーマンス，つまり正しい記述が見られた。特に，光合成をする試験管を選択し，光合成の仕組みについて記述することについては他の群との差が顕著であった。その一方で，L群に属する受検者は光合成も呼吸もしない試験管についてはわかるが，それ以外に関する記述は少なかった。また，M群に属する受検者とL群に属する受検者からは，「光が当たらない試験管の中のオオカナダモはしおれてしまう」といった興味深い解答が見られた。これらの結果から，第1学年のパフォーマンス課題としては第2分野の項目のほうがふさわしいと言える。

第2学年の客観式調査において，全問正答者は6名，全問誤答者は0名であったため，尺度値θの

推定が不可能な受検者は6名であった。この6名を除いた際の尺度値 θ の平均値は -0.02，標準偏差 0.96，最小値 -1.98，最大値1.76であった。θ について，0.98 ～ 1.76と全問正答者をH群（N=12），-0.43 ～ 0.74をM群（N=11），-1.98 ～ -0.50をL群（N=11）と3群に分割した。

第2学年のパフォーマンス課題は，第1分野の項目が電力に関する項目であった。この項目は，2種類の回路のいずれかと2種類の抵抗のいずれかの組み合わせにより得られる4つの選択枝から，最も豆電球が暗くなる回路を選択し，電力の考え方を用いて電流と電圧に着目し，部分抵抗や全体抵抗の計算等の結果から選択の根拠を示すことを想定したものであった。選択枝の正答率はM群が最も低く，H群とL群はほとんど同じ正答率であった。選択理由については，全体的に見ると回路について着目して解答した割合がどの群においても高くなった。H群については，電流，抵抗，回路といった多くの観点に着目して選択枝を選択した受検者が多く見られた。正答したH群とL群の受検者の多くは電圧に着目しており，電圧が正しい選択枝を選択する要因となったように見える。この結果から，解答に際して根拠となる電流，抵抗，電圧，回路の4つ観点に対してどのように着目したかによって受検者の評価ができる可能性があると言える。

一方，第2分野の項目は，オタマジャクシとカエルの違い，カエルの飼育について文章のみならず表や図を用いて表現するものであった。この項目では，いずれの小問においてもほとんどすべての受検者が望ましいパフォーマンス，予想されるパフォーマンスやそれ以上のパフォーマンスを行っていた。したがって，学力群の間にパフォーマンスの差が生まれず，本研究で用いる課題としては適切ではないと言える。

予備調査を終えて，信頼性係数についての課題が残った。予備調査で使用した15項目からなる客観式調査のクロンバックの α 係数は，第1学年の調査では0.73，第2学年の調査では0.69であった。このような数値であっても，テストとして十分な性能を備えているものも実際に存在する。しかし，「全県学力調査」では25項目からなる第1学年，第2学年の調査ともにクロンバックの α 係数は0.8であったことから，客観式調査においては項目数が少ないことが影響して信頼性が高いとは言えない結果となったことがわかる。後述する本調査では，この課題を解決するために，第1学年と第2学年の客観式調査項目を同一尺度上に位置付け，両年度の項目を用いることで項目数を増やし，その結果としてテストの信頼性を担保することとした。この手続きを行うにあたり，調査対象を第2学年に絞り，パフォーマンス課題は第1分野の項目を使用することとした。本調査のために，予備調査の問題冊子から第2分野の項目を削除し，A4版1枚分に選択枝と選択理由を書けるようにした。また，問題文の下方の空欄には計算欄を設けた。

3　本調査

予備調査の結果を踏まえて，宮城県内B市立b中学校の協力を得て，平成27年1月に本調査を行った。調査対象者は第2学年6学級216名であった。予備調査と同様に，客観式調査（理科）とパフォーマンス課題の2つの問題冊子を配布し，解答してもらった。本調査における客観式調査（理科）は，第1学年の14項目と第2学年の11項目を合わせて25項目で構成し，同時尺度調整法によって同一尺

度上に位置付けた。この手順においては，新潟県における平成18年度「全県学力調査」のデータ（受検者：中学1年生22035名，中学2年生21520名，実施：平成18年1月）と本研究で得たデータを用いた。受検者パラメタ（尺度値 θ）の推定は最尤法による。なお，これらのパラメタ推定にはEasyEstimation（熊谷，2009）を使用した。その結果，識別力母数の等化前後の相関係数は0.995，困難度母数の相関係数は0.998となった。

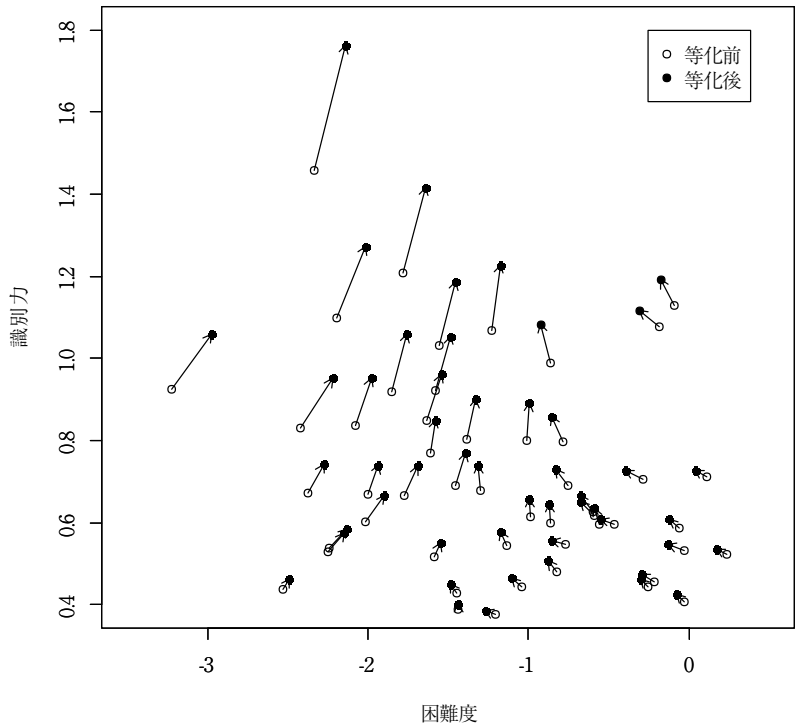

図1　等化前後の項目母数の推定値の変化

図1は等化前後の識別力と困難度の推定値の変化を表したものである。この図から，識別力が大幅に上がっている項目が存在することがわかる。これは，本来，第2学年の生徒にとっては難易度の低いはずの第1学年の項目に対して誤答したために，能力分布が広がり，結果として識別力が上がったものと考えられる。

本調査の客観式調査（理科）において，全問正答者が4名，この4名を除いた尺度値 θ の平均値は-0.350，標準偏差0.972，最小値 -3.470，最大値2.331であった。尺度値 θ について，全問正答者を含めた216名を72名ずつ値の低い順にL群，M群，H群と3群に分割し，パフォーマンス課題の解答を分析した。全体の傾向としては，学力群のレベルが上がるごとに選択枝の正答率は上昇し，観点の着目率が上昇した。

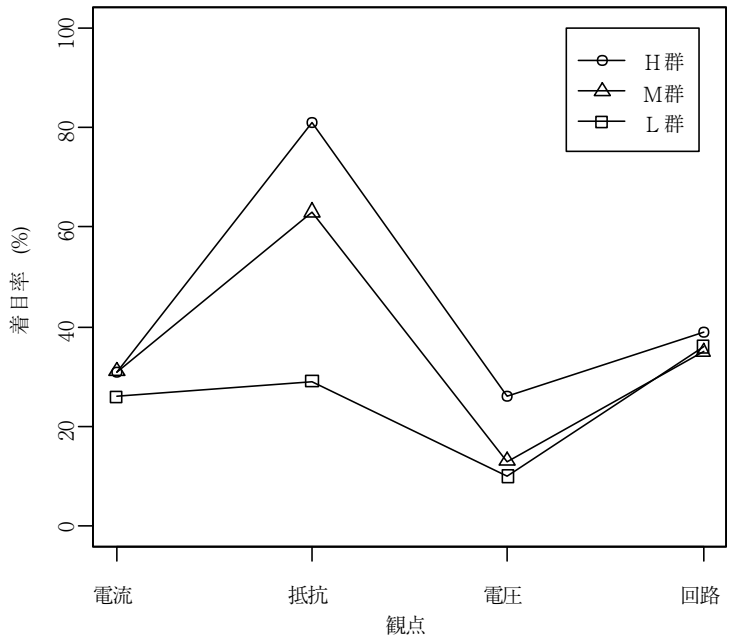

図2　学力群ごとの観点への着目率

図2から，電流と回路については学力群の間に大きな差は見られないが，抵抗への着目には大きな差が見られることがわかる。その一方で，学力群ごとの特徴も見られた。特にM群では回路と豆電球の明るさの関係について「直列回路は豆電球を2つつなげたとき，あまり光が出なかったから」といった実験などで得た直接的な経験をもとに主観的な解答する生徒が目立った。また，予備調査を行ったa中学校と比較をすると，2点の大きな違いが見られた。1つは，観点への着目率と正答率の関係である。a中学校では，観点への着目率が高い受検者ほど正答率が高いという結果であったが，b中学校のH群においては抵抗の観点のみで正しい選択枝を選択している受検者が多く見られた。その他の点として，抵抗への着目の仕方が挙げられる。a中学校では抵抗について計算する受検者が多く見られたが，b中学校では計算によってではなく，グラフの読み取りによって状況を把握している受検者が多く見られた。

4　考察

本研究より，IRT 尺度値を利用することによって，学力群が高いほど，多面的な視点で解答を考える割合が増加すること，並びに学力群とパフォーマンスの特徴を関連付けることが示せた。しかしながら，a中学校との比較からわかるように，学校の違い，あるいは指導者の違いによって思考の過程が異なっていた。先に述べた2点の違いのうち，観点への着目の仕方は，着目した観点の数を示す軸とそれぞれの観点についてどれほど深く考えたかを示す軸の2つの異なる軸についての情報を与えた。また，抵抗への着目の仕方については，数式を用いて解答を導く数学的な思考のパターンと，資料を読み取るという資料活用能力を用いた解答パターンが見られたことで，理科以外にど

のような教科と関連づけて学習を行っているかが明らかになった。これらのことは，同様の調査を指導者の異なる受検者を対象にした際，指導者の学習展開に影響を受けた解答が得られる可能性を示している。言い替えれば，ルーブリックの作成には個別の生徒集団の性質や教師の指導方法・指導方針などへの配慮が必要となると指摘できる。

【付記】

本研究はJSPS科研費25380867の助成を受けたものである。

【文献】

AERA（2014）．Standards for educational and psychological testing.

中央教育審議会（2008）．幼稚園，小学校，中学校，高等学校及び特別支援学校の学習指導要領等の改善について（答申）．

中央教育審議会（2010）．児童生徒の学習評価の在り方について（報告）．

Doran, R., Chan, F., Tamir, P. & Lenhardt, Carol.（2002）．Science Educator's Guide to Laboratory Assessment.（古屋　光一（監訳）（2007）．理科の先生のための新しい評価方法入門――高次の学習を育てるパフォーマンス課題，その実践例．北大路書房．）

遠藤貴広（2012）．教育評価改革の持続可能性をめぐる実践上の論点―ニューヨーク州テスト政策に対抗する草の根の取り組みを事例に―．福井大学大学院教育学研究科教職開発専攻（教職大学院）『教師教育研究』，5，255-263.

Hart, D（1994）．Authentic Assessment A Handbook for Educators. Dale Seymour Publications.（田中　耕治（監訳）（2012）．パフォーマンス評価入門　「真正の評価」論からの提案．ミネルヴァ書房．）

池田央（1994）．現代テスト理論．朝倉書店．

石井英真（2010）．Ⅳ教育目標と教育評価の関係7ルーブリック．（田中耕治（編）（2010）．よくわかる教育評価　第2版．ミネルヴァ書房．）

石井英真（2011）．第1章パフォーマンス評価の理論第3節パフォーマンス評価をどう実践するか．（田中耕治（編）（2011）．パフォーマンス評価　思考力・表現力・判断力を育む授業づくり．ぎょうせい．）

加藤健太郎・山田剛史・川端一光（2014）．Rによる項目反応理論．オーム社．

岸本実（2010）．Ⅶ学力評価のさまざまな方法11パフォーマンス評価：パフォーマンス課題とそのつくりかた．（田中耕治（編）（2010）．よくわかる教育評価　第2版．ミネルヴァ書房．）

Lane, S. & Stone, C. A.（2006）. Performance Assessment.（Robert L. Brennan（Ed.）（2006）. Educational Measurement（American Council on Education/Oryx Press Series on Higher Education）. 4th ed.）

松下佳代（2007）．パフォーマンス評価．日本標準．

文部科学省（2008）．小学校学習指導要領．

文部科学省（2008a）．中学校学習指導要領．

文部科学省（2008b）．中学校学習指導要領解説理科編．

村木英治（2011）．項目反応理論．朝倉書店．

新潟県教育委員会（2005）．平成16年度「全県学力調査」報告書．

新潟県教育委員会（2007）．平成18年度「全県学力調査」報告書．

西村和雄・戸瀬信之（2004）．アメリカの教育改革．京都大学学術出版会．

西岡加名恵・田中耕治（編）（2009）.「活用する力」を育てる授業と評価　中学校――パフォーマンス課題とルーブリックの提案. 学事出版.

西岡加名恵（2010）. Ⅶ学力評価のさまざまな方法1学力評価の方法の分類.（田中耕治（編）（2010）. よくわかる教育評価　第2版. ミネルヴァ書房.）

野口裕之・大隅敦子（2014）. テスティングの基礎理論. 研究社.

佐藤誠子・柴山直（2013）. IRT モデルにもとづく学力評価ルーブリック作成手法の試み――面積比較課題を例として. 日本教育心理学会第55回総会論文集, 38.

田中耕治（編）（2002）. 新しい教育評価の理論と方法　第Ⅰ巻　理論編. 日本標準.

田中耕治（編）（2010）. よくわかる教育評価　第2版. ミネルヴァ書房.

田中耕治（編）（2011）. パフォーマンス評価　思考力・表現力・判断力を育む授業づくり. ぎょうせい.

豊田秀樹（2012）. 項目反応理論〔入門編〕（第2版）. 朝倉書店.

Welch, C.（2006）. 第13章パフォーマンステストの問題作成（Steven M. Downing and Thomas M. Haladyna Eds.（2006）. Lawrence Erlbaum Associates, Inc）（池田央（監訳）（2008）. テスト作成ハンドブック. 教育測定研究所.）

Yen, W. M. & Fitzptrick, A. R.（2006）. Item response theory.（Robert L. Brennan（Ed.）（2006）. Educational Measurement（American Council on Education/Oryx Press Series on Higher Education）. 4th ed.）

Interpretation of the Performance of the Junior High School Science using IRT Scale Scores:

A performance task for “Electricity” unit

Tadashi SHIBAYAMA
（Professor, Graduate School of Education, Tohoku University）
Yoko CHIBA
（Graduate Student, Graduate School of Education, Tohoku University）

Recently performance assessments are adopted as methods of evaluating complex-achievement such as higher-order thinking, problem solving and so on. But this type of methods relies on raters’ subjective judgement so that its reliability is not so high. This article tried to refer to the scales based on IRT in interpreting junior-high school students’ performance of science. In practice, students answered objective test about science and a performance task about a unit of this subject which is called as “electric power”. Using the data, the items of objective test was carried out equating to position the first grade and second grade items on the common scale by a concurrent calibration method. And analyzing the results of the performance task with IRT scale scores, there was a tendency to increase the number of points of view of interest as the level of academic achievement increases.

Key Words：Performance assessment, Junior high school science, IRT, Equating, Rubric