様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月 ２９ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３８０８６７

研究課題名（和文）パフォーマンス評価におけるIRT尺度を利用した信頼性向上のための基礎研究

研究課題名（英文）Improving reliability of performance assessment with IRT scores

研究代表者
柴山　直（SHIBAYAMA, Tadashi）

東北大学・教育学研究科（研究院）・教授

研究者番号：７０２４０７５２

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：　高度で複雑な学力を評価するための手法としてのパフォーマンス評価は，評価者の主観に左右されるという本質的な問題を抱えている。逆に心理計量・教育測定の分野で発展してきた項目反応理論は学力を単一のパラメーターないし尺度値で表現するため信頼性・客観性の担保に優れている反面，多様性にあふれた現実の学力を捨象しすぎるという限界がある。
　本研究ではこの２つの相反する利点と欠点をもったアプローチを，言語学で発展してきたCan-Do-Statementsを媒介することによって結びつけ，パフォーマンス評価の信頼性・客観性および指導への有用性の向上をはかった。

研究成果の概要（英文）：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 study 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.

研究分野：教育測定学

キーワード：パフォーマンスアセスメント　IRT　項目反応理論　等化　垂直尺度　ルーブリック　理科　信頼性

１．研究開始当初の背景

一般的に何かを評価するためには，「誰が」「何を」「どのような規準で」「何の目的のために」を明確にする必要がある．一方，これからの時代に必要なスキルないし能力として喧伝されている「問題解決能力」「論理的思考力」「クリティカル・シンキング」あるいは「ジェネリック・スキル」等々は，これらをいざ心理学的構成概念として明確に定義しようとするとその作業は困難を極め，それに応じて上記の諸点のうち特に「何を」と「どのような規準で」の２つは，評価する側のいわば暗黙知に依存した抽象的なレベルの記述にとどまらざるを得ない．

そこで，対象とする能力ないしスキルに習熟している人間が評価者となり，具体的な所産物・行為等から評価対象者の能力・スキルを判定する，いわゆるパフォーマンス評価が採用される．高度で複雑な学力を評価するための手法として，いまやパフォーマンス評価は必須の評価方法となっている．しかしながら，パフォーマンス評価は人間が直接関わるが故に，つねに評価者側の信頼性・一貫性が問われ，かつ，その結果の主観性から本質的に逃れることができないという宿命を背負っている．

翻ってOECD/PISAやヨーロッパ言語共通参照枠(Common European Framework of Reference for Language)などを視野にいれると，そこでは習熟度レベル(proficiency level)という概念を導入し，IRT モデルなどの計量心理学的なモデルを通して尺度表現できたあるスコアの値をもつ受検者が実際にどのようなことができるのかの記述(Can-Do Statements)を行うことで，その受検者のもつ能力・スキルの質を可能な限り客観的に保証しようとしていることがわかる．パフォーマンス評価においては，いわば価値判断である評価とその判断のもとになる測定とが渾然一体となる一方，後者の心理計量的なアプローチは測定と評価を厳密に区別しているとも表現できる．

本研究の目的は，パフォーマンス評価とIRTモデルに基づく心理計量的なアプローチの両者を Can-Do Statements の精緻化によってさらに密接に接合し，前者の信頼性・客観性を向上させる一方，後者のスコアから得られる情報の多様性を確保するための基本的な方法論を開発することにある．具体的には，ある尺度に位置する児童・生徒が実際の「思考力」「論理的判断力」を試される場面でどのようなことができるのかを，詳細に実験・観察・記述し，それを形成的に評価していくためのツールとしてのルーブリックの作成を試み，さらには，評価のみならず，その児童・生徒が躓いているポイントを見つけ出すし具体的な指導に活かせるノウハウを確立することにある．

２．研究の目的

高度で複雑な学力(Higher and Complex Achievement)を評価するための方法論として，今や必須となったパフォーマンス評価は，その一方で評価者の主観に左右されるという本質的な問題を抱えている．逆に心理計量・教育測定の分野で発展してきた項目反応理論(Item Response Theory ; IRT)はターゲットとする学力を単一のパラメーターないし尺度値で表現するため信頼性・客観性の担保に優れている反面，多様性にあふれた現実の学力を捨象しすぎるという限界がある．
本研究ではこの２つの相反する利点と欠点をもったアプローチを，言語学で発展してきた Can-Do Statements を媒介することによって結びつけ，パフォーマンス評価の信頼性・客観性および指導への有用性の向上をはかることを目的とした．

３．研究の方法

本研究は以上の経緯と着想を受け，以下の 3 点を主たる課題とした．これらの課題は具体的ない実践の場としては，協力校として内諾を得ている小学校ならびに中学校において算数/数学ならびに理科をフィールドとして展開した．

１）IRT 尺度の構成　上記学力データにもとづき算数/数学と理科の全項目の IRT パラメーターの推定ならびに尺度の構成をおこなう．ただし，調査時期以降学習指導要領の改訂があったため，新学習指導要領の範囲外の問題項目(数個)は削除し，使用対象外とした．

２）IRT 尺度にもとづく Proficiency Level の設定とパフォーマンス課題の作成　上記の尺度にもとづき，Proficiency Level を実証的に設定し，そのレベルに対応したパフォーマンス課題を作成した．その際，作成されたパフォーマンス課題を評価するためのルーブリックの設定を「○○ができる(Can Do)」の形式でおこなった．

３）IRT 尺度とパフォーマンス課題の接合　パフォーマンス課題の結果と IRT モデルによって得られた学力特性値の相関分析による両者の接合を試みた．

４．研究成果

協力自治体で過去に実際に実施された算数／数学および理科の学力テストデータの貸与を受け，データクレンジング作業をおこない，その全項目 IRT パラメータの推定ならびに尺度の構成を行った．また，それと併行して，研究体制の確立，国内外における資料収集および先行研究の調査，実データの IRT 分析による問題項目のデータベースの作成，パフォーマンス課題の試作を行った．さらに

理論的検討・データ収集デザイの検討・パフォーマンス課題の試作においては整理したデータにもとづき IRT 分析を行い，今後の展開に必要な算数／数学および理科の問題項目に関する項目困難度や識別力の推定値を取得した．推定のためのプログラムとしては EasyEstimation(熊谷，2009)を用いた．協力校(宮城県内，山形県内)の教員とともに算数/数学および理科に関するパフォーマンス課題の試作を行った．

算数については 23 問からなる学力テストを課す一方，平行四辺形を題材に，パフォーマンス課題として等積変形課題２問，等周長変形課題２問を課した。その結果を学力層別に描くと下図に示すとおりになった。理由分類やデータ分析の結果を検討しルーブリックが例示できた。

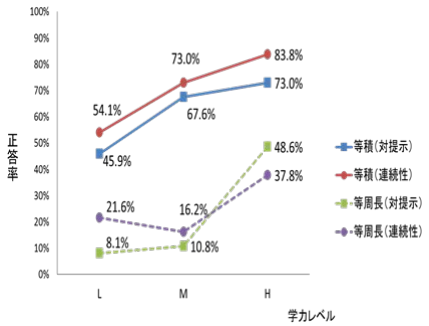

図１　学力レベルごとの面積課題の正答率

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を測る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，まず，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った．一方，予備調査の結果によって設定されたルーブリックの観点では本調査の調査対象者のパフォーマンスを捉えきれない場合があったことも見いだされた．

副次的な研究成果としては、近年発展してきた拡張 IRT モデルを用いた妥当性の検証方法を提案できた．具体的には拡張 IRT モデルである線形ロジスティックテストモデル（LLTM）を用い，カリキュラム等に基づいて作成されたテストに関する「本質的な側面の証拠」を見出す手法である．

また、異なる学年間でのＩＲＴ等化（２母数ロジスティックモデル）を行った場合，下図に示すとおり，困難度においては縮退，識別力においては困難度が相対的に低い項目群での識別力の向上がみられた。この現象を実データで見いだせたことは，今後達成度尺度を作成する際などに大きな知見をもたらすものと期待できる。

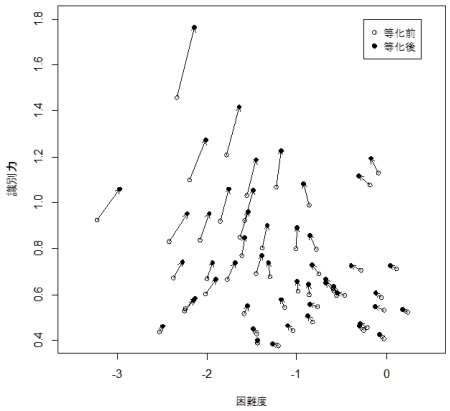

図２　等化前後の項目母数の推定値の変化

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

①柴山直・千葉陽子(2015) 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－，東北大学大学院教育学研究科研究年報，査読無，第 63 集第 2 号，Pp.213-221

〔学会発表〕（計４件）

①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 回総会発表論文集，p.111.

②坂本佑太朗・柴山直(2014) 拡張 IRT モデルによる日本型テストの計量的分析の試み，日本テスト学会第 12 回大会発表論文抄録集，Pp.120-121

③佐藤誠子・柴山直 (2014) 算数学力評価ルーブリックの妥当性検討の試み―面積学習領域を例として―，日本教育心理学会第 56 回総会発表論文集，発表番号 PE038.

④柴山直・千葉陽子 (2015) IRT 尺度値を利用した理科のパフォーマンスアセスメント結果の解釈について，日本教育心理学会第 57 回総会発表論文集，発表番号 PF070.

〔図書〕（計　　件）

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：

出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
柴山　直（SHIBAYAMA, Tadashi）
東北大学大学院教育学研究科・教授
研究者番号：７０２４０７５２

(2)研究分担者
清水禎文（SHIMIZU, Yoshifumi）
東北大学大学院教育学研究科・助教
研究者番号：２０２３５６７５

佐藤誠子（SATO, Seiko）
石巻専修大学・人間学部・助教
研究者番号：２０６３３６５５

(3)連携研究者
（　　）

研究者番号：