

パフォーマンス評価における項目反応理論を利用した アカデミック・ライティング力の測定

斎藤 有吾

1. 問題と目的

昨今、高等教育において問題解決能力や論理的思考力といった汎用的で高次の能力を育成することが求められている。またそのような学習成果を評価する具体的な手法として、2014年の中央教育審議会の答申「新しい時代にふさわしい高大接続の実現に向けた高等学校教育、大学教育、大学入学者選抜の一体的改革について」において、複雑な課題に知識やスキルを活用して探究し表現することを求める「パフォーマンス評価」、そうした複雑な課題の達成度を数段階に分け達成度を判断する基準を示す「ルーブリック」の活用を高等教育機関においても積極的に推進していく旨が明記された。それにより、高次の能力の学習成果に関する議論はますます活発になっていくことが予想される。さて、学習成果の評価には大きく分けて直接評価と間接評価の2つがある(山田, 2012)。松下(2014)によると、直接評価とは学習成果を学生の知識や能力の表出(何ができるか)によって評価したものであり、例として各科目の試験やレポート、卒業研究などの評価があげられる。他方、間接評価とは学生の自己認識の報告(何ができているか)によって間接的に評価するものである。パフォーマンス評価やルーブリックは、まさに高次の能力の学習成果を直接評価するものであるといえる。以後本研究では、直接評価によって当該学習成果を数値化したものを直接的指標、間接評価によって数値化したものを間接的指標と表現する。

これまでの日本の高等教育研究では、アメリカのカレッジインパクト研究を主な理論的基盤としながら、自己報告式アンケートを用いた学生調査によって学生自身に学習の達成度や成長の度合いなどを回答させ、それを学習成果とみなし、その学習成果に対して学習環境や授業経験、デモグラフィック要因、サイコグラフィック要因などがどのような影響を与えるのかといった実証的研究が積み重ねられてきた(例えば森・山田, 2009; 山田・森, 2010; 山田, 2012)。問題解決能力や論理的思考力といった高次の能力の学習成果に関しても、このような間接的指標が用いられている場合が多い。ただし間接的指標である以上、「何ができているか」という側面から学習成果を捉えようとしているということに留意する必要がある。Pike(1996, 2011)によると、このような学習成果の間接的指標と直接的指標は弱～中程度の相関であり、間接的指標は、「何ができるか」を示す直接的指標の代理として使用できるものではないことが指摘されている。間接的指標は簡便に多くの学生の学習成果を得点化でき、プログラムレベル

のインパクトを検討しやすいという利点があることを認めつつも、もし、「何ができると思っているか」ではなく、「何ができるか」という側面から学習成果を捉えたい場合には、やはり直接的指標を使用することが望ましいといえよう。

ただし、高次の能力を直接評価によって信頼性・妥当性を担保して測定・評価することは容易ではない。松下・小野・高橋 (2013) によると、パフォーマンス評価では表面的妥当性は担保しやすい反面、信頼性は担保しにくく、ルーブリックを用いたとしてもそれだけで信頼性が担保されるという保証はない。事実、松下他 (2013) の教員 3 名によるルーブリックを用いたアカデミック・ライティングのパフォーマンス評価の一般化可能性研究によると、教員 3 名の場合の信頼性係数は 0.62 であり、もし教員 1 名の場合、信頼性係数は 0.4 未満と見積もられるということから、やはりルーブリックそれのみで信頼性の担保が保証されるとは言い難い。心理測定学的には信頼性は妥当性の必要条件に対応するものと位置づけられており、信頼性が低いのであれば妥当性も低いということになるため、信頼性の担保は妥当性を論じる前提として重要である。

しかし、松下 (2012) によると、パフォーマンス評価やルーブリックは、心理現象を数値化することによって成り立つ心理測定学とは異なるパラダイムを基礎としており、そのような評価方法を代替・補完するようにデザインされたものである。すなわち、質的なデータから質的な評価基準のもとで学習の質を捉えることを重視するものであり、その評価結果は主に文章などによって質的に表現され (数値で表わされる場合もある)、形成的な評価として使用されることが多い。そのため、パフォーマンス評価において評価結果を数値化した場合のみを取り上げ、それを心理測定学に依拠して信頼性や妥当性を検討して良し悪しを判断することはオルタナティブな評価方法としての利点を損なう危険性をはらんでいる。しかしながらそのような評価において心理測定学的な枠組みを用いることで、複数観点の類似性や、水準の設定の適切性、複数評価者の評価の一致度や各評価者の評価能力、さらには当該評価がどのようなレベルの学生に向いているのかといったことを検討し、より良い評価にするための手がかりを得ようとした先行研究はいくつか存在する。特にそのような検討で有効な手法として近年用いられているのは、項目反応理論 (Item Response Theory; IRT) である。

項目反応理論は古典的テスト理論とは異なり、評価やテストに含まれる項目の困難度や識別力といった特性と受験者の能力を別々に推定し、同じ物差しの上で評価することが可能である (加藤・山田・川端, 2014)。例えば、あるパフォーマンス評価において、複数観点の分析的ルーブリックで得点づけをした場合の、各観点の困難度や識別力に加え、観点全体の測定の精度などを受験者集団に依存せずに検討できる。他にも、同じ構成概念を測ろうとしているが異なる課題で構成された評価間の結果の比較を容易にしたり、測定精度を保ったまま項目や問題を減らしたりといったことが可能になる (豊田, 2002)。

レポートや小論文、オープンエンド型の論述式課題による評価はパフォーマンス評価になる。そのような評価への項目反応理論の応用を試みた研究としては、小論文の評価データに項目反応理論の段階反応モデルをあてはめ推定の精度を検討したもの (渡部・平井, 1993)、小論文の評価における受験者と評価者のバイアスを同時に評価するための項目反応モデルの提案を行ったもの (宇佐美, 2010)、国語の様々な問題形式や評価方法に対し、重複テスト分冊法や

項目反応理論を適用するための方法を検討したもの(柴山他, 2012), レポートのピアアセスメントにおける評価者の評価能力を考慮し, より精度の高い能力推定をするための項目反応モデルの提案を行ったもの(宇都・植野, 2015)などがあげられる。しかし, ルーブリックを用いたパフォーマンス評価における心理測定学的検討から, 各観点の難易度や識別力といった特徴を明らかにし, 改善に有用な情報を得たりするといった研究はまだ少ない。昨今の高等教育への要請を考えれば, そのような知見を積み重ねていく必要があるだろう。

松下他(2013)ではアカデミック・ライティング力を評価するためのパフォーマンス評価の信頼性を一般化可能性理論によって検討している。しかし, 各観点はすべて同列に扱われており, 各観点の困難度や識別力に関する検討は行われておらず, 観点の数や評価者の数をどの程度にすれば一定の信頼性が担保されるかといったことを検討するにとどまっている。そこで本研究では, 松下他(2013)と同じ授業実践のパフォーマンス評価(ただし, 実施年度は異なる)を事例とし, そのデータを項目反応理論によって分析することで, (a) ルーブリックの各観点や観点全体の測定論上の特徴を明らかにすること, また, (b) 項目反応理論を用いて推定した場合の学生の能力値と, 全観点の得点の単純な合計点との関連を見ることで, 実用上の示唆を得ることを目的とする。

2. 方法

新潟大学初年次教育科目「大学学習法」におけるパフォーマンス評価

2013年度と2014年度で新潟大学歯学部において開講された初年次教育科目「大学学習法」のパフォーマンス評価のデータを使用する。よって調査対象者は「大学学習法」を履修した新潟大学歯学部の初年次生である。コースの詳細は松下他(2013)を参照されたい。ただし, 評価課題とルーブリックは松下他(2013)から変更されている。このコースではレポートの書き方やプレゼンテーションの仕方などについて講義・演習を行い, 学生には講義・演習で身につけた知識やスキルをもとにレポートを作成する課題が与えられる。レポートでは「医療や科学の進歩がもたらした諸問題」に関して学生自身で話題を設定し, 主体的に調査・学習を行い, 自身の考えを3000字程度で記述することが求められる。課題の詳細を図1に示す。

医療や科学の進歩にはめざましいものがあります。少し前までは治療することが難しかった病気にも対処できるようになりました。また、便利な道具も世の中に溢れており、日々、私たちは恩恵を受けて暮らしています。しかしその一方で、進歩がもたらしたさまざまな問題も抱え込んでしまいました。皆さんもテレビや新聞で見聞きしていることでしょう。その解決のために、多くの議論がなされていますが、どの問題もしかるべき解決策が示されていないのが現状ではないでしょうか。
そこで、与えられたテーマ「医療や科学の進歩がもたらした諸問題」から話題を設定し、主体的に調査・学習を行い、自分の考えをレポートにして提出してください。

図1 2013・2014年度 大学学習法 パフォーマンス課題

そして, この課題に対する学生のレポートから学生のアカデミック・ライティング力を評価するためのルーブリックを表1に示す。「背景と問題」「主張と結論」「根拠と事実・データ」「対立意見の検討」「全体構成」「表現ルール」の6つの観点からなる。このルーブリックは, 松下他(2013)の「大学学習法」において使用されたルーブリックをもとにしている。2011年度に

「大学学習法」担当教員の一人が考えた素案に、他の担当教員との協議の上での修正を加え、さらに2012年度に教育評価の専門家との協議によって「問題解決能力」や「論理的思考力」の定義に理論的根拠を与え観点を整理するなどして修正が加えられた。また2012年度のルーブリックに関しては一般化可能性理論による評価の信頼性の検討が行われており、一定の信頼性が担保されていたものの、水準数を増やすとともに各水準の記述語を修正する必要性などが示唆され、それらを反映して修正したものが本研究で使用するルーブリックである。このように担当教員と教育評価の専門家によって数回の修正が加えられており、表面的妥当性や内容的妥当性は担保されたものといえるだろう。2012年度までのルーブリックの詳しい開発の経緯や理論的根拠は松下他(2013)を参照されたい。

課題の提示と同時に、学生へのルーブリックの提示と評価基準の解説が行われた。提出されたレポートの評価は、すべての学生と観点に関して、大学学習法において特にライティング指導に関与した教員3名によって行われた。この3名は評価に先立ってルーブリックの記述語とその意味するところを確認して共通理解を目指し、さらに前年度の学生のレポートのうち特徴的な3本を採点事例として共有するなどしてキャリブレーション(calibration)を行っている。それぞれ「レベル3」から「レベル1」、「レベル1」に満たないものは「レベル0」の4段階で評価し、レベルに応じて得点化した。本研究では3名の教員の中央値をそれぞれの学生の当該観点の得点とした。例えば、Aという学生の「背景と問題」の観点の評価に関して、教員aが1、教員bが2、教員cが3だった場合、その学生の「背景と問題」の得点は2点となる。同様に、「主張と結論」の観点の評価に関して、教員aが3、教員bが3、教員cが2だった場合、その学生の「背景と問題」の得点は3点となる。本研究において中央値の算出の際、タイ順位はこのように単純に扱うこととする。このような手続きで、これら6つの観点から得点化されるものを本研究において「アカデミック・ライティング力」と操作的に定義する。本研究では、6つの観点の単純な合計点と、項目反応理論によって推定される能力値 θ の2つの「アカデミック・ライティング力」を取り扱う。

分析方法

まず、項目反応理論においては観点の一次元性が前提となるため、固有値のスクリープロットと、ポリコリック相関係数に基づいた因子分析によって各観点の因子負荷量を確認し、一次元性を検討する。その後、項目反応理論の段階反応モデル(Samejima, 1969)によって各観点の境界母数や識別力母数、すなわち項目母数を推定し各観点の特徴を捉える。同時にアカデミック・ライティング力を示す学生の特性値 θ を推定し、その分布を確認する。さらに観点全体の情報量、すなわちテスト情報量から、本研究におけるパフォーマンス評価によるアカデミック・ライティング力の測定精度を考察する。最後に、2つのアカデミック・ライティング力と総合印象点との相関係数を算出し、それぞれの関連を検討する。なお、項目反応理論による分析はExametrika(Shojima, 2007)を使用した。

表 1 アカデミック・ライティングルーブリック

観点	論理的思考					文章表現
	問題解決	主張と結論	根拠と事実・データ	対立意見の検討	全体構成	
観点の説明	与えられたテーマから自分で問題を設定する。	設定した問題に対し、展開してきた自分の主張を関連づけながら、結論を導く。	自分の主張を支える根拠を述べ、根拠の真実性を立証する事実・データを明らかにする。	自分の主張と対立する意見を取り上げ、それに対して論駁(問題点の指摘)を行う。	問題の設定から結論にいたる過程を論理的に組み立て、表現する。	研究レポートとしてのルール・規範を守り、適した文章と言い回しを用いてレポートを作成する。
レベル3	与えられたテーマから問題を設定し、論ずる意義も含め、その問題を取り上げた理由や背景について述べている。	設定した問題に対し、展開してきた自分の主張を関連づけながら、結論を導いている。結論は一般論にとどまらず、独自性を有している。	自分の主張の根拠が述べられており、かつ根拠の真実性を立証する信頼できる複数の事実・データが示されている。	自分の主張と対立するいくつかの意見を取り上げ、それらすべてに対して論駁(問題点の指摘)を行っている。	問題の設定から結論にいたる論理的な組み立て、記述の順序、パラグラフの接続が整っている。概要は本文の内容を的確に要約している。	・研究レポートとして適した文章と言い回しを用いてレポートを書いている。 ・引用部分と自分の文章の区別を明示し、引用部分については、レポートの最後に出所を確認できる形で参考文献を記載している。 ・概要、本文ともに字数制限が守られている。
レベル2	与えられたテーマから問題を設定し、その問題を取り上げた理由や背景について述べている。	設定した問題に対し、展開してきた自分の主張を関連づけながら、結論を導いている。	自分の主張の根拠が述べられており、かつ根拠の真実性を立証する信頼できる事実・データが少なくとも一つ示されている。	自分の主張と対立する少なくとも一つの意見を取り上げ、それに対して論駁(問題点の指摘)を行っている。	問題の設定から結論にいたる論理的な組み立て、記述の順序、パラグラフの接続がおおむね整っている。	・3つの条件をすべて満たす場合は「レベル3」、2つの場合は「レベル2」、1つの場合は「レベル1」とする。>
レベル1	与えられたテーマから問題を設定しているが、その問題を取り上げた理由や背景の内容が不十分である。	結論は述べられているが、展開してきた自分の主張との関連づけが不十分である。	自分の主張の根拠は述べられているが、根拠の真実性を立証する信頼できる事実・データが明らかにされていない。	自分の主張と対立する意見を取り上げているが、それに対して論駁(問題点の指摘)がなされていない。	問題の設定から結論にいたるアウトラインはたどれるが、記述の順序やパラグラフの接続に難点のある箇所が散見される。	
レベル0	レベル1を満たさない場合はゼロを割り当てること。					

3. 結果と考察

要約統計量

上述の 2013 年度と 2014 年度の「大学学習法」を履修した学生のうち、パフォーマンス評価の結果を使用できた 116 名を分析に使用した。なお、2013 年度は 60 名、2014 年度は 56 名であった。まず、ルーブリックによる得点の要約統計量を表 2 に示す。

表2 ルーブリックによる得点の要約統計量

変数名	平均値	中央値	標準偏差	最小値	最大値
背景と問題	1.64	2.00	0.53	0.00	3.00
主張と結論	1.22	1.00	0.51	0.00	2.00
根拠と事実・データ	1.12	1.00	0.53	0.00	3.00
対立意見の検討	0.48	0.00	0.80	0.00	3.00
全体構成	1.05	1.00	0.57	0.00	3.00
表現ルール	1.46	1.00	0.70	0.00	3.00
全観点平均点	1.16	1.00	0.42	0.33	2.67
全観点合計点	6.97	6.00	2.50	2.00	16.00

各観点の平均値と中央値をみると、「背景と問題」が最も高く、「対立意見の検討」が最も低い。「背景と問題」の中央値は 2 であり、およそ 60% の学生 (74 名) がレベル 2 に該当していた。「対立意見の検討」の中央値は 0 であり、およそ 70% の学生 (80 名) がレベル 0 に該当しており、床効果が発生していると考えられる。これらの傾向を考慮すると、観点同士の相関を検討するのであれば、ポリコリック相関係数を用いることが妥当と考えられる。また、「主張と結論」は最大値が 2 となっており、レベル 3 に該当する学生はいなかった。

ポリコリック相関係数をもとにしたカテゴリーカル因子分析

全観点の一次元性の確認のため、全観点から得られる固有値のスクリープロットを図2に示す。また、平行分析・MAPとカテゴリーカル因子分析結果を表3に示す。第1因子の固有値は2.91、寄与率は48.48%であり、第1因子で全観点のおよそ半分近くを説明できているといえる。また、MAP(最小平均偏相関)は第1因子で最小となり、対角SMC平行分析は対角SMCの固有値が乱数データの対角SMCの固有値を第1因子の時点で逆転しており、どちらも1因子構造を支持している。堀(2005)に従えば、1因子構造、すなわち一次元性が満たされているとみなして差し支えないと考えられる。

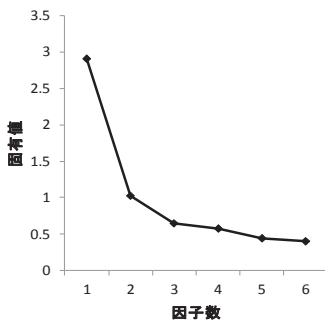


図2 スクリーンプロット

表3 平行分析・MAPとカテゴリーカル因子分析結果(1因子・重み付き最小二乗法)

	固有値	累積寄与	対角SMC	乱数対角SMC	MAP
Factor1	2.91	48.48	2.29	0.43	0.05
Factor2	1.03	65.59	0.28	0.29	0.12
Factor3	0.65	76.37	-0.05	0.11	0.27
Factor4	0.58	86.00	-0.12	-0.02	0.51
Factor5	0.44	93.38	-0.15	-0.14	1.00
Factor6	0.40	100.00	-0.19	-0.27	-

項目	Factor1	共通性
主張と結論	.93	.86
全体構成	.91	.82
根拠と事実・データ	.79	.63
対立意見の検討	.73	.53
背景と問題	.53	.28
表現ルール	.53	.28
因子寄与	3.40	

次に、1因子を指定した場合のポリコリック相関係数に基づくカテゴリーカル因子分析の結果を見ると、因子負荷量はいずれの観点も.50を超えており、削除すべき観点は特にないといえる。なお、全観点合計点(全観点平均点)の α 係数は.76であった。これら6つの観点を使用した項目反応理論による分析を行い、その心理測定学的特徴を検討する。

項目母数の推定と解釈

項目反応理論(Samejimaの段階反応モデル)による項目母数の推定結果を表4に示す。また、これらの項目母数をもとに描いたカテゴリ特性曲線を図3に示す。 a は識別力母数、 b は境界母数を示す。各観点の識別力母数 a は0.89-1.31の範囲であり、すべての観点が豊田(2002)が最低ラインとしている0.2を上回っていた。また、過剰に高い推定値も出ていない。さらに、値のばらつきは小さく、いずれも似たような識別力を有しているといえよう。

境界母数 b は観点によってその様相が異なることがうかがえる。以下では、特徴的な傾向を示した観点を確認していく。なお、「主張と結論」はレベル3に該当する学生がいなかったため、 b_3 が推定されなかった。

まず「背景と問題」に関して、 b_1 が-3.45、 b_2 が-0.74、 b_3 が3.83であった。この場合、もし能力値 θ が-3.45より上であればレベル1以上になる確率が、そうでない確率(すなわちレベル0になる確率)よりも高くなることを示している。同様に、 $\theta > -0.74$ であればレベル2以上に、 $\theta > 3.83$ であればレベル3となる確率がそうでない確率よりも高くなる。これらの b の間隔が広いということがこの観点の特徴といえる。それほど能力がなくてもレベル1、平均的な能力があればレベル2にはなれるが、かなり高い能力がないとレベル3

にはなれない。さまざまな能力の集団に使用できる観点といえよう。

表4 各観点の識別力母数と境界母数				
	a	b_1	b_2	b_3
背景と問題	0.89	-3.45	-0.74	3.83
主張と結論	0.95	-2.42	0.87	—
根拠と事実・データ	1.06	-2.22	1.31	2.88
対立意見の検討	1.17	0.60	1.36	2.84
全体構成	1.06	-1.81	1.48	2.58
表現ルール	1.31	-2.86	0.27	2.07

それに対し、「対立意見の検討」は b_1 が 0.60, b_2 が 1.36, b_3 が 2.84 であった。 $\theta > 0.60$ という平均的よりも上の能力でないと、レベル 1 以上になる確率がそうでない確率を上回らない。すなわち困難度が高めの観点であるといえる。また間隔が狭く、この観点は能力が高めの集団に敏感に機能するものであるといえよう。ただし、カテゴリ特性曲線を見ると、レベル 1 とレベル 2 の曲線の交差状況から、レベル 1 まで出来しまえばレベル 2 も出来る可能性が高いということがわかる。「対立意見の検討」のレベル 1 とレベル 2 の違いは自分の主張と対立する意見に論駁をしているかどうかであるが、今回の結果から、対立する意見を取り上げることができるほどのアカデミック・ライティング力を有していれば、論駁もしている可能性が高いということが示された。これらを総合的に解釈すると、対立意見を取り上げることができるかどうかがこの観点にとって重要な境界となっており、それには平均的よりも上のアカデミック・ライティング力を要すると考えられるため、高いアカデミック・ライティング力の学生にうまく機能する観点であると考えられる。上述のように、今回の調査対象者は初年次生であり、「対立意見の検討」ではおよそ 70% の学生 (80 名) がレベル 0 に該当していることから、この解釈が支持されよう。

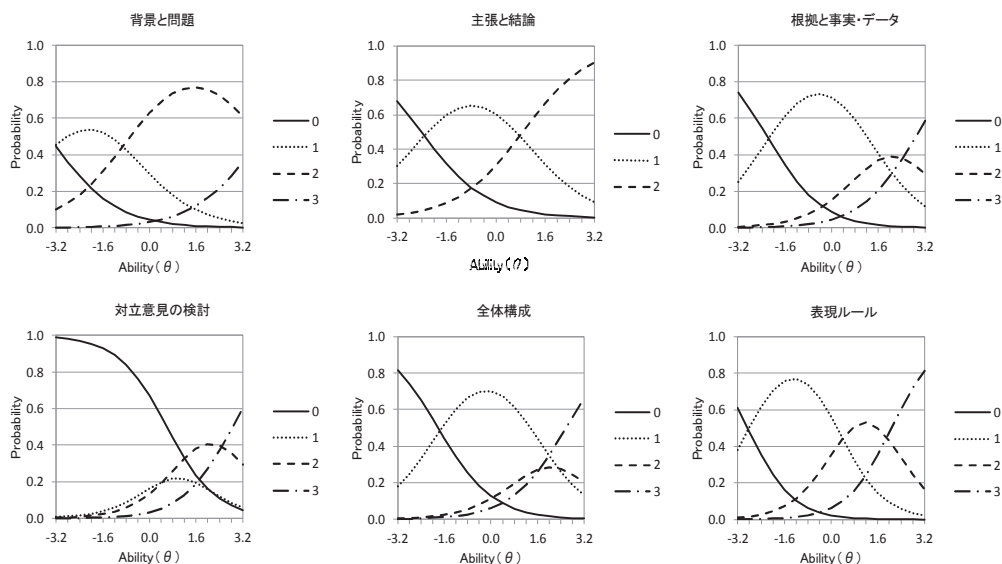


図 3 カテゴリ特性曲線

期待得点とテスト情報関数

多値型 IRT モデルを使用することで、能力値が θ である受験者のテスト全体に対する期待得点 (今回の場合、特定の能力値 θ の受験者におけるループリック全観点の合計得点の期待値) を求めることが可能となる。また、古典的テスト理論における信頼性係数に相当する、テストの測定精度を評価するためのテスト情報量を算出することが可能になる。ただし、テスト情報関数は古典的テスト理論における信頼性係数と異なり、受験者集団に依存しないテスト固有の測定精度に加え、受験者ごとの測定精度を検討できる。図 4 にテスト情報曲線を示す。能力値 $\theta = 1.6$ 付近で情報量が最大であり、測定精度が最も高いことがわかる。ただし、 θ によって大幅な違いは生じていない。テスト情報量の平方根の逆数は、当該 θ における真のテスト得点からの誤差となるが、 θ によって情報量に大幅な違いが生じていないため、誤差の大きさにも違いはさほどないといえる。全受験者の平均的な標準誤差は 0.79 (範囲は 0.73-0.84) であり、受験者によって測定精度の大きな違いは見られなかったこともそれを裏付けている。

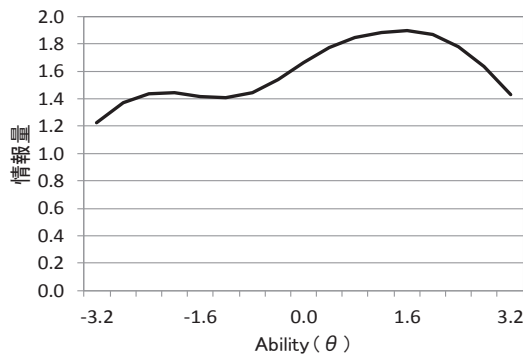


図4 テスト情報関数(TIF)

次に、推定された能力値 θ のヒストグラムと要約統計量を図 5 に示す。なお、ヒストグラムの階級数はスタージスの式により 8 とした。受験者の分布に標準正規分布を仮定して推定を行うため、推定された能力値も同様の分布となることが一般に望ましいとされる。図 5 より、平均値は 0 に、標準偏差は 1 に近い値であるが、歪度と尖度が正の値であり、左に偏り尖った分布となっていることがわかる。歪度と尖度の検定であるジャック・ベラ検定は 5%水準で有意 ($\chi^2(2) = 6.139, p > .046$) であり、また正規性の検定 (コルモゴロフ・スミルノフ検定) も 5%水準で有意 (KS 統計量 = 0.13, リリフォースの補正 $p > .05$) であるため、標準正規分布から有意に乖離しており、解釈には慎重になる必要がある。しかし受験者が初年次生であったことを考慮すると、このように θ が 0 よりも左に偏った分布になることはある程度自然なことといえよう。

また、図 6 にテスト特性曲線を示す。テスト特性曲線は特定の能力値 θ の受験者のテストの素点の期待値を表す。すなわち、特定の能力値 θ の受験者における、全観点の得点の単純な合計点の平均的な値といえる。 $\theta = 0$ 付近の受験者を見ると、期待得点は約 7 点であり、ループリックの各観点が平均的にレベル 1 程度であることがわかる。また、テスト特性曲線の急になったり緩慢になったりすることがあまり見られず、全ての領域で識別力が同程度であることがわ

かる。

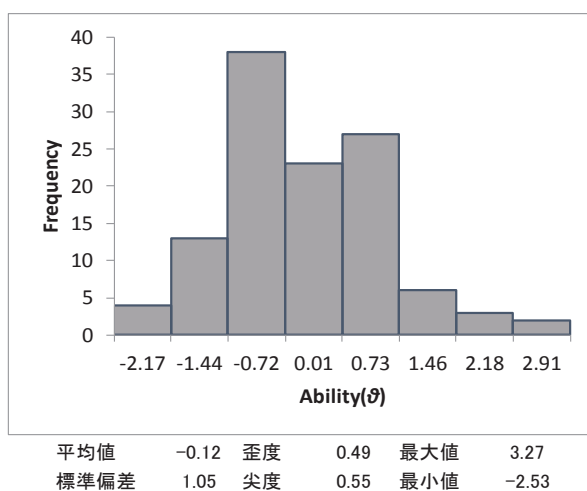


図5 能力値 θ のヒストグラムと要約統計量

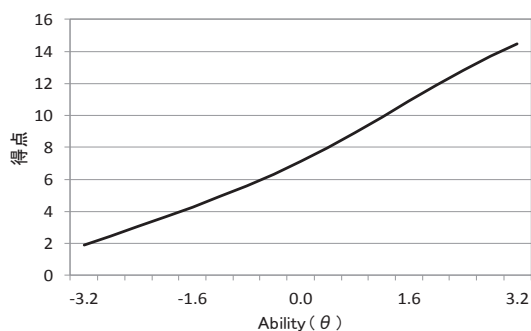
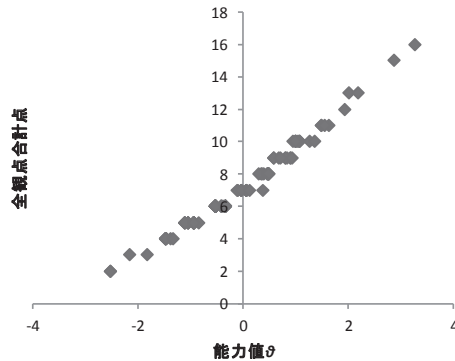


図6 テスト特性曲線(TCC)

最後に、項目反応理論を用いて推定した場合の学生の能力値 θ と、全観点の得点の単純な合計点との散布図を図7に示す。散布図からわかるように、能力値 θ と全観点合計点はほぼ完全な線形関係となっており、2つの変数の相関係数が $r = .99$ であることもそれを裏付けている。よって、能力値 θ と全観点合計点に実質的な差はほぼないといえるだろう。

図7 能力値 θ と全観点合計点の散布図

4. 結論

本研究では、新潟大学歯学部「大学学習法」のアカデミック・ライティング力を評価するためのパフォーマンス課題とルーブリックから得られたデータを項目反応理論によって分析し、ルーブリックの各観点や評価全体の測定論上の特徴を明らかにすることに加えて、項目反応理論を用いて推定した場合の学生の能力値と、全観点の得点の単純な合計点との関連を見ることで、実用上の示唆を得ることを目的とした。

まず、ルーブリックの全観点の一次元性を確認した後、Samejima の段階反応モデルによって項目母数や学生の能力値の推定を行なった。どのくらいの能力値をもった学生の評価に有効に機能するのは観点ごとに異なった傾向が確認された。学生には課題作成前にルーブリックが提示されているため、学生はルーブリックを参照しながらレポートを作成することができた。それでもなお「対立意見の検討」のように高いアカデミック・ライティング力の学生にこそ機能すると考えられる観点は、今回のような初年次生ではなく3年次生、4年次生の能力を識別することに適しているのかもしれない。また、水準がうまく段階的になっていない可能性も見出され、特定のレベルの記述語を変更する必要があるかもしれない。課題の中に「対立意見を取り上げ検討しなさい」などの指示を直接的に加え、本研究と同様の分析をして項目母数の変化をみることで、このような観点がどの能力段階の学生に機能するかに関して、より詳細な検討が可能となるだろう。このように項目母数を推定することによって特徴を観点ごとに検討したり、どのような能力段階の学生に当該評価がうまく機能するのかといったことを検討することで、評価の改善の方向性を検討することができる。このようなことを目的にする場合には、項目反応理論による検討は有用だと考えられる。

ただし、項目反応理論により推定した場合の学生の能力値 θ と、全観点の単純な合計点との関連を確認したところ、能力値 θ と全観点合計点に実質的な差はほぼなかった。本研究のように項目反応理論により各観点の特徴を検討したり、他にも経年調査で使えるように課題を等化するための情報を得たりするといった目的でなければ、大きなサンプルサイズや複雑な計算が必要な項目反応理論ではなく、古典的テスト理論の枠組みによる検討で特段問題がないと考えられる。

本研究においても、項目反応理論による分析をする際に一般的に必要とされるような十分なサンプルサイズを満たしているわけではない。また、あくまで一つのパフォーマンス評価の事例であり、サンプルも限られたものであるため、一般化には慎重になる必要があるだろう。サンプルと事例を増やし、推定値の再現性が得られるかどうかを検討することを今後の課題としたい。

5. 参考文献

- 中央教育審議会 (2014). 新しい時代にふさわしい高大接続の実現に向けた高等学校教育, 大学教育, 大学入学者選抜の一体的改革について (答申)
- 堀啓造 (2005). 因子分析における因子数決定法-平行分析を中心にして. *香川大学経済論叢 / 香川大学経済研究所編*, 77(4), 35-70.
- 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.
- 松下佳代 (2012). パフォーマンス評価による学習の質の評価—学習評価の構図の分析にもとづいて—. *京都大学高等教育研究*, 18, 75-114.
- 森朋子・山田剛史 (2009). 初年次教育における協調学習が及ぼす効果とそのプロセス: 学生同士の〈足場づくり〉を中心に. *京都大学高等教育研究*, 15, 37-46.
- Pike, G. R. (1996). Limitations of using students' self-reports of academic development as proxies for traditional achievement measures. *Research in Higher Education*, 37(1), 89-114.
- Pike, G. R. (2011). Using college students' self-reported learning outcomes in scholarly research. *New Directions for Institutional Research*, 150, 41-58.
- Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 17.
- 柴山直・佐藤喜一・熊谷龍一 (2012). 全国規模の学力調査における重複テスト分冊法の展開可能性について (平成 23 年度文部科学省委託研究「学力調査を活用した専門的課題分析に関する調査研究」研究成果報告書) 国立大学法人東北大学.
- Shojima, K. (2007). Neural test theory. *DNC Research Note*, 07-15.
- 松下佳代・小野和宏・高橋雄介 (2013). レポート評価におけるルーブリックの開発とその信頼性の検討. *大学教育学会誌*, 35(67), 107-115.
- 豊田秀樹 (2002). 項目反応理論入門編 朝倉書店..
- 宇佐美慧 (2010). 採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル—mcmc アルゴリズムに基づく推定. *教育心理学研究*, 58(2), 163-175.
- 宇都雅輝・植野真臣 (2015). ピアアセスメントの低次評価者母数をもつ項目反応理論. *電子情報通信学会論文誌*, J98-D (1), 3-16.
- 山田礼子 (2012). 学士課程教育の質保証へむけて: 学生調査と初年次教育からみえてきたもの 東信堂.
- 山田剛史・森朋子 (2010). 学生の視点から捉えた汎用的技能獲得における正課・正課外の役割. *日本教育工学会論文誌*, 34(1), 13-21.
- 渡部洋・平井洋子 (1994). 段階反応モデルによる小論文データの解析. *東京大学教育学部紀要*, 33, 143-150.

謝辞

本研究を進めるにあたり、貴重な授業実践のデータを提供してくださった新潟大学歯学部の小野和宏教授に感謝致します。また、授業実践にかかわられた他の先生方、学生の皆様にも感謝致します。

(日本学術振興会特別研究員 高等教育開発論講座 博士後期課程3回生)

(受稿 2015 年 9 月 1 日、改稿 2015 年 11 月 4 日、受理 2015 年 12 月 24 日)

パフォーマンス評価における項目反応理論を利用したアカデミック・ライティング力の測定

斎藤 有吾

昨今、大学教育において問題解決能力や論理的思考力といった汎用的で高次の能力を育成することが求められている。またそのような能力を評価する具体的な手法としてパフォーマンス評価やルーブリックに注目が集まっている。そのような評価の測定精度に関する研究はまだ発展途上であり、検討を積み重ねていく必要がある。本研究ではそのような評価を行った授業実践のデータを項目反応理論によって分析し、(a) ルーブリックの各観点や評価全体の測定論上の特徴を明らかにすること、また、(b) 項目反応理論を用いて推定した場合の学生の能力値と、全観点の得点の単純な合計点との関連を見ることで、実用上の示唆を得ることを目的とした。その結果、本研究の評価はどの学生にでも同じような能力の推定精度を持っていた。また、項目反応理論による能力の推定値と単純な合計点とには 0.9 以上の相関が見られ、他の指標との相関も同程度であることが明らかとなり、実用上は合計点を使用して問題がないことが示唆された。

Measurement of Academic Writing Abilities Using Item Response Theory Analysis Based on Performance Assessment

SAITO Yugo

A key responsibility of universities is to foster of high-level, general-purpose cognitive abilities, including problem-solving abilities and logical thinking skills. The focus is now on performance assessments and rubrics. Research on the accuracy of these types of assessments is still in the developmental stages, and accumulation of data is needed. This study used item response theory (IRT) to analyze data on such assessments performed in test classrooms. The purposes were to (a) elucidate rubrics from a variety of perspectives as well as theoretical aspects of measurement for overall assessments, and to (b) observe relationships between students' skill scores as predicted by IRT and the actual total scores as added together for all perspectives, to obtain practical propositions. The results showed similar skill prediction accuracies for all students assessed in our study. A correlation greater than 0.9 was found between skill scores as predicted by IRT and the actual score totals, suggesting that the use of total scores for practical purposes is appropriate.

キーワード： パフォーマンス評価, ルーブリック, アカデミック・ライティング, 項目反応理論

Keywords: Performance Assessment, Rubric, Academic Writing, Item Response Theory