# 識別力を重視したライティングルーブリック開発の試み
## ―分散分析を用いた特異項目機能分析―

石　川　勝　彦　　児　島　功　和

**要約**　小論文評価は正解が一意に決まる多肢選択型の試験と異なり、評価者が評価基準を参照しながら採点を行うため最終的には評価者が主観的に行わざるを得ない。したがって評価基準を構成する観点や項目はできるだけ測定論的に優れたものでなければならない。ではどのような評価項目が良い項目なのだろうか。本稿では 2 つのルーブリックを構築し、ルーブリックに基づいて定量評価を行った結果を分析し、評価項目の識別力を評価した。

研究 1 では、授業で練習すれば無理なくマイルストーン（最高点）に到達できるレベルに設定されたルーブリックの識別力を検討した。具体的には授業の中で行われたコンテストにおいて優秀賞を受賞した論文と受賞しなかった小論文を識別できるかどうか、という視点から行った。分析の結果、受賞論文のみならず非受賞論文も比較的容易にマイルストーンに到達し予想通り天井効果が生じることが示された。

研究 2 では受賞論文と非受賞論文を識別しうる項目を新たに追加し、評価者を 2 名に増やしたうえで、識別力がどのように変動するか特異項目機能分析（DIF）の枠組みから確認した。分析の結果、多くの項目は良好な識別力を示したが、「文の簡潔さ」「独自な視点・発想」の 2 つの観点は受賞 / 非受賞を識別する力が不十分であること、「問いの魅力」は不均一 DIF の観点から識別性に問題が生じることが分かった。

## はじめに

### 1. 研究の目的

本稿の課題は、小論文執筆を主な教育目標とする初年次ゼミ（基礎演習Ⅰ）で用いたルーブリックの有効性に関する検討を行なうこと、加えて、識別精度の高い項目のリストを作成することである。

基礎演習Ⅰは山梨学院大学の 1 年生の多くが履修するゼミ形式（一クラスで 20 名前後）の授業であり、入学したばかりの学生が大学生活に慣れること、そして小論文（1200 字程度）を書きあげることを目標として設定している。履修者は共通テキストを使用しながら小論文の書き方を学んでいく。授業は履修学生全員が小論文を書きあげて終了となる。その後、各クラスで最もよい出来の小論文を選出し、更にそこから学科の代表作といえる小論文を選出し表彰する「小論文コンテスト」が開催される。

2016 年度の基礎演習Ⅰでは筆者の一人である児島が作成したライティング・ルーブリックを指導に運用した。12 項目からなりゼミ教員と学生が執筆方針を確認し、小論文の出来を評価できるように作られた。このルーブリックは授業で教わる「書き方」を一覧にした科目ルーブリックとして設計された。

特にレベル設定に特徴があり、マイルストーン（最上位の基準）は授業で教わるなかで無理なく到達できる水準に設定された。ベンチマー

ク（最下位の基準）はいわゆる「不可」のレベルに設定され平均的な水準をクリアするよう配慮された。

こうした設計意図ゆえに、コンテストにかけられる優れた小論文と平均的な小論文をどれくらい識別するかを確認すること、識別できない場合は新たな項目プールを開発する必要がある。

研究 1 では 2016 年に運用されたルーブリックが優れた小論文（コンテストの受賞論文）と平均的な小論文（非受賞論文）をどのように識別するかを検討した。研究 2 では、優れた小論文と平均的な小論文をより敏感に識別する項目の探索を行った。

## 2. ルーブリックの評価方法

ではルーブリック項目の良し悪しはどのように検討すれば良いのだろうか。

ルーブリック評価は多肢選択式のテストのように回答が一つに定まるものではなく、評価者の主観によって評価結果が異なる。したがって、項目を作るだけでは不十分で、実際に採点に使って評価データを生成し、データからルーブリックの信頼性･妥当性を確認する必要がある（宇佐美，2012；Pink, 1996/2011）。具体的には項目そのものの信頼性を向上させるアプローチ（木村，2004；秋山，2011；梶井，2001；平，1995；松下・小野・高橋，2013；宇佐美，2011）、そして測定やスコアリングの手続きを工夫するアプローチ（庄司・野口・金澤・青山・伊東・迫田・春原・廣利・和田，2004；松下・小野・高橋，2013；斎藤，2016）の 2 つがある。本論は前者の、項目の信頼性を統計的に評価するアプローチをとる。

項目の信頼性は識別力の観点から考えることができる（Lord, 1980）。識別力の低い項目は、受験者能力が大きく異なってもスコアがあまり変化しない項目であり、識別力の高い項目は受験者能力が変化するに応じて敏感にスコアが変化する項目である。識別力が高い項目はその項目の得点がテストで測定している能力をよりよく表し、受験者の能力をよく区別しているとみなされる（加藤・山田・川端，2015）。

本論では、特に特異項目機能分析（Differential Item Functioning：以下 DIF）の手法を応用する。DIF は「テストが測定しようとしている特性・能力が等しいにも関わらず、所属する下位集団によって正答率が異なる状態」と定義される（渡辺・野口，1999）。つまり同じテストを同じような能力分布を持った 2 つの集団で測定したときに評価結果に変動がないことは、良いテストの条件であると言える。本研究では 2 名の評価者が同一の受験者集団を同じルーブリックで評価するものであり、DIF 検出法が厳密に適用できる。

本論では均一 DIF（uniform DIF）と不均一 DIF（nonuniform DIF）の両方に関心がある。複数の評価者が同一の小論文を同一の項目で評価して、評価の厳しさに差が生じた項目は均一 DIF があると考える。2 名の評価者の平均値に差がある、という状況に近い。一方、不均一 DIF は受験者特性の高低と二人の評価者の評価が交互作用を示すことを意味する（Chan, 2000）。適切なルーブリックであれば、誰がルーブリック採点を担当しようとも、優秀な小論文は高得点、平均的な小論文は平均的な得点が与えられなければならない。不均一 DIF が生じている状況とはこうした整合的な採点状況が崩れてしまっている状況である。同じルーブリックで同一の受験者集団を評価しているにもかかわらず、評価者 A が高得点をつけた小論文に評価者 B が低得点をつけている状況が不均一 DIF である（野口・熊谷・脇田・和田，2007）。本論では、研究 2 において受賞論文／非受賞論文という区別を軸に、不均一 DIF の枠組みで分析を行う。様々な DIF 検出方法が

あるが（Holland & Thayer, 1988：田崎，2007；熊谷，2003；Su & Wang, 2005）、受賞する／受賞しないという 2 値の変数が含まれるデータセットなのでこれに適合する分散分析を用いた方法（Cleary & Hilton, 1968）を利用する。

## 研究 1

### 1. 目的

2016 年度に運用されたルーブリックの困難度および識別力を検討することを目的とする。当該ルーブリックは授業の最小限の到達目標を表示すること、および授業テキストと対応が取れていることを狙いとして作成されており、様々なレベルにある小論文を広く識別することを目的に開発されたものではない。そのため、測定値は天井効果を生じること、そして相対的に受験者能力の低い小論文に対し高い識別精度を発揮すると予想される。そうした結果が得られた場合、授業を通じてルーブリック上の到達目標が多くの小論文において実現している、と解釈することが可能である。

### 2. 方法

#### ルーブリック

2016 年度のルーブリックは、授業および授業テキストと連動することを最優先に開発された。特にレベル設定については、いずれの観点も授業を通じて課題をこなしていけば、マイルストーンに無理なく到達できるように設定にした。ベンチマークは、「不可」である記述を載せることで学生に到達目標を明示するよう設計した。

以上のことから、現行のルーブリックを用いて採点を行った場合、天井効果が生じる可能性がある。このことは項目記述が不適切であることを意味するのではなく、ルーブリックの設計意図が実現している、採点対象の小論文が授業の到達目標を実現していると解釈することができる。

研究 1 で用いられたルーブリックを Table.1 に示した。なお、Table.1 からは削除したが、授業で運用されたルーブリックには「小論文タイトル、自分の名前、学部・学科、学籍番号が書かれている」「本文字数が規定字数（1200 字）のおよそ 1 割前後（1080～1320 字）」の 2 項目が掲載されていた。前者は、個人情報保護の観点から除外した。字数は大学に提出する小論文の形式要件を満たすうえで重要であり、小論文の質に一定の影響を与えることが示唆されている（宇佐美，2011）が、指導教官がチェックした時点でほぼ満たされると予想されるためルーブリックから除外した。

#### 評価対象・評価者・評価方法

2016 年基礎演習 I で執筆された小論文 78 編を評価対象とした。うち 12 編は学内で開催された小論文コンテストで最優秀賞もしくは優秀賞を受賞したものである。

ルーブリック評価は、小論文の評価手法の研究経験のある大学教員 1 名に依頼した。原本から学籍番号・氏名・所属学科を削除したうえで通し番号を振りなおしたファイルを作成し、データをメールで送信した。評価者には受賞論文と非受賞論文の区別はマスクされた。評価の依頼文には「①各論文について、10 の観点すべてについて評価してください、②いずれの観点も『1～3』の 3 件法で評価してください、③評価結果は評価表シートに数字を打ち込む形で進めてください」と記した。評価後は小論文のファイルを破棄するよう依頼した。

#### 分析

まず天井効果の有無を確認した。次に項目の

**Table.1　形式的・構造的側面に焦点化した小論文採点ルーブリック**

| 評価観点／評価レベル | 観点 | 良（3点） | 可（2点） | 不可（1点） |
|---|---|---|---|---|
| 日本語表現・ルール | 誤字脱字 | 誤字・脱字がない | 誤字・脱字が3個以下である | 誤字・脱字が4個以上ある |
| | 文の簡潔さ | 一文が簡潔にまとめられ、文章もわかりやすい | 一文が長く、一部の文章にもわかりにくい個所がある | 一文が長く、文章全体もわかりにくい |
| | 段落分け | 全体的に段落わけが適切になされている | 適切な段落わけが一部だけである | 全体として段落わけが十分になされていない |
| | 言い回し | 正しい日本語の言い回しができている | 日本語の言い回しに関するおかしな点が3個以下である | 日本語の言い回しに関するおかしな点が4個以上ある |
| 体裁 | 注の適切さ | 全体として注が適切につけられており、図書・雑誌・インターネット記事等の資料情報についても適切な形式で書かれている | 注が一部不適切につけられている、または図書・雑誌・インターネット記事等の資料情報について一部不適切な形式で書かれている | 全体として注のつけかたが不適切もしくは十分につけられておらず、または図書・雑誌・インターネット記事等の資料情報についても全体的に不適切な形式で書かれている |
| 構成 | 問い・主張・理由 | 問い・主張・理由がきちんと書かれている | 問い・主張・理由が書かれているものの、わかりづらい | 問い・主張・理由が書かれていない、あるいは非常にわかりづらく書いてある |
| | 序論・本論・結論 | 序論・本論・結論が適切に書かれている | 序論・本論・結論という形式では一部書かれていない | 序論・本論・結論という形式でほとんど書かれていない |
| 説得性 | 証拠の信頼性 | 理由をサポートする証拠（例：数値で示されるデータ等）が信頼できるものである | 理由をサポートする証拠（例：数値で示されるデータ等）の一部が若干信頼できるか疑わしい | 理由をサポートする証拠（例：数値で示されるデータ等）が信頼できるか非常に疑わしい |
| | 証拠の十分さ | 理由をサポートする証拠について十分に調べてある | 理由をサポートする証拠について少ししか調べていない | 理由をサポートする証拠についてほとんど調べていない |
| | 問いと主張の対応 | 問いと主張と理由がしっかりつながっている | 問いと主張と理由のつながりが弱い | 問いと主張と理由がきちんとつながっていない |

識別力を確認するため、コンテスト受賞論文（12編）と非受賞論文（66編）を識別するかどうか、項目ごとに *t* 検定を実施した。最後にルーブリックの測定精度の特性を把握するためテスト情報曲線の算出を行った。これによりどのようなレベルにある小論文を高い精度で評価できるの

か、どのレベルに合わせたルーブリックとなっているのか明らかにする。

## 3. 結果と考察

### 受賞論文と非受賞論文の識別力

天井効果の有無を確認する。平均値に 1SD を加えた値が項目上限値の 3 を超えるかどうか検討した（天井効果の指標）ところ、受賞ありではすべての項目、受賞なしでは「言い回し」「問い・主張・理由」の 2 項目を除く 8 項目で 3 を超えていた。これらの項目で天井効果が確認された。

受賞の有無別に平均値、標準偏差、受賞の有無を独立変数とする *t* 検定の結果を整理した（Table.2）。「文の簡潔さ」はすべての採点対象に 3 がつけられ分散が生じていないため検定は行わなかった。主効果が有意だったのは「言い回し」（$t$=3.69, $d$=1.15）、「注の適切さ」（$t$=2.11, $d$=.66）、「問い・主張・理由」（$t$=3.86, $d$=1.20）の 3 項目だった（$ps$<.000）。その他の項目では有意性が検出されなかったため、受賞論文と非受賞論文の間で差が見られない観点であると解釈できる。

このことから、このルーブリックが開発目的（最低ラインをマイルストーンに設定する）を充足していること、併せて多くの小論文が到達目標に到達していることが示唆され、授業運営が学習成果に結びついていると推察できる。

### テスト情報曲線

最後に、当該ルーブリックが全体として、学生の能力の個人差をどのように識別するのか確認する。具体的にはテスト情報曲線を算出し、どのレベルの学生の個人差を敏感に検出するのか確認する。

テスト情報曲線を計算する前提として、対象となる項目群は同一の構成概念を測定している、すなわち尺度が一元性をもっていることが求められる。一元性が実現していない場合、テスト情報曲線の推定精度は著しく損なわれる。

因子数の決定について、対角 MSC と MAP 分析がともに 1 因子を提案した。因子数を 1 に指定して主成分分析を行ったところ、「文の簡

**Table.2　受賞の有無を独立変数とする *t* 検定**

| | 受賞あり（N=12） | | 受賞なし（N=66） | | *d* | *t* 値 | *p* 値 |
|---|---|---|---|---|---|---|---|
| | 平均値 | SD | 平均値 | SD | | | |
| 誤字脱字 | 3.00 | 0.00 | 2.71 | 0.52 | .59 | 1.91 | .06 |
| 文の簡潔さ | 3.00 | 0.00 | 3.00 | 0.00 | – | – | – |
| 段落分け | 2.92 | 0.29 | 2.86 | 0.39 | .14 | 0.45 | .65 |
| 言い回し | 2.92 | 0.29 | 2.11 | 0.75 | 1.15 | 3.69 | .00 |
| 注の適切さ | 2.83 | 0.39 | 2.35 | 0.77 | .66 | 2.11 | .04 |
| 問い・主張・理由 | 3.00 | 0.00 | 1.88 | 1.00 | 1.20 | 3.86 | .00 |
| 序論・本論・結論 | 3.00 | 0.00 | 2.79 | 0.41 | .55 | 1.77 | .08 |
| 証拠の信頼性 | 3.00 | 0.00 | 2.68 | 0.59 | .58 | 1.87 | .07 |
| 証拠の十分さ | 3.00 | 0.00 | 2.97 | 0.17 | .19 | 0.60 | .55 |
| 問いと主張の対応 | 2.50 | 0.52 | 1.58 | 0.80 | 1.19 | 3.82 | .16 |

Note.「文の簡潔さ」はすべての採点対象に「良」がつけられたため検定は行わない

潔さ」「注の適切さ」「序論・本論・結論」の主成分負荷量が .40 を下回った。当該 3 項目を除外して再度主成分分析を行った結果を Table.3 に整理した。主成分負荷量は十分な値を示し、$\alpha$ =.71、$\omega$ =.81 と信頼性係数も良好な値を示した。この 7 項目の一元性が確認できたため、7 項目を用いてテスト情報曲線の算出に進む。

Figure.1 にテスト情報曲線を示した。受験者能力 $\theta = -1.3$ あたり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。また $\theta$ によって情報量が大幅に変動することもみてとれる。情報量が高いエリアは受験者能力 $\theta$ が −3.0〜0.0 のエリアであり、その両側との間に情報力の差が生じていることが見て取れる。当該ルーブリックは主に小論文が相対的に不出来な受験者群に対して高い識別精度を発揮していることが確認できた。

このことは、ルーブリックのレベル設定の狙いが実現していたこと、授業運営が学生をマイルストーンまで引き上げることに成功したことを裏付けるものである。

**Table.3　ルーブリックの因子パターン**

| 項目 | F1 | 共通性 |
|---|---|---|
| 問いと主張の対応 | .74 | .54 |
| 誤字脱字 | .68 | .46 |
| 問い・主張・理由 | .68 | .46 |
| 段落分け | .59 | .35 |
| 言い回し | .57 | .33 |
| 証拠の信頼性 | .55 | .30 |
| 証拠の十分さ | .54 | .29 |
| 因子寄与 | 2.731 | |

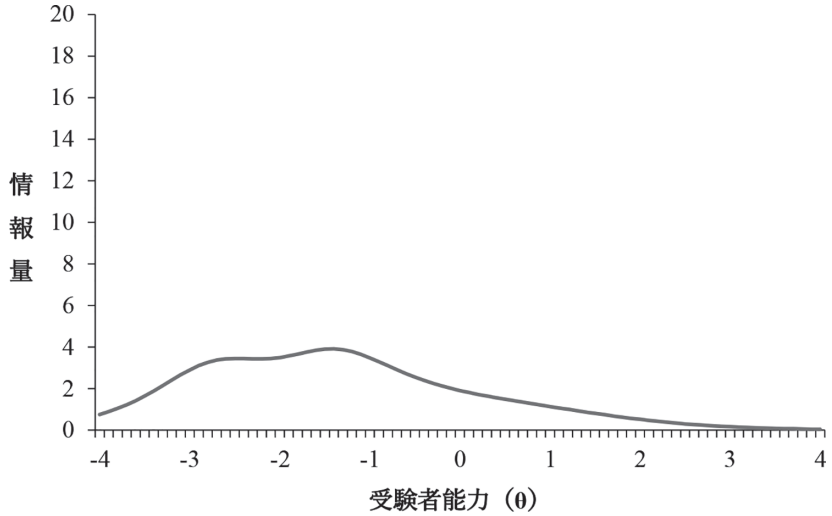
## 研究 2

### 1. 目的

研究 2 は、優れた小論文（受賞論文）と平均的な小論文（非受賞論文）を識別できる項目プールを得ることを目的とする。具体的には受賞論文と非受賞論文を識別する項目を探索することとする。

均一 DIF として、各項目に 2 人の評価者の間で評価の厳しさに違いがあるか検討する。次に不均一 DIF として受賞論文と非受賞論文を

**Figure.1　ルーブリック（7 項目）のテスト情報曲線**



識別する項目が、2 名の評価者の間で同じかどうかを検討する。もし 2 名ともある項目によって受賞論文と非受賞論文を識別していれば、当該項目の識別力には問題がない。一方、評価者 A はある項目によって受賞論文と非受賞論文を識別しているが、評価者 B は識別していないような項目が見つかった場合、その項目は評価者が異なれば異なった働きをしていることになる。評価者が誰であるかによって、おなじ小論文を評価しているのに得点が高かったり低かったりするような項目は識別力に問題を抱えているといえるだろう。

## 2. 方法

### 評価項目の設定

研究 2 で用いたルーブリック項目を Table.4 に整理した。「問い・主張・理由」「序論・本論・結論」「段落分け」「文の簡潔さ」の 4 項目は研究 1 で用いられたルーブリックから引き継いだ。「問いの焦点化」「問いの魅力」「発見的な面白さ」「論証の説得力」「文章の躍動感」「独自な視点・発想」の 6 項目を新たに設定した。新たな項目の設定に当たっては、採点対象となる小論文を読み込んだうえで、国内外のライティング・ルーブリックを参考（宇佐美, 2011；ReadWriteThink, 2004；turnitin, 2012；West virginia department of education, 2008；Woman in defense, 2015）にし、協議によって決定した（Table.4）。

### 評価対象・評価者・評価方法

評価対象は研究 1 と同様の小論文である。2016 年基礎演習 I で執筆された受賞論文 12 編、非受賞論文 66 編を評価対象とした。評価者は 2 名とした。評価者 A は心理学を専攻する博士課程の大学院生、評価者 B は心理学を専攻する大学教員である。評価方法について、小論文のデータファイルの作成方法や評価者とのデータのやり取りに関する手続きは研究 1 と同様であった。依頼内容に 1 点違いがあり、3 件法ではなく 5 件法での評価を依頼した（不均一 DIF の検出精度を高めることが重要であることからこのような処置を行った）。連動して依頼文にも一定の変更が生じた。評価の依頼文には「①各論文について、10 の観点すべてについて評価してください、②いずれの観点も『1〜5』の 5 件法で評価してください、③評価結果は評価表シートに数字を打ち込む形で進めてください」と記した。評価後は小論文のファイルを破棄するよう依頼した。

### 分析

2 名の評価者による複数評価データであるので、まず評価者間一致率を検討し、一致率の低い項目が存在しないかどうか確認する。次に、均一 DIF および不均一 DIF を検討する。統計処理として、評価者（2）×受賞の有無（2）を独立変数、評価スコアを従属変数とする 2 要因分散分析を行う。もし評価者の要因が有意なら均一 DIF を生じていると解釈できる。交互作用が有意になれば、その項目は不均一 DIF を生じていると考えられる。最後にテスト情報曲線を算出し、ルーブリック全体がもっている識別精度の特性を把握する。

## 3. 結果と考察

### 評価者間信頼性

まず 2 名の評価の一致率を検討した。Kendall の $W$ 係数を算出したところ、おおよその項目で .60 の周辺に値がばらついた（Table.5）。全体で .52〜 .63 の範囲に分布しており、著しく一致率が低い項目はないといえるだろう。相対的に値が低い項目は「独自な視点・発想」（$W$=.52）であった。他の項目に比べ評価者間

**Table.4　研究 2 で検討するルーブリック**

| 評価観点／評価レベル | 観点 | 良（各 5 点） | 可（各 3 点） | 不可（各 1 点） |
|---|---|---|---|---|
| 構成 | 問い・主張・理由 | 問い・主張・理由がきちんと書かれている | 問い・主張・理由が書かれているものの、わかりづらい | 問い・主張・理由が書かれていない、あるいは非常にわかりづらく書いてある |
| | 序論・本論・結論 | 序論・本論・結論が適切に書かれている | 序論・本論・結論という形式では一部書かれていない | 序論・本論・結論という形式でほとんど書かれていない |
| | 段落分け | 全体的に段落わけが適切になされている | 適切な段落わけが一部だけである | 全体として段落わけが十分になされていない |
| 問い | 問いの焦点化 | 問いが十分に焦点化されている | 問いがある程度焦点化されているが，ややあいまい | 問いがあいまいで焦点化できていない |
| | 問いの魅力 | 問いが「答えを知りたい」と思わせるものである | 問いがある程度「答えを知りたい」と思わせる | 問いが「答えを知りたい」と思わない |
| 論証 | 発見的な面白さ | 論証の内容に発見的な面白さがある | 論証の内容は堅実だが面白味みかける | 論証の内容が平板でつまらない |
| | 論証の説得力 | 論証には「なるほど」と思わせる説得力がある | 論証は多少説得力がある | 論証はあまり説得力がない |
| 文章表現 | 文章の躍動感 | 新鮮で躍動感のある文章で書かれている | ある程度新鮮で躍動感を感じさせる文章で書かれている | メリハリのない文章で書かれている |
| | 文の簡潔さ | 一文が簡潔にまとめられ、文章もわかりやすい | 一文が長く、一部の文章にもわかりにくい個所がある | 一文が長く、文章全体もわかりにくい |
| オリジナリティ | 独自な視点・発想 | 書き手の独自な視点・発想が盛り込まれている | 書き手の独自な視点・発想が，ある程度みられる | 書き手の独自な視点・発想がなく，新しみがない |

の一致率が低い項目であるが、項目そのものを削除すべき水準ではないと判断し後の解析にも含めることとした。

### 分散分析による不均一 DIF の検討

Table.6 に平均値を整理した。まず天井効果の有無を検討した。平均値 + 1SD が上限値の 5 を超えるか検討したところ、評価者 A は「段落分け」、評価者 B は「問い・主張・理由」「序論･本論･結論」「段落分け」で一部 5 を超えた。研究 1 から引き継いだ 4 項目のうち 3 項目に再び天井効果が表れたことから、これらの項目は

マイルストーンに容易に到達できる項目であると言える。新たに追加した 6 項目に天井効果は見られなかった。

受賞の有無（2）×評定者（2）を独立変数、評価の粗点を従属変数とする 2 要因分散分析を項目ごとに行った。Table.7 に 2 要因分散分析の検定統計量を整理した。

分散分析の検定結果を確認する。まず受賞の有無に注目する。主効果は「文章の簡潔さ」（$F$=2.54, $\eta_p^2$=.02, *n.s.*）「独自な視点・発想」（$F$=1.50, $\eta_p^2$=.01, *n.s.*）を除く 8 つの項目で有意だった。多くの項目が受賞論文と非受賞論文を識別しうることがわかった。

次に均一 DIF として評価者の主効果に注目する。検定結果をみると「文の簡潔さ」の主効果が有意となり（$F$=35.29, $\eta_p^2$=.19, $p$<.01）、残る 9 つの項目では有意とならなかった。このことから 9 つの項目は 2 名の評価者の採点の傾向が一致していると解釈できる。「文の簡潔さ」については平均値に差が見られており 2 名の評価者の評価の厳しさにズレが生じている項目である。

最後に不均一 DIF を見るため交互作用に注目する。「問いの魅力」（$F$=6.76, $\eta_p^2$=.04, $p$<.05）で受賞の有無×評価者の交互作用が有意となり、残りの 9 つの項目では有意にならなかった。単純主効果を確認すると、評価者 A では受賞の有無によって平均値に差はみられない（$t$（152）=0.48, $d$=.20, *n.s.*）が、評価者 B では受賞論文のほうが非受賞論文よりも平均値が高か

**Table.5　評価者間一致率（Kendall の *W*）**

| | Kendall の *W* |
|---|---|
| 問い・主張・理由 | .60 |
| 序論・本論・結論 | .61 |
| 段落分け | .63 |
| 問いの焦点化 | .60 |
| 問いの魅力 | .63 |
| 発見的な面白さ | .61 |
| 論証の説得力 | .61 |
| 文章の躍動感 | .65 |
| 文の簡潔さ | .57 |
| 独自な視点・発想 | .52 |

**Table.6　項目別・評価者別・受賞の有無別の平均値**

| | 評価者 A | | | | 評価者 B | | | |
|---|---|---|---|---|---|---|---|---|
| | 受賞論文 | | 非受賞論 | | 受賞論文 | | 非受賞論 | |
| | 平均値 | SD | 平均値 | SD | 平均値 | SD | 平均値 | SD |
| 問い・主張・理由 | 4.17 | 0.58 | 3.32 | 0.73 | 4.17 | 0.83 | 3.61 | 1.01 |
| 序論・本論・結論 | 4.17 | 0.58 | 3.64 | 0.72 | 4.50 | 0.52 | 3.91 | 0.89 |
| 段落分け | 4.67 | 0.65 | 4.38 | 1.00 | 4.58 | 0.67 | 3.88 | 0.97 |
| 問いの焦点化 | 3.75 | 0.87 | 3.30 | 0.76 | 3.92 | 0.67 | 3.20 | 0.85 |
| 問いの魅力 | 3.17 | 0.39 | 3.06 | 0.49 | 3.58 | 0.90 | 2.67 | 0.87 |
| 発見的な面白さ | 3.25 | 0.45 | 3.03 | 0.58 | 3.42 | 0.79 | 2.71 | 1.03 |
| 論証の説得力 | 3.67 | 0.78 | 3.02 | 0.79 | 3.33 | 1.30 | 2.64 | 0.99 |
| 文章の躍動感 | 3.08 | 0.29 | 2.85 | 0.59 | 3.42 | 0.51 | 2.86 | 0.86 |
| 文の簡潔さ | 3.25 | 0.45 | 3.09 | 0.72 | 4.42 | 0.51 | 4.02 | 0.94 |
| 独自な視点・発想 | 3.17 | 0.39 | 3.09 | 0.63 | 3.50 | 0.80 | 3.15 | 0.95 |

**Table.7　受賞の有無（2）×評定者（2）を独立変数とする 2 要因分散分析（項目別）**

| | 問い・主張・理由 | | 序論・本論・結論 | | 段落分け | | 問いの焦点化 | | 問いの魅力 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\eta_p^2$ | $F$ | $\eta_p^2$ | $F$ | $\eta_p^2$ | $F$ | $\eta_p^2$ | $F$ | $\eta_p^2$ | $F$ |
| 受賞の有無 | .08 | 13.75** | .07 | 10.61** | .04 | 5.58* | .07 | 10.77** | .07 | 10.76** |
| 評価者 | .00 | .57 | .02 | 3.10+ | .01 | 1.93 | .00 | .03 | .00 | .01 |
| 受賞の有無×評価者 | .00 | .57 | .00 | .03 | .01 | .98 | .00 | .59 | .04 | 6.76* |

| | 発見的な面白さ | | 論証の説得力 | | 文章の躍動感 | | 文の簡潔さ | | 独自な視点・発想 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\eta_p^2$ | $F$ | $\eta_p^2$ | $F$ | $\eta_p^2$ | $F$ | $\eta_p^2$ | $F$ | $\eta_p^2$ | $F$ |
| 受賞の有無 | .04 | 6.55* | .07 | 10.83** | .04 | 6.47* | .02 | 2.54 | .01 | 1.50 |
| 評価者 | .00 | .18 | .02 | 3.02+ | .01 | 1.26 | .19 | 35.29** | .01 | 1.29 |
| 受賞の有無×評価者 | .01 | 1.80 | .00 | .01 | .01 | 1.05 | .00 | .47 | .00 | .62 |

**$p$<.01, *$p$<.05, +$p$<.10

**Table.8　因子パターン（カテゴリカル因子分析）**

| 評価者 A | | | 評価者 B | | |
|---|---|---|---|---|---|
| 項目 | F1 | 共通性 | 項目 | F1 | 共通性 |
| 問いの魅力 | .89 | .79 | 論証の説得力 | .90 | .81 |
| 問い・主張・理由 | .87 | .76 | 文章の躍動感 | .89 | .80 |
| 発見的な面白さ | .86 | .75 | 問いの焦点化 | .89 | .80 |
| 序論・本論・結論 | .86 | .75 | 発見的な面白さ | .87 | .76 |
| 論証の説得力 | .81 | .66 | 序論・本論・結論 | .85 | .73 |
| 独自な視点・発想 | .79 | .63 | 問い・主張・理由 | .82 | .67 |
| 問いの焦点化 | .77 | .60 | 段落分け | .81 | .66 |
| 文の簡潔さ | .72 | .52 | 問いの魅力 | .79 | .62 |
| 文章の躍動感 | .72 | .51 | 独自な視点・発想 | .73 | .53 |
| 段落分け | .50 | .25 | 文の簡潔さ | .71 | .50 |
| 因子寄与 | 6.21 | | 因子寄与 | 6.88 | |
| 乖離度 | 2.91 | | 乖離度 | 1.15 | |
| α係数 | .88 | | α係数 | .93 | |

った（$t$(152) =4.16, $d$=1.30, $p$<.01）。「問いの魅力」は評価者 B においては受賞を左右する要因であるが、評価者 A においては受賞を左右する要因ではない。この項目は評価者が異なれば受賞／非受賞の識別に効果を持ったり持たなかったりするため、識別力に問題を抱えているといえるだろう。

### テスト情報曲線

続いて、テスト全体の情報量を評価した。尺

度の一元性を確認する。評価者 A と評価者 B において、順に対角 SMC が 3 因子と 2 因子、MAP が 2 因子と 1 因子を提案した。スクリープロットを確認したところ、固有値の落ち込みは一因子と解釈可能な形状をなしていたため、一因子を指定してカテゴリカル因子分析を実施した（Table.7）。いずれの評価者においても因子負荷量はどの項目でも .40 を超え十分な水準にあったためすべての項目を採用した。評価者 A では $\alpha$ = .88、$\omega$ = .88、評価者 B では $\alpha$ = .93、$\omega$ = .93 と十分な内的一貫性を示した。なお、以降の分析では、乖離度がより小さく因子寄与が大きい評価者 B の評価結果を用いることとした。

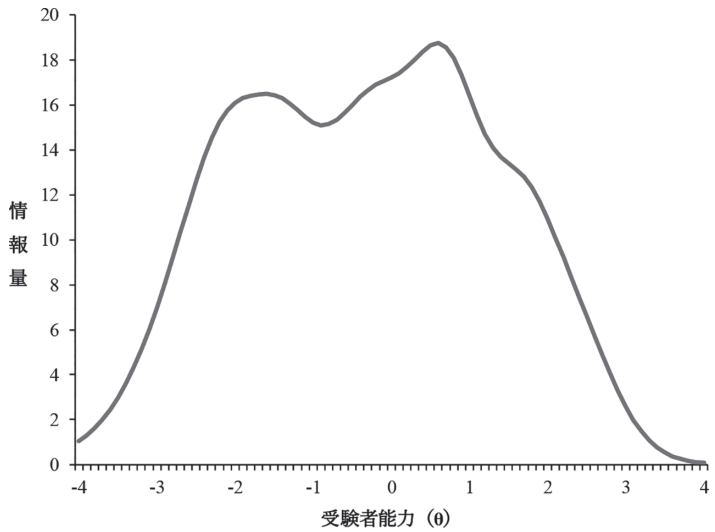
テスト情報曲線を Figure.2 に示した。$\theta$ =1.0 付近に目立ったピークがみられ、受験者能力 $\theta$ の－3.0～2.0 のエリアで高い情報量を示した。その一方、$\theta$ =－3.0 より低い受験者群や $\theta$ =2.0 より高い受験者群に対しては急激に情報量が低下している。このことから、本ルーブリックは平均的な小論文を中心に広い範囲の小論文を適切に識別することができる一方、極めて優れた小論文や不適切な小論文については識別精度が低く、うまく評価できない可能性があることがわかった。

## 総合考察

研究 1 で用いたルーブリックは、レベルを設定する際、テキストをこなし授業に参加していれば、無理なくマイルストーンに到達できるよう設計された。検討の結果、8 項目が天井効果を示し、受賞／非受賞を識別したのは 10 項目中 3 項目だった。科目ルーブリックとしての性能を十分に備えている一方、クラスの代表に選ばれる優れた小論文と平均的な小論文の差別化には新たな項目やレベル設定が必要であることも見えてきた。

研究 2 では、新たに 6 項目案出し、研究 1 から 4 項目を引き継いで計 10 項目でルーブリックを作成した。受賞／非受賞論文を識別するか、2 名の評価者の間で不均一 DIF を生じていないかどうか項目ごとに検討した。受賞／非受賞の識別力に問題があったのは「文の簡潔さ」「独自な視点・発想」の 2 項目だった。不均一 DIF を生じたのは 10 項目中 1 項目で「問いの魅力」

**Figure.2　テスト情報曲線**



だった。テスト情報曲線を見てみると、平均的な小論文を中心に広いレベルの小論文を識別しうることが示唆された（$\theta$ =−3.0〜3.0 に分布）。ところが極端に出来の良くない、あるいは出来の良い小論文に対しては急激に識別力が落ち込むことも示された。

新たに加えた 6 項目についてまとめる。いずれも天井効果を生じず、因子の一元性および内的整合性の点で問題はなかった。受賞／非受賞の識別精度の観点からは、「独自な視点・発想」が識別力の不足を示し、「問いの魅力」が不均一 DIF を生じ識別力が不安定であることが示唆された。その原因としては、リッカートスケールで回答を求めたため、各水準の定義が不明確であったことが寄与している可能性、あるいは項目文の運用において主観が入り込みやすい項目であった可能性が考えられる。この 2 つを除く項目は、評価者が変わってもおおよそ同じように運用され類似した評価結果を導いていた。識別力と再現性の高い項目群であると考えられる。

識別力の検討を終えた項目プールを持つことは、ライティング科目の授業案や科目シラバスを作成する際に補助資料として有益である。授業を作る際に到達目標を検討するわけだが、このことは評価の観点を定めることと深く関係する。例えば「オリジナリティのある小論文を書かせる」という目標が可能かどうかは、そもそもオリジナリティを定義できるか、担当教員の間でその定義に合意できるか、合意できたとして実際に信頼性のある評価ができるか、ということと切り離せない。評価できなければ到達度を把握できないからだ。

こうした評価の可能性を探るには、実際に学生が書いた小論文を評価項目を立てて評価し、項目の信頼性をチェックすることでかなりの程度検討できる。信頼性の有無を実証的に検討したうえで結果を担当教員間で共有するなどすることで、想定や思い込みに基づく議論に決着をつける一助となる可能性がある。

今後の研究上の方向として、識別力の検討方法にさらなる改善が可能である。本研究では小論文が受賞／非受賞に分類されているという情報を利用して、分散分析を用いた不均一 DIF を検討した。この方法で観察できる識別力は、少数の優秀な小論文と平均および平均以下の小論文の識別を中心としている。不出来なもの、平均的なもの、優秀なものを広く識別するかどうかは、項目反応理論を用いた検討が別途必要である。

授業で運用する際の課題として、3 点検討課題がある。第 1 にスケールの問題である。研究上の必要から 5 件法で評価を実施したが、授業でのピア評価や受賞論文審査にあって 5 件法での運用がベストとは限らない。5 件法では水準が多すぎてつけにくいとか、評価が負担だということが出てくることも考えられる。教室で運用する場合には、2 件法や 3 件法が良いかもしれない。またスケールが縮んだ場合に、本研究で示された識別力が再現されるかどうかは検討してみなければわからない。第 2 に形成的評価に耐えうるか、という問題がある。本調査では最終成果物を教員が評価するという局面でデータを元に検討を行ったため、総括的評価を行ううえでは一定程度の妥当性と信頼性が保たれると思われる。しかし、授業内で学生が小論文を改善していくプロセスを支援する道具、つまり形成的評価の道具として本ルーブリックがどれだけ実用的かは検討の余地がある。第 3 に小論文の形式的な側面ではなく「内容の良さ」を評価できる項目を探索していきたい。基礎演習 I では小論文のテーマは自由であるためどのようなテーマの小論文でも評価できるルーブリックでなければならない。テーマや立論の多様性に対応するため抽象的、構造面への偏りが生じがちである。採点者へのヒアリングや協議を通じ

内容面を評価できる項目を案出していく必要がある。

謝辞　採点対象の小論文をご提供いただいた山梨学院大学青山貴子先生に感謝いたします。なお本研究は山梨学院大学学習・教育開発センターの研究事業である。

## 引用文献

秋山朝康　2011　教員採用試験における模擬授業テストの公平性：ラッシュモデルによる評価者バイアス（bias）の分析．英語英文学 **38**，3-20.

Chan, D. 2000 Detection of differential item functioning on the kirton Adaptation-Innovation Inentory using multiple-group mean and covariance structure analyses. Multivariate Behaviral Research **35** 169-199.

Cleary, T.A. & Hiton, T.L. 1968 An investigation of item bias. Edudation and Psychological Measurment **28** 61-75.

Holland, P.W. & Thayer, D.T. 1988 Differential item performance and the Mantel-Haenzel procedure. In H. Wainer & H. Baum（Eds.）Test validity. 129-145. Hillsdale, NJ: Lawrence Erlbaum.

梶井芳明　2001　児童の作文はどのように評価されるのか．教育心理学研究 **49**（4），480-490.

加藤健太郎・山田剛史・川端一光　2015　R による項目反応理論．オーム社.

木村真治　2004　主観的テスト採点トレーニングにおける κ 係数、カッパ係数、多相ラッシュモデルの利用．言語と文化 **7**，27-36.

熊谷龍一・脇田貴文　2003　特異項目機能検出方法の比較 -BILOG-MG と SIBTEST を用いた検討 -．心理発達科学 **50**，83-90.

Lord, F.M. 1980 Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

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**（1），107-115.

野口裕之・熊谷龍一・脇田貴文・和田晃子 2007 日本語 Can-do-statements における DIF 項目の検出．日本言語テスト学会研究紀要 **10**，106-118.

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.

ReadWriteThink 2004 Writing Rubric. https://www.ramapo.edu/fa/files/2013/04/Writing-Rubric-3.pdf（2016 年 11 月 20 日閲覧）

齋藤有吾　2016　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．京都大学大学院教育学研究科紀要 **62**，427-439.

Su, Y. H., & Wang, W. C. 2005 Efficiency of the Mantel, Generalized Mantel-Haenszel, and logistic discriminant function analysis methods in detecting differential item functioning for polytomous items. Applied Measurement in Education **18** 313-350.

庄司惠雄・野口裕之・金澤眞智子・青山眞子・伊東祐郎･迫田久美子，春原憲一朗，廣利正代･和田晃子　2004　大規模口頭能力試験における分析的評価の試み．日本語教育 **122**，42-51.

平直樹　1995　物語作成課題に基づく作文能力評価の分析．教育心理学研究 **43**（2），134-144.

田崎勝也　2007　文化的自己感は本当に「文化」を測っているのか―平均構造・他母集団同時分析を用いた特異項目機能の検証―．行動計量学 **34**（1），79-89.

turnitin 2012 COMMON CORE STATE STAN-

DARDS WRITING RUBRICS Grades 9-10.
http://www.schoolimprovement.com/docs/Common%20Core%20Rubrics_Gr9-10.pdf（2016 年 11 月 20 日閲覧）

宇佐美慧　2011　小論文評価データの統計解析―制限字数を考慮した測定論的課題の検討―．行動計量学 **38**（1），33-50.

宇佐美慧　2012　論述式テストの運用における測定論的問題とその対処　日本テスト学会誌 **9**（1），146-164.

渡辺 直登・野口 裕之　1999　組織心理測定論―項目反応理論のフロンティア―．白桃書房

West virginia department of education 2008 West virginia writing rubric.
https://wvde.state.wv.us/teach21/writingrubrics/（2016 年 11 月 20 日閲覧）

Woman in Defense 2015 Essay Rubric. http://wid.ndia.org/about/Documents/WID_EssayRubric.pdf（2016 年 11 月 20 日閲覧）