

国語科の評価問題における記述式問題の採点過程に関する研究

－ 採点基準と採点答案の分析を通して －

【研究者】 教科教育部 指導主事 河原 宜央

【研究協力校】 広島県立広島観音高等学校

研究の要約

本研究は、国語科の記述式の評価問題における採点過程について調査・分析を行い、採点の妥当性の観点から課題を明らかにし、その解決の在り方を提言するものである。

研究の手順として、まず記述式問題の評価について、先進的に研究が進められてきた心理学や統計学の成果、及び国語科教育における評価に関する論考を概観し、採点の妥当性の担保に国語科教育の立場からアプローチをするという本研究の方向性を明らかにした。次に、実際の採点過程を把握するための調査を行い、その分析結果から、評価対象とする能力の重み付けや文章の表現面の評価観において採点者間の差異が発生しやすいことが分かった。そして、そうした差異の発生を抑えるための方法について、国内外の学力調査の事例を参照し、適切な採点基準づくりのポイントや、採点技術向上のための研修の在り方について提案をした。

キーワード： 記述式問題 採点 妥当性

I 研究の目的

本研究は、国語科の評価問題における記述式問題について、採点者がどのように採点基準を作成し、その基準に基づいてどのような採点を行うのかを調査・分析することによって、その課題を明らかにし、課題の解決に向けた方策の考案を試みるものである。

高大接続システム改革会議の「最終報告」（平成28年、以下「最終報告」とする。）では、大学入学選抜改革の具体的方策として、「学力の3要素」について、多面的・総合的に評価すべく、現行の大学入試センター試験を改め、平成32年度（2020年度）から「大学入学共通テスト」を実施するとしている。（「最終報告」段階は「大学入学希望者学力評価テスト」という名称であったが、その後「大学入学共通テスト」とされた。本稿では以下、「共通テスト」とする。）また、個別の大学における入試においては、小論文、プレゼンテーション、推薦書等の多様な評価方法を用いるなどの改善が求められた⁽¹⁾。

このうち「共通テスト」では、複数の情報を統合し構造化して新しい考えをまとめる思考・判断の能力や、その過程を表現する能力の評価のため、記述式問題が導入される。

この「共通テスト」への記述式問題の導入及び、個別の大学入試における改善をめぐって、国立大学協会（平成28年）は、「記述式・論述式問題に評価すべき能力をいかに構造化できるかは、根源的な課題である。」¹⁾とし、現状において国立大学全体にも大学入試センターにも、能力の構造化のための知識やノウハウの蓄積が十分でないことを指摘している。

また、個別の大学入試の改善について、高等学校の教員から、趣旨には賛同するものの評価方法の客観性や公平性に対して不安の声が挙がっている⁽²⁾。

こうした大学入学選抜改革の実現上の課題は、ことに記述式問題という点に特化してみると、さほど新しい課題ではない。記述式の問題は、これまでも個別の大学入試では、国語の試験をはじめとして出題されており、また小論文のように解答の自由度が比較的高く、採点者の主観が影響しそうな試験も各種試験において広く実施されてきているからである。さらにいえば、学校内で実施される定期考査などの評価問題においても、記述式問題は日常的に出題されているものである。したがって、記述式問題が抱える、評価すべき能力の構造化や採点の妥当性などに関する課題は以前からもあったといえる。

心理学や統計学の分野においては、小論文等の論

述型の試験における評価の方法や、結果の分析方法、評定者によるバイアスの影響などが研究されてきた。ただ管見の限り、そのようなテーマについてこれまで国語科教育の側からはほとんど取り上げられることがなかった。もちろん国語科教育においても学力調査や評価に関する研究はあるが、多くは評価の対象となる能力の措置をめぐるものであり、記述式問題に起こり得る採点方法の妥当性などが考察されているわけではない。

このように、これまでの国語科教育で潜在的には存在しながらもクローズアップされてこなかった、記述式問題のもつ諸課題が議論され出した背景には、一つに大学入試センター試験に代わって行われる「共通テスト」が、受験生や高等学校の授業内容に与える影響力の大きさがある。また、「共通テスト」の在り方が検討される中で、PISAやTOEFLで用いられている項目反応理論（以下、「IRT」とする。なお、IRTはItem Response Theoryの頭文字を取ったものである。）が話題になる⁽³⁾など、国語科教育以外の分野の概念や研究成果が取り沙汰された影響もあると考えられる。

そこで本研究は、国語科の評価問題における記述式問題の諸課題について、心理学や統計学等の研究成果と国語科教育のおかれている現状を踏まえつつ課題を整理し、その克服に向けた評価問題の在り方を模索したい。

Ⅱ 記述式問題の定義

菅千索（1988）は、記述式テストについて「○×式とも呼ばれる客観テスト（正誤法、選択肢法、組み合わせ法など）以外のテスト法の総称」と定義し、さらに記述量に応じて「論文形式、小論文形式、論述形式、短答式に、おおよそ区分される。」²⁾としている。

文部科学省は、記述式問題について明確な定義をしているわけではない。しかし、大学入試センターが示した「『大学入学共通テスト（仮称）』記述式問題のモデル問題例」（平成28年、「仮称」とあるのは例が公開された当時のママ）では、出題形式を次のように分類している³⁾。

- | | |
|----------|---------|
| ① 選択式 | ② 短答式 |
| ③ 条件付記述式 | ④ 自由記述式 |
| ⑤ 小論文 | |

大学入試センターによる出題形式の分類

このうちモデル問題例は、②～④に該当している。②の短答式というのは、テキスト中の比喻表現の内容を答えたり、抽象的・難解な表現の内容を答えたりするものである。また、③の条件付記述式は、テキストの内容の範囲で答えるものであり、④や⑤はテキストの内容を踏まえつつ、発展的に自分の考えを述べる問題として位置付けている。

記述式という表現があるのは③と④であるが、⑤の小論文は出題の内容によっては④の自由記述式の一つにも捉えられるし、高校入試に出題されるようなレベルのものであれば、③の条件付記述式の延長にあるともいえ、③～⑤は明確な区分がしにくい。

菅と大学入試センターの分類を踏まえると、広義には上記②のような、解答が短いものでも記述式だといえるが、Iで述べた課題が発生すると考えられるのは④や⑤のような問題の場合であろうと想定されることから、本研究においては、②～⑤の出題形式を記述問題とし、特に④や⑤の形式の問題に焦点を当てて扱うこととする。

なお、参照した文献によっては、「論述式」や「論述試験」と称していることもある。典拠の引用や要約部分では出典通りの表記とする。

Ⅲ 記述式問題に関する先行研究

1 心理学や統計学の分野における論考

(1) 論文試験の採点上の課題

国公立大学入試の二次試験や私立大学入試で小論文試験が導入され出したころ、心理学者であり、実際に自身の在籍する大学の小論文試験の実施に関与した石井巖（1981）は、小論文試験の採点について、「小論文試験の採点に当っては、試験の信頼性・妥当性・公平性のために、特に配慮が必要である。採点者の頭の中にある基準が、出題の意図と合致するようにしなければならない。」⁴⁾と述べ、その具体的方策として、採点者に出題の意図や採点基準などについて事前に説明をし、その後基準について採点者間で討論をし、さらに各自の基準がいくらか安定するまで採点の練習をすることが望ましいと述べている。また、採点のカテゴリー（段階）については田中良久（1965）の理論を援用し⁴⁾、「現実的には、5段階から7段階が妥当であろうと思われる」⁵⁾と述べ、さらに採点者間の採点の甘辛の違いを調整するための計算式を提示している。

このように、記述式問題の採点の公平性の問題は、小論文試験が導入され出した当初から指摘さ

れ、その対応が検討されていた。

(2) IRTを用いたバイアスの評価

宇佐美慧（2010）は、採点者の主観や複数の採点者の間の甘辛などの差からくる評定の信頼性の問題や受験者側の特性による影響を考慮したデータ分析を行うために、IRTを用いた測定モデルを提案している⁽⁵⁾。

IRTは、テストの項目（試験でいう設問のこと）に対する反応（受験者の解答の正誤など）について、事前に設定した学力などの尺度を用いて測定する理論や分析の体系である。このモデルのメリットは、採点者間の判断の個人差や受験者の文字の美醜が及ぼす影響などを多角的に評価できることにある。IRTは異なる問題であっても比較的簡便に能力を測定ができることから、様々なテスト開発の場面で適用されてきた。

(3) 測定論的問題とテスト関係者の課題意識

宇佐美（2015）はIRTによるモデルを開発したが、一方で、測定の信頼性・妥当性、バイアスなどの測定論的問題を抱える論述試験の場合、IRTを運用していく以前に、測定論的な問題そのものの解決を図ることや、測定論的問題の深刻さを教育測定の専門家、実務者、テストの作成者の間で共有すべきであるとの見解を示している。なぜなら大学入試のように大規模な試験の場合、IRTを用いた試験の開発・導入に、多大な時間的・人的・経済的コストがかかるためである。

宇佐美は共有すべき測定論的な問題の論点として、次のような点を挙げている⁽⁶⁾。

- ① 項目数と測定論的問題の関係
- ② 信頼性を高めるための方法—項目数・採点者数—
- ③ 採点訓練・採点方法
- ④ 回答内容の自由度と測定する能力・特性の関係

①と②の点について、宇佐美は項目数（設問数）が多いほど信頼性が高まるとしている。また、採点者も複数の方が信頼性の点でよいが、4名を超えると頭打ちになるということ自身の研究結果から述べている。

③④の点については、十分な研究蓄積がないとして、今後の研究・検証の求められる点だとしている。特に④の点は、測定論的に見て信頼性や妥当性に重要な影響力をもっており、「テストの作成時に

は、測定を意図した能力・適性の定義について明確化するとともに、それに対して不必要に回答内容の自由度が高い課題内容や回答方法になっていないか吟味することは重要である」と指摘している。

(4) 分散調整済み平均の使用とその価値

阿久津洋巳（2017）は、小論文の採点について、多くの場合、同一答案に対して複数の採点者によって出された素点を単純に平均してその答案の評定点としていることの問題点を指摘し、各採点者の分散を調整して平均得点を出す方法を提示している⁽⁷⁾。

阿久津の提示した方法は、エクセルでも計算可能であるため、統計や測定に関する専門知識がなくても、比較的導入しやすいが、阿久津自身もこの方法を「小論文試験全体を見たときには実は小手先の技術にしかすぎない。」⁽⁸⁾として、まずは妥当性のある問題を作ることが重要であると述べている。

(1)～(4)のような心理学・統計学等の分野における研究を概観すると、記述式問題における公平性の問題は早くから取り上げられ、採点者の個人差を調整したり、個人差に極大影響されないための測定方法、算出方法が提示されたりしてきたといえる。一方で、問題そのものの信頼性・妥当性を高めることや採点者の熟練の必要性、そしてそうした課題を専門家だけでなく問題作成者、試験の実施者、採点者とも共有していくことが指摘されてきた。

2 国語科教育学における論考

国語科教育においては、記述式問題という形式そのものを研究の対象とはせず、PIISAの結果や学習指導要領の改訂によって、新たに求められるようになった国語の能力に目が向けられている。

(1) 学力論の視点による様々な指摘

鶴田清司（2010）は、PIISAや全国学力・学習状況調査（以下、「全国学力」とする。）について言及する中で、「『学力』は学習者の多様な能力の一部であり、学力調査は何をどのように測るのかという点で、きわめて限定された能力を限定された方法で調査するものであるという認識が前提である。」⁽⁹⁾と述べている。

また、足立幸子（2010）は、海外の学力調査と「全国学力」を比較した上で、諸外国の読む能力に関する調査は、読む能力の理論的枠組みを示しているのに対し、日本の場合は評価の観点が学習指導要領の文言とほぼ同じで「どのような読書の過程を想定し、どのような力を測定しているのかが分からない」⁽⁸⁾と指摘している。

足立の指摘については、平成 29 年 11 月に実施された「共通テスト」のプレテストでは「問題のねらい」として、学習指導要領の内容の横に「主に問いたい資質・能力」と該当する小問の概要が具体的に示されており⁽⁸⁾、見取る能力を具体化する動きがみられる。

鶴田の指摘は、国語科に限らずいえることであり、その問題を克服すべく、ペーパーテスト以外の評価方法や、「真正の評価」の重要性が各方面で謳われてきていることに通じている。しかし、各学校の実践レベルならそれは可能であるが、入試のように限られた時間に多数の受験者の学力を評価するためには、C B T の導入も検討はされているものの、現状ではやはりペーパーテストが有力な評価方法になると考えられる。

(2) 「生産的な読み」の能力の評価

間瀬茂夫・守田庸一・宮本浩治(2014)は、P I S A や「全国学力」の B 問題について「従来は育成のための教育的営みはされていても、客観的な評価が難しいため、ペーパーテストという方法では測定しようとしなかったような能力について、課題を工夫し、何らかの基準を設定して評価しようとする試みである」⁹⁾とその意義を述べている。

間瀬は、文章に書かれていることを理解するだけでなく、「既存の知識や情報と結びつけて推論したり、批評したりしながら読書行為を行う『生産的な読み手』を育成することは、我が国の国語教育とりわけ中等教育段階の読みの学習指導が取り組むべき課題である。」¹⁰⁾として、その育成のための授業モデルと評価問題及び評価指標(ループリック)を開発した。例えば、評論文の評価問題の一つとそのループリックは表 1 のようになっている⁽⁹⁾。

表 1 間瀬らによる評価問題とループリックの一例

問いの枠組み		読みの構えを問う設問
問い		あなたが見出しを見て予想したことと、実際に文章に書かれていたこととは、どのような違いがありましたか。説明してください。
ループリック	3	自分の行った予想と本文との違いが具体的に述べられているとともに、文章に書かれていることの把握が的確である。
	2	自分の行った予想、もしくは文章に書かれていたことのいずれかが具体的に書かれているが、どちらか一方は具体的に述べられてい

		ない。もしくは、文章把握が的確ではない。
	1	自分の行った予想と書かれていたことの一 致もしくは不一致の判断や感想のみが書かれて いる。
	0	無解答等。

この研究で開発されているループリックは実際の高等学校の生徒に評価問題に解答してもらい、その答案から作成したもので、いわゆる採点基準とは異なる。

この評価問題では、「読みの構え」という、いわば読み手が文章にどう向かっているかという、読みのプロセスを最初から問おうとしている点で斬新である。問いはこの他に、P I S A や「全国学力」にみられるような、文章の表現の効果を評価する問い(「書き手と読み手の関係を問う」設問)や、読み手が書き手の伝えたいことを受け取り、自らの既存の認識に照らして、その認識を更新するべきかどうかの判断を求める問い(「テキスト世界と現実世界の関係を問う」設問)などがあり、「生産的な読み」の能力を見取ろうとしている。

こうした評価問題とループリックは、従来のペーパーテストでは測定されなかった(しにくかった)範囲の国語の能力の評価をペーパーテストという手段の範囲で試みるものである。

(1)で鶴田が「限定された能力」と批判した点について、間瀬らの開発した評価方法は、評価する能力の範囲を拡張するものになっているといえる。一方で、このループリックは採点基準ではないため、これをそのまま本研究で取り上げている採点の妥当性の課題に当てはめることはすべきではないが、仮にこのループリックを使って採点をする場合には、もう少し詳細な基準が要るであろう。例えば、表 1 の 3 段階の「具体的」とはどの程度か、「的確」の度合いはどこまで許容するのかなどである。

IV 先行研究を踏まえた本研究の位置付け

国の方針や各分野の先行研究を踏まえると、国語の記述式問題をめぐっては、おおよそ次のような現状にあるといえる。

- ① 「学力の 3 要素」の特に思考・判断・表現に関わる能力を入試においてもより多面的・総合的に評価するため、従来のペーパーテストでは測らなかった能力を測定するようになってき

た。

- ② ①に伴って、出題形式は記述式で、解答の多様性を認める問いを増やす方向に向かっている。
- ③ ②によって、問題や採点の妥当性など、従来から指摘されていた測定上の課題が一層深刻になると予想される。
- ④ 心理学・統計学の見地から③の課題にアプローチもできるが、運用にはコストがかかる上、問題の妥当性確保や採点者の熟練などの課題までは解決できない。
- ⑤ ④の克服のためには、各分野の専門家、テストの作成者、実施者、採点者の連携が必要である。

大きく区分すると、どのような国語の能力が求められるかという問題、統計学などからみた測定上の問題、入試システムの問題が相互に絡んでいるといえる。

本研究は教育センターという立場から、上記③の採点の妥当性の向上や、④の採点者の熟練という部分に焦点を当て、課題の解決方法を探る。その理由は、採点基準作りと採点業務は、どの教員も行うことであり、まず採点の在り方を検討することが全体的な信頼性の向上につながると考えられるからである。心理学分野からの指摘にあるように、熟練した採点者が採点基準に共通理解をもって採点に当ること、採点者間の個人差を抑えることができよう。

個別の大学入試の採点者は基本的にその大学の教員であろうし、「共通テスト」については、現段階では採点を民間業者に委託するということが示されているが、国語科の教員の採点の妥当性に対する関心が高まれば、採点に携わる大学や業者に一定のメッセージを送ることも可能になるであろう。

もっとも採点は、問題がどのような能力を問い、その能力はどのような構造になっているのかということとも関連している。しかし、大槻和夫（1988）や益地憲一（2013）が指摘するように⁽¹⁰⁾⁽¹¹⁾、国語科で育成する能力の構造は他教科に比べて複雑で曖昧な部分が多いため、問題の根本的な妥当性を検討するにはさらに議論を待たねばならない。したがって本研究では、問われている能力についての一般化は避け、具体的な問題に即して、採点をする上で関係するものを取り上げる。

1 調査の概要

本研究においては、採点の妥当性を高める採点方法や訓練の在り方を考察するため、国語科教員による実際の採点過程を調査した。調査対象及び手順等は表2のとおりである。

表2 調査対象及び手順等

調査対象	【採点者】 広島県立広島観音高等学校の教諭（国語科） 3名 【答案】 同校 全日制第1学年生徒1学級（41名）分
調査時期	平成29年7月
調査手順	① 研究者から採点者に調査の趣旨、方法を説明。 ② 第1学年の1学級の生徒41名に、研究者が用意した評価問題に解答してもらう。 ※生徒には、教育センターによる調査であること、学校の成績には無関係であることを伝えた。 ③ 41名分の未採点答案を3部ずつ複写。また、41名のうち、抽出した8名分の未採点答案をさらに3部複写。 ※抽出に際しては、普段の国語の成績層に照らして概ね偏りなく抽出するように依頼した。 ④ 【一次採点】 3名の採点者に各自で採点基準を作成してもらい、抽出した8枚の答案を各自で採点してもらった。 ※基準作成時及び採点中は、他の採点者とあえて相談しないように依頼した。 ⑤ 【二次採点】 3名の採点者で採点基準のすり合わせを行い、統一した採点基準を作成してもらい、その基準に基づいて41名分の答案を採点してもらった。 ※一次と同様に採点中の相談はせず、また、一次に採点した抽出答案は、2回目に評価が変わっても、1回目の採点の修正しないように依頼した。 ⑥ 研究者が全ての採点・未採点答案と問題用紙を回収し、採点に要した時間や採点しにくかった点などを聞き取った。

V 採点過程の調査・分析

研究協力校の広島県立広島観音高等学校（全日制）は、1 学年240名規模の総合学科の学校である。部活動が盛んで、卒業後は大学に進学する生徒が多い。そのような学校規模・実態から、複数の採点者を確保でき、検証可能な答案が一定数得られると想定されると判断し、研究協力を依頼した。

採点をしたのは、それぞれ教職経験年数が異なる3名の教諭である。（以下、採点者A、B、Cとする。）

調査に用いた評価問題は、大分県教育委員会が作成し、平成28年度の公立高校入試において実施された問題を改題したものである。

問題は、電車内でお年寄りに席を譲ろうとしたが断られた男子高校生に関する投書を読み、自分が男子高校生の立場ならどうするか、別の資料を用いて200 字程度で自分の考えを書くものである。調査問題の概要は、図1 のとおりである。

電車内で男子高校生が、七十代の男性に席を譲ろうとしたところ断られ、きまり悪そうにしていた様子を見た中学生が、「席を譲るという行為はなかなかむずかしいものだ」と述べる投書。	設問のリード文及び条件
--	--------------------

図1 調査問題の概要

図1 中の「設問のリード文及び条件」は次のようにした。

【新聞の投書】中の傍線部分に <u>席を譲るという行為はなかなか難しいものだ</u> とあるが、あなたが【新聞の投書】中の高校生の立場で席を譲るとしたら、どのようにしますか。あなたの考えを、次の条件1～3に従って書きなさい。 条件1 自分が【新聞の投書】中の高校生の立場だったらどのようにするか、という想定については、そのように考える理由も書くこと。 条件2 【資料】の内容を踏まえて書くこと。 条件3 内容に応じて段落を変え200字程度で書くこと。

設問のリード文と条件

問題選定に当たっては、解答の自由度が一定程度ある自分の考えを述べる問題とし、かつ高校一年生にとって無理のない難易度であることや公平性の観点から、他県の公立高校入試問題より選定した。設問のリード文や資料の誤読、及び資料の関連付けの仕方の違いによって、解答の質に極端なばらつきが出ることを防ぐため、資料の数やリードの問い方については改題をした。

解答例は次のとおりである。採点者に対しては、問題と解答例のみを示し、配点が10点であることは共通にしてもらうよう伝え、採点基準は各自で作成してもらった。

私なら同じように譲り、断られたら「失礼しました。」とさりげなく言ってそのまま離れた所で立っておく。なぜなら、たとえその男性が座らなくても、後からその席を必要とする人が乗ってくるかもしれないからである。

資料から、七十代の人よりも妊娠中や幼児連れのの方が席に座りたいと思っている割合の高いことが分かる。断られたとしても、その席を必要とする人が他にもいるかもしれないという思いで譲るべきだと考える。

(197 字)

採点者に提示した解答例

この設問で問われている能力としては、大分県教育委員会が公表した設問の意図⁽¹²⁾を基に、改題を踏まえて次に示す点に整理した。なお、この能力の整理は採点者には示していない。

- a 投書やグラフの内容を正確に把握する。
- b 自分ならどうするかという結論を明確に述べる。
- c b の理由を論理的に述べる。
- d 資料から得た情報を c の根拠や補強として用いる。
- e 構成に応じて段落分けをして書く。
- f 漢字・語句を適切に用いる。

調査問題で評価する能力

2 一次採点（抽出）

(1) 一次採点の採点基準

各採点者が作成した採点基準について、概ね類似している観点ごとに整理したものを表3に示す。

3名の採点基準を比較すると、観点については結論（自分が高校生の立場ならどうするか）とその理由、資料の関連付け、語法の適切さ、字数という点で一致していた。段落構成については、採点者B・

表3 各採点者が一次採点前に作成した採点基準（○数字は配点，合計10点）

観点／採点者	A	B	C
結論の明示	どのようにするか③ (記述がある②) (適正である①)	自分が投書中の高校生の立場だったらどうするか②	・投書中の高校生の立場で席を譲るとしたら、どのようにするか、考えを明確に述べている③ ・投書中の高校生の立場で席を譲るとしたら、どのようにするか、考えを述べているが、明確でない② ・自分の意見を述べていない④
理由	理由③ (記述がある②) (適正である①)	上記の理由② *ただし、根拠を資料に求めている場合は加点しない。	・上記の理由を述べており、説得力がある③ ・上記の理由を述べているが、説得力がない② ・理由を述べていない①
資料の自説への関連付け	資料への言及③ (記述がある②) (適正である①)	・資料に触れており、その内容が自分の主張と関係している④ ・資料の内容に触れている②	・資料の内容を効果的に用いている③ ・資料の内容を用いているが、効果的でない② ・資料の内容を用いていない④
一貫性	全体としてのまとまり①		
段落構成		段落を変えている①	・正しい段落構成が出来ている① ・段落構成ができていない④
字数	160字未満は全体④	200字程度で書いている①	字数不足 減点①
言語事項	誤字脱字 減点① 不適切な係り受け 減点①	誤字脱字／不適切な日本語 一種類につき減点①	誤字脱字／言葉の誤用／文のねじれ 減点①

Cが観点としている。採点者Aのみが設定したものに、全体としてのまとまり（一貫性）がある。

これらの観点のうち、結論・理由・資料の関連付け・段落構成・字数については、設問のリード文と条件に明示されていることである。また、採点者A・Bは解答例に基準を決める書き込みがあり、解答例からも基準を作成していることが分かる。

一方配点について、結論・理由・資料の関連付けは、各観点を段階に分けて加点する方法で一致した。採点者Aは要素に分解しているが、「適正である」は「記述がある」ことを前提にしていると考えられるので、実質「0点－2点－3点」の三段階とみなせる。また、採点者Bは、理由に資料の内容を挙げただけの場合は、理由としてみなさず、あくまで資料を根拠にした上での理由付けを「理由」とみなしていたり、採点者Cは資料の用い方が効果的かどうかを基準としていたりするなど、「理由」や「資料の関連付け」の内実については採点者間に差

異がある。誤字脱字や用語、文の適切さについては、減点式をとっている点も一致している。しかし、どの程度減点するかは採点者によって異なっている。

(2) 一次採点の結果

各自が採点した結果と各採点者間の相関係数（ピアソンのr）は表4のとおりである。なお、答案の番号は抽出した8枚の答案に識別上の番号を付したものであり、生徒の出席番号とは無関係である。

三者の得点が一致したのは2枚であった。三者とも異なる得点を出したものは3枚ある。得点差の多くは1，2点だが、最大で4点になった答案もある。採点者間の相関係数は0.6～0.7であった。

三者の評価の観点は類似しているものの、配点の仕方が異なるため、自ずと得点の不揃いが生じている。しかし、採点された答案を見ると、誤字脱字や語法、文のねじれなど同じ評定ができると思われる部分についても、見落としがあったり、話し言葉を

不適切とみなすかどうかには差があったりして、得点差が生じていることが分かった。

事後の聞き取りで採点の難しかった点として、誤字脱字や表現面をどこまでみるか、条件にある「資料を踏まえて」というのをどの程度で線引きをするかという点が挙げられた。

表 4 一次採点の結果（得点）と相関係数

答案	A	B	C	点差
①	7	8	6	2
②	6	6	5	1
③	6	6	6	0
④	10	6	6	4
⑤	10	9	8	2
⑥	10	10	10	0
⑦	5	6	5	1
⑧	8	6	9	3
平均	7.8	7.1	6.9	
標準偏差	1.92	1.54	1.76	
相関係数	A－B 0.60 A－C 0.70 B－C 0.65			

3 二次採点（全答案）

(1) 採点基準の統一

一次採点の後、三者で協議をして統一した採点基準を作成してもらった。その採点基準は表 5 のとおりである。

表 5 統一した採点基準

観点	基準・配点
主張	・ 明確である ④ ・ 不明確 ② ・ 書いていない 0
理由	・ 理由が明記され、資料を引用しており、理由と論理的に関わりのあるもの ⑤ ・ 理由が明記され、資料も引用しているが、論理的整合性に乏しいもの ③ ・ 理由が明記されているが、資料の引用、関わりのないもの ② ・ 理由のないもの 0
字数	・ 8 割以上 ①
言語事項	・ 誤字脱字：一種類ごとに減点① ・ 係り受け：一種類ごとに減点①

聞き取りによると、統一基準づくりに際しては、

まず各自が一次の採点に用いた基準と答案を出し合い、特に得点差のあった答案を中心に基準の見直しをしていったという。結果的に観点は絞り込む方向で整理することになったという。

一次の各自の採点基準と比較すると、観点が整理され、特に「資料の関連付け」は採点者 B の一次の基準に近い形で「理由」の中に段階的に位置付けられている。抽出答案では、理由の根拠として資料を取り上げている解答が（不十分なものも含めて）多く、そうした点を踏まえて理由と資料を独立した採点とするのではなく、どう関連付けているかが評価のポイントになったのであろう。設問の条件では並列的になっていた「理由」と「資料の関連付け」が構造的な関係にあると判断されたといえる。

また、三者で基準の割れた段落に関する観点はなくなり、言語事項は、誤字脱字と係り受けについて各一種類ごとに減点するように整理された。抽出答案では明らかに不適切な段落分けが見られなかったこと、一次の採点で言語事項は適切さの度合いが主観的になりやすいと感じたことなどが影響していると考えられる。

(2) 二次採点の結果

統一した採点基準を用いた二次採点（全答案）の結果を表 6 に、一次で採点した抽出答案についての二次との比較を表 7 に示す。

表 6 二次採点の結果

	A	B	C
平均点	8.1	7.5	8.0
標準偏差	1.95	1.53	1.4
相関係数	A－B 0.32 A－C 0.68 B－C 0.63		

表 7 抽出答案の比較（一次－二次）

	一次採点				二次採点			
	A 1	B 1	C 1	差	A 2	B 2	C 2	差
①	7	8	6	2	2	7	6	4
②	6	6	5	1	10	8	8	2
③	6	6	6	0	7	7	7	0
④	10	6	6	4	7	8	7	1
⑤	10	9	8	2	9	8	9	1
⑥	10	10	10	0	10	10	10	0
⑦	5	6	5	1	4	7	5	2
⑧	8	6	9	3	8	9	8	1
Av.	7.8	7.1	6.9		7.1	8.0	7.5	

SD	1.92	1.54	1.76		2.67	1.00	1.50	
相関係数	A－B 0.60				A－B 0.70			
	A－C 0.70				A－C 0.86			
	B－C 0.65				B－C 0.83			

また、各採点者の評定点を度数分布で表すと次の図2～4のようになる。

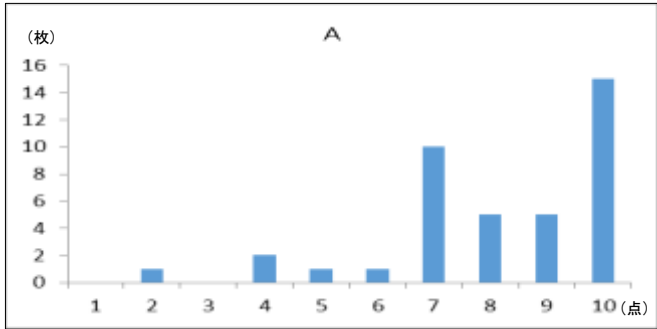


図2 採点者Aの評定点の度数分布

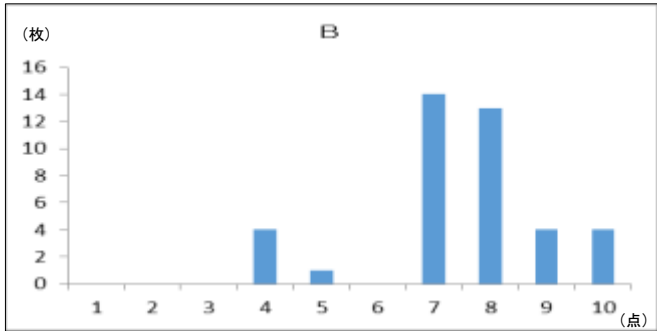


図3 採点者Bの評定点の度数分布

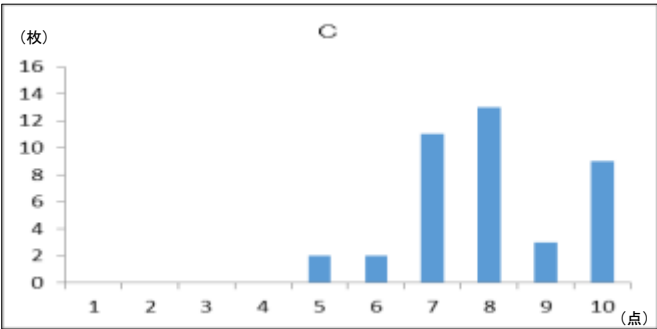


図4 採点者Cの評定点の度数分布

二次採点の結果、抽出答案8枚については、三者の平均点差は縮まり、それぞれの採点者間の相関係数も平均でほぼ0.8と高くなったが、全答案では三者の相関係数は0.32－0.68となった。度数分布をみると、採点者Aは満点（10点）を多く与えているが、採点者Bは満点が少なく、平均点付近に集中している。採点者Cも平均点付近への集中がみられるが、相対的に低い評定点にならない傾向があるとみ

られる。

全答案（41枚）中、三者の評定した得点が一致したものは9.7%、1点差までを含めると46.3%にとどまった。また、抽出答案の得点差は、小さくなったものが3枚、大きくなったものが3枚、不変が2枚であった。

採点者間の相関は抽出答案では高まったが、全体としては低い値となった。度数分布に表れるように、協議を経て作成した採点基準に基づいて採点を行っても、採点者間に差が発生した。

(3) 採点者間の差異の分析

採点基準に沿った採点のどのようなところで採点者間の差異が発生しているのかを把握するため、観点ごと（字数の観点を除く）の各採点者の評定点を確認した。すると、三者の評定点の一致状況は表8のようになった。

表8 観点ごとの評定点の一致状況（枚）

	三者一致	二者一致	不一致
主張	37	4	0
理由	15	24	2
言語事項	14	21	6

主張については、ほとんど評定点が一致しているのに対し、理由と言語事項は一致している答案が全体の半分以下にとどまった。

ア 理由の採点の分析

理由について、三者の点数が分かれた2枚の答案（識別上⑪、⑫とする）と、それぞれの得点を次に示す。

〔答案⑪〕

私が「新聞の投書」の中の高校生の立場だったら、その人と同じように席を譲ります。

理由は、例えば足が悪い高齢者の人がもしも電車やバスなどに乗っていて、席に乗っていない時に急に止まってしまうと、もっと大きなけがをしまい、危ないと思うからです。そして、優先席対象者の人達も、席を譲ってほしいと思っている人が多いからです。

このような事から、私も投書中に出てきた高校生の立場だったら同じようにします。

A：5点　B：2点　C：3点

〔答案⑫〕

僕がもしこの高校生の立場なら相手が受け入れやすいような声かけをしようと思う。例えば、もうすぐ降りる

のでどうぞとか、自分が譲ろうと思った理由や男性の心境をよみてみたりすることだ。資料をみてみても座りたいと思うのには理由がある。

この場合だと高校生は一言で席を譲っている。これでは男性も気をつかってしまう。ので前に書いたことをふまえて声をかけてみれば相手も受けいれやすくなるのではないかなと思う。

A：3点 B：0点 C：2点

表現は答案のママ。

採点基準における理由の採点ポイントは、理由が明記されていることと、資料の引用が理由に論理的につながられているかということである。

答案⑪の理由は二つあり、資料と関係するのは「そして」以降の二つ目である。おそらく答案⑪は、他にも席を譲ってもらいたい人がいるであろうから、断られても譲るべきだと主張したかったのであろう。ただ、そのつながりは明示されていない。この明示されていないつながりを汲み取って採点したと思われるのが採点者Aで、採点者B、Cは明示されていないことは採点の対象としなかったと考えられる。

答案⑫は、おそらく妊婦や乳幼児連れなど、それぞれの人に応じた座りやすい声かけをしたらよいと主張したいのであろうが、そもそも資料には座りたいと思う理由は記されていない。採点者Bは0点としているが、これは根拠とした資料の読み取りが正確でないとして、理由が成立しないと判断した考えられる。一方、採点者A、Cは、資料の読み取りについて、いくらか解答者の拡大解釈を許容したと推測される。

全答案の「理由」についての平均点をみると、採点者Aが4.0点、Bが3.1点、Cが3.6点であった。採点者Bは、一次の採点基準でも資料と理由の関連付けの正確さを重視しており、そうした傾向が二次の採点にも表れたといえる。

このように「理由」については、採点基準にある「論理的整合性」について、解答者の非明示的なロジックや拡大解釈をどこまで許容するかで採点者間の差が出ているといえる。

イ 言語事項の採点の分析

理由に比べて客観的な採点がしやすいと思われた言語事項が最も不一致の多い結果となった。これについても減点の度合いが三者で特に大きく異なった2枚の答案(⑬, ⑭とする。)において、各採点者がどこで減点したのかを表9、10に示す。

表9 答案⑬の言語事項の減点状況

〔答案⑬〕 下線は稿者による。表現は答案のママ。				
もし、私がこの男子高校生だったら、まず行き先を聞 くだろう。そしてそれが何駅も先なら即譲るし、近くて も、「どうぞ」といって譲るだろう。この男子高校生は 断わられて恥かしそうにしているが、席をたってはなれ てみるなど方法はあったはずだ。				
また、資料を参照しても、9割方が座りたいと思って はいるという結果がでたので、譲って、相手も自分も不 自然なく、気持ちいい移動が良いと思ったので私は以上 のように席を譲るだろう。				
減点箇所	減点理由	A	B	C
断われ	送り仮名×		×	
恥かし	送り仮名×		×	×
9	漢数字でない		×	
譲って	マス目はみ出し		×	
相手も……気持ちいいと思った	受け係り×	×		
不自然なく	不適切な用語	×		

各採点者欄の×は、その採点者が減点対象としたことを表す。

表10 答案⑭の言語事項の減点状況

〔答案⑭〕 表現は答案のママ。下線及び () 内は稿者による。				
僕が高校生の人の立場なら、「いいえ、私は結構で す。」と言われても、その男性が座るまでねばり続け る。高齢(←実際にはつくりの令が間違い)者なのでずっと 立っておくことは、足こしにふたんがかかるから少しで も楽をしてもらいたいから。				
70代の方は、特に座りたいと思わないと思う人は、 全体の19%しかいないので、座はゆずるべきだと思 う。遠慮して、結構ですと言ってしまう人も多くいると 思うけど、高齢者にはゆずるべきだ。				
減点箇所	減点理由	A	B	C
齢	誤字	×		
から	受け係り×		×	×
70	漢数字でない		×	
人は…人は…	「は」の重複	×		
思うけど	話し言葉	×		

各採点者欄の×は、その採点者が減点対象としたことを表す。

言語事項の採点状況をみると、基準は「誤字脱字」と「係り受け」となっているが、採点者Aは不適切な表現、採点者Bは原稿用紙の使い方も減点の対象としている。また、漢字の間違いや送り仮名の間違いのように客観的に判別できるものについても、採点者間で差があった。

客観的に判別できる誤字脱字などに差が出ることは、見落としがあるためだと考えられるが、一人の採点者が内容面を評価しながら同時に誤字脱字もチェックすることの困難さを表しているともいえる。また、誤字脱字と係り受け以外の減点が少なからずあることについては、一次採点の結果と同様に採点者の主観に左右されることが大きいといえる。本来基準にないことなので、減点対象としないはずだが、それでも減点に相当すると判断している点は、個々の採点者がそれだけ重視している点だと考えられる。

(4) 言語事項の減点をもたらす影響

観点ごとの採点状況を分析する中で、主張と理由といったいわゆる内容面では高得点となりながら、言語事項の減点によって大きく減点されるというケースがかなり発生していることが分かった。

図5は、全答案について、主張と理由面の各採点者の評定点の平均による順位（内容順位とする）と、言語事項による減点を行った後の評定点の平均による順位（総合順位とする）の分布を示したものである。例えば、図の最も左の縦軸 15～20 の間のプロットに当たる答案は、内容順位では1位だったが、総合順位で17位になった答案である。

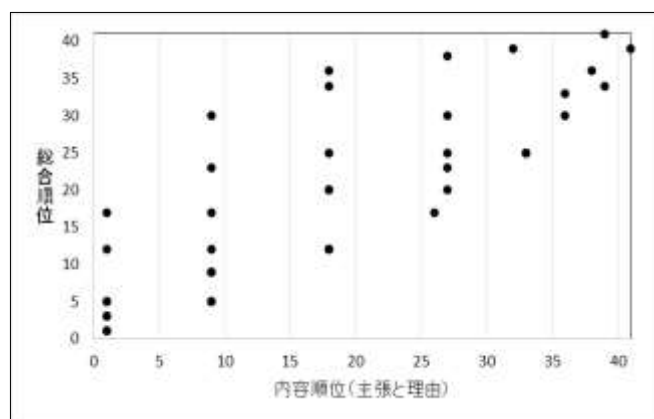


図5 内容順位と総合順位の分布

図5をみると、概ね正の相関がみられるのは内容順位 35 位以下であり、それ以上の順位では最大で 20 位程度の変動がある。つまり、主張と理由は良くても、誤字等が多ければ順位が下がる（その逆もある。）という逆転現象が起きている。10点刻みの評定点なので同点者が多く、1点の違いが大きな順位変動につながってしまうことを考慮しても、言語事項についての1つの減点が重みをもっているといえる。

こうした順位の逆転現象は、国語の能力として何を重視して評価するかということに大きく関連している。採点後の聞き取りでは、結果的に内容的に良いが得点の低い答案、内容的にさほど良くなくても得点の高い答案が出てきたことで、言語事項のようなところをどこまで厳密に採点するかについて疑問をもったという。基準の上では内容面を重視しているが、拙い表現も看過しにくいというのが採点者の実際の思いであろう。

4 調査結果に基づく考察

この調査結果からいえることはおよそ次のようなことである。

- ① 採点者は各自で採点基準を作成する際、設問のリード文、条件設定、解答例に基づいて観点を設定し、その結果、観点の区分はほとんど一致した。
- ② 統一基準の作成においては、採点にバラつきのあった観点について、協議を経て、評価する能力（項目）の関係が整理された。
- ③ 資料と理由のつながり方については、非明示的なロジックや、解答者の拡大解釈によって論が進められている部分をどこまで汲むかによって採点に差が出る。
- ④ 誤字脱字や表現上の適切さについては、見落としが発生したり、採点者の主観が入り込んだりしやすい。
- ⑤ ④による影響は、配点の仕方によって、内容面における評価を逆転させるほどになる。

①について、詳細な採点基準を採点者が作成するケースとしては、例えば本県の公立高校入試における一般学力検査などが考えられる。設問のリード文や条件を、問いたい能力を測定できるようなものにすることは、解答者（受検者）に問題の趣旨を伝えるだけでなく、採点者にも評価すべきポイントを伝える効果があるといえる。

②は、協議の結果、評価する観点間の関係について整理が図られた点で注目に値する。設問の条件では、対等に並んで示されていた「理由」と「資料の関連付け」であるが、この両者には密接な関係があり、このまま独立して採点すると、採点者によってどちらの観点に重きを置くにバラつきが生じてしまう。各採点者が実際に採点した実感を基に協議して観点を整理したことで、各採点者が同じ観点で評

価かできるようになったといえる。

③は、②のように観点の整理をして基準を作成しても、各基準についての採点者ごとの許容範囲の違いが依然として残ることを示している。採点者による解釈の幅が極力生じないような基準の示し方を検討する必要があると考えられる。また、本調査では調査内容の関係上、採点中に採点者間での相談は行わないようお願いしたが、聞き取りでは自分の採点内容について、他の採点者と協議したくなる場面が多々あったという。基準を設定しても、採点のズレが生じる場合は、その基準の捉え方について採点者間で適宜協議をして共通理解を図る必要があろう。

④の見落としについては、複数の採点者でチェックをすることなどによって改善できる。ただ、言葉遣いのような表現上の適切さは、採点者それぞれの感じ方にもよるので、チェックをしても採点者間で合意をするのが難しいことが予想される。例えば、誤字は少ないが、ほとんど漢字を使わずに書いた答案をどう評価するかなどである。

また、④の問題は、⑤の問題にもつながる。例えば、今回の調査では、言語事項による減点が答案によって3点になった答案もある。これは理由の観点でいうと、資料の引用が理由と論理的に整合しているか（5点）、引用が理由に無関係か（2点）の違いに相当する。言語事項で3点減点されることと、理由の論理的整合性が不十分で3点減点されることが、国語の能力に対する評価として等価に扱われてよいかという問題である。

これに対しては、言語事項を評価の対象としないことも考えられるが、例えば理由と引用の論理的整合性を説明する表現が拙いために、その整合性の読み取りに支障をきたす場合もあり、言語事項を完全に切り離して採点することも難しい。こうした難点は、たとえ出題意図が明確に示された設問であったとしても発生し得るので、採点基準を作成・修正する段階で、関連する国語の諸能力とどの能力をどの程度評価するかを整理を行うことが必要だと考えられる。

以上の考察から、国語教育の観点から解決を図るべき課題を、次のように整理した。

- 基準の捉え方や採点上の許容範囲に採点者による違いが生じる。
- 解答に必要な能力について、どの能力をどの程度重視して評価するかを考慮した基準や配点にしなければ、高次な能力に対する評価

が点数に反映しない。

VI 記述式問題の採点上の課題解決に向けて

1 採点基準と採点技術

前章で整理した、採点上の課題を解決していくために、本研究では適切な採点基準づくりと採点技術の向上という点について、国内外の事例を参照しながら、現実的な在り方を検討していく。

日本テスト学会（2007）が示している「テスト・スタンダード」（テストの開発、実施、利用、管理にかかわる規準）では「採点基準が妥当なものであり、適切な指導方法でトレーニングを受けるならば、ある程度安定した採点基準を共有することができる。」¹¹⁾とされている。そうした適切な採点基準づくりと採点技術の向上は、国語科の教員が直接取り組むことができるという点でも現実的で有効な改善策であろう。

2 適切な採点基準づくり (1) 全米学力調査の採点基準

アメリカでは、全国的な学力調査の一環として作文テストが実施されている。平井洋子（2008）は、その評価システムを紹介している。

表 11 は、平井が Persky, H. R. ら（2003）の報告書を基に示した全米学力調査の作文の評価基準（12学年用、説得型課題）の一部である¹³⁾。

表 11 全米学力調査の作文の評価基準（抜粋）

6点 優れている (Excellent)
・ 明確な立場をとり、一貫してよく吟味された理由や例を用いて自分の立場を固めている。論じるなかで、説得的な方略を用いることがある。
・ 焦点がぼやけず、効果的に展開してよく構成されている。
・ 全体にわたって文の構造が変化に富み、言葉の選択も正確である。
・ 文法、スペル、句読点の誤りがほとんどなく、文章の理解を妨げない。
4点 目的にかなう (Sufficient)
・ 明確な立場をとり、大体において重要な理由や例を用いて自分の立場を固めている。
・ 全般的に構成されているが、部分間でのつながりが乏しい。
・ 文の構造が単純で変化がない。言葉の選択はおおむね正確である。
・ 文法、スペル、句読点の誤りはあるが、文章の理解を妨げない。

2点 不十分な (Insufficient) : 以下のいずれかの特徴がみられる。
・ ある立場をとっているが、ほとんど展開がない。
・ 大部分において、構成がみられなかったり焦点がぼやけていたりする。もしくは明確だが簡潔すぎる。
・ 文の区切りや構造がほとんどできていない。不正確な言葉の選択が多い。
・ 文法、スペル、句読点の誤りによって文章の理解が妨げられる。

実際の評価基準は、1点～6点の6段階になっている。特徴としては、複数の観点で一定の基準を満たすかどうかで段階が決まるようになっていることである。このような基準は、作文に要するどの能力をどの程度評価するかということが、各段階で設定されており、採点者は諸能力を総合的にみて段階（水準）を評定することになる。したがって、採点者によってどの観点到に重みをおくかという違いが小さくなると考えられる。例えば「目的にかなう」と評価される4段階においては、自分の立場は明確にして、その立場に対する重要な理由が必要なのに対し、いわゆる誤字脱字のような間違いは「文章の理解を妨げない」程度であればよいことになっている。

なお、評定をこのような段階で数値化することで、IRTを用いた分析が可能になり、異なる問題でも経年比較ができるようになるというメリットもある。

このように利点のある基準でも「一貫してよく吟味された」「大体において重要な理由」など、採点者によって解釈に幅が出そうな余地がある。その点について、平井は、表11の基準はあくまで一般的な到達度水準であり、実際の採点では具体的な問いに即した基準の詳しい注釈と段階ごとの解答例が作成されると紹介している⁽¹⁴⁾。確かにそうすることで、採点の精度は上がると考えられるが、人的・時間的コストがかかるのが難点であろう。

(2) 「全国学力」の解答類型

「全国学力」の場合、いわゆる採点基準は公表されていない。文部科学省が示している「採点作業の流れ」では、委託を受けた民間業者が、答案を国立教育政策研究所が作成した「解答類型」（実施後に公表される。）に沿って分類し、採点することになっている⁽¹⁵⁾。したがって、採点基準はおよそ「解答類型」に準じたものになっているとみなしてよいであろう。

表12に、例として平成29年度「全国学力」の問題B 1 三の解答類型を示す。

表12 平成29年度「全国学力」中学校国語 問題B 1 三の解答類型⁽¹²⁾

類型	条 件
1	次の条件を満たして解答している。 ① 〈心に残った一文〉に、【本の一部】から比喩を用いた表現が含まれる一文を抜き出して書いている。 ② 〈感想〉に、①で取り上げた表現について、「誰（何）」の、「どのような」様子なのかを明確にして書いている。 ③ 〈感想〉に、①で取り上げた表現について、感じたことや考えたことを具体的に書いている。
2	①、②を満たし、③を満たずに解答している。
3	①、③を満たし、②を満たさずに解答している。
4	②、③を満たし、①を満たさずに解答している。
9	上記以外の解答
0	無解答

この問題は、小説の一部から自分の心に残った比喩表現を含む一文を抜き出し、その比喩が表す内容を明示した上で、自分の考えを具体的に書くものであり、設問の趣旨は「表現の仕方を捉え、自分の考えを書くことができるかどうかをみる。」とされている。

「全国学力」の場合は、基準の表現が問いの具体に即しているため、採点者による解釈の入る余地はかなり狭くなっていると考えられる。また、基準の文末が「～書いている」となっていることから、非明示的な解答者の意図を採点者が斟酌することは少ないであろう。

ただ逆に具体的な文言であるがために、一般的なレベルでどのような能力をみようとしているのかが明確でない。「平成29年度 全国学力・学習状況調査報告書」に示されている正答例では、①、②は比喩表現についての理解をみており、③は比喩の効果として考えられることを問うているようであるが、特に③の測定意図は伝わりにくいのではないかと考えられる。また、①～③の軽重の関係も曖昧である。

(3) 適切な採点基準をつくるポイント

以上の参照及び研究協力校における調査から、本研究では、適切な採点基準をつくるポイントを次のように整理した。

- ① どのような知識や能力を評価の対象とするか明確化する。
- ② ①で明確化した知識・能力間の関係（構造や重み）を整理して採点の観点とする。

〈例〉

観点

立論 … 主張－理由付け－根拠となる事実

資料の引用

構成 … 段落の役割－事実と意見の区別

言語 … 漢字／語法／受け係り

- ③ 下の例のように、②の観点をその設問に即して5段階程度に分け、それぞれが各到達段階において満たすべき水準を言語化する。
- ＊言語化をする際には、「～を書いている」「～を明示している」のように、採点者の斟酌が入りにくい表現にする。

〈例〉 1, 2, 4段階の記述は省略

段階	観 点		
	立論	構成	言語
5	明確な主張が示され、その主張を支えるための根拠として適切な資料を引用し、根拠と主張をつなぐ理由付けが明示されている。	意見と事実が区別され、結論、例、根拠などの段落の役割を明確にして書いている。	適度な漢字使用で、誤字脱字・不適切な受け係りもいくらかあるが、読み取りに支障はない。
4	…	…	…
3	主張が示され、その主張を支えるための根拠として資料を引用しているが、根拠と主張をつなぐ理由付けが明示されていない。	意見と事実は区別されているが、結論、例、理由などの段落の役割が不明確な叙述になっている。	適度な漢字使用で、誤字脱字、不適切な受け係りが複数あり、読み取りに支障がある。
2	…	…	…
1	…	…	…

- ④ 各観点の軽重に応じて、配点を決める。
- ⑤ プレ採点として、④の基準を用いて同一答案数枚を複数の採点者で採点し、採点にブレの生じている点を中心に協議して基準の修正を行う。

適切な採点基準をつくるポイント

このポイントの特徴は、②で、評価する知識・能力の関係を整理し、それを③でどの程度できればいいのかを段階で示すことにある。

一般に、漢字間違いや不適切な受け係りは、減点式で採点されることが多いと推測されるが、それは一方で一切ミスのない完璧な文章を求めていることにもなる。また、その減点が、立論や構成の観点と等価に扱われると、場合によって、立論はよくできたが漢字がほとんどできない解答者と、立論はできないが漢字は全て正しく書いた解答者の能力を同じとみなすことになる。

採点基準作りにおいては、まずこの観点の整理と到達度の設定が重要であろう。

また、③の＊に挙げたように、基準に示す表現は、できるだけ採点者の解釈が入り込む余地を狭めた方がよい。〈例〉では、特定の問題に即した記述になっていないが、実際には「全国学力」のように実際の問題に即した具体的な内容にすることが望ましい。

なお、5段階程度としているのは、次のような理由である。池田央（1992）は、評定の段階が多くなるほど採点者間の一致度も高くなると述べている⁽¹⁶⁾。しかし、渡部洋、平井洋子（1993）は、小論文の採点の実験結果から、5段階を超えるとかえって識別力の下がる採点者が出るため、5段階で相関は頭打ちになることを示している⁽¹⁷⁾。実際、採点に要する時間や労力を考慮すれば5段階程度が現実的だといえる。

I R Tの活用の可能性などを考えれば、全米学力調査のように、複数の観点をまとめて総合的に段階分けをする方がよいが、国語の設問の場合、妥当性のある区分をするのが難しいことが予想される。例えば、立論は5段階レベルだが、構成が3段階レベルと評定される答案をどう扱うかといった問題である。

また、⑤の作業も重要である。先述の「テスト・スタンダード」では、論述試験のように多様な回答が得られることの多い試験では、事前に定型化した基準をつくることには限界があり、「実際には採点前にある程度の採点基準を定め、提出された回答を概観した後に、必要に応じて採点基準を追加・修正することになる。」¹³⁾としている。採点基準は万能ではないと捉え、採点をしながら採点者間で協議を行い、修正をしていくことも必要である。

3 採点技術の向上

(1) 採点基準の作成までを含めた採点技術の向上

2の適切な採点基準づくりは、目下の採点の妥当性向上に直結するものだが、採点技術の向上は中・長期的にその妥当性を高めようとするものである。

採点の妥当性の向上のために、先述の全米学力調査においても、また論述試験の歴史の古いフランスのバカロレア試験においても、規模や期間の違いはあるが、採点者を対象とした研修を行うシステムが確立されている。

ただ、そこで実施される研修は、主に解答を基準に照らして識別したり、既成の採点基準を答案の実態に応じて微調整したりすることができるようにするものである。

それに対して、国語科の教員は、採点者であると同時に、所属する学校の定期考査については、作問者でもあり、さらにはその評価の対象となる能力を育成する学習の指導者でもある。基準と照合して解答を識別する技術だけでなく、評価しようとする能力の到達度を適切に判定できる基準の作成能力も重要になると考えられる。なぜなら、先述の「適切な採点基準をつくるポイント」の②において、どの観点を重視するかは、定期考査の場合、とりわけその評価と対応する学習指導の内容と一致させる必要があるからである。

そこで、本研究では採点技術の向上には、採点基準に照らして解答を識別するだけでなく、採点基準を、どのような設問においても作成できるようになることまでも含めて捉える。

(2) 採点技術の向上を図る研修

ア 法定研修の一環としての実施

採点技術の向上は、法定研修のような悉皆の研修の一環として実施することが望ましい。複数の採点者間の差異を減らすという点から、採点についての考え方は全ての教員が共通理解をした方がよい。また、小規模校などで教科担当者が一人である場合は、主観的な採点をしても表面化しにくいという点からも研修の場で、他校の教員と協議を行って採点基準を作成することに意義があるといえる。

研修の内容としては、ワークショップ型で、例えばある記述式問題について、本研究で提示した「適切な採点基準をつくるポイント」に沿った基準の作成や採点後の協議、基準の修正といった一連の採点作業を体験し、望ましい採点基準作成について学ぶといったことが考えられる。

イ 学校内でできること

採点に係る研修は時間と研修参加者の労力を要することや、同一校に複数の国語科教員がいても、担当学年・科目の違いから共通の評価問題についての採点をする人数は限られることもあり、基本的には校外研修などの場で行う方がよいと考えられる。

校内では、なるべくその年度の最初の定期考査までに採点基準の作成方法や採点の進め方について教

科内で認識を共通させ、実際の採点業務に臨むというように、妥当性のある採点を実行する場とすることが現実的であろう。

教科内で共通の問題、採点基準について検討するのであれば、校内の研究授業の機会を利用することが考えられる。例えば、研究授業の担当者はその指導内容を範囲に含む評価問題（記述式）も作成しておき、授業後の協議の場において、その日の授業の内容だけでなく、どのような採点基準で評価することが望ましいかを協議することが可能であろう。

VII 成果と今後の課題

本研究における成果は次の点である。

- 国語科の記述式問題をめぐる国語科教育学と他の学問分野の研究の現状を概観し、国語科教育サイドからの課題解決の可能性を示した点。
 - 調査・分析から、国語科の記述式問題の採点基準及び採点において、採点者間の差異が発生しやすい事項を整理し、その解決の方向性及び採点基準作成のポイント、採点者研修の在り方を提示した点。
- また、今後の課題としては次のようなことがあげられる。
- すでに指摘されているように、評価問題そのものの妥当性についても検討の対象とすることが必要がある。国語の能力を規定することは容易ではないが、個別具体の設問の範囲であれば、評価の対象とする能力を規定し得ると考えられる。
 - 測定論の立場から、適切な採点基準の作成や、採点技術の向上が採点の妥当性向上にどの程度寄与するのかを、統計的な手法によって検証することが期待される。
 - 国語科教育の立場から、適切な採点基準の作成や、採点技術の向上を国語科の学習指導の改善にどのような形でつなげていくかを検討していくことが求められる。

おわりに

本研究を進めるに当たり、採点過程の調査に御協力いただいた広島県立広島観音高等学校の皆様から感謝を申し上げます。なお、本研究における調査は、分析をするために特別な条件を設定して実施したものであり、分析の結果が採点を担当した教諭の日常的な採点を直接的に反映しているものではないことを申し添えておきたい。

【注】

- (1) 「高大接続システム改革会議『最終報告』」(平成28年) p.8 による。
- (2) 河合塾全国進学情報センター(2016):「特集 高大接続改革」『Guideline 2016.7・8』に詳しい。河合塾全国進学情報センターがモニターの高等学校教員に調査したアンケート結果(2016年4・5月実施)による。
- (3) 「高大接続システム改革会議『最終報告』」(平成28年) p.26 による。
- (4) 田中良久(1965):『心理学測定法』東大出版会を援用している。
- (5) 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』pp.163-175 に詳しい。
- (6) 宇佐美慧(2015):「論述式テストの測定論的問題再考—主要な論点の整理とその現実的解決のために—」日本テスト学会第13回大会におけるレジュメによる。
- (7) 阿久津洋巳(2017):「論述式テストの研究(2):小論文採点の集計法」『岩手大学教育学部附属教育実践総合センター研究紀要 第16号』pp.61-70に詳しい。
- (8) 大学入試センターが WEB 上で公表している「大学入学共通テスト導入に向けた試行調査(プレテスト)」(平成29年11月実施)の「問題のねらい、主に問いたい資質・能力及び小問正答率(速報値)等」による。
- (9) 間瀬茂夫、守田庸一、宮本浩治(2014):「高等学校における評論の読みの学力評価—学力調査による分析—」第127回全国大学国語教育学会レジュメ p.80 による。
- (10) 大槻和夫(1988):「国語科における評価」『現代教育評価事典』金子書房 p.243 に詳しい。
- (11) 益地憲一(2010):「国語科評価論に関する研究の成果と展望」『国語科教育学の成果と展望2』p.39 に詳しい。
- (12) 「平成28年度大分県立高等学校入学者選抜学力検査(第一次)の分析結果」(平成28年)を基にした。
- (13) 平井洋子(2008):「作文の評価」荒井克弘、倉元直樹編『全国学力調査 日米比較研究』金子書房 p.88 に詳しい。
- (14) 平井洋子(2008):前掲書 pp.88-89 に詳しい。
- (15) 文部科学省(平成26年):「全国学力・学習状況調査を実施するための委託事業」スライド資料による。
- (16) 池田央(1992):「テストの科学—試験にかかわるすべての人に」教育測定研究所 初版本は絶版。本稿では、2007年発行のオンデマンド版 p.36 によった。
- (17) 渡部洋、平井洋子(1993):「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 通号 33』

に詳しい。

【引用文献】

- 1) 国立大学協会(平成28年):「大学入学希望者学力評価テストの実施時期等に関する論点整理 〜とくに国語系記述式試験の取扱いについて〜」
- 2) 菅千索(1988):「記述式テスト」『現代教育評価事典』金子書房 p.126
- 3) 大学入試センター(平成29年):「『大学入学共通テスト(仮称)』記述式問題のモデル問題例」p.43
- 4) 石井巖(1981):「論文試験とその評価」『行動計量学 8 巻 1 号』日本行動計量学会 p.26
- 5) 石井巖(1981):前掲書 p.27
- 6) 阿久津洋巳(2017):前掲書 p.68
- 7) 鶴田清司(2010):「学力論・方法論の視点から国語学力調査を検証する」全国大学国語教育学会『国語学力調査の意義と問題』明治図書出版 p.17
- 8) 足立幸子(2010):「国語学力調査の比較研究」全国大学国語教育学会 前掲書 p.152
- 9) 間瀬茂夫・守田庸一・宮本浩治(2014):前掲書 p.78
- 10) 間瀬茂夫(平成27年):「中等国語科における生産的な読み手育成のための読解力・授業力診断評価システムの開発」『科学研究費助成事業 研究成果報告書』日本学術振興会 WEB ページ
- 11) 日本テスト学会(2007):『テスト・スタンダード—日本のテストの将来に向けて』金子書房 p.88
- 12) 国立教育政策研究所(平成29年):「平成 29 年度全国学力・学習状況調査 報告書」pp.71-72
- 13) 日本テスト学会(2007):前掲書 p.87

【参考文献】

- 別府正彦(2015):『「新テスト」の学力測定方法を知る I R T 入門 基礎知識からテスト開発・分析までの話』河合出版
- 間瀬茂夫(平成27年):「高等学校における高次読解能力の評価のあり方—読解力評価問題の活用—」『国語教育研究 第56号』広島大学国語教育会
- 細尾萌子(2010):「フランスのパカロレア試験における評価観:問題作成と採点に関する議論の歴史的検討を通じて」京都大学学術情報リポジトリ