Science Journal of Kanagawa University **30** ：33-38 (2019)

■原　著■　2018 年度神奈川大学総合理学研究所共同研究助成論文

論述式試験に関する諸問題と採点支援システムの研究動向

後藤智範 [1, 2]　永松礼夫 [1]

Studies on Essay-type Examination and Automated Essay Scoring System

Tomonori Gotoh[1, 2] and Leo Nagamatsu[1]

[1] Department of Information Sciences, Faculty of Science, Kanagawa University, Hiratsuka City, Kanagawa 259-1293, Japan
[2] To whom correspondence should be addressed. E-mail: gotout01@kanagawa-u.ac.jp

Abstract: It is expected that description type examinations (DTEs) are able to evaluate abilities for logical inference and analytial thinking. DTEs have already been introduced in some faculties and departments for entrance examinations for universities. As a part of "the national common test for university entrance qualification", large-scale DTEs are planned for subjects of Japanese language and mathematics. The scoring process for DTEs differs from that for mark sheet-type examinations, requiring human support for scoring. In the field of pedagogy, some empirical studies on DTEs have been reported and many problems have been pointed out. In the U.S.A., to scorr DTE, some ESSs (Essay scoring Systems) have been in operation for over 20 years. In this paper on DTEs, we reviewed problems and solutions highlighted by recent study results based on the theory of education. Also, based on summarized trends of ESS studies in the U.S.A. and Japan, from the viewpoint of problems with DTEs, we discussed features and properties required for better ESSs.

Keywords: automatic essay scoring, exam, essay-type examination, machine learning, domain knowledge

序論

現在の形式での大学入試センター試験は本年度（2020 年 1 月）の実施を最後に廃止され、2020 年度からは新形式の「大学入学共通テスト」が実施される。「大学入学共通テスト」では論述式問題が導入される予定であり [1]、2017 年秋には試行調査が行われ、国語、数学で実施された [2, 3]。

論述式試験問題は大学入試における試験問題の形態の 1 つとして特に国立大学では、既に 10 年以上前から学部、学科単位では個別に導入されており、採用されている科目等についての調査研究がある [4]。学部、学科単位での実施は、それぞれ試験問題が異なっており、1 つの問題に対する受験者数は多くても 10^3 程度と想定される。「大学入学共通テスト」において、国語、数学で実施されれば前者では 10^5 となり、学部・学科個別での実施とは比べると様々な影響を与えると想定される。

論述式問題は、マークシート方式に代表される選択式問題などと比較すると、主な相違点として以下が挙げられる。

(1) 答案に対する正解／不正解が明確に区分困難
(2) 機械的な採点が事実上利用できず、人手による採点が必要

(1) については、教育学、特に教育測定分野での研究があり、様々な観点から問題が指摘されている。(2) については、前述したように問題数×受験者数に応じた採点者を揃える必要があり、これを支援するための教育工学、自然言語処理、機械学習を応用した研究がされつつある。

本研究は、論述式問題について試験形態としての位置づけ・特徴を明確にし、(1) については教育学、心理学、特に教育測定論の研究動向を概観し、(2) については、教育工学、自然言語処理、機械学習の最近の研究動向における主な論点について解説する。さらに、両者の観点からの問題点と解決の方向につ

いて論ずるものである。

方法

出題形式による試験問題の分類

現在、様々な試験問題形式が採用されている。試験問題の形式全般については、宮本の研究[4]、論述式問題の特徴については大野木[5]があり、本節では、これらの研究に基づいて、多面的な観点から分類を試みる。

(1) 正解の解答形式の分類

(a) 客観的試験問題、 (b) 主観的試験問題

(a) 客観的試験問題

正解が問題作成時に決定され、答案に対して採点者による正誤判断が生じない問題である。客観的問題は、正解の提示の有無により下記の形式がある。

(a) 選択式、(b) ○×式、(c) 記入式

(a) はマークシート方式に代表される試験で、受験者は提示されている解答群から正解を選択するという形式で、正解は解答群中に含まれる。(b) はあるテーマについて、複数の説明文が提示され、その内容に対する正誤を判断する問題である。(a)、(b) 共に正解あるいは不正解は提示されるのに対し、(c) は受験者自身が正解と想定される語句(文を構成しない)・数値・記号などを記入する問題であり、穴埋め式と呼ばれる形式もこれに属する。

(b) 主観的試験問題

客観的な正解が無く、したがって答案と正解との照合による機械だけによる採点は事実上できない。個々の答案に対して採点者による評価が必要とされ、答案に対する解釈、評価の観点など、採点者により採点結果が異なり得るという懸案が生じる。後述する論述式出題形式はこれに属する。

論述式出題形式の分類

論述式出題形式は、解答の記述量および解答への課題・テーマに提示の仕方からさらに分類される。

(1) 記述量

記述量について語数または字数の制限があり、その量（長さ）により 2 つに分類される。

(a) 小論文 (essay)、 (b) 短答式

(a) は英文では 200 語、日本語文では 400 字程度の制限が課された文章で記述する問題である。国公立大学の学部／学科で採用されている「論述式」とは呼ばれる問題形式がこれに相当し、通常正解（文）は無いと考えられている。一方、(b) は数十字から長くても 200 字程度での記述量で、「記述式」とも呼ばれる出題形式である。通常、採点に先立って正解文もしくは模範文が用意される。

両者の本質的な相違は字数というよりも文の数であり、前者は数文以上から構成され、後者は多くても 2 文程度である。この相違は、後述する論述式出題形式の解答に対する評価（採点）基準に密接に関わる。

(2) 課題の提示形式

通常、論述式出題形式では、記述すべき内容に密接に関連する課題が示される。課題の提示形式により、以下に挙げる 3 種類のカテゴリーがある。

・課題小論文、・素材小論文、・データ小論文

「小論文」と記されているが (1) の分類での (a) だけでなく短答式も含まれる。(a) は複数の文（全体で数百字前後の長さ）からなる課題文（prompt, 素材文とも呼ばれる）が、設問に先立って提示される形式である。(b) は課題文を提示せずに、長くても 1 文あるいは短い語句でテーマ（記述するための素材）が指定される形式である。(c) は文章だけではなく、図や表が提示され、これらに記載されるデータの内容に基づいて論述する形式である。

上述の分類によれば、2017 年秋に行われた試行テストの国語の論述式問題 2) は、課題小論文であり、その設問の一部が短答式として分類される。

論述式出題形式の評価能力

論述式問題により受験者のどのような能力が評価（得点の大小による）できるかという問題である。客観式問題では評価できないあるいは困難であるが、論述式問題が評価可能な能力として以下が挙げられている。

(a) 表現力、(b) 構成力、

(c) 読解力、(d) 独創性、

(e) 知識

(a)、(b) および (d) は受験者が記述した文章に対する評価であり、前節で挙げた他の問題形式では評価不可能な能力であることは自明であろう。(c) は他の問題形式であっても、説明文が (c) が必要とされる長さ（語数、字数）を満たしていれば、当該能力を測定することは可能である。一方、論述式であっても前節で挙げた素材小論文形式では、複数の文から構成される課題文が無いため、(c) を評価することはできない。

上記の評価能力と関連するが、以下に挙げる側面も評価対象として考慮される[6]。

・分析的思考、 ・批判的思考、

・問題発見能力、・問題解決能力

論述式出題形式の諸問題

論述式出題形式は、他の試験形式と比較して、教育

測定論上の問題が提起されている。具体的には、以下の 3 項目がある [7, 8]。

(a) 妥当性、(b) 信頼性、(c) バイアス

以下では、(a)、(b) について採り挙げる。

(a) 妥当性

当該試験問題に対する個々の受験者の採点結果、得点は、出題者の意図した能力の測定という観点から正しく反映しているものかどうか、という問題である。教育測定論では、妥当性をさらに以下の 5 項目に分類している [8]。

(1) 内容的妥当性、(2) 基準連関妥当性、
(3) 因子的妥当性、(4) 交差妥当性、
(5) 結果妥当性

以下では、(1) ～ (3) についてその内容について採り挙げる。

(1) は客観的試験問題では、教科目標を構成する項目に対する設問内容が妥当であるか、言い換えれば不備や偏向が有無という観点から判断可能である。一方、論述式問題の場合には、能力測定の対象が特に前節で挙げた分析的思考など測定対象が抽象的であるため妥当であるかどうか検証が困難である。

(2) は当該試験問題と外部の試験問題の関連について得点データを比較する、具体的には確率統計学的分析をすることで判断される妥当性である。

(3) は、得点データに対し因子分析を適用し、因子すなわち背後にある能力を明らかにすることで判断される妥当性である。

(2) および (3) の観点から、論述式試験問題と他の形式・科目と比較とについて最近の研究としては荒井らの調査研究がある [9]。この研究では、2 つの小論文と大学入試センター試験の 8 科目を同一被検者（213 名）に対して実施し得点データに対して、因子分析を含む統計解析をし、以下の結果を得た。

・小論文課題と他科目との相関は低く 0.3 程度
・因子として小論文、文系／理系科目の 3 因子

上記結果により、小論文試験は、他の科目とは異なる能力を測定しており、試験として有用であるという結論を得ている [9]。

一方、小論文の得点（素点）と分散調整した得点の両者の平均の相違を明らかにするために数値シミュレーションを用いた、阿久津らの研究がある [10]。結果として素点を用いることにより、得点順位が変わる受験者が非常に多い（90%）ことを明らかにし、小論文試験問題を、大学入学試験に導入することに否定的な見解を示している

(b) 信頼性

客観的試験／主観的試験問題に対する信頼性については、教育測定論において、統計学的な指標があるが、前節で述べた各種客観式試験、論述式試験 (= 主観的試験) において、信頼性についての統計学的指標のためのパラメータが異なる。

客観的試験：受験者数、設問数
論述式試験：受験者数、設問数、採点者数

論述式試験の場合、採点者に関する 2 つ問題が信頼性を低下させる大きな要因となることが報告されている [9, 10]。この問題は、(a) 採点者内相関、(b) 採点者間相関という 2 つの指標で評価される。(b) は答案に対し複数採点者で得点が異なる度合いを示す指標で、(a) は 同一採点者の採点結果の一貫性を示す指標である。これら 2 つの指標に与える要因として、採点者および設問の仕方に関してそれぞれ以下の要因が指摘されている [6, 11]。

採点者：評価の観点、答案に対する解釈
設問：答案記述の字数制限、知識

上記要因は、説明文、設問文と密接に関連し、個々の試験問題個別であるため、詳細には論及されていない。一方、これら 2 つの指標の値を高める、言い換えれば信頼性を高めるための全般的な方策として以下が提案されている。

(a) 採点者数と (b) 設問数の増加、
(c) 採点のカテゴリー化（例：5 段階、 7 段階）

これらの方策について、一般化可能性理論に基づく宇佐美らの実証的研究は以下の結果を得ている。

・採点者数は 4 名以上では効果が低下する。
・(a) よりも (b) の方が効果が大きい、
・(a) に対する (b) の最適な数値
・：5 段階が適切；（離散値による情報損失を抑制)

上記は、試験終了後の実施すべき信頼性向上のための方策であるが、試験開始前の方策として以下が提案されている。

(a) 採点基準の事前作成
(b) 採点基準に関する事前協議

前者は、答案の評価に対する採点者共通の制約を設定するものである。阿久津らの的研究では、採点基準の有無により採点間の相違について有意な差があったと報告している [6]。

小論文（essay）を試験に課している米国の全国学力調査（NAEP: National Assessment of Educational Progress）では、(b) を実施している。国内において (b) の有無による採点者間相関の実証的研究では、相反する結果が報告されている [10]。

論述式試験支援システム

序論で記した論述試験を支援するシステムの研究は、

米国では Essay Scorning System (ESS) なる名称で、1960 年代から行われている [12]。米国で開発された ESS の対象とする論述式試験は、前章の分類における小論文（essay）であり、短答式ではない。すなわち、指定されたテーマを主題とする複数の文から構成されるパラグラフとも捉えられる。

ESS の構造

ESS は主に 3 つのモジュールから構成される。

(1) 自然言語解析、(2) 特徴解析、(3) 評価予測

(1) で、論述式問題を構成する素材文、設問文、および小論文 (解答) に対し、品詞辞書、統語規則データを用いて形態素解析、構文解析、パラグラフ解析を行い、構成単語、句、文、文間のつながりを同定する。(2) では、(1) の結果に基づき、以下に挙げる特徴を同定する。

語の意味、語の使用、構文構造（文体）

文間のつながり（段落構造）

この段階で、綴りの誤り、語の使用傾向、文体などが明らかになる。(3) では、(2) の結果をもとにシステムに設定された評価基準に基づき対象小論文の得点を予測する。個々のシステムにより、評価基準、および得点予測の手法が異なる。(3) において、小論文に使用された用語の妥当性を評価するために、素材文の内容に応じて、百科事典、専門書などから用語（とその出現頻度データ等）が収集される。得点予測は、(2) において特徴が多変量であることから、(3) では評価手法として心理学で用いられる重回帰分析を使用するシステムが多い。また、(3) の処理おいては、人間の採点者との相関を高めるために、過去の大量の採点データが用いられる [12]。

米国の状況

米国では、1990 年代にいわゆるビジネススクール入学試験 GMAT で、小論文が課されており、この採点に初期の実用的な ESS の 1 つである”e-rater（Electronic Essay Rater）が採点支援ツールとして用いられた。1990 年代に開発された米国の個々の ESS の特性、特に評価手法について詳細な説明が石岡のレビューに解説されている [12]。

表 1 に米国における現在の主要な ESS とその特徴を掲載する [13]。表 1 に示されるように、ESS で採用している評価手法の多くは、1990 年代に開発されたシステムで用いられている確率・統計学的手法であ

表 1. 米国で現在運用されている ESS(13) の表 1 で末尾の JESS を削除)

評価システム	開発	評価基準	評価手法	特記事項
AutoScore	American lnstitutes for Research （AIR）	意味概念／段落間の意味的つながり／語の多様性／文法エラー	統計的手法	採点基準は論題依存
LightSIDE	カーネギーメロン大学	内容／文体／構造／態	教師あり機械学習	オープンソース
Bookette	CTB/McGraw-Hill	構造／文法／意味／技巧	ニューラルネット	90 の特徴量
E-raterT	ETS	構造／組織化／内容	重回帰モデル	12 の評価指標
Lexile Writing Analyzer	MetaMetrics	語彙使用の多様性／繰り返し使われる語彙の出現度合／文章としての流ちょうさの抑制	統計的手法	学年（grade），ジャンル，論題，句読法（punctuation）によらない
PEG	Measurement Inc.	構造／組織化／形式／技巧／独創性	重回帰モデル	意味理解に着手中
IEA	Pearson Education	内容／文体／技巧	潜在的意味解析 (LSI)	論理構成／語の出現順を評価しない
CRASE	Pacific Metrics	アイデア／文章の流ちょうさ／組織化／態／語彙選択／慣習／プレゼンテーションのうまさ	機械学習＋統計(ベイズアプローチ)	Java 言語で実装
lntelliMetric	Vantage Learning	一貫性／内容／構成／文章の複雑さ／アメリカ英語への適応	ルール発見	論題ごとに大量のデータが必要

ることがわかる。IEA が採用している LSI（潜在意味索引、Latent Semantic Indexing）は情報検索の古典的検索モデルの 1 つであるベクトルモデルの発展形である。また今世紀に入り開発されたシステムでは機械学習、ニューラルネットが用いられている。

日本の状況

米国の ESS と同じく小論文を対象として評価をするシステムとして 2002 年に石岡、亀田によって開発された JESS がその嚆矢として挙げられる [14]。JESS は採点基準、評価手法共に e-rater に準拠して開発されている。

JESS の開発以後、小規模の ESS の研究開発がされていれた。これらの研究は、以下のような運用状況での運用を想定している。

(a) 少人数の受験者（高校、塾の教室規模）[15]
(b) 問題出題者、採点者の支援 [16-18]

(b) を主眼とした、三重大学のグループの一連の研究は、e-Learning システムの機能の拡張を意図したものである。具体的には、講義の過程において、web 上で教員（出題者かつ採点者）が短答式試験を課し学生が解答し、リアルタイムで解答・模範解答を表示するという利用を想定し開発されている [17]。このため、問題の提示、素材文、模範文中のキーワード、などについての表示インターフェースに主眼がおかれている [17]。さらに、当該システムを実際の講義で使用し、システムとしての短答式試験の評価もされている [18]。

一方、序論でふれた「大学入学共通テスト」で論述式試験が導入される予定との報告を受けて、米国の ESS の日本語版の開発を指向した研究・開発が現在されつつある。但し、米国版とは異なり、評価対象は小論文ではなく、短答式問題である。

上述の小規模の ESS ではなく、大量の試験結果を対象とした ESS のプロトタイプが石田、亀田らによって開発された [19]。このシステムは、予測手法として機械学習アルゴリズムの一種のランダムフォレストを用い、理科／社会の 8 問の短答式問題について評価実験を行っている。

寺田らの研究 [20] では、複数の機械学習アルゴリズム、SVM (Support Vector Machine)、CNN (Convolutional Neural Network) 等を評価手法として用い、7 問の短答式問題 (解答の平均文字数：約 32 字) に対して実験した結果、87 ～ 98% という高い精度を得ているが、評価は正誤の 2 値である。

上述の 2 つの研究での評価は、それぞれ個別の試験問題に対してであり、また採点法も異なっている。岡山大学の研究グループは、以下の 2 項目についてデータ収集を目指して研究している [21-23]。

(a) 共通のシステム試験条件
(b) 汎用的な用語データと重みの計算手法

(a) はシステムの評価で使用される問題、模範解答、答案データはシステム個別で、同一の問題に対してではない。他の研究者が利用できる共通の問題・模範解答・答案データの構築を目指したものである。(b) は、論述式試験が今後多くの科目で導入されると仮定した場合、網羅的な分野の用語データがあることが望ましい。Wikipedia をコーパスとして利用し、問題の分野に応じた計算手法を提案している。

討論

前章で述べたように、米国では小論文、日本では今後短答式である。論述式試験形式の諸問題で引用した実証的研究で対象とされた出題形式も小論文であり、短答式形式に対する教育測定論分野での研究が期待される。

教育測定論的アプローチの限界

前章の論述式出題形式の諸問題で取り上げた問題、特に信頼性の問題および、信頼性向上のための方策は、短答式に対しても適用可能であるが、試験問題作成後に対するものである。

論述式試験問題は、以下の 3 種類の文から構成される。

(1) 課題文、(2) 設問文、(3) 模範文

(2) は解答文についての記述内容を指示するが、指示の仕方によって、解答文の内容は変更しうる。この側面からの研究としては安永らの研究 [24] があるのみであり、言語表現の観点からの上記の 3 つの関係についての研究が必要とされる。

評価基準の問題

小論文は複数の文から構成されるため、構造的な特性、個々の文の妥当性などの観点からの評価が必須で複数の基準から評価される。表 1 に挙げる評価基準はこのことを示している。一方、事実上 1 文だけからなる短答式問題は、これらの評価基準のほとんどが適用できない。短答式問題の主要な評価基準は、以下とされる。

正解文（模範文）と解答文の意味的同義性

短答式問題に対し、上記基準に基づく多くの実証的な研究が必要される。

謝辞

本研究は研究課題 2018 年度神奈川大学総合理学研究所共同研究助成に対する「記述式解答の自動採点

に向けた日本語文解析手法と採点方式の研究」（RIIS 201802）を受けて行った。記して感謝する。

文献

1) 大学入試センター (2017) 大学入学共通テスト実施に向けた検討状況. [http://www.dnc.ac.jp/daigakunyugakukibousya-gakuryokuhyoka_test/progress.html].
2) 大学入試センター (2017) 大学入学共通テスト・平成29年度試行調査・問題，正解表，解答用紙等. [http://www.dnc.ac.jp/daigakunyugakukibousya-gakuryokuhyoka_test/pre-test_h29_01.html].
3) 平成29年度試行調査．大学入試センター (2017) [https://www.dnc.ac.jp/sp/daigakunyugakuki-bousyagakuryokuhyoka_test/pre-test_h29.html].
4) 宮本友弘，倉元直樹 (2017) 国立大学における個別学力試験の解答形式の分類．*日本テスト学会誌* **13**: 69-84.
5) 大野木裕明 (1994) *テストの心理学.* ナカニシヤ出版.
6) 阿久津洋巳，菊池　梢，鈴木安澄，鈴木　光，渡邊愛枝 (2006) 論述式テストの研究 (1) －採点者間の一致度－．*岩手大学教育学部付属教育実施総合センター研究紀要* **5**: 115-122.
7) 宇佐美慧 (2012) 論述式テストを通した評価と選抜の信頼性に関する諸要因の影響力についての定量的比較検討．*日本教育工学会論文誌* **36**: 451-464.
8) 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**: 145-164.
9) 荒井清佳，石岡恒憲，宮埜壽夫 (2013) 大学入学者選抜における小論文試験と教科・科目試験との関連について．*日本テスト学会誌* 9: 27-36.
10) 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．*岩手大学教育学部付属教育実施総合センター研究紀要* **16**: 61-70.
11) 平井洋子，渡邊　洋 (1994) 小論文評点のカテゴリ化に関する測定論的考察．*計量行動学* **21**: 21-31.
12) 石岡恒憲 (2016) 記述式テストにおける自動採点システムの最新動向．*行動計量学* 31: 67-87.
13) 石岡恒憲 (2016) コンピュータ上で実施する記述式試験―エッセイタイプ，短答式，マルチメディア利用について―．*電子情報通信学会誌* 99: 1005-1011.
14) 石岡恒憲，亀田雅之 (2002) コンピュータによる日本語小論文の自動採点システム，*電子情報通信学会技術研究報告* No.TL2002-40: 43-48.
15) 篠田有史，中山弘隆，松本茂樹 (2007) 文の構造を利用した記述式問題の自動採点．*コンピュータ&エデュケーション* **22**: 41-44.
16) 高瀬治彦，川中普晴，鶴岡信治，森田直樹 (2013) 記述式小テストの解答群の分析手法 ―解答群からのキーワード自動抽出―．*コンピュータ&エデュケーション* **22**: 46-49.
17) 大庭知也 (2014) 記述式小テスト支援システム - キーワードの用いられ方の可視化．*PC Conference*: 54-57
18) 大庭知也 (2015) 多人数クラスにおける記述式小テストを支援するシステム ―学生の理解状況をすばやく把握するためのインターフェイス―．*コンピュータ&エデュケーション* **39**: 86-91.
19) 石岡恒憲，亀田雅之，劉東　岳 (2016) 人工知能を利用した短答式記述採点支援システムの開発．*電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション.* pp.87-92.
20) 寺田凜太郎，久保顕大，柴田知秀，黒橋禎夫，大久保智哉 (2016) ニューラルネットワークを用いた記述式問題の自動採点．*第22回言語処理学会年次大会発表論文集*．pp. 370-373.
21) 泉仁宏太，竹内孔一，大野雅幸，田口雅弘，稲田佳彦，飯塚誠也，阿保達彦，上田　均 (2017) 小論文採点支援のための関連文書取得法の考察．*電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション*．pp.47-51.
22) 竹内孔一，大野雅幸，泉仁宏太，田口雅弘，稲田佳彦，飯塚誠也，阿保達彦，上田　均 (2017) 小論文の自動採点に向けたオープンな基本データの構築および現段階での自動採点手法の評価．*言語処理学会第23回年次大会発表論文集*．pp.839-842.
23) 大野雅幸 (2018) 小論文自動採点データ構築と理解力および妥当性評価手法の構築．*言語処理学会第24回発表論文集* . pp.368-371.
24) 安永和央，石井秀宗 (2011) 国語読解テストにおける設問文中の単語の難しさが能力評価に及ぼす影響－具体例を回答させる設問の検討．*名古屋大学大学院教育発達科学研究科紀要（心理発達科学）*．**58**: 105-112.