小特集
Special Feature

［AI 採点システム］

3 AI 採点システムが変える大学入試

石岡恒憲　（独）大学入試センター・研究開発部

自動採点は可能か

一般に記述式問題は多肢選択式に比べ測りたいことを直接的に測っているとされ，適切かつ短時間に採点ができるのなら，利用価値が大きいと考えられている．文部科学省も「思考力・表現力・判断力等」を測定するため，大学入学共通テストでの導入を検討してきた．

これに対し，同省の「大学入試のあり方に関する検討会議」は 2025 年以降の共通テストの国語・数学への記述式問題導入は実現困難だが，個別大学の入学者選抜では出題を推進するよう提言した．同省は記述式問題の充実も含めて，選抜の改善に取り組む大学を支援しようとしている．

こうした流れの中で，今後，国内の大学入学者選抜でも AI 自動採点を使用する動きが出てくる可能性がある．入試では限られた時間とマンパワーで適切な採点を行うことが求められるからだ．本稿では「自動採点は可能なのか，課題はないのか，個別入試で活用して問題ないのか」について論じたい．そのために，AI による自動採点の現状，海外の先行事例，国内の取り組み状況，自動採点導入時の課題について紹介したい．

AI による自動採点の現状

一言で AI 自動採点といっても，取り扱うべき対象が 800 字程度のエッセイなのか，30 字から 120 字程度の短答記述なのか，数式かなどによって用いる技術も，その実用性の程度も大きく違ってくる．

(1) エッセイ採点

エッセイ自動採点システムが盛んに開発されていた 2000 年初頭は「作文能力（Writing Ability）」を評価することが主流であった．たとえば「富と名声とどちらが重要か？　あなたの経験に基づいて述べなさい」といった論題に答えるもので、そこに正解はない．このような作文能力の評価ツールとしての採点システムなら，教育訓練を受けた人間評価者と同等であるとの多くの実証結果 [1] が出ている．E-rater, IntelliMetric, CRASE などの実用システム [1] はすでに商用として広く利用されている．日本語なら筆者が開発した Jess[2] があり，いまや誰でもフリーで利用できる．これらエッセイ自動採点システムは，システムによって多少の違いはあるが，主に修辞（文法，語彙など），論理構成，内容の適切性などを評価する．

全米最大のテスト機関である ETS が開発した e-rater は，現在 TOEFL iBT や GRE（アメリカのビジネススクール入学のための共通試験）の採点に人との併用として使用されている．その一方で，e-rater は 2016 年以降，SAT（アメリカの大学入学のための共通テスト）エッセイの採点システムとして採用されていない[☆1]．これは当時の SAT のエッセイの問いかけ形式が「明確な正解を求める形式で

[☆1] SAT エッセイは 2021 年 6 月以降提供されないことが決定した

あって単なる自らの論理展開をするだけでは済まない」ためだと思われる．つまり素材文の正しい読解（＝理解）と分析（＝解釈）が SAT エッセイには求められており，e-rater はまだそれに対応できていないためだろう．実際, e-rater の開発者であるジル・バーシュタイン（Jill Burstein）博士によれば，「大学レベルの作文能力における技能や知識を検査することをほとんどの研究はできていない」[3] と述べている．現代のエッセイでは「出題者が意図する正解を求める」形式の問題が出されるようになってきており，この適切性への評価はいまなお難しい．

作文能力を評価する自動採点については人間採点と遜色のない妥当性を持つと評価される一方で多くの批評や反論がある．自動採点の批評家として有名なマサチューセッツ工科大学のレス・ペレルマン（Les C. Perelman）教授によれば，測定できない本質に以下があるとしている [4]：

- 書かれている内容の確かさ（accuracy）
- 論法（reasoning）
- 証拠の適切性（adequacy of evidence）
- 良識（good sense）
- 倫理的スタンス（ethical stance）
- 説得力（convincing argument）
- 意味のある組織化（meaningful organization）
- 明瞭性（clarity）
- 誠実さ（veracity）

自動採点がエッセイの本質を評価できないというこのような批判は相当に的確である．自動採点の利用は，結局，実用的に十分だという妥当性や正当性（validity）を是とするのか，それとも人間のみが可能な本来測りたいものを直接的に測る真正性や正統性（authenticity）を是とするのか，といった哲学的な問題に帰着するといってよいだろう．他方,「正解を求めるタイプ」の問題への自動採点システムはいまだ実用的に十分ではない．

(2) 短答式記述採点

一般の人の直感とは異なり，短答式記述採点は作文能力を評価するエッセイ採点よりはるかに難しい．模範解答との意味的一致性や含意などを評価する必要があるためだ．このためユーザが準備した任意の試験問題を採点するための実用的な商用のシステムはおそらくまだない．ただこれについては 2019 年に米グーグルが開発した言語モデルである BERT により性能が格段に向上している．先頭文字の B は **B**idirectional の略で，文章を双方向（文頭と文末）から学習することによって「文脈を読むこと」を実現している．

代々木ゼミナールの高校「国語」模試記述テストの解答はこの BERT を利用して採点されたが，正しくタグ付けされた解答データ（解答文字列のどの部分がどの採点基準に合致して何点が付与されたかという情報を含むタグを解答文字列に埋め込んだもの）が各設問で千件程度あれば，人間並みの採点は可能だとしている．ただ実際の採点現場において，このタグ付け作業は，かなりの手間といわざるを得ない．通常の個別入試で千件という数字は受験者数に比肩する相当の数だろうから，タグ付けして残りの答案を自動採点するというスタイルは現実的ではない．ユーザフレンドリーなインタフェースを持つタグ付けシステムが別に構築されていてもなお，そうだろう．また，機械学習した採点ルールに載らない解答は，ある一定の確率で必ず出る [5]．一般には十分であるかもしれない 96％の一致率では大学入学者選抜の採点としては恐らく許容してもらえない．

BERT では，解答文字列のどの部分が採点に寄与したかの色付けは可能である．この注目部分はアテンションと呼ばれる．図 -1 は，代ゼミ「記述式を AI 採点する現代文トレーニング」の出力例である．A, B, C, D の 4 つの採点基準に対し，解答例において対応した記述部分に色付けがされている．これは採点根拠の提示になるだろう．

(3) 数式認識

数式を記入させる問題においては，解答者が電子データで解答を入力できれば自動採点も可能だ

が，数式入力のための記述言語である MathML や LaTeX での入力を求めるのは非現実的である．数式エディタの利用も限られた試験時間の中では難しく，結局，タブレット等による手書き数式認識が実際的だろう．ただ複雑な数式の認識は最新の技術をもってしても難しい．東京農工大学・中川正樹特任教授の研究グループは 2016 ～ 2017 年に実施した大学入試センター共通テストの試行調査での数式解答を対象に，筆者らとの共同研究のもとで認識実験を行った．彼ら [6] によれば，ノイズや太さ，ぼやけがない対象の認識エンジンに，試行調査のぼやけやノイズのある正解ラベルなし数式解答画像約 1 万パターンにより転移学習を施し，1 千のテストパターンで 47.2％の正解率を報告している．

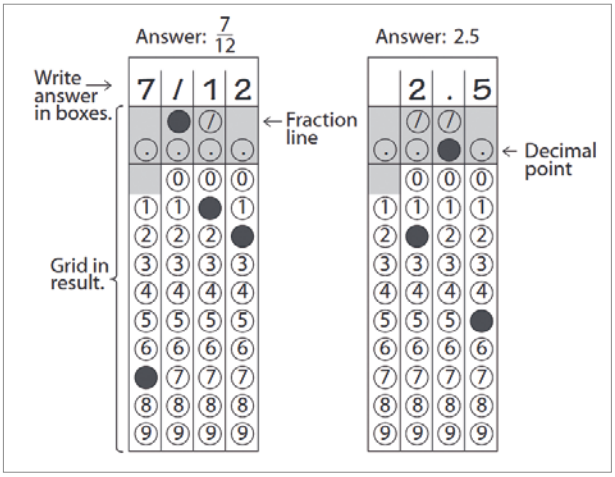
実際の試験で数式入力を行った場合に，認識結果をその場で確認できることはだぶん必要だろう．解答者の意図通りに機械が認識できればそのままでよいが，認識が間違っていた場合には，数式全部を書き直すのでなく，一部のみを消しゴムで消すようなイメージで修正したのち再認識という手順が好ましい．ただこの一連のユーザ操作については，標準的な手順や規格というものが現時点ではない．せいぜいあるのはペン入力のマーク付け言語 InkML（Ink Markup Language）の規格のみである．この規格はペンの動きのトレース，ペンの角度や圧力，そのコンテクストといったものを XML としてマーク付けするものである．

数式認識が難しいなら，いっそのことマークシートの記入の仕方を工夫するのも 1 つの方法だろう．現在の大学入学共通テストでも数学はマークシートで解答しており，一般の多肢選択以上の多様な形式の解答を可能にしている．またアメリカの共通テスト SAT の数学では，一部の解答は受験者が自ら作る Student-Produced Response という形式がある（図 -2）．グリッド・イン（Grid-ins）とも呼ばれるこの形式では，数字に加えてスラッシュと小数点をマークすることができる．図 -2 は 7/12 と 2.5 を示す 2 つの解答例である．小数の 2.5 は 5/2 と解答しても正解となる．多肢選択に比べ，解答のバリエーションは飛躍的に多くなる．

図 -1　代ゼミ「記述式を AI 採点する現代文トレーニング」の出力例 (https://prtimes.jp/main/html/rd/p/000000005.000072274.html)

海外の先行事例

多様なタイプの記述問題の自動採点を行っている最も有名な例の 1 つが TOEIC（Test of English for International Communication）ライティングテストであろう．TOEIC とは英語を母語としない者を対象とした，英語コミュニケーション能力を測定する世界共通のテストである．聞く・読む力を測る「リスニング & リーディングテスト」と，話す・書く力を測る「スピーキング & ライティングテスト」がある．本稿では，このうちライティングテストを取り上げる．ライティングテストにおけるテストの構成は表 -1 に示す通り 3 つのタイプの問題が出題される．



■図 -2　SAT における数学の解答例（SAT 公式ガイドより抜粋）（https://satsuite.collegeboard.org/media/pdf/sat-practice-test-1.pdf）

これら 3 つの記述解答については機械によって採点される．このうち「E メール作成問題」と「意見を記述する問題」については，いわゆるエッセイタイプの問題として e-rater が採点する．残りの「写真描写問題」は，正解のあるいわゆる短文記述式解答であるが，これについて自動採点を可能ならしめる仕組みは語彙指定である．実例を挙げる．図 -3 は TOEIC ライティングテストのサンプル問題である．写真の絵を見て airport terminal と so の指定

■表 -1　TOEIC ライティングテストの構成（https://www.iibc-global.org/toeic/test/sw/about/format.html より抜粋の上，一部転載）

内容	問題数	解答時間	課題概要	評価基準	採点スケール
写真描写問題	5	5 問で 8 分	与えられた 2 つの語（句）を使い，写真の内容に合う一文を作成する	• 文法 • 写真と文章の関連性	0 ～ 3
E メール作成問題	2	各問 10 分	25 ～ 50 語程度の E メールを読み，返信のメールを作成する	• 文章の質と多様性 • 語彙 • 構成	0 ～ 4
意見を記述する問題	1	30 分	提示されたテーマについて，自分の意見を理由あるいは例とともに記述する	• 理由や例を挙げて意見を述べているか • 文法 • 語彙 • 構成	0 ～ 5

■図 -3　TOEIC ライティング写真描写問題；airport terminal / so の 2 語を指定（Copyright © 2018 Educational Testing Service. www.ets.org. Reprinted by permission of Educational Testing Service, the copyright owner.）

された 2 語を用いて写真の内容を描写する．“There are so many cars parked at the airport.” と解答すると満点の 3 点が得られる．

写真描写問題を含む TOEIC ライティングテストの問題数と評価基準は表 -1 に示す通りであるが，これを見ると写真描写問題で求める観点はきわめてシンプルで，文法と写真との関連性のみであることに気づく．文法の誤りについては多くの英文チェッカーや英文校正ツールがすでにあり，写真との関連性については潜在的意味解析やトピックモデルなどの既存の技術が利用できる．この自動採点を可能とするのはひとえに 2 語の語彙指定のおかげだろう．これにより内容の正確性や妥当性を，単にその指定語句が含まれているか否かで代替できるようになる．

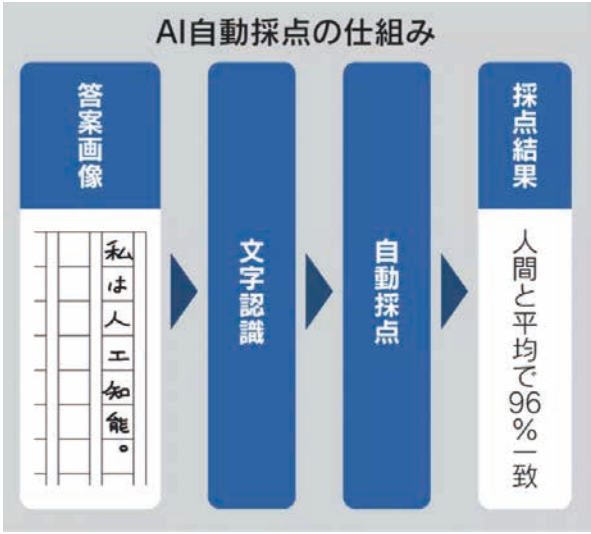
国内の取り組み状況

自動採点を大学入試に取り込むことを検討している国内の事例はまだまだ少ないが，ここでは 3 つの事例を取り上げる．

岡山大学では岡山大学運営費交付金機能強化経費「小論文，エッセイ等による入学試験での学力の三要素を評価するための採点評価支援システムの開発導入」を受けて，2015 年から研究を開始した．平成 31（2019）年度に全国初の小論文自動採点を組み込んだ入試（マッチングプログラム型入試，岡大 AO）を企画するも，単なる作文力を評価するのではない正解を求めるタイプの出題であったために，十分な精度が出ず，入試に採用するには至らなかった．そこでは 4 つの講義「グローバリゼーションの光と影」，「自然科学の構成と科学教育」，「東アジア経済の現状」，「批判的思考とエセ科学」を受講し，それぞれに対して設定された各 3 つの設問に答える．調査に用いた試験問題や解答データは「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」として現在，公開されている．このデータを用いた自動採点研究については岡山大学・竹内孔一准教授らの研究グループが引続き精力的に行っている．

代々木ゼミナールでは，東北大学・理化学研究所の乾健太郎教授らとの共同研究により，過去の代ゼミ現代文の模擬試験から，評論・随筆を中心に 9 題をピックアップし，問題集として公開している．記述式問題は AI 自動採点を行い，その結果を解答者にフィードバックする．

筆者と東京農工大学・中川特任教授との研究グループでは手書き文字解答から自動採点までを一気通貫で行い，人手を一切用いない方法 [5] での性能を評価している（図 -4）．通常の AI システムでは汚い文字や消し跡汚れに起因する文字認識の誤りを人手で修正したり，採点時の細かな判断基準とその適用をタグ付けして事前に機械に教え込む．こうした「理解の補助輪」を使うことで採点精度が上がるが，採点時間に制限のある大規模試験では現実的ではない．筆者らは補助輪なしの実運用で平均 96%，最低でも 93%の一致率を確保している [5]．言語モデルには標準的な BERT を用いている．



■図 -4　AI 自動採点システムの仕組み（筆者寄稿，日経新聞，2021 年 12 月 21 日，教育面掲載）

AI 採点導入時の課題

受験者の解答入力の課題

入力をどのようにするのか，CBT（Computer Based Testing）にするのか，その場合キーボード入力にするのか，タブレットにするのか，PBT（紙と鉛筆）か，によって存する問題はさまざまである．導入に際しては，その課題を知っておく必要があるだろう．

(1) キーボード入力

3 点を指摘しておく．1．キーボード入力のできない，あるいは苦手な学生は多い．多くの学生はスマートフォンのフリック入力に慣れているからだ．2．横書きでよいのか．国語の問題は今後も縦書きで出題されるだろう．古文や漢文では余計にそうである．解答を横書きで書くことの妥当性は問われるし，不自然でもある．3．一斉実施なら機器の台数が限られる．一斉実施でなければ問題漏洩が懸念される．

(2) タブレットによる手書き入力

かなり安価な入力装置が期待できるが，文字の認識率は求めるほどには高くない．たとえば認識率 96％はオフィス業務では十分かもしれないが，大学入学者選抜の試験においては 100 字解答で 4 個の誤字は実用には耐えない．

(3) 鉛筆による手書き入力（紙に記入し，スキャンして文字認識）

消しゴムによる消し跡の汚れが誤認識を招く．このためタブレット入力に比べ認識率をより低下させる．30 字を超える（場合によっては 100 字程度までの）文字解答を消しゴムなしに一発で解答できる受験生はほとんどいない．

(4) 手書き文字認識共通の問題

3 点を指摘しておく．1．試験では字は丁寧に書かれない．これが認識率の低下を招く．2．間違った文字も正しい文字に変換してしまう可能性がある．たとえば「完璧」→「完璧」,「酒落」→「洒落」,「三味」→「三昧」などがそうである．普段仕事で使う実用上のシステムではこれで構わないが，テストにおいては間違ったものは間違ったものとして認識されなければならない．3．数字への誤認識が目立つ．最新の言語モデルは著作権の問題や大量データの必要性から Wikipedia で構築することが多い．そこでは数字が多用されるために縦書きの音引きを数字の 1 に，平仮名の「る」を数字の 3 に誤認識する例が散見される．

字数制限の課題

アメリカではエッセイに字数制限はない．時間制限 30 分程度はある．このためアメリカでは書いた単語数が採点に大きく寄与する．作文力を測定したいならこの方が適切だろう．日本ではどのようにすべきか検討の余地がある．

短答記述では字数制限は設定されるだろう．80～120 字の解答は表現のバリエーションが多く，膨大な採点済みの解答データがルール学習には必要だ．数千では不足する．BERT を用いた場合，この程度のサンプルサイズでは学習は十分に収束しないことが，我々の 6 万人規模の試行調査データ結果から確認されている [5]．

採点を機械に任せて大丈夫か

エッセイにせよ，短答記述にせよ，採点に丁寧な時間をかけたプロの専門の評価者に現在の AI はなかなかかなうものではない．ただ採点の枚数が増えると，人間も疲れてしまい，なかば機械的に評価してしまう．機械的に評価するなら機械にまかせればいいし，その方がミスもなくなる．要は利便とリスクのバランスということなのだろう．

エッセイ自動採点システム e-rater の公的試験における運用では，2 人の人間の評価者が独立に採点する代わりに，一方を機械が採点し，もし判定に大きなずれが生じたなら，最終点を第 3 の別の「人間」

が判定し，最終得点を出す．最終判断は機械でなく人間ということを担保しないと，受験者の納得は得られないだろう．

また自動採点を行ったら，その採点の論拠を示す透明性が求められるだろう．米国教育研究協会，米国心理学会（APA），全米教育測定評議会（NCME）の共同出版である試験基準（Testing Standards）の最新版（2014 年版）によれば，自動採点に関して以下の記述がある．

基準 4.19：複雑な被験者解答を評価するために自動採点が使われる場合，各スコアレベルにおける特徴や特性値が，アルゴリズム使用の理論および経験に基づき文書化されていなければならない．

基準 6.8：テスト採点の責任者は採点の手順（プロトコル）を確立しなければならない．複雑な応答をコンピュータが採点したなら，アルゴリズムやプロセスの正確性が文書化されていなければならない．

日本ではこれに相当するテスト基準の項目はないが，今後要求されることは必然である．自動採点が個別の大学入学者選抜試験に適用されるためには模範解答やあらかじめ用意された採点基準から，人手による採点済データなしに適切な採点を可能にする必要がある．つまり模範解答と採点基準のみから答案を自動的に採点するのだ．そのための研究はすでに始まっている．

参考文献

1）Shermis, M. D. and Burstein, J.（Editor）: *Handbook of Automated Essay Evaluation : Current Applications and New Directions,* Routledge, 1st edition（2013）.
2）Ishioka, T. and Kameda, M. : Automated Japanese Essay Scoring System based on Articles Written by Experts. *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics（Coling-ACL 2006）*, pp. 233-240（2006）.
3）Burstein, J., McCaffrey, D., Klebanov, B. B. and Ling, G. : Exploring Relationships between Writing & Broader Outcomes with Automated Writing Evaluation, *Proceedings of the 12th Workshop on Innovative Use of NLP or Building Educational Applications,* pp.101-108（2017）.
4）Perelman, L. C. : Critique（Ver. 3.4）of Mark D. Shermis & Ben Hammer, Contrasting State-of-the-Art Automated Scoring of Essays: Analysis（2013）.
https://graphics8.nytimes.com/packages/pdf/science/Critique_of_Shermis.pdf
5）Oka, H., Nguyen, H. T., Nguyen, C. T., Nakagawa, M., and Ishioka, T. : Fully Automated Short Answer Scoring of the Trial Tests for Common Entrance Examinations for Japanese University, *Artificial Intelligence in Education,23rd International Conference, AIED 2022*, Durham, UK, Proceedings, Part I, Springer, pp.180-192（July 2022）.
6）Ung, H. Q., Nguyen, H. T., Nguyen, C. T., Ishioka, T. and Nakagawa, M. : Visual Constraints for Generating Multi-Domain Offline Handwritten Mathematical Expressions, *IEICE Technical Report* PRMU2021-69（2022-03）, pp.54-59（2022）.

（2023 年 1 月 10 日受付）

■石岡恒憲　tunenori@rd.dnc.ac.jp

（独）大学入試センター研究開発部部長・教授（現職）．1992 年博士（工学）．2000 年文部省長期在外研究員（カーネギーメロン大学，Language Technologies Institute）．2012 ～ 2016 年東京工業大・社会理工学研究科連携教授．2022 年～現在，東京農工大・客員教授（兼任）