# 英語教師が知っておきたいICTとテストの話

石井　雄隆

## 1．はじめに

本稿では，ICTを活用した言語テストの現状と課題について考えたいと思います．具体的には，TOEICやTOEFLなど身近な英語試験で使われている項目反応理論の考え方や現在盛んに研究が進められている自動採点研究の仕組みを題材に，英語教師が知っておきたいICTとテストのことについて考えていきたいと思います．2020年度から実施される大学入学共通テストでは，民間の試験を活用しながら，これまでのように「読む」，「聴く」だけではなく，「書く」，「話す」能力を含んだ4技能を総合的に評価すると言われていますが，それらはコンピュータを用いたテストになると言われています．ICTを活用した言語テストには，どのような利点と欠点があるのでしょうか．本稿では，項目反応理論が用いられるコンピュータ適応型テスト，自動採点を事例として考えていきたいと思います．

## 2．コンピュータ適応型テストの現状と課題

コンピュータ適応型テストとは「コンピュータを使用し，項目応答理論によって事前に特性値が算出されているテスト項目を，各受験者の応答を適時判断しながら出題し，効率よく受験者の能力推定値を算出するテスト」(中村，2007)を指します．

コンピュータ適応型テストを理解するためには，項目応答理論について理解する必要があります．項目応答理論とは，「テストを構成する項目やテストの受験者集団に依存することなく，項目の困難度や識別力などの項目特性値やテスト受験者の能力値を算出するために，テスト項目に対する受験者の応答(例えば，正答か誤答かなど)とテストが測定しようとする能力や特性を表現する潜在特性尺度との関係に確率モデルを導入するテスト理論」(野口・大隈，2014)です．説明の際に，よく例として用いられるのは視力検査です．視力検査では，世界標準で決められた視力の定義を基に，徐々に検査者の視力に合った大きさのランドルト環に近づけていきます．視力検査のように，項目の困難度を事前に決定し，学習者の能力に合わせて，問題を出題していくのがコンピュータ適応型テストです．

コンピュータ適応型テストの強みはなんでしょうか．それは，いつ誰が受けても正確に学習者の能力を測定することができることです．アメリカのように西海岸と東海岸で時差が3時間あるような国では，日本のように全員が一斉に同じテストを受験することは不可能であるため，こういったテストはとても重要となります．

それでは，コンピュータ適応型テストの問題点はどんな点でしょうか．木村(2016)はとても面白い視点を提示しています．コンピュータ適応型テストは，アダプティブに学習者のレベルに合った出題をするため，学習者の正答率は50%に近くなります．日本人大学生を対象としたアンケートデータから，コンピュータ適応型テストを受験した学習者は「難しい問題が出ても問題数が少ないテスト」よりも「問題数が多くても易しい問題が出るテスト」を望んでいるということが明らかになりました．これは従来のテストと比べると，コンピュータ適応型テストは，受験者の動機づけや自己効力感などに好ましくない影響を与えるという可能性を示唆しました．

またコンピュータ適応型テストは，学習者の能力を測定するためのアイテムバンクの構築に大きな労力がかかることも指摘されています．

本節では，コンピュータ適応型テストの強みと弱みについて述べました．多くの学習者の能力を測定するのに，有用なコンピュータ適応型テストですが，テスト受験者の学習や動機づけという観点から考えると，好ましくない場合もあるということに留意する必要があると思います．

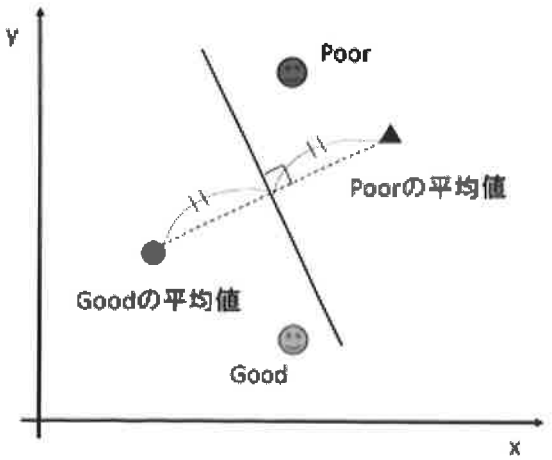
## 3．自動採点研究の現状と課題

### 3．1．自動採点研究の現状

本節では，自動採点研究の現状と課題について言及します．そもそも自動採点とは，どのような仕組みになっているのでしょうか．ここでは石井(2015)の例を基に説明していきたいと思います．図1は，自動採点の仕組みを簡単に図示したものです．例えば，ここでは二つの文法的誤り(x は動詞の語彙に関するエラー，y は語順に関するエラー)を基に Good, Poor の二段階でライティングを評価するという前提で考えてみます．

はじめに，人間の評定者が評価したデータを基に評価モデルを作成します．Good と Poor それぞれの評価を人手で付与されたライティングに，二つの文法的誤りがどれくらい含まれているかを計算し，x と y の平均値を求めます(図では●と▲で表されています)．その後，Good と Poor の平均値を図のように線で結び，垂直二等分線を引きます．これで二つの文法的誤りからライティングを二段階に評価する自動採点システムとなります．その後，インプットされたライティングが垂直二等分線のどちら側に配置されるかによって，ライティングの評価が決定します．単純なモデルですが，自動採点はこういった仕組みで行われます．

図1．自動採点の仕組み(石井，2015)




自動採点研究の利点とはなんでしょうか．大きく分けて二点あると思います．一点目は，大量の学習者のパフォーマンスデータを一斉に評価することができることです．人手による評価は，とても時間がかかるため，その問題を解決することが可能となります．

もう一点は，一貫した評価が行えるということです．石岡・亀田(2003)によると，測定において生じる誤差要因には下記のようなものが存在すると言われています．

書き手，題目，形式，制限時間，テスト状況，評定者
文字の巧拙
評定の系列的効果
課題選択
その他(書き手の性別，人種など)

人間による評価において生じやすいこういったさまざまな誤差要因を自動採点は防ぐことができます．これら二点が自動採点研究の大きな利点と言えます．

### 3．2．ライティング自動採点研究の現状

本節ではライティング自動採点の現状について論じます．本稿では Educational Testing Service で開発された e-rater を紹介したいと思います．

e-rater は，下記12個の変数を基に作文を自動評価します．

1. 総語数に対する文法エラーの割合
2. 総語数に対する語の使用法についてのエラーの割合
3. 総語数に対する手順のエラーの割合
4. 総語数に対するスタイルについてのエラーの割合
5. 必要とされる談話要素の数
6. 談話要素における平均語数
7. 作文を6点法で採点する際に語彙の類似度が一番近い点数
8. 最高点を取った作文との語彙の類似度
9. Type-Token Ratio
10. 語彙の困難度
11. 平均単語長
12. 総語数

(Burstein, Chodorow, & Leacock, 2004)

これらの変数をもとに重回帰分析と呼ばれる手法を用いて，ライティングを自動で評価します．重回帰分析とは，複数の変数を基に別の一つの変数を予

測する統計的な手法で，ここでは上述した12個のカテゴリーの言語的特徴からライティングの評価という一つの変数を予測します．

e-raterはどれくらいの精度で学習者のライティングを評価することができるのでしょうか．Burstein et al.(1998)によると，専門家とe-raterの評価の一致率は，87%から94%であったと報告されています．またe-raterを用いているCriterionというサービスでは，英語学習者のライティングを自動で採点し，診断的なフィードバックを返すことも可能になっています．学習者のライティング能力を自動で採点するシステムは，精度が高くなってきており，テストで運用される日も近くなってきています．

### 3．3．スピーキング自動採点研究の現状

ライティングに対して，スピーキング自動採点はどうなっているのでしょうか．スピーキング自動採点の現状について近藤・石井(2017)を例に挙げながら検討してみましょう．

スピーキングの自動採点を考える際に重要になってくるのが，音声認識です．Zechner, Higgins, Xi, and Williamson(2009)が報告しているSpeechRaterでは，学習者の単語の約半分程度しか認識できないと報告されていました．　しかしながら，TOEFL Juniorで利用しているSpeechRaterは発話が制限されているということもあり，タスクによって認識率は異なりますが，約70%から90%の精度で学習者の音声が認識できるということです(Evanini, Heilman, Wang, & Blanchard, 2015)．

近藤・石井(2017)では，日本人英語学習者の発話を自動で採点するためのシステムを構築し，英語教育プログラムにおける自動採点の実現可能性について検討しました．学習者の発話をある程度制限するために，談話完成タスクにおける自動採点を検討しました．談話完成タスクとは，下記のようなタスクを指します．

You (A) want to end your conversation. What would you say in the conversation below?
A: (　　).
B: See you.

（近藤・石井，2017）

本研究では，サポートベクターマシーンとナイーブベイズ分類機を用いました．詳細は近藤・石井(2017)を参照して頂きたいですが，どちらも図1で示したような分類のアルゴリズムです．

人間の評価者による点数とシステムが算出した点数は74%が一致しており，また本システムの単語認識率は71%でした．この結果から，発話の自由度をある程度制限することで，比較的高い精度で実用可能なレベルの発話自動採点が行えるということがわかりました．

### 3．4．自動採点研究の課題

上記で記したように自動採点は大きな可能性を秘めています．しかしながら，大きく分けて二つの問題点があると思います．

一点目は，コンピュータは学習者のパフォーマンスを理解することはできないということです．コンピュータは人間が評価をする際に用いている全ての要素を自動で計算することはできないでしょう(Xi, Higgins, Zechner, & Williamson, 2008)．二点目は，大量のデータが必要となるということです．システムを構築する際に，大量のデータをもとにモデルを作成する必要があるので，少数のデータからモデルを構築できるようになることが今後の課題と言えるでしょう．

英語教師が自動採点について考える際に重要な点は，人間にできること，できないこととコンピュータにできること，できないことを正確に理解することではないかと思います．

## 4．おわりに

本稿では，ICTを活用した言語テストの現状と課題について，コンピュータ適応型テスト，自動採点研究を事例として紹介しました．

最後に，これからの展望について言及します．Cope and Kalantzis (2016)は，これまでの評価は，学習プロセスの外に位置付けられてきましたが，学習者のビッグデータが大量に，そして自動で蓄積されるようになることで，評価は学習に埋め込まれるようになるのではないかと主張しています．言い換えると，総括的評価だけではなく，形成的評価がこれまで以上に根付くのではないかということを予測しています．

具体例としてはポートフォリオなどが挙げられるでしょう．ポートフォリオとは，文部科学省の定義によると，「学生が，学修過程ならびに各種の学修成果(例えば，学修目標・学修計画表とチェックシート，課題達成のために収集した資料や遂行状況，レポート，成績単位取得表など)を長期にわたって収集し，記録したもの」を指します．これをオンライン上で実装したものがｅポートフォリオと呼ばれます．森本(2015)によると，ｅポートフォリオは二つの役割を持つと言われています．一つ目は，「学習者の学習・評価を促進させるためのツールとしての役割」であり，二つ目は「学習者の学習成果を引証づけるためのエビデンスとしての役割」です．こういったｅポートフォリオなども今後ますます発展していくことが考えられます．

コンピュータ適応型テスト，自動採点，そしてｅポートフォリオの発展などは教室における英語指導に大きく影響を与えると思います．しかしながら，それぞれの利点と欠点を把握し，それらを基に適切な教室における英語指導を行うことが重要ではないでしょうか．

**参考文献**


Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine, 25* (3), 27-36.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. D. (1998). Automated scoring using a hybrid feature identification technique. In B, Christian., & Whitelock, P. (Eds.), *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (pp. 206-210). Burlington, MA: Morgan Kaufmann Publishers.

Cope, B., & Kalantzis, M. (2016). Big Data Comes to School: Implications for Learning, Assessment, and Research. *AERA Open, 2* (2), 1-19.

Evanini, K., Heilman, M., Wang, X., & Blanchard, D. (2015). Automated scoring for the TOEFL Junior® Comprehensive Writing and Speaking test (Research Report No. RR-15-09). Princeton, NJ: Educational Testing Service.

石井雄隆(2015).「データマイニングの手法を用いた英語ライティングへのアプローチ―日本人英語学習者のエッセイ評価に影響を与える文法的誤りパターンの検討―」『EIKEN BULLETIN』27, 28-39.

石岡恒憲・亀田雅之(2003).「コンピュータによる小論文の自動採点システムJessの試作」『計算機統計学』16(1), 3-18.

木村哲夫(2016).「コンピュータ適応型テストの心理的側面：目標正答確率を調整するシステムへの日本人大学生の反応」『統計数理研究所共同研究レポート』356, 47-56.

近藤悠介・石井雄隆(2017).「英語学習者の発話自動採点システムの開発と英語教育プログラムへの導入可能性の検討」『Language Education & Technology』54, 23-40.

森本康彦(2015).「ｅポートフォリオとしての教育ビッグデータとラーニングアナリティクス」『コンピュータ＆エデュケーション』38, 18-27.

中村洋一(2007).「コンピュータ適応型テストの可能性」『日本語教育』148, 72-83.

野口裕之・大隅敦子(2015).『テスティングの基礎理論』研究社.

Xi, X., Higgins, D., Zechner, K., & Williamson, D. (2008). Automated Scoring of Spontaneous Speech Using SpeechRater v1.0 (Research Report No. RR-08-62). Princeton, NJ: Educational Testing Service

Zechner, K., Higgins, D., Xi, X., & Williamson, D. M. (2009). Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication, 51,* 883-895.


(早稲田大学 助手)