「話す」,「書く」能力の評価と自動採点の可能性

近藤　悠介

1．はじめに

ここ数年で日本の英語教育は大きく変わろうとしています．2020 年より大学入学共通テストが導入され，英語に関しては「読む」,「聞く」,「書く」,「話す」の 4 技能が評価されるようになります．小学校においては 5，6 年生で英語が教科化され，2021 年度からの中学校の英語の授業は原則英語で行うよう学習指導要領で示されており，さらには，高校入試においても，「話す」能力の評価の導入が検討されているところもあるようです．こうした一連の教育政策によって，「話す」,「書く」といった産出能力に注目が集まっています．

本稿では，「話す」,「書く」能力の評価における問題点を指摘し，その解決策のひとつと考えられる自動採点システムの導入の可能性を検討します．

2．「話す」,「書く」能力の評価における問題点

「話す」,「書く」といった産出能力の評価は，「読む」,「聞く」といった受容能力と大きく異なります．産出能力の評価は，受容能力を評価する場合に比べ，かなりのコストがかかります．受容能力の評価では，多肢選択式など採点にあまり時間がかからない項目を用いても妥当性の高い試験を作成することが可能であるのに対し，産出能力の評価においては，採点を効率化できるような項目の作成は困難です．

産出能力の評価には時間がかかり，特に「話す」能力の評価では，受験者が実際に話している時間に加え，多くの場合，受験者の発話は録音され，試験後に評定者はその録音を複数回聞いて評価することになります．それに加えて，「話す」能力は他の能力と異なり，一斉に試験を行うことが困難です．基本的には，試験官が 1 人あるいは 2 人の受験者を面接し，個別に評価します．同じ産出能力でも「書く」能力は，問題用紙と解答用紙を一斉に配布し一度に数百人，数千人の試験を実施することが可能です．「話す」能力の試験の実施にかかる人的，時間的コストは他の能力を評価する試験に比べ膨大です．また，受験者が試験を受けて，その結果が返されるまでにはある程度の時間がかかり，短い期間で返却される試験の結果と比べると，そのフィードバックの効果は低いと言えるでしょう．

さらに，産出能力の評価においてはその評価の信頼性を保つことが重要になります．産出能力の評価は，いわゆる客観式と言われる多肢選択式の項目のみで構成された試験のように，誰が採点をしても同じ点数になることが保証されていません．産出能力の試験においては，多くの場合，2 人以上の評定者が受験者の発話あるいは作文を評価します．これらの評定者の評価が一致しなかった場合，最終的な評価はどのように決定すればよいでしょうか．しばしば用いられる方法として，まず初めに 2 人の評定者が評価を行い，その 2 人の評価が一致しなかった場合のみ 3 人目の評価者が評価を行い，多数決で評価を決める方法があります．3 人の評価がまったく一致しないことは考えにくいですが，できるだけ一致した評価が行えるよう評定者を訓練する必要があります．この訓練にもかなりのコストがかかることは容易に想像ができます．さらには，訓練を受けた評定者であっても 1 日に何十人もの評価を行えば，疲労による影響がまったくないとは言えません(Ling, Mollaun, & Xi, 2014)．また，信頼できる評定者を訓練によって育成できたとしても，試験実施機関がそのような評定者を長期的に確保することも簡単ではなさそうです．

「話す」能力の評価を行う評定者には，かなり高い要求がなされます．適切な評価を行う際には，ある程度の診断情報が必要な場合があります．例えば，「自己紹介」というタスクを大学生にやってもらったとしましょう．このタスクでは，それほど熟達度の高くない受験者であっても “My name is ... ” や

"I am majoring in Social Science." などといった定型文をある程度の流暢さで発話することが可能です．さらに定型文が多く含まれているので文法的な間違いも少なく，語彙の選択もそれほど間違えることはないかもしれません．何より定型文が多く含まれ，流暢さが保たれることで，評価の観点にもよりますが，熟達度の高くない受験者でも高い評価を受ける可能性があります．このタスクでは，より複雑な文を正確に発話できる学習者とそうでない受験者を見分けることができません．そのため，詳細な評価を与えるためには，試験官は受験者集団の平均的な受験者が，ある程度の流暢さを保って行えるタスクをはじめに出題し，そのタスクの出来具合によって，より簡単な／難しいタスクを出題し，受験者の「話す」能力を即座に診断的に評価しなければなりません(もちろん，「話す」能力の評価においても合格／不合格のみを判定する試験においてはこのような診断情報は必要ありません)．これは受験者が紙に解答を記入することで測定することが可能な他の能力とは，大きく異なる点です．例えば，文法的知識を問うような試験では，難易度の異なる問題を出題することができるので，正解／不正解の数がそのまま診断情報になります．

このように産出能力を評価することには，乗り越えなくてはならない問題がいくつもあります．大学入学共通テストでは，英語の「話す」，「書く」能力が評価されるようになりますが，この試験は民間の検定試験などが利用されることになっており，受験者は任意の試験を複数回受験することが可能で，最も良い点数を申請することになるようです．現在行われている大学入試センター試験は約 50 万人が受験しており，その多くが外国語の教科として英語を受験します．2020 年度以降もこの受験者数および傾向は大きく変わらないことが予想されます．英語の試験として複数の検定試験が利用可能になる予定ですが，高い信頼性を保って数万人の「話す」，「書く」能力の評価を行うことは，ここまで指摘した問題を考慮すると，かなり困難なことです．これほどの規模でなくても，1 クラス 40 人の学級が 5 クラスある高校の生徒全員の「話す」能力の評価を行おうとした場合でも，教員への負担は相当なものになることは明らかです．

これまで述べてきた産出能力の評価に関する問題を整理します．

1. 試験の実施，評価に人的，時間的コストがかかる．
2. 評価の信頼性を保つことが困難である．
3. 評定者の訓練に人的，時間的コストがかかる．
4. 訓練した評定者を長期間確保することが困難である．
5. 「話す」能力を評価する評定者には高い能力が必要である．

次節ではこれらの問題の解決策としての自動採点について考えます．

3．解決策としての自動採点

「話す」，「書く」能力の評価における問題の解決策のひとつとして昨今自動採点の導入に注目が集まっています．自動採点システムの構築は，評価がすでに付与されている学習者の作文，発話を大量に集め，コンピュータが測定することができ，かつ評価を予測する特徴量を見つけます(例えば，語数や無音ポーズの長さなどが用いられます)．評価が付与されていない作文や発話において，すでに見つけた評価を予測する特徴量を測定し，この特徴量を用いて評価を決定するというのが自動採点システムの仕組みです(自動採点の仕組みについては，本誌 84 号の「英語教師が知っておきたい ICT とテストの話」(石井，2017)も参照してください)．このような仕組みの自動採点システムで人間による評価との一致度がかなり高いものがすでに運用されています(例えば，Yoon & Zechner, 2017)．

自動採点システムを導入することによって，上述の 1 から 5 の問題を解決できるのでしょうか．1 に関しては，コンピュータが採点するので人的，時間的コストはかなり減少します．ですが，受験者 1 人につき 1 台のコンピュータが必要になるのでそのコストが必要です．スマートフォンなどで受験できるような試験にすれば，そのコストも下げることができますが，その場合，顔認証による本人確認のシステムやその他の不正行為を防止する仕組みも実装しなければなりません．2 に関しては，コンピュータは常に一定の点数を算出するのでこの問題は解決すると言っていいでしょう．3，4，5 に関しては，

問題となっている評定者はコンピュータなので，一度構築してしまえば，構築の際にかかるコストはありますが，その後訓練の必要もありません．しいて言えば，コンピュータの保守，点検作業がコストとしてかかるので，このコストが完全になくなるとは言えません．自動採点システムはデータを収集したり，予測の方法を検討したりと構築の際にかなりのコストがかかりますが，一度構築してしまえば，人間が評価を行う場合に比べてかなり少ないコストで試験が行えます．

4．自動採点の問題点

ここまで自動採点の良い部分を述べてきましたが，当然問題点もあります．先述のとおり，自動採点システムの構築には，評価がすでに付与されている学習者の作文，発話が必要でした．ある特定のタスクを自動採点にするにはそのタスクの作文，発話が必要で，別のタスクを自動採点にするためにはそのタスクの作文，発話が必要です．例えば，写真 A を口頭で描写するタスクを自動採点するシステムを構築し，写真を A から B に変えようとした場合，その写真 B で口頭描写をした受験者の発話が必要で，さらにその発話に評価を付与する必要があります．つまり，タスクごとに使用される単語や予測に使用される特徴量が異なるので，あるタスクで構築した自動採点システムは他のタスクでは使用できません．同じ形式のタスクでもタスクの難しさや期待される回答によって評価を予測できる特徴量は異なる可能性が高いことが考えられます．あるタスクで評価を予測できる特徴量でも，他のタスクではそうではない場合がしばしばあります．

これは実際の試験に自動採点を導入する場合において大きな問題です．同じタスクを繰り返し出題しても良い試験であれば問題ではありませんが，入学試験，検定試験などでは，公平性を考慮して，同じタスクを繰り返し出題しないものもあります．この場合，新たなタスクを作成するたびに自動採点システムを構築しなければなりません．自動採点システムは，人間の評定者に比べて評価が安定するという利点はありますが，新たなタスクを作成するたびに新たなシステムを構築するのでは人間の評定者を毎回訓練するのと同じようなことになってしまい，自動採点システムの利点が活かせません．

5．おわりに

本稿では，産出能力の評価に関する問題点を指摘し，その解決策のひとつとして自動採点システムの導入を検討しました．しかしながら，自動採点システムの導入においても問題があり，産出能力の評価をすべて自動採点で行うことは簡単ではないことを示しました．今後の自動採点システムの研究の課題として，未知のタスクにおいても信頼性の高い評価が得られるシステムの構築が望まれます．しかし，これは人間にとってもかなり難しいことです．「話す」能力の評価を行ったことはあるが，扱ったことがないタスクを用いて評価を行うことは人間にとってもかなり難しいことです．この問題を解決することによって自動採点の研究は次の段階に進みます．また，この研究の過程で人間の評定者がある評価を行う際の意思決定についても，さまざまなことがわかるのではないかと思います．

参考文献

石井雄隆(2017)．「英語教師が知っておきたい ICT とテストの話」『CHART NETWORK No.84』，17-20.

Ling, G., Mollaun, P., & Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking responses. *Language Testing, 31*, 479-499.

Yoon, S-Y., & Zechner, K. (2017). Combining human and automated scores for the improved assessment of non-native speech. *Speech Communication, 93*, 43-52.

（早稲田大学　准教授）