Flucher, G. (2010). *Practical Language Testing*. London: Hodder Education.

Chapter 7.1-7.2

7.1 Scoring items

■ テストの項目が作られると同時に、どのようにそれを採点するのかも決定する必要がある。採点方法や基準を取り決めるのには困難が伴うことも少なくない。例えば、記述問題において、どの程度までスペリングのミスを許容するかといったことはしばしば議論になる。

■ このような採点に関する困難性は、一般的に思われている以上にテスト作成時に発生する。採点について検討することはテスト開発の核心に位置するので、これは軽視できない問題である。

■ 例として、文整序問題をとりあげる (p.199 参照)。この問題では、生徒はバラバラに提示された複数の文を正しい順番に並び替えることを求められる (Alderson et al., 2000)。

下に示すものは最近起こったストーリーです。5 つの文の順番が変えられています。正しい順番になるように、各文に 1~5 の番号をつけましょう。ただし、1 番目は既にマークしてあります。

_____ (a) She said she'd taken the computer out of the box, plugged it in, and sat there for 20 minutes waiting for something to happen.
__1__ (b) A technician at Compaq Computers told of a frantic call he received on the help line.
_____ (c) The woman replied, 'What power switch?'
_____ (d) It was from a woman whose new computer simply wouldn't work.
_____ (e) The tech guy asked her what happened when the pressed the power switch.

正解: (b)→(d)→(a)→(e)→(c)

1→ 2 → 3 → 4 →5

■ ここで最初に浮上する問題は、ある文に不正解の順番をつけると、その文だけでなく、別の文も同時に不正解になってしまうことである

→(例) 本来 2 番目の文に 3 番をつけると、その文だけでなく、本来 3 番目の文を正答することもできなくなる。

■ さらに、2~3 文を正しい「順序」に並べることができても、誤った「場所 (slot)」に配置してあれば、たとえ文間のつながりを理解できていたとしても全て不正解になってしまう。

→(例) (c)→<u>(d)→(a)→(e)</u> ←下線部の順序は合っているが、正しい場所に入っていないため全て不正解

■ 最も深刻なのは、上記の指摘からも分かるように、正しく並べる能力を使えても正解になるとは限らないため、本来整序問題の核心的な構成概念であるはずの「正しい並び方」を考慮できていない点である。このように、一見上手く構成概念を測れているかに思えるテストでも、採点方法の不備によって台無しになっている場合はしばしばある。

■ この文整序問題の採点問題に対し、Alderson et al.は次の 4 つの採点方法を提案している。

→①<u>*Exact match*</u>: 順序と場所の両者が正しい場合にのみ点数を与える。最も単純で、採点が容易であることが利点。しかし上で見た問題点を孕む。

→②<u>*Classic*</u>: exact match に点を与えることに加え、たとえ場合が誤っていても順序が正しい 2~3 文のペアや、最

後の文 (edge score) にも点数を与える。

→③*Added value*: 正しい順序で並べられた 2~3 文ペアや、edge score に点数を与える (場所は考慮しない)。

→④*Full pair*: 正しい順序で並べられた 2~3 文ペアに点を与える (edge score に点数を与える)。

■ この中では、classic が最も良い採点方法で、その後 added value, full pair と続き、最も良くない採点方法が exact match だと仮定される。Alderson et al. はこれらの採点方法を 4 つのタスク条件を設定して検討した。

Task 1: Short story (上記と同様)

Task 2: Shot story (上記と同様) だが、1 番目の文が何かが示されていない。

Task 3: 7 文からなるテキスト

Task 4: 原文を読んだ後、それを要約したテキストを順番に並べる

■ 主な結果は以下の通りであった。

・1 番目の文が示されていなければ、採点方法に関わらず、タスクは難しくなった

・検証を通して exact match による点数が最も低く、classic が最も高かった

→順序が考慮されることで、受験者は多くの点数を獲得できる

・テキストが短いやタスクが簡単な場合、タスクとテスト全体の成績の相関は、1 文目が示されていないときの方が高かった

→1 文目を与えない方が弁別力が良く、みたい構成概念を確実に測定している

・全体的にみると、exact method は、他の採点方法よりもテスト全体のスコアとの相関が高い傾向にあった

・4 つの採点方法間の相関は、Task4 (原文提示→要約文を並び替え) において高かった

・4 つの採点方法による結果は高い相関をみせており、同じ構成概念に sensitive であることが示唆された

■ 順序を考慮した方が構成概念をより良く反映した採点方法になると考えられるが、問題は人の手で考えられる全ての順序を採点することは非常に困難であるということである。これに対処するためには、computer-based の採点を考慮する必要がある。

■ この節で我々が学ぶべき教訓は、採点方法はテスト項目や仕様書の作成と同時並行して考えなければならない問題だということである。どのように項目が採点されるのか、どのような場合に点数が与えられる、もしくは与えられないのかという点が不明瞭ならば、その項目による点数は有意義な情報を与えるではない。

■ 次節では、テストが実施される環境下における、より実用的な採点可能性 (scorability) の観点をみる。

7.1 Scorability

■ 採点可能性とは、特定の実施環境で (e.g., 紙ペース, コンピュータベース) テストの採点がどの程度容易かを指す。Lado (1961) は望ましいテスト要素の中にこれを含めており、紙ベースのテストを例にとってみる。

■ 解答が closed な項目でも、答えが紙面上散らばっていれば、採点に要する時間が長くなり、それとともに採点ミスを起こす可能性も高くなることが考えられる。採点可能性がこのように脅かされる場合は、別の解答用紙を利用することで、採点時間を短縮し、ミスを犯す危険性を減少させられる。

■ 過去、採点の容易化のために用いられていた道具が型板 (stencil) である。これを解答が書かれた紙に合わせるだけで、問題文を参照することなく採点することができる。

■ 今日の教師も、迅速な採点のために、採点のテンプレートを作成している。技術はほとんど変わっていない。アセテートシート (プロジェクターと一緒に使われるもの) を解答用紙に重ねれば、簡単に正解を数えることができ

る (p.202 の画像に透明のシートがあるが、おそらくこういったものを解答用紙に重ねて迅速に採点するのだと考えられる)。

- 閉じた回答項目 (closed response items) を使うもう一つの利点はコストの低さである。型板を用いてもコストを減らすことができるが、採点に専門知識があまり必要ではなく、事務職員や学生のよっても採点できるので、人的コストを節約することも可能である。
- コンピュータは、採点可能性に対する最大の解決策を提供してくれる。1938 年に IBM は、多肢選択式問題の自動採点コンピュータを公開した (p. 203, fig. 7.2)。操作者の技量によるが、1 枚の解答用紙につき 150 項目を、1 時間当たり 800~1000 枚採点することができた。
- このコンピュータは当時、大きなテスト機関にとっても高額過ぎたため、採点の迅速さを考慮しても合理的とはいえなかった (Lado, 1961)。しかし今日では、安価で携帯可能な採点機器も利用できる。
- このような背景を受け、コンピュータベースのテストは大いに普及してきた。コンピュータによるテストは採点を効率化し、スコアを直ちに知らせることができる。コンピュータベーステストの黎明期において、たびたび話題となった問題は、コンピュータと紙のテストで結果が異ならないか、というものである (Fulcher, 1997)。この懸念は writing の分野で特に関心が高かった。なぜならば、解答をペンではなく、タイプさせる形式は、受験者とって不利な要素になる可能性が考えられたからである (Russel & Haney, 1997)。
- また、EFL 環境にコンピュータベースのテストが導入される 1998 年までには、マウスとキーボード操作のテストはこの層の受験者に困難性を与えるか否かを検証する研究が多くなされてきた (Kirsch et al., 1998)。さらに、スクリーン上で、スクロール操作をしながらテキストを読む場合と、紙のテキストを読む場合とでは、読解プロセスに違いがあるのかについても盛んに検証されてきた (Sawaki, 2001)
- このような検証は、コンピュータベースのテストはある層の受験者にとって不利なのではないかという懸念から出発したものであったが、検証の結果、コンピュータが受験者に与える影響は、実用上は問題ないと言えるほど小さなものであることが示された。
- 初期の TOEFL には練習のためにセクションが設けられていたが、すぐにこれは削除された。今後、手書きよりもタイピングを好む受験者が増えると思われるが、これもコンピュータテストの隆盛を考慮すると、当然のことと考えられるだろう。
- 現在のコンピュータベーステストにおける大きな課題は、テストのデザインにある。特に、インターフェイスデザイン (注. 機械の使いやすさに関わる、用語、色彩、形状など様々なデザインのこと) に起因する要素によって、テスト結果が歪んでしまうことは避けねばならない。
- 採点に関しては、受験者による回答にコンピュータがどのように反応するかが重要になる。ここでは 3 つのオプションを紹介する。

→①Linear tests

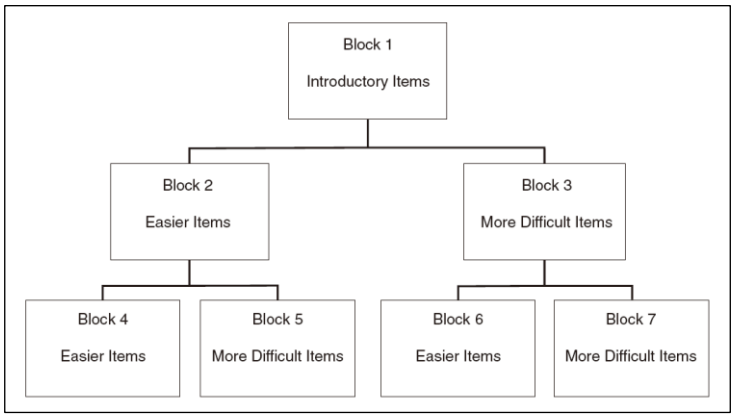
受験者は紙ベースのテストと同様に、決まった項目を決まった順番で提示される。受験者の反応は 1 か 0 の値でデータベースに記録され、素点が算出される。

→②Branching tests

・受験者の回答に応じて項目のまとまり (i.e., ブロック) が変化し得る。たとえば、最初は同じ項目ブロックが提示されるが、受験者が特定の問題数を連続して正答すれば、難易度の高い項目ブロックに移行する。逆に間違えれば難易度の低い項目ブロックに移る。

・この形式の利点は受験者の回答に応じた難易度の問題が出されるため、難易度が高くなりすぎず、やる気をそがれることが少ないことにある。問題点は、異なる問題を解いた生徒のスコアを比較することがで

きないことである。これには、最終的に正答できた項目ブロックのレベルに受験者を割り当てるという対策がとられる。例えば、Block1 の問題ほとんどに正答→Block3 の問題にも正答→Block7 の問題に苦戦→Block6 の問題に正答、というルートを経た受験者のレベルは 7 と評価される。



→③Adaptive tests

・この形式では、受験者の各項目への回答に応じてコンピュータが難易度を調節する。つまり、受験者がある項目に正答すれば、コンピュータは次の項目としてより難易度の高いものを選出するということである。

・ゆえに、このテストでは**項目プールが十分ならば**同じセットの項目を解く受験者はいないことになる。しかし、十分な量の項目プールを確保することは、深刻な問題である。抜群に出来る生徒、もしくは出来ない生徒の能力に応じて項目を提供するためには、莫大な項目プールが必要になり、作成者への負担が極めて大きくなる。

■ Adaptive tests の困難性を p. 207 の Figure 7.4 によって解説する。これは項目プールの分散を示している。コンピュータベースの adaptive testing では、Rasch model (Bond & Fox, 2007) によって、全受験者に能力スケールが、全項目に難易度スケールがそれぞれ割り当てられる。

■ この図から、低学力層に対応する項目が不足していることが見て取れる。よって、この項目プールは当該受験者層にとって難し過ぎるとが考えられる。この項目プールが adaptive testing で機能するためには、不足している項目を作成者が大量に追加しなければならいない。

■ このような性質のため、adaptive testing は、high-stakes なテストを実施する大規模な教育プログラム等でのみ実現可能である。

■ コンピュータベースの adaptive testing は、安全性を高めたり、個に対応したテストを提供できる点から、言語テストにおける万能薬の如く捉えられていた時期もあった。現に最初のコンピュータベース TOEFL は adaptive testing を採用していた。しかしながら、大量の項目プールを用意するにあたっての資金的、人員的負担から TOEFL iBT は linear testing を採用した。Fulcher (2005) は、この事実を「大規模国際言語テストにおける adaptivity 時代の終焉」と述べている。

■ adaptive testing の居場所が全く無いわけではない (国家規模のテストでは活用できる) が、この議論から我々が学んだことは、linear testing は何の問題もなく使え、実用性の面から強く勧められる形式だという点である。昔ながらの tried-and-tested が再び浮上したといえる。

☆コンピュータベーステストについてコメント

■本章で議論されるコンピュータベーステストについて興味を抱いたので、現在どのようなものが利用されているのかを調べた。

■英検が実施している BULATS (Business Language Testing Service) について述べる。これはケンブリッジ ESOL と提携して開発・実施されたオンラインでのコンピュータテストである。四技能を、実際のビジネスシーンで使用される英語の観点から総合的に評価することを目標としており、通常の英検との相違点として、原則的に法人や団

体 (受験者数 10 人以上) による受験のみ受け付けている。リーディングとリスニングテストでは adaptive 形式が採用されており、受験者の能力が確定されるまでテストが継続される (通常 60 分〜85 分程度) 仕組みになっている。テスト結果は CEFR と ALTE (The Association of Language Testers in Europe: 英語をはじめとしたヨーロッパ諸言語のテストを統括する組織) による指標が提供され、通知は 10 営業日以内と迅速であるため、コンピュータテストの利点をうまく活用しているといえる。

■我が国の英検とケンブリッジ英検の提携のもと誕生したコンピュータテストということもあり、精緻にデザインされている印象を受けた。HP 上のサンプル問題を見ると、同じくビジネスシーンを対象とする TOEIC よりもバラエティに富んだ内容 (e.g., 業務の解説が与えられ、それに最適な職を選択する問題) で、対策もやりがいのあるものになると思われる。留意点としては、個人で自由に受験できないことがあげられる。法人・団体受験への限定実施は adaptive testing を採用しているためだと考えられる。私は adaptive なテストを受験したことがないので、具体的に感想を述べることは出来ないが、一般的に考えても受験者のレベルに応じた問題のみが出題されることの利点はいくつかあげることができる。まず、問題数を減少させることができる。受験者にとって極端に難しい問題や、簡単な問題は、対象の能力や構成概念の測定項目に適うとは言い難い。また、そのような問題は受験者のモチベーションを削ぐ要因にもなり得る。ゆえに、1 問 1 問の回答に応じた難易度が選択されることで、テストを効率化することが可能になると考えられる。また、コンピュータベースのテストでは、試験用紙を印刷する手間や、紙資源の浪費を回避することが可能である。これはテスト実施側にとって有利な点で、運営上の手順を効率化したり、資源を節約したりできれば、その分実施の回転を速めることが可能になると考えられる。

■BULATS から今後のコンピュータテストの可能性を再認識できた。重要なことは、すべてのテストをコンピュータやインターネットベースにするのではなく、この形式が長所を発揮できる点を見定めることにあるだろう。今後は項目プール数などの課題がいかに解決されるのかに注目したい。