Sec. 3 Scoring constructed response tasks

Methodology 4: Scaling descriptors

■ CEFR 作成のために考案された

1. 評価尺度を出来るだけ多く特定・収集する（North, 1993）
2. 評価尺度を切り離し、2,000 の記述子（descriptor: 学習者の各レベルを説明する記述）のプールを作成する
3. 専門家の判断により、記述子はコミュニケーション活動のタイプ別に分類され（North, 2000）、分類に不足が感じられたら新たな記述子を追加する。
4. 教師集団が記述子から成績の上中下を記述するのに利用できるものを分類し、さらに各レベルを 2 つに分ける。→ 計 6 レベルに分類される。
5. 別の教師集団が選定された記述子を難易度順に並べ、その難易度をラッシュモデルで分析する。

■ この方法は、典型的な生徒集団を評価するのに教師が経験から抽出した記述子に感じられる難易度データに基づいており、実証的だといえる。教師集団、言語使用場面に関わらず使用されるものだけが、最終的な評価尺度として残る。

■ この方法は理論に根差しておらず、元は別のところで使用された尺度をラッシュモデルで新尺度に統合したものである（North, 1995）。ゆえに実際の言語使用・獲得の過程を説明してはいないという点で課題が残る。

Methodology 5: Performance decision trees (PDTs)

■ methodologies 2, 3 を合わせ、二者択一・データに基づいた尺度を作成する（Fulcher et al., 2011）。

■ Mills (2009) のタスク（pp. 150-154）の採点を例に挙げると、まず実際の service encounter のデータを収集し、談話を分析する。service encounter のスクリプトには sale intention (SI), sale request (SR), greeting (G), side sequence (SS) がある（p. 214）。

■ side sequence は"relational management"と呼ばれ、rapport を構築する（Gremler & Gwinner, 2000）。実際にトレーナーは販売員に relational management talk により、顧客の購入体験を向上させるよう指導する。

■ p. 214 の会話の成功は笑顔、定期的なアイコンタクトなどの non-verbal な面にもある（Gabbot & Hogg, 2000）。これらの要素は、コミュニケーション能力のモデル（4 章）における discourse and pragmatic competence に相当する。上記の分析に基づく尺度を図 7.6 (p. 215) に示す。

1. 必須要素（e.g., SI, SR）の産出の有無
2. relational management の産出の有無 → yes なら 2 点が与えられ、さらに rapport 構築の尺度（各 1 点）へ
3. 他に discourse management の 5 つの尺度によって点数が加算される（全体で 0~20 の尺度になる）。

■ 点数の加算が二値で行われる点で、data-based methodology と EBB methodology が組み合わされている。

Sec. 4 Automated scoring

■ 近年 speaking / writing を中心に自動採点への関心が高まっており（Wresch, 1993）、特に writing の自動採点がもっとも成功している。

(ex.) e-rater: 米国の大規模テストや TOEFL iBT (second rater) にも利用されている。語、テキストの長さ、語彙などの統語的特徴を分析する。discourse marker や語彙集合（lexical set）を実際のエッセイと比較し、談話構造やトピックとの関連度において人の採点と相関が高い採点ができる（Lee et al., 2008）。

■ 一方 speaking の自動採点は難しい。最初に開発されたのは PhonePass (Bernstein, 1999a) だった。これは PC で音読、繰り返し、反意語、短文応答問題を採点する。スコアが listening / speaking 能力に敏感なことで妥当性が主張された。またポーズの長さ、発音、タイミング、リズムを評価することで、人間の採点と中程度の相関が得られた。

■ 自動採点の妥当性は主に人間の採点と相関を持たせることで追及されている（criterion-oriented approach)。この方法では人間の採点を gold standard とし、自動採点と高い相関があれば同じ構成概念を測定していると想定している（e.g., Bernstein et al., 2000)。そうした自動採点を second marker として使用した場合、人間の second marker と同程度 first marker と高い関連性があるとされている（Atali, 2007)。

■ しかし自動採点について、筆記テストにおける語彙、文法、読解の成績が自動的に同等の listening / writing 能力を表すわけではないことに注意が必要である（correlation fallacy の問題: Kaulfers, 1944)。人間が採点するのと同様に微妙なニュアンスの違いを機会が採点できるかは疑問が残る。また、教師が与えるような feedback が機械には与えられないといった問題も存在する（Scharber et al., 2008)。

Sec. 5 Corrections for guessing

■ Lado（1961）はテスト受験者の推論効果を修正する公式を提案した。ただしこれは自分が推測したかわかっていない（十中八九推測していない）場合、厳しくなりすぎてしまう。他に項目応答理論を利用して推測の変数を推定する方法も考案されたが、推測の修正はあまり勧められない。

$$\text{スコア} = \text{正答数} - \frac{\text{誤答数}}{\text{各問の選択肢数} - 1}$$

※ 水量による正解を認めるべきか?

→ 集団基準準拠テストと目標基準準拠テストとで異なる。

→ 完全な当て推量と知識を使った推量を区別できればいいが、現状では不可能。

Sec. 6 Avoiding own goals

■ 本章では言語テストの測定要素に触れてきた。テスト受験者の成績の評価方法は、test specification や prototyping とともにデザインされるべきである。理想的には、prototyping や piloting の間に十分に議論するのが望ましい。これは特に writing / speaking の採点に当てはまる。

■ また採点し易さと豊富な記述子の作成の間にも対立がある。単純な採点方法は早く効率的な一方、複雑な採点方法は非常に高い精度が必要になる。複雑な採点方法を確立させることで得られる利益とコストを考慮する必要がある。

■ テスト得点にはよく特別な意味が付与される（e.g., 大学受験資格)。これは cut score の設定や基準の確立で行われている（8 章参照)。