22/03/2016
JOPT2016シンポジウム
キャンパスプラザ京都

外国語口頭能力測定の基礎
―分析モデルを中心に―

野口　裕之

発表の流れ

1. 最近の大規模言語テスト
   ーパフォーマンス測定
2. パフォーマンス測定の分析モデル
   ー般化可能性理論
   多相ラッシュ・モデル
3. 解釈基準の設定

1. 最近の大規模言語テスト

1) コミュニケーション能力を測定する、すなわち、パフォーマンス測定へ重心が移動、
2) より精密な議論や検討ができる項目応答理論 (Item Response Theory ; IRT) をベースにした得点化システムを採用する、
3) 測定結果の解釈規準として「～ができる」という Can-do statements を用意する、
4) CEFRとの関連づけを行なう 、

という大きな潮流がある。

パフォーマンスの測定

- 実際に「話す」「書く」など「やってみる」テスト
- 学習者が実際に遭遇する現実場面に近い状況で測定するということで，真正性 (authenticity) という概念が重要
- パフォーマンスの測定には，採点者，評価者の主観の影響が大きい．客観性，信頼性をどうやって確保するのかが問題となる

  → 機械採点がどのような形で導入可能か？
- 受験者に対して提示できる課題を多くはできず，そういう意味で妥当性の面でも十分ではないこともある．

- パフォーマンステストにおいては、評価者が受験者のテスト課題に対するパフォーマンスを観察して信頼性と妥当性の高い適切な評価を下す必要がある.
- 評価の方法には，包括的評価と分析的評価の2つがある.

- 包括的評価　パフォーマンスを要素や側面に分けることなく総体として評価

　客観性？　評価の観点を評価者間で統一
　評価者内でのぶれを最小化

- 分析的評価　パフォーマンスを要素や側面に分けて個別に評価
　評定尺度を複数用意する、など

　妥当性？　要素への分割の仕方
　要素得点の総合化
　の適切さが重要になる

- 包括的評価を採用している試験として,「日本語OPI (Oral Proficiency Interview) 」 → 鎌田先生
- 分析的評価を採用している試験として「日本語口頭能力試験（試行版）」（庄司・野口・金澤他(2004), 安高(2015)）
  → 大規模試験の中で実施することを意図しているという制約。
  パーソナル・コンピュータで課題提示、
  受験者の発話を記録、
  訓練された採点者が
  チェックリスト評定 (言及事項の量的評定),
  査定基準評定 (質的評定) をWebベースで実施.
  外在基準との相関　JLPT2級　0.674 ,
  日本語OPI 0.64程度

評価者（評定者）に関する信頼性

- **評価者内信頼性**
  評価者の採点に関する安定性
- **評価者間信頼性**
  評価者間の採点の一致度
- 信頼性の高い測定を実施するため、採点の観点や得点化の基準などを厳密に定めた「**採点基準**」を用意して、評価者（評定者）に対して十分な研修を実施する。
- 実際のパフォーマンス・テストでは、評価者間一致度と評価者内一致度を示す指標を用いて採点結果のぶれの程度を検証しておく必要がある。

2. パフォーマンス測定の分析モデル

2. パフォーマンス測定の分析モデル

- パフォーマンス測定の場合、通常の言語テストと異なり、評定者（評価者）が介在するために、その誤差についても分析モデルに組み入れる必要がある。
- テストのデータが、受験者×評定尺度（評価の観点）×評定者という3次元構造を持つ
- そのため、通常のテスト分析と異なるモデルが用いられる
- よく用いられるモデル
  一般化可能性理論
  多相ラッシュ・モデル

2.1 一般化可能性理論

2.1 一般化可能性理論

- 各受験者のテスト得点は評定尺度得点の合計で表わすとして、このテスト得点の信頼性は、古典的テスト理論の枠組で想定されていた、受験者×項目という2次元構造のものとは異なる。
- これは、問題項目に関係する誤差に加えて、「評定者」に起因する誤差、すなわち「評定者の信頼性」も問題になるからである。

- 一般化可能性理論は古典的テスト理論で「偶然誤差」として包括的に扱っていたものを、
  問題項目に起因する誤差、
  評定者に起因する誤差、
  など複数の誤差要因に分けて取り扱い、
  それらがテスト得点の分散の中でどの程度の大きさを占めるかを評価。
  逆に、評定者を何名配置すれば所与の精度を持った測定が可能か、などについて検討。

2.2.1 Generalizability 研究

- テストの精度に影響する特定の要因がどの程度の大きさになるかを推定する段階
  例えば、
  異なる評定者が採点することにより生ずる誤差の大きさを推定、あるいは、テスト得点の分散の中でどの程度の割合を占めるのかを推定する。
- 具体的には、分散分析モデルを適用して各誤差要因の分散を推定したり、受験者間の能力差に起因する分散の大きさを推定する。これらの推定された分散の大きさから、「一般化可能性係数」が推定され、実際場面で用いられる。

2.2.2 Decision 研究

- G研究で得られた結果をもとに、実際のテスト場面で特定の要因から生ずる誤差を一定限度以内に押さえるにはどうするか、
  例えば、
  　評定者を何名配置する必要があるか、
  　課題をいくつ用意するか、
  などの具体的な問題に解答を得る。
- 古典的テスト理論の枠組では測定精度が信頼性係数で定義され、スピアマン・ブラウンの公式を用いて一定の信頼性を保証するのに必要な項目数が計算されたことに対応している。

2.2.3 2相完全クロス・デザイン

- 最も基本的な「2相完全クロス・デザイン」の「変量モデル」について取り上げる。
- このモデルでは、
  例えば日本語学習者のスピーキング能力を測定するのに、
  受験者（N名）×課題（nコ）×評定者（r名）のすべての組み合わせでデータが得られている場合を想定する。
  すなわち、
  すべての評定者が、すべての受験者のすべての課題に対する回答（解答）を評定する、という状況。

  さらにこれらの受験者、課題、評定者には母集団が想定でき、そこからのランダム・サンプルであることを仮定。

（データの構造が理解しやすいように、次のスライドに受験者10名、課題3コ、評定者3名の場合のデータ例を示しておく。）

受験者 10 名×3 課題×評定者 3 名　の場合の観測データ例

受験者	課題 1				課題 2				課題 3				平均
	評価者 1	評価者 2	評価者 3	平均	評価者 1	評価者 2	評価者 3	平均	評価者 1	評価者 2	評価者 3	平均	
1	X_{111}	X_{112}	X_{113}	$X_{11\cdot}$	X_{121}	X_{122}	X_{123}	$X_{12\cdot}$	X_{131}	X_{132}	X_{133}	$X_{13\cdot}$	$X_{1\cdot\cdot}$
2	X_{211}	X_{212}	X_{213}	$X_{21\cdot}$	X_{221}	X_{222}	X_{223}	$X_{22\cdot}$	X_{231}	X_{232}	X_{233}	$X_{23\cdot}$	$X_{2\cdot\cdot}$
3	X_{311}	X_{312}	X_{313}	$X_{31\cdot}$	X_{321}	X_{322}	X_{323}	$X_{32\cdot}$	X_{331}	X_{332}	X_{333}	$X_{33\cdot}$	$X_{3\cdot\cdot}$
4	X_{411}	X_{412}	X_{413}	$X_{41\cdot}$	X_{421}	X_{422}	X_{423}	$X_{42\cdot}$	X_{431}	X_{432}	X_{433}	$X_{43\cdot}$	$X_{4\cdot\cdot}$
5	X_{511}	X_{512}	X_{513}	$X_{51\cdot}$	X_{521}	X_{522}	X_{523}	$X_{52\cdot}$	X_{531}	X_{532}	X_{533}	$X_{53\cdot}$	$X_{5\cdot\cdot}$
6	X_{611}	X_{612}	X_{613}	$X_{61\cdot}$	X_{621}	X_{622}	X_{623}	$X_{62\cdot}$	X_{631}	X_{632}	X_{633}	$X_{63\cdot}$	$X_{6\cdot\cdot}$
7	X_{711}	X_{712}	X_{713}	$X_{71\cdot}$	X_{721}	X_{722}	X_{723}	$X_{72\cdot}$	X_{731}	X_{732}	X_{733}	$X_{73\cdot}$	$X_{7\cdot\cdot}$
8	X_{811}	X_{812}	X_{813}	$X_{81\cdot}$	X_{821}	X_{822}	X_{823}	$X_{82\cdot}$	X_{831}	X_{832}	X_{833}	$X_{83\cdot}$	$X_{8\cdot\cdot}$
9	X_{911}	X_{912}	X_{913}	$X_{91\cdot}$	X_{921}	X_{922}	X_{923}	$X_{92\cdot}$	X_{931}	X_{932}	X_{933}	$X_{93\cdot}$	$X_{9\cdot\cdot}$
10	$X_{10,11}$	$X_{10,12}$	$X_{10,13}$	$X_{10,1\cdot}$	$X_{10,21}$	$X_{10,22}$	$X_{10,23}$	$X_{10,2\cdot}$	$X_{10,31}$	$X_{10,32}$	$X_{10,33}$	$X_{10,3\cdot}$	$X_{10,\cdot\cdot}$
平均	$X_{\cdot 11}$	$X_{\cdot 12}$	$X_{\cdot 13}$	$X_{\cdot 1\cdot}$	$X_{\cdot 21}$	$X_{\cdot 22}$	$X_{\cdot 23}$	$X_{\cdot 2\cdot}$	$X_{\cdot 31}$	$X_{\cdot 32}$	$X_{\cdot 33}$	$X_{\cdot 3\cdot}$	$X_{\cdot\cdot\cdot}$

モデル式

- この時、受験者iの課題jに対する回答を評定者kが採点した結果を、$X_{ijk}\ (i = 1, \cdots, N; j = 1, \cdots, n; k = 1, \cdots, r)$ として、その母分散$\sigma^2(X_{ijk})$ は、

$$\sigma^2(X_{ijk}) = \sigma_i^2 + \sigma_j^2 + \sigma_k^2 + \sigma_{ij}^2 + \sigma_{ik}^2 + \sigma_{jk}^2 + \sigma_e^2$$

と分解して表わされる。
右辺は順に、受験者の能力に起因する分散成分（σ_i^2）、
課題の困難度に起因する分散成分（σ_j^2）、
評定者の厳しさに起因する分散成分（σ_k^2）、
受験者と課題の交互作用（相性）に起因する分散成分（σ_{ij}^2）、
受験者と評価者の交互作用（相性）に起因する分散成分（σ_{ik}^2）、
課題と評価者の交互作用（相性）に起因する分散成分（σ_{jk}^2）、
そして、残差（誤差）成分（σ_e^2）である。

- G研究ではこれらの大きさを推定するが、実際の測定場面で評価者の厳しさ（甘い辛い）の影響が小さいことが望ましい場合には、モデル式の右辺で、$\sigma_k^2, \sigma_{ik}^2, \sigma_{jk}^2$ の値が小さければ、そのような測定が実現できていることになる。
- G研究では実際に観測された得点X_{ijk} $(i = 1, \cdots, N;\ j = 1, \cdots, n;\ k = 1, \cdots, r)$が得られている時に、受験者iの平均$\overline{X}_{i..}$、課題jの平均$\overline{X}_{.j.}$、評定者kの平均$\overline{X}_{..k}$、受験者iと課題jの組み合わせの平均$\overline{X}_{ij.}$、受験者iと評定者kとの組み合わせの平均$\overline{X}_{i\cdot k}$、課題jと評定者kとの組み合わせの平均$\overline{X}_{.jk}$、全ての観測データ（受験者と課題と評定者の組み合わせ）の平均$\overline{X}_{...}$が計算でき、それらをもとに　次のスライド　に示す平均平方を求め、さらに各変動要因の分散推定量が計算できる。

G研究（２相完全クロス計画）における各変動要因の平均平方と

それらから導かれる分散推定量

変動要因	平均平方	分散推定量
i　受験者	$MS_i=nr\sum_i(\bar{X}_{i..}-\bar{X}_{...})^2/(N-1)$	$\hat{\sigma}_i^2=\left[MS_i-MS_{ij}-MS_{ik}-MS_{ijk}\right]/nr$
j　課題	$MS_j=Nr\sum_j\left(\bar{X}_{.j.}-\bar{X}_{...}\right)^2/(n-1)$	$\hat{\sigma}_j^2=\left[MS_j-MS_{ij}-MS_{jk}-MS_{ijk}\right]/Nr$
k　評定者	$MS_k=Nn\sum_k(\bar{X}_{..k}-\bar{X}_{...})^2/(r-1)$	$\hat{\sigma}_k^2=\left[MS_k-MS_{ik}-MS_{jk}-MS_{ijk}\right]/Nn$
ij	$MS_{ij}=r\sum_i\sum_j\left(\bar{X}_{ij.}-\bar{X}_{i..}-\bar{X}_{.j.}+\bar{X}_{...}\right)^2/(N-1)(n-1)$	$\hat{\sigma}_{ij}^2=\left[MS_{ij}-MS_{ijk}\right]/r$
ik	$MS_{ik}=n\sum_i\sum_k(\bar{X}_{i\cdot k}-\bar{X}_{i..}-\bar{X}_{..k}+\bar{X}_{...})^2/(N-1)(r-1)$	$\hat{\sigma}_{ik}^2=\left[MS_{ik}-MS_{ijk}\right]/n$
jk	$MS_{jk}=N\sum_j\sum_k\left(\bar{X}_{\cdot jk}-\bar{X}_{.j.}-\bar{X}_{..k}+\bar{X}_{...}\right)^2/(n-1)(r-1)$	$\hat{\sigma}_{jk}^2=\left[MS_{jk}-MS_{ijk}\right]/N$
e 誤差	$MS_e=\sum_i\sum_j\sum_k\left(X_{ijk}-\bar{X}_{ij.}-\bar{X}_{i\cdot k}-\bar{X}_{\cdot jk}+\bar{X}_{i..}+\bar{X}_{.j.}+\bar{X}_{..k}-\bar{X}_{...}\right)^2$	$\hat{\sigma}_{ijk}^2=MS_{ijk}$

- D研究では、G研究で得られた各変動要因の分散推定量をもとに、具体的なテスト実施場面における条件を決めて行く。

例えば、評定者数を2名とするか3名とするか、
評定者数を増やすのと課題数を増やすのとどちらを優先させるか、
などの決定をする。

2.2.4 一般化可能性係数と信頼度指数

- この時、測定の性能を表わすのに、
  一般化可能性係数(generalizability coefficient)と
  信頼度指数(index of dependability)が用いられる。
  → 古典的テスト理論で信頼性係数に相当する指標
- 両者の違いは「誤差」をどのように考えるかの違い。
  一般化可能性係数では、D研究で得られた受験者得点（評定者がつけた得点の平均値）の相対的位置（順位）がユニバースにおける得点の順位付けにどの程度一致するか、という「相対誤差」を問題にするのに対して、信頼度指数では、相対的位置だけではなく、得点の絶対的な差異も問題にする「絶対誤差」を取り上げるため、評定者の厳しさも誤差に影響する。

- 相対誤差の分散を$\sigma^2(\delta)$、絶対誤差の分散を$\sigma^2(\Delta)$と表わすと、

一般化可能性係数Gは

$$G = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\delta)}$$

信頼度指数Φは、

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\Delta)}$$

で定義される。

そして、これらの各項は

$$\hat{\sigma}^2(\tau) = \hat{\sigma}_i^2$$

$$\hat{\sigma}^2(\delta) = \frac{\hat{\sigma}_{ij}^2}{n'} + \frac{\hat{\sigma}_{ik}^2}{r'} + \frac{\hat{\sigma}_{ijk}^2}{n'r'}$$

$$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(\delta) + \frac{\hat{\sigma}_j^2}{n'} + \frac{\hat{\sigma}_k^2}{r'} + \frac{\hat{\sigma}_{jk}^2}{n'r'}$$

で推定でき、上記3式右辺の各項は、G研究で既に分散推定量として得られているため、一般化可能性係数Gおよび信頼度指数Φが推定できる。

- ここで、n′およびr′は、当該のD研究で用いる課題数および評定者数を表わす。
- n′とr′の一方あるいは両方を増やすと、$\hat{\sigma}^2(\delta)$および$\hat{\sigma}^2(\Delta)$が小さくなり、従って、一般化可能性係数Gおよび信頼度指数Φの値が大きくなる。
- 逆に実際の測定場面で、一定の精度の得点を得たい場合に、課題数もしくは評定者数をいくつにすればよいかを知ることができる。

2.2.5 適用に際しての注意

- 一般化可能性理論を実際に適用するには、分散成分の推定には必ず誤差が混入すること、G研究は少ない標本数で実施せざるを得ないことがほとんどで、そのような場合には標本誤差の影響を大きく受けることに注意が必要である。

- また、一般化可能性理論ではここで述べた完全クロス・デザイン以外に現実に合わせた多様なデザインがある。

- 詳しくは、池田(1994)、平井(2007)
  Brennan(2001)などを参照されたい。

2.2 多相ラッシュ・モデル

2.2 多相ラッシュ・モデル

- 通常テストの結果は、

  受験者の能力　→　正答数得点

  当該受験者が正答した項目数

  項目の困難度は　→　正答率（通過率）

  受験者全体で当該項目に正答した受験者の比率

  で表される。

- 正答数得点は受験者の能力が解答した項目群に依存して表示され、通過率は、項目の困難度がその項目に解答した受験者集団の能力に依存して表示されるという限界がある。
- 同じ受験者でも、易しい項目から構成されたテストでは高い得点を示すが、難しい項目から構成されたテストでは低い得点にとどまり、同一の能力水準を表現する得点が受験したテストにより異なることになる。
- 従って、受験者固有の能力値や項目に固有の困難度の値で表示することができない。

これに対して、

1) 受験者の能力を解答した項目と独立した値で表現できないか？
2) 問題項目の困難度を解答した受験者集団と独立して表現できないか？

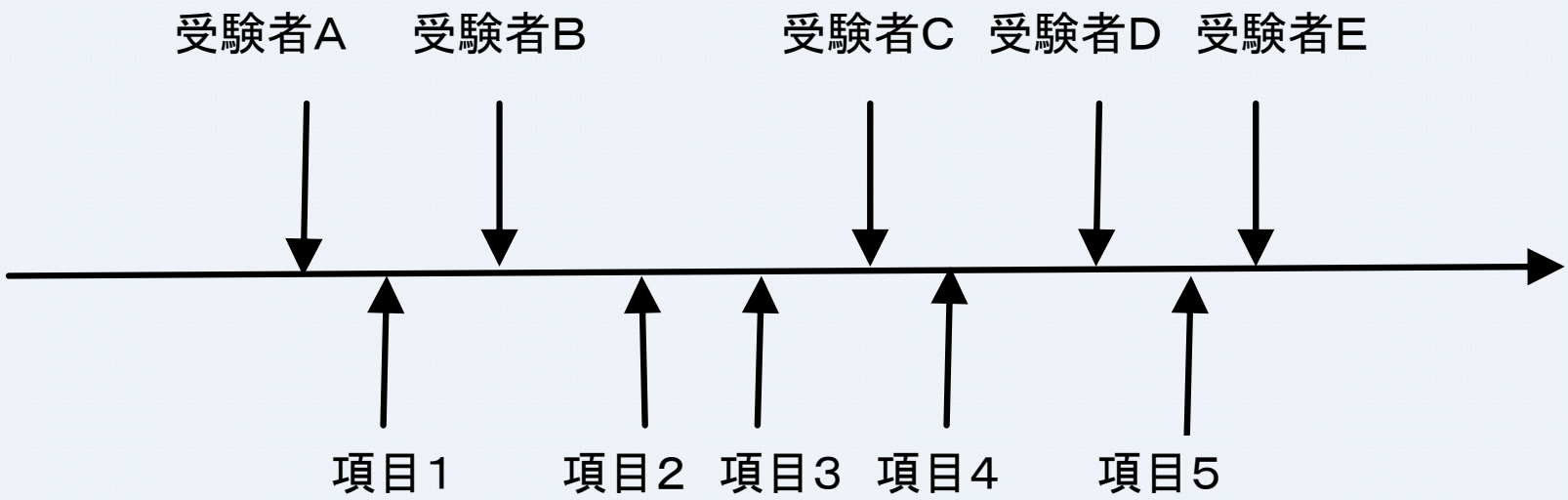
という問題

2.2.1 ラッシュ･モデル

- デンマークの数学者、Rasch,G.が受験者の能力を表わす値（パラメタ）と項目の困難度を表わす値（パラメタ）を分離して独立に表現するモデルを提案した(Rasch, 1960)。

- その後、シカゴ大学の Wright, B. を中心に研究および普及活動が進められ、欧州や豪州では言語テストの分析モデルとして標準的なモデルとなっている。

ラッシュ･モデルでは、受験者の能力水準と項目の困難度水準とが同一の尺度上に位置づけられて表現される



能力水準の高低に応じて5名の受験者が、また困難度水準の高低に応じて5つの項目が順に直線上に配置されている。

直線の右方向にある受験者ほど能力が高く、右方向にある項目ほど難しいことを表わしている。

- 直線上の位置で、
  受験者－項目の距離が0より大きいほど、
  当該受験者がその項目に容易に正答でき、
  0より小さいほどその項目に正答することが
  困難である。

受験者Eは項目2に容易に正答できるが、受験者Aが項目2に正答するのは難しい。

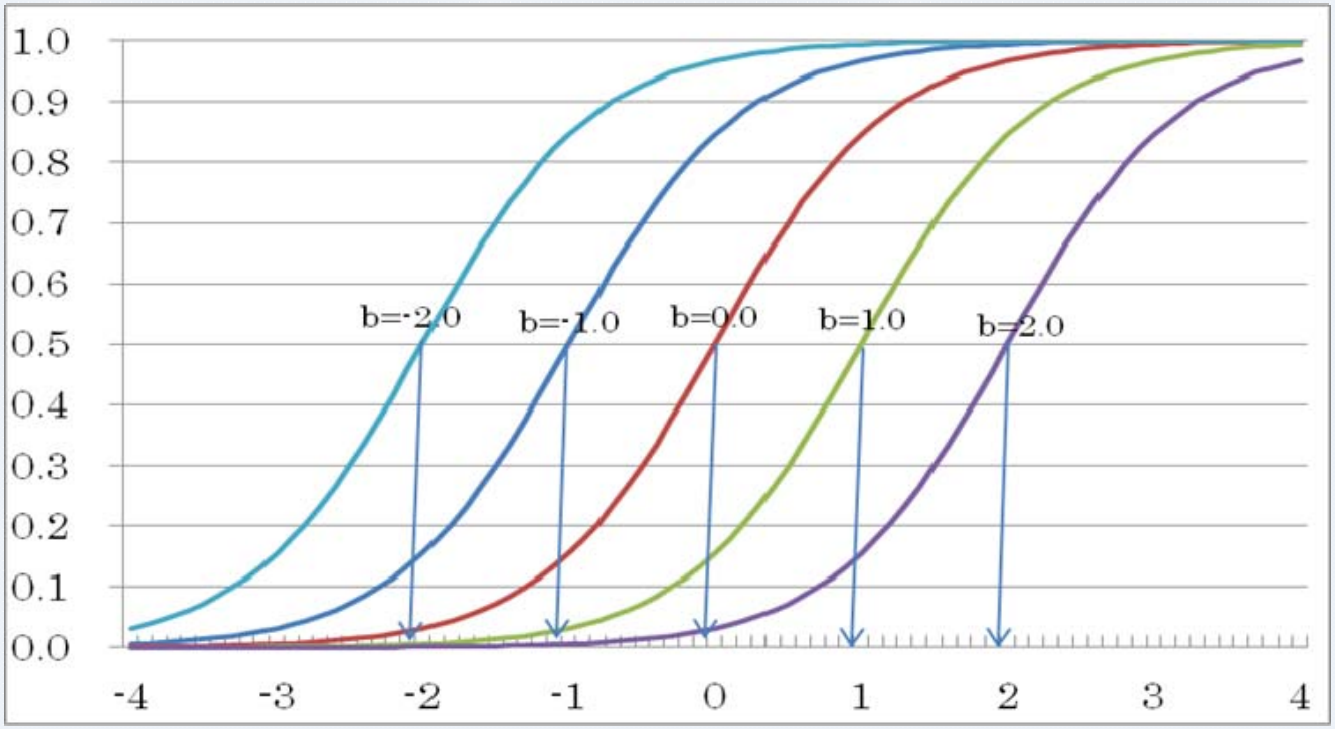
- 実際にはこの正答確率を表わす関数を具体的に定義する必要がある。当初ラッシュが想定した関数とは異なり、現在では、

$$\mathrm{Prob}(u_{ij}=1 \mid \theta_i, b_j) = \exp(\theta_i - b_j)/\{1+\exp(\theta_i - b_j)\}$$

が用いられている。

θが特性尺度（能力）値、bが項目困難度

この関数の表わす曲線を項目特性曲線と呼ぶが、次のスライドに$b_j = -2, -1, 0, +1, +2$ の5つの場合について示した。



横軸がスライド32の直線に相当し、**項目の困難度および受験者の能力尺度値を表わす座標軸**

縦軸が確率を示す座標軸で、項目特性曲線により横軸の値に対応する**正答確率**を読み取ることができる。

2.2.2 多相ラッシュ・モデル

- パフォーマンス測定の結果は、言語テストの場合、多相ラッシュ・モデルにより分析されることが多い（例えば、McNamara,1996 参照）。
- ラッシュ・モデル　受験者×項目の二相からなる矩形のデータ行列を分析
- 多相ラッシュ・モデル　受験者×項目（評価の観点）×評価者の三相からなる直方体状のデータを分析するなど、相の数が3以上のテスト・データを分析するモデルであることを表わしている。

- 例えば、受験者のスピーキング能力を評価するのに、ある課題を与えてそれに対する受験者の発話を評定して得点化する、という状況を想定する。
- 評定者はA，B，Cの３名
- 評定の観点は、１）即応性と滑らかさ、２）発音のわかりやすさ、３）語彙のわかりやすさ、４）構造のわかりやすさ、
- 各観点について、1,2,3,4,5の5段階で評定者が評定しているとする（庄司ほか(2004)を参照して例示用に作成）。

この場合には、受験者×評価の観点という矩形のデータ行列が、評価者3名分あり、仮に受験者10名の場合には、10×4×3の直方体状にセルが並んだデータが得られることになる。
論文や報告書等の中で実際に立体形式にデータを表現することは難しいため、以下に示すような表現がとられることが多い。

受験者	評価者A				評価者B				評価者C			
	即応性	発音	語彙	構造	即応性	発音	語彙	構造	即応性	発音	語彙	構造
1	4	4	4	5	4	4	4	4	4	4	4	5
2	2	3	2	2	3	4	2	2	3	3	2	2
3	3	4	2	3	4	4	3	3	4	4	3	3
4	1	3	1	1	2	3	2	2	1	3	2	2
5	3	3	3	3	4	3	3	2	3	3	3	3
6	1	3	2	2	3	3	2	3	2	3	2	2
7	1	3	2	2	2	2	2	3	2	2	2	2
8	4	3	2	2	4	4	3	4	4	3	3	3
9	3	3	3	1	4	3	3	2	3	3	2	2
10	3	2	1	2	4	3	2	3	4	3	1	2

- このような場合に、
  受験者の能力(ability)、
  項目（評価観点）の困難度(difficulty)、
  評価者の厳しさ(severity)、
  をパラメタとして、ラッシュ・モデルを拡張することができる。
  → 多相ラッシュ・モデル

多相ラッシュ・モデルは実用水準で活用され、

- 評価者トレーニングが評価者個人の厳しさの変動を小さくする効果はあるが、評価者間の厳しさの違いを小さくする効果はあまりない（例えば、McNamara(1996) p.235）
- 言語能力記述文を多数用意して、学習者に対する教師の評価をデータにこれらの言語能力記述文を単一の尺度上に配置することによって言語能力水準を記述（North and Schneider(1998), North(2005)）

などの成果が得られている。

詳しくは、例えば、野口・大隅(2014), 野口(2015)などを参照されたい。

2.3 多相ラッシュ・モデルと一般化可能性理論

- 「一般化可能性理論 (Generalizability theory)」のモデルでは、古典的テスト理論で偶然誤差として包括的に扱っていた測定誤差を、問題項目に起因する誤差、評価者に起因する誤差など複数の誤差要因に分けて取り扱い、それらがテスト得点の分散の中で占める大きさ(割合)を「分散分析モデル」を用いて推定する。評価者要因に関しては、評価者間の厳しさの違いを評価者集団全体での測定誤差として取り上げる。
- これに対して、多相ラッシュ・モデルでは評価者個人の厳しさを表わすパラメタをモデルに組み込んで、評価者毎の厳しさを個別に推定するという点が異なっており、そのためモデルに適合しない評価者を具体的に検出することも可能である。

3. 解釈基準の設定

3. 解釈基準の設定

- 受験者の得点からその受験者の能力水準を判断するためには、得点の解釈基準が必要になる。どのような解釈基準が必要であるかは試験の測定目的によって異なる

- 到達度を見るのか、プロフィシェンシーを見るのか
- 「できること」を示した解釈基準か相対的な位置を示した解釈基準か

など

日本語能力試験の場合

- 得点を具体的に解釈する基準は現段階では用意されていない。
- 日本語Can-do statements尺度が開発され（三枝ほか,2004）、日本語能力試験との関連などが検討され、一定の成果を得ることができた。
- 2010年の改定後も得点の解釈基準を設定するには至っていない

Can-do statements

- The ALTE(The Association of Language Testers in Europe) Framework　では
  Can-do statements　を開発し、それを扇の要として欧州各国で実施されている各言語テストにおけるレベル認定基準の相互比較を可能にしている 。
- 欧州評議会では、外国語テストをCEFRに関連付ける(aligning/relating)ためのマニュアルを整備。

CEFR準拠のテストとは？

- 欧州評議会の言語政策部門では、言語テストをCEFRに関連付けるために必要な手続き、および、その理論的根拠、技術的側面をまとめた文書をマニュアルとして出版している（Council of Europe. 2009）。

具体的な言語テストをCEFRに関連づけるためには、

1)Familiarization（パネル参加者のCEFRに対する習熟）、

2)Specification（テスト内容とCEFRの整合性評価）、

3)Standardization（判断の標準化）、

4)Empirical Validation（経験的妥当化）

という、相互に関連した4つの手続きを経ることが要請されている。

- マニュアルの試行版が2003年の終わりに配布されたのに続いて、2005年初頭にReference Supplement（補遺）が出された。
- 基準設定、古典的テスト理論、テストの妥当化に関する質的方法、一般化可能性理論、因子分析、そして、項目応答理論へのアプローチなどの広範囲な議論が述べられている。
- このreference supplement ではマニュアルそれ自身よりも、より特化した技術的な情報が提供されている。現在は2009年に出版されたマニュアルおよび補遺が最新の内容になっている（Council of Europe. 2009）。

しかし、・・・

- CEFR「準拠」でないといけないのか？
- CEFRのレベルで能力が表わされると、
  解釈に広範囲な共通性が生まれる→　便利！
- CEFR「参照」テストもあり得る
- どのようなテストにするかは、開発者（機関）の考え方による　→「考え方」は明示する！
- テストの性能の検証はevidence based に、公開性、透明性を持つべきである！

文　献

1)安高紀子(2015).「第10章　コンピュータによる日本語口頭能力テスト」, 李在鎬 編 (2015).『日本語教育のための言語テストガイドブック』., くろしお出版.

2) Brennan,R. (2007). Generalizability Theory. Springer.

3) Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching,Assessment (CEFR). Manual* . Strasbourg, France: Language Policy Division.

4) 平井洋子 (2007). 主観的評定における評定基準, 評定者数, 課題数の効果について－一般化可能性理論による定量的研究. 首都大学東京人文学報, 380, 25−64.

5) 池田央 (1994). 現代テスト理論. 朝倉書店.

6) McNamara (1996). *Measuring Second Language Performance*.
London and New York: Addison Wesley Longman.
7) North,B. (2005). The Development of a Common Framework Scale of Descriptors of Language Proficiency Based on a Theory of Measurement. *System,* 23, 445-465.
8) North, B. & Schneider, G. (1998). Scaling descriptors for language proficiency scales, *Language Testing*, 15, 217-262.
9) 野口裕之・大隅敦子 (2014).『テスティングの基礎理論』. 研究社.
10) 野口裕之 (2015).「第11章 大規模言語テストの世界的動向」, 李在鎬 編 (2015).『日本語教育のための言語テストガイドブック』., くろしお出版.
11) Rasch,G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmarks Paedogogiske Institut.（内田良男 監訳1985『心理テストの確率モデル』 名古屋大学出版会）

12) 三枝　令子ほか　2004　日本語Can-do-statements尺度の開発, 科学研究費補助金研究成果報告書.
13) 庄司惠雄・野口裕之・金澤眞智子・青山眞子・伊東祐郎・迫田久美子・春原憲一郎・廣利正代・和田晃子　(2004). 大規模口頭能力試験における分析的評価の試み. 日本語教育, 122, 42-51.
14) Tannenbaum,R.J. and Wylie,E. (2008). Linking English-Language Test Scores Onto the Common European Framework of Reference: An Application of Standard-Setting Methodology, RR-08-34, ETS, Princeton, NJ.

ご清聴ありがとうございます。