

第三の評価者が多相ラッシュモデルのパラメタ推定に及ぼす効果の検討

—英語スピーキングテストの評価データを用いて—

○光永悠彦 (名古屋大学)
神澤克徳 # (京都工芸繊維大学)

羽藤由美 # (京都工芸繊維大学)

キーワード：パフォーマンステスト、項目反応理論

背景と目的

京都工芸繊維大学において行われている、英語発話能力を測定するためのテスト(KIT test, Hato et al, 2017) は CBT (computer based test) として行われており、受験者の発話内容が2名の評価者によって、二つの評価観点(Task Achievement 及び Task Delivery) から評価される。パフォーマンス評価においては、2名の評価結果の間に著しい乖離があった場合、第三の評価者が評価しない方法が一般的である。KIT test においても同様の手続きが行われ、第三の評価者による評価結果がデータセット内に部分的に存在している。

宇佐美 (2011) では、一事例の研究ではあるものの、論述式テストデータを用いた一般化可能性理論による分析で、採点者数が概ね4名を超えると一般化可能性係数の効果が頭打ちになることを指摘している。第三の評価者による採点法は部分的に採点者が増えることで評価の信頼性を高める効果があるといえるが、そのことによって推定されるパラメタに差異がみられたならば、評価者を増やしたことが項目特性などの評価に影響することにつながる。本研究では、KIT test のデータを題材に、多相ラッシュモデル (many-facet Rasch model, MFRM; Linacre, 1994) を当てはめる際、第三の評価者の結果を考慮に入れた場合とそうでない場合で、採点者間信頼性や推定されるパラメタにどのような違いがあるかを検討する。

方 法

2017年12月に実施されたKIT testによる567名のデータを用いた。9項目からなるテスト版を3種類用意し、受験者(大学1年次生)には3種類の版のうちランダムに一つを割り付けた。また77名のモニター受験者に対し、3種類のテスト版すべてを解答させた。解答はすべて英語発話で行わせ、その内容を録音しておき、一つの解答につき評価者2名(英語ネイティブとノンネイティブ)を割り当て、独立して採点させた。評価観点ごとに6件法で評価させ、評価結果に2段階以上の差が生じた評価については、第三の評価者による評価結果を収集した。

モデルは「評価者」「評価観点」「項目」「受験者」の4相を仮定したMFRMとし、分析はFacets 3.71.4を用いて行った。ただし、評価者の属性を考慮し、評価者と評価観点との間に交互作用項を仮定した。

結果と考察

7182個の解答に対し、第三の評価者が下した評価の数は521個(約7.3%)であった。これらの評価を考慮したモデルとしないモデルでの項目困難度を27項目(9問×3版)についてプロットした結果をFigure 1に示した。また2項目について、第三の評価者を考慮した場合、困難度パラメタの推定値の標準誤差が0.1程度小さくなった。

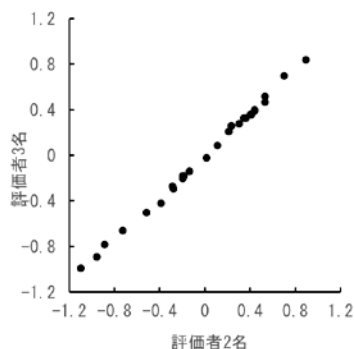


Figure 1 第三の評価者を入れた場合と入れなかった場合での項目困難度の推定結果の比較。

評価観点や評価者の評価の厳しさ、カテゴリ境界パラメタについては、モデル間で大きな差はみられなかった。これらの結果から、第三の評価者を入れることは項目困難度や他のパラメタの推定結果に大きく影響しないことが指摘できる。ただし、項目数が少ない場合や評価の不一致がより多い場合において、更なる検討が必要であろう。

引用文献

- Hato, Y., Kanzawa, K., Tsubota, Y., Mitsunaga, H., Shimizu, Y. & Edmonds, G. (2017). Developing a CBT Speaking Test of English as a Lingua Franca: The Evolution of Rating Scale. JACET 56th International Convention.
Linacre, J.M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.
宇佐美 慧 (2011). 小論文評価データの統計解析—制限字数を考慮した測定論的課題の検討—行動計量学, 38, 33-50.

付 記

本研究は科学研究費補助金(基盤研究(B)、課題番号16H03448)の助成を受けて行われた。