公益財団法人　日本英語検定協会

2014 年度　英語教育研究センター　委託研究

ICT 等を活用した評価についての調査・研究

報告書

2015 年 3 月 31 日

池田　央・村木英治・大友賢二

中村洋一・法月　健

公益財団法人　日本英語検定協会
2014 年度　英語教育研究センター　委託研究

ICT 等を活用した評価についての調査・研究

報告書

＊＊＊＊＊

はしがき

公益財団法人　日本英語検定協会　英語教育研究センターの中には、特定のテーマごとに委員会を構成することになっている。その委員会は、リーダー、アドバイザー、さらに研究協力者として外部のサブコミティーメンバーで構成することになっている。また、その委員会の意思決定・作業手順の原則は、委員の中で相談の上、リーダーがこれを決定することになっている。

本委託研究は、日本英語検定協会理事長　松川孝一と常任審議委員　大友賢二の間で平成 26 年 6 月 1 日に以下のとおり締結された契約書に基づくものである。

研究題目：　「ICT 等を活用した評価についての調査・研究」
アドバイザー：池田　央、村木英治
リーダー：大友賢二
サブコミティーメンバー：中村洋一、法月　健
研究期間：平成 26 年 6 月 1 日から平成 27 年 3 月 31 日

本報告書の 1 は、進捗状況報告書として執筆された、大友賢二、中村洋一、法月健による報告書である。

本報告書の 2 は、2015 年 1 月 16 日に常任審議委員会で行なわれた進捗状況報告会に用いられた Powerpoint の資料である。

本報告書の 3 は、最終報告書「Mixture Rasch Model の考察と展望」に関する池田　央、大友賢二、中村洋一、法月　健によって執筆されたものである。

以上の研究報告書が公益財団法人 日本英語検定協会 英語教育研究センターのさらなる発展に少しでも貢献できれば、この上ない喜びである。

大友賢二
2015 年 3 月 31 日

委託研究　進捗状況報告書：1．1.
＊再考：言語テストの規準設定＊

大友賢二(筑波大学名誉教授)

1.　なぜ、規準設定の方法を究明する必要があるのか：

この研究は、ICT 等を活用した評価についてのものである。ICT(Information and Communication Technology)というのは、 情報通信技術を表す IT に、コミュニケーションの概念を加えた言葉である。つまり、コンピュータやインターネットに関連する情報通信技術のことである。それなどを活用した評価に関するものがこの研究の主題である。これを評価の角度から考察、また情報通信技術の中で評価の課題を取り上げるとすれば､どんなことが検討されなければならないかということになる｡ここでは、その評価に関する多くの分野の中でも、課題の中心を「規準設定」(standard setting) と結びつけて考えることとする。しかし、規準設定とは何かを明確にしておくことが必要である。規準設定というのは、ごく簡単に言えば、a process by which a standard or cut score is established. ということである。たとえば、ABC という 3 つの外国語習得水準があるとした場合、ある受験者の能力水準が C ではなくて、B と判断できるには、どんな条件が必要かということである。また、どんな状態であれば、B ではなくて A と判断できるかという課題でもある。

1.1.　CEFR と CAN-DO statements：

なぜ、「規準設定」が、「ICT 等を活用した評価」という領域に関連するのであろうか？その第 1 の理由は、現在の英語教育の流れと切り離すことができない課題のひとつ：CEFR および CAN-DO statements　のもっている重要な課題と「規準設定」は密接に結びついているからである。

我が国の外国語教育に関する動向の流れの中で､平成 26 年 2 月から 9 月にかけて、「英語教育のあり方に関する有識者会議」が 9 回行われ、その審議のまとめとして、「今後の英語教育の改善・充実方策について～グローバル化に対応した英語教育改革の 5 つの提言～」がまとめられたことは、(向後(2015；10-15))からも知ることができる。提案された 5 つの概要は、改革 1：国が示す教育目標・内容の改善、改革 2：学校における指導と評価の改善、改革 3：高等学校・大学の英語力の評価及び入学者選抜の改善、改革 4：教科書・教材の充実、そして改革 5：学校における指導体制の充実である。このなかで、とくに「規準設定の方法」に関連するものは、改革 2：学校における指導と評価の改善、3：高等学校・大学の英語力の評価および入学者選抜の改善、

を取り上げなければならない。コンピュータなどでの外国語教育の中でも、指導と評価の方法は、さらに検討されなければならない。しかし、目標に到達しているかどうかは、どうしたら分かるであろうか？この課題を解く第一歩は、目標設定のための要素とは何かをまずとりあげなければならない。目標が明確でなければ、当然、その目標が到達されたかどうかは、検討の仕様もないからである。その目標設定とその目標に到達したかどうかの判断は、どうしたら可能になるかということである。

現在、世界の言語テスト界で広く利用されている CEFR に対して、わが国でも広く議論されているが、建設的な批判もあることは理解しておかなければならない。たとえば、中津原(2010)や野口・大隈(2014:165)でも指摘されているように、「CEFR で測られているのは、学習者の能力ではなく、教師や評価官がその能力をどう受けとったか、である。」(North, 2000: 573)や「CEFR は実証的に妥当性が示されている言語能力の記述、あるいは言語学習過程のモデルに基づいたものではない」(Fulcher, 2003)や、「CEFR の言語能力記述尺度における'can-do statements' の困難度は、実際の言語行動場面の文脈に依存して変動するが、そのことが十分に考慮されていない」(Weir, 2005)などもあるということである。Can-do statements は、規準設定(standard setting)に深く関連するもので 、ICT 等を活用した評価の中でも取り上げられなければならない重要な課題のひとつである。

1.2. Standard Setting 研究結果の流れ：

なぜ、「規準設定」が、「ICT 等を活用した評価」という領域に関連するのであろうか？その第 2 の理由は、現在の英語教育の流れと切り離すことができない課題のひとつ：「規準設定研究結果」の流れが、じつに様々であり、可能であれば、一つの方向を求める必要がありと考えられるからである。

わが国ではあまり論点とはなっていないようであるが、この規準設定法に関する研究は、今までの外国語教育界で行われてきていないわけではない。その中の一つには、規準設定の方法を 4 つに分類している Hambleton & Pitoniak (2006 ; 440) がある。(1) methods that involve review of test items and scoring rubrics, (2) methods that involve review of candidates, (3) methods that involve looking at candidate work, (4) methods that involve panelist review of score profiles. がそれである。(1)の分類に該当するものとしては、Angoff Method, Extended Angoff and Related Methods, Ebel Method, Nedelsky Method, Jaeger Method, Bookmark and Other Item Mapping Methods, Direct Consensus Method などがある。(2)の分類に該当するものとしては、Borderline Group Method や Contrasting Groups Method などがある。また、(3)の分類に該当するものとしては、 Item-by-Item Approaches, Holistic Approaches, Hybrid Approaches を上げることができる。そして(4)の分類に該当するものとしては、Judgemental Policy Capturing method, Dominant Profile Method, Item Cluster Method, また、Compromise Methods としては Hofstee Method,

Beuk Method などを上げることができる。

こうした多くの規準設定法に関して、これまで検討された結果を概観すると、以下のような(1)否定的見方、(2)中立的見方、(3)肯定的見方などがある。「ICT 等を活用した評価の研究」の中で、規準設定法を取り上げなければならない理由は、これまで提案された規準設定法の研究に対する反応が、以下のように、否定的、中立的、肯定的というものがある。そこで、いったいどの方向を取り上げるのが適切であるかを究明する必要があるからである。ちなみに、(1)の否定的見方は、Kaftandjieva, F.(2004；31), AERA,APA&NCME (1999；53), Jaeger and Mills (2001；314)などに見ることができる。(2)の中立的見方は、Cizek and Bunch (2007；320), Zieky, Perie & Livingston (2008；197) などに見られる。(3)の肯定的立場は、これまで開発された規準設定法は恣意的考察にすぎないという見方に対する反論でもある。例えば、Nicholes, Twinge, Mueller and O'Malley (2010；14-24), Peterson, Schulz & Engelhard (2011；3-14)などがある。ここでは、特に、NAEP standard setting などでは、Angoff Method より Bookmark Method を強く支持している方向が理解できる。しかし、どの方向がより妥当であるかを見極め、それを ICT 等で活用することができるかを究明することは極めて重要な課題である。

1.3. 大学入試と段階別表記：

なぜ、「規準設定」が、「ICT 等を活用した評価」という領域に関連するのであろうか？その第 3 の理由は、最近話題になっている「段階別表記」と深い関わりを持っているからである。

最近の中教審答申案に関する課題は、いたる所に見られる。例えば、中央教育審議会高大接続特別部会配布資料(10/24/2014)では、大学入試に関して、「一点刻み」の客観性にとらわれた評価から脱し、各大学の個別選抜における多様な評価方法の導入を促進する観点から、大学及び大学入試希望者に対して、「段階別表示」などによる成績提示を行う」と述べている。なるほどと、考えさせられる答申案と受け止められる場合もあろう。そして、この根底には、テスト得点の解像度や測定誤差の課題が見えてくる。

例えば、「体重計が、A さんの体重を 65kg と報告したとき、私たちは、A さんの体重が 64kg でもなく、66kg でもなく、まさに 65kg であると信じることができる。次の日に測定したら、値は微妙に変わるが、A さんのまさにその瞬間の体重は、65kg であると私たちは常に疑わない。しかし、ある英語テストで A さんの英語力が、65 点と評価されたとき、私たちは、その得点が 64 点でもなく 66 点でもなく、まさに 65 点であると信じることができるであろうか。」(荘島(2010；83))。ここでは、テストは、同じぐらいの学力の持ち主の違いを見分けるほどの解像度が高い測定道具ではないと言う。テストは、学力をせいぜい 5−20 段階に見分けるくらいの分解能力しかないと言う。

これを後押しするかのように、「暫定入学」という見方が、入試方法の話題のひ

とつとして取り上げられていることに気づく。「国際教養大学には、入試の点がわずかに合格ラインに届かなくとも、「暫定入学」を認める制度がある。1 年間、正規学生と並んで勉強し、一定の成績をとれば、正規学生になれる。大学によると、毎年、数人がこの制度で入学し、ほとんどが正規学生になっているという。」(浅倉(「朝日新聞：/4/1/2014」))

しかし､こうした現象に対し､大学入試における段階別表示を行うことの問題点は、以下のように数多く指摘されていることも事実である。(1)情報量の減少、(2)段階への分け方の恣意性、(3)段階内での個人差や、個人内変化の無視と段階間で差や変化の誇張、(4)選抜における測定結果の多様な活用を阻害(南風原(2014；50))。

こうした選抜のための測定に関する規準設定に関して、ICT 等を活用した評価の視点では、どう取り上げるのが適切か、極めて大きな課題である。

1.4. Mixture Rasch Model の考察：

これまでの多くの規準設定法に向けられている批判は多様である。その中には、たとえば、psychophysical method, JND (just-noticeable difference)などと呼ばれている視点がある。しかし、そうした課題を乗り越えた客観的・統計的解決法はないのであろうか？規準設定の方法を再考しなければならない第 4 の理由は、そこに存在する。それは、一体何であろうか？また、それを求めることに意味があるのであろうか？ICT の中でそれを探ることは可能であろうか。

規準設定法の多くは、古典的テスト理論の連続尺度を仮定している。また、項目応答理論の theta というもののうえに、この潜在的な連続尺度を仮定しているものもある。例えば、Bookmark Method などがそうである。しかし、この視点を変えた、潜在的な順序尺度を仮定している考え方があるということである。学力を段階別表示するためのテスト理論に関連する「潜在的な順序尺度」の存在である。こうした中で、規準設定のための客観的データ取得を目指す方法の一つとして注目を浴びているのが、いわゆる、Mixture Rasch model (略称 MRM) と言われるものである。この解決に向けて論じられているのは、純粋な順序尺度に基づく「潜在ランク理論」(Latent Rank Theory: LRT)に､連続尺度の精度向上をめざす Rasch model の特性を補充することで、より明確な、より実用的な基準設定法を見出して行こうというものである。

Templine & Jiao (2012; 387)による説明によると、この MRM の研究は、Kelderman & Marcredy (1990); Mislevy & Verhelst (1990); Rost (1990) などにその発端を見ることができる。ここでは、Rasch model に加えて、LRT を基盤としている Latent Class Analysis (LCA) を統合して、テストデータを分析することであるとしている。これに関する考察は、第 3 章「課題の内容と展望」で、さらに詳しく述べることとする。

1.5. 異なるデータを用いた規準設定：

規準設定に関しては、選抜を目的とした場合と学習を目的とした場合との 2 つの視点から、それぞれ異なった角度の考察が必要である。この視点を混同すれば、さらなる混乱をきたすことになることを、ここで、確認しておかなければならない。ICT 等を活用した評価の中で規準設定を検討しなければならない第 5 の理由は、こうした異なるデータを用いた規準設定に関する課題である。

段階別表示を目標とすれば、古典的テスト理論や項目応答理論による連続尺度ではなく、それにふさわしい順序尺度を用いることの意味を確認しておかなければならない。高校入試、大学入試などの英語能力測定においては、段階別表示をよしとする視点は、先に述べたように、「暫定入学」など、一見して妥当という印象を与えるかもしれない。しかし、「暫定入学」の中に潜んでいる基本的視点は、観察されるテスト得点は、真の得点と誤差とから構成されているという考え方と関係があるのかもしれない。つまり、テスト得点には、誤差があるという考え方がある。したがって、テスト得点自体に絶対の信頼をゆだねることは適切ではないという考え方かもしれない。そうであれば、測りたい特性を反映する成分、つまり、真の得点の構成を大きくして、テストの項目の妥当性を高めるように努力をしなければならない。

その一方で、考えなければならない最も重要な問題は、古典的テスト理論、項目応答理論で求めた連続尺度のデータを取り出して、それをニューラルテスト理論や潜在ランク理論などで求めた順序尺度データとみなしている点にある。もっと明確に言えば、学力を段階別表示するためのテスト理論によらないで作成したデータを用いて、それをもって学力を段階別表示する作業を進めているとすれば、それは問題であるということである。つまり、連続尺度のデータを用いて段階別表示を行うということ、の矛盾を認識しなければならない。段階別表示で 5 段階を構成するために、連続尺度のデータをそのまま用いているというのであれば、それは極めて無理であると言わなければならない。

規準設定は、2 つの目的を達成するために存在することに、その意味がある。1 つは、資格判定や入試合否判定のための目的である。もう 1 つは、外国語学習など、指導と学習に関連した目的である。資格判定や入試に関する基準設定は、どちらかといえば、項目応答理論等で求めた連続尺度を用いる方向をとるのが適切であろう。しかし、指導と学習に関しては、どちらかといえば、潜在クラス分析等で求めた順序尺度を用いる方向をとるのが適切であろう。もし、一つ上の級への合格可能性を指導するのに役立てることを重視するとしたら、順序尺度に加えて連続尺度も活用できる手立てがあれば、より適切と考えられる。広い視野に立つ規準設定に関するさらに建設的な検討結果のひとつとしては、Tannenbaum & Cho (2014)など、ごく最近の発言にその一端を見ることもできる。

Rasch model と latent class analysis を用いた Mixture Rasch Model を有効に活用するために、どのような目的を達成するのには、どのような手順を設定するのがもっとも適切であるかを十分検討しなければならない。Mixture Rasch Model におけるそのためのさまざまな考察に関しては、のちに、第 3 章で述べることとする。

参考文献

AERA, APA & NCME (1999). *Standards for Educational and Psychological Testing*, 53. AERA

Cizek, G.J. and Bunch, M.B. (2007). *Standard Setting, A Guide to Establishing and Evaluating Performance Standards on Tests.* 320. Sage

Fulcher, G. (2003). *Testing second language speaking.* New York, NY: Pearson Longman.

Hambleton, R.K. & Pitoniak, M.J. (2006) Setting Performance Standards. In Brennan, R.L. (ed.) *Educational Measurement* (Fourth Edition). 434-470, ACE & Praeger Publishers.

Jaeger, R.M. and Mills, C.N. (2001). An Integrated Judgment Procedure for Setting Standards on Complex , Large-Scale Assessments. In Cizek, G.T. (ed.) *Setting Performance Standards*. 314. Lawrence Erlbaum Associates, Publishers.

Kaftandjieva, F. (2004). Section B: Standard Setting., *Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR.* 31. Council of Europe.

Kelderman, H. & Marcready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement,* 27, 307-327.

Mislevy, R.J., & Verhelst (1990). Modeling item responses when different subjects employy different solution strategyes, *Psychometrika,* 55(2), 195-215.

Nicholes, P., Twing, J. Mueller, C.D. and O'Malley, K. (2010). Standard-Setting Methods as Measurement Process, *Educational Measurement: Issues and Practice,* 29(1), 14-24.

North, B. (2000) *The development of common framework scale of language proficiency*. New York: Peter Lang.

Peterson, C.H., Schulz, E.M., and Engelhard, Jr. G. (2011). Reliability and Validity of Bookmark-based Methods for Standard Setting, *Educational Measurement: Issues and Practice*. 30(2). 3-14.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271-282.

Tannenbaum, R.J. & Cho, Y. (2014). Critical Factors to Consider in Evaluating Standard-Setting Studies to Map Language Test Scores to Frameworks of Language Proficiency. *Language Assessment Quarterly*, 11:3, 233-249.

Templin, J. & Jiao, H. (2012). Applying model-based approaches to identify performance categories. In G.J. Cizek (Ed.) *Setting performance standards*. (Second Edition). 379-397. New York, Routledge.

Weir, C.J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22,281-300.

Zieky, M.J., Perie, M. & Livingston, S.A. (2008). *Cutscore: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. 197. ETS

浅倉拓也(2014).「教育 2014　学歴は変わるか」. 『朝日新聞』(4/1/2014).

南風原朝和(2014).「入試選抜の測定問題」『大学入試センターシンポジューム 2014：大学入試の日本的風土はかえられるか 11/29/2014』、50, 大学入試センター.

向後秀明(2015).「グローバル化に対応した今後の英語教育改革–小・中・高における英語教育改善の方向性」　『英語展望 No.122』.10-15. ELEC.

中央教育審議会高大接続特別会(2014).「学力評価のための新たなテスト(仮称)」. 中教審 10/24/2014 配布資料.

中津原文代(2012).「能力基準としての Can-do statements とテストの妥当性を検証する Socio-cognitive Framework について」『第 3 回言語教育評価フォーラム：Can-do statements をどう活用するか』主催：桜美林大学・国際交流基金　言語教育評価共同研究所：9/15/2012

野口裕之・大隈敦子(2014).『テステングの基礎理論』. 165, 研究社.

荘島宏二郎(2010).「ニューラルテスト理論–学力を段階評価するための潜在ランク理論」.植野真臣·荘島宏二郎著『学習評価の新潮流』.83, 朝倉書店.

2. 主な先行研究資料としては、どのようなものがあるか？

「再考：言語テストの規準設定」に関する先行研究資料としては、すでに述べた 1. 2. Standard setting 研究結果の流れ、および、1. 4. Mixture Rasch Model　の考察等で述べたものがある。それに加えて、見逃すことができない多くの先行研究資料の中でも、特に取り上げたい資料を、2. 1. 言語テスト一般 と、2. 2. 専門雑誌による Mixture Rasch model とに分ければ、次のようになる。この 2. 2. では、数ある研究の中のもっとも新しい論文を取り上げている。また、とくに注目に値するものの幾つかを、2. 3. 先行研究:　abstract と寸描として示すこととする。

2.1.　言語テスト一般の先行研究：

Fulcher, G. (2012). Scoring performance tests. In Glenn Fulcher and Fred Davidson (Eds.). *The Routledge Handbook of Language Testing*, 378-392. Routledge.

Jiao, H., Lissitz, R.W., Macready, G. Wang, S. & Liang, S. (2011). Exploring levels of performance using the mixture Rasch Model for standard setting. *Psychological Test and Assessment Modeling*, Vol. 53, 2011 (4). 499-522.

Lissitz, R. W. (2013). Standard Setting: Past, Present and Perhaps Future. In M. Simon, K. Ercikan and M. Rousseau (Eds.). Improving Large-Scale Assessment in Education. 154-174. Routledge Taylor & Francis Group.

Pitoniak, M.J. and Morgan, D.L. (2012). Setting and Validating Cut Scores for Tests. In Charles Secolsky and D. Brian Denison (Eds.) *Handbook on Measurement, Assessment, and Evaluation in Higher Education*. 343-366. Routledge Taylor & Francis Group.

Way, W. D. and Mcclarty, K.L. (2012). Standard Setting for Computer-based Assessments. In Gregory J. Cizek (Ed.). *Setting Performance Standards: Foundations, Methods, and Innovations*, Second Edition, 379-398. Routledge Taylor & Francis Group.

2.2. 専門雑誌による **Mixture Rasch Model** 関係の先行研究：

2.2.1. ETS RESEARCH REPORT SERIES:

There are 45 results for: Mixture Rasch Model in ETS Research Report Series:

Vol. 1987, Issue 2, Dec. 1987, pp.i-52, Robert. J. Mislevy, and Norman Verhelst

Vol. 1992, Issue 1, June 1992, pp.i-30, Eiji Muraki

Vol. 2014, Issue 1, June 2014, pp.1-48, Patrick. C. Kyllonen, Anastasiya. A. Lipnevich, Jeremy Burrus and Richard D. Roberts/

2.2.2. PSYCHOMETRIKA:

There are 121 results for MIXTURE RASCH MODEL in Psychometrika.

Stefano Noventa, Luca Stefanutti, Giulio Vidotto (2014). An Analysis of Item Response Theory and Rasch Model Based on the Most Probable Distribution Method.

2.2.3. APPLIED PSYCHOMLOGICAL MEASUREMENT：

Results of 84 found for Mixture Rasch Model (all words) in Full Text in selected journals: Applied Psychological Measurement.

Hong Jiao, George Macready, Junhui Liu, and Youngmi Cho (2012). A Mixture Rasch Model-Based Computerized Adaptive Test for Latent Class Identification, *Applied*

Psychological Measurement, September 2012: Vol. 36, 6: 469-493.

2.2.4. JOURNAL OF EDUCATIONAL MEASUREMENT:

There are 44 results for: Mixture Rasch Model in Journal of Educational Measurement.

Lianghua Shu and Richard D. Schwarz (2014). IRT-Estimated Reliability for Tests Containing Mixed Item Formats, *Journal of Educational Measurement*, Vol.51. Issue 2, 163-177.

2.2.5. EDUCATIONAL MEASUREMENT: ISSUES AND PRACTICE:

There are 6 results for: Mixture Rasch Model in Educational Measurement : Issues and Practice.

Laine Bradshaw, Andrew Izak, Jonathan Templin and Erik Jacobson (2014). Diagnosing Teachers' Understanding of Rational Numbers: Building a Multidimensional Test Eithin the Diagnostic Classification Framework. *Educational Measurement: Issues and Practice*. Vol. 33, Issue 1, 2-14.

2.2.6. LANGUAGE TESTING:

Results of 12 found for Mixture Rasch Model (all words) in Full Text in selected journals: Language Testing.

Thomas Eckes (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*. Vol.31. 1: 39-61.

2.2.7. LANGUAGE ASSESSMENT QUARTERLY:

Displaying of 14 results for Mixture Rasch Model in Language Assessment Quarterly.

Khaled Barkaoui (2013). Using Multilevel Modeling in Language Assessment Research: A Conceptual Introduction, *Language Assessment Quarterly,* Vol. 10, Issue 3, 241-273.

2.3. 先行研究：abstract と寸描：

2.3.1. Cizek, G.J. and Bunch, M.B. (2007)

Cizek, G.J.は、「規準設定法」に関するもっとも著名な研究者の一人である。あの有名な NCME (National Council on Measurement in Education) の President (2012-2013)

の大役も務めている。Cizek, G.J. (Ed.) (2001), *Setting Performance Standards*. Lawrence Erlbaum Associates, Cizek, G.J. (2006), Standard Setting. In Downing and Haladyna (Eds.). *Handbook of Test Development*, Lawrence Erlbaum Associates, Cizek, G.J. (Ed.)(2012). *Setting Performance Standards* (Second Edition).Routledge などは、その重要な文献である。本書における彼の次の発言は、きわめて印象的である。According to Segal, "A man with a watch knows what time it is. A man with two watches is never sure." Because there is no equivalent of an atomic clock in the field of standard setting, our recommendation is simply for practitioners to invest in a single watch of greatest quality given available resources. (p.320)

2.3.2. Jiao, H., Lissitz, R.W., Macready, G., Wang, S., and Liang, S. (2011)

Exploring levels of performance using the mixture Rasch model for standard setting というこの論文は、*Psychological Test and Assessment Modeling*, Vol.53, 2011(4)に掲載されたものである。この論文は、University of Maryland の Department of Measurement , Statistics and Evaluation の Associate Professor である Hong Jiao が中心になって書かれたものである。その要約では、Mixture Rasch Model の優れている点を 4 つほど上げているが、その第 1 として指摘している次の点は、注目に値する。There is a lack of a data-driven statistical validation step in the current widely used standard setting methods. The proposed method based on the mixture Rasch model provides empirical evidence related to the validity of the number of proficiency levels set by policy based on both qualitative and quantitative data. (p. 514)

2.3.3. Jiao, H., Macready, G., Liu, J., and Cho, Y. (2012)

This study explored a computerized adaptive test delivery algorithm for latent class identification based on the mixture Rasch model. Four item selection methods based on the Kullback–Leibler (KL) information were proposed and compared with the reversed and the adaptive KL information under simulated testing conditions. When item separation was large, all item selection methods did not differ evidently in terms of accuracy in classifying examinees into different latent classes and estimating latent ability. However, when item separation was small, two methods with class-specific ability estimates performed better than the other two methods based on a single latent ability estimate across all latent classes. The three types of KL information distributions were compared. The KL and the reversed KL information could be the same or different depending on the ability level and the item difficulty difference between latent classes. Although the KL information and the reversed KL information were different at some ability levels and item difficulty difference levels, the use of the KL, the reversed KL, or the adaptive KL information did not affect the results

substantially due to the symmetric distribution of item difficulty differences between latent classes in the simulated item pools. Item pool usage and classification convergence points were examined as well.

2.3.4. Mislevy, R.J. and Verhest, N. (1990)

MRM (Mixture Rasch Model) 関係の論文に、必ず掲載されているものの一つは、1987 年の論文と全く同じ内容であるこの 1990 年のものである。どちらかといえば、1990 年の Modeling Item Responses When Different Subjects Employ Different Solution Strategies, *Psychometrika*, Vol.55, No.2, 195-215 が、多く引用されている。これは、Educational Testing Service の Robert J. Mislevy, と CITO (National Institute for Educational Measurement) Arnhem, The Netherlands の Norman Verhelst によるものである。筆者は、8/2-8/4, 1993 , Queen's College, Cambridge, UK, 8/5-8/7, 1993, Hotel Haarhuis, Arnhem, Netherlands で行われた Language Testing Research Colloquium 1993 で、CITO を訪れたことがある。現在、Program Manager PISA 2018 Framework Development at Pearson plc である John H.A.L. DeJong は、当時、Head Lang Testing Unit , CITO の要職にあった。

2.3.5. Peterson, C.H., Schulz, E.M., Engelhard, Jr. G. (2011)

Standard Setting の方法に関して、否定的な見方をしている研究者に対し、この論文は、肯定的な考えを示すものの一つである。Reliability and Validity of Bookmark-Based Methods for Standard Setting というこの論文でもっとも力点を入れているのは、Bookmark-Based Methods についてである。NAEP(National Assessment of Educational Progress)で長い間用いられてきた Angoff-Based Methods は、2005 年に再検討がなされた結果、Bookmark-Based Methods の方が、妥当性、信頼性においても、より優れていることが認められた。その結果、規準設定のための結論に達するまでの検討時間、その費用などの観点からも、NAEP の基準設定には、Angoff-Based Methods よりも、Bookmark-Based Methods がより適切であろうとの結論に達している。

2.3.6. Rost, J. (1990)

この論文は、mixture Rasch model 開発の重要な先行研究の一つである。この研究結果が引用されている論文は、極めて多い。その数は、32にも及んでいることは、注目に値する。その概要は、以下のとおりである。A model is proposed that combines the theoretical strength of the Rasch model with the heuristic power of latent class analysis. It assumes that the Rasch model holds for all persons within a latent class, but it allows for different sets of item parameters between the latent classes. An estimation algorithm is outlined that gives conditional maximum likelihood estimates of item parameters for each class. No a

priori assumption about the item order in the latent classes or the class sizes is required. Application of the model is illustrated, both for simulated data and for real data。

2.3.7. Templin, J. and Jiao, H. (2012).

Applying Model-Based Approaches to Identify Performance Categories というこの論文で、Mixture Rasch Model-Based Standard Setting Methods という見出しの中に示されている内容で、特に注意を引く一節は、次のものである。The MRM (Mixture Rasch Model) assumes that multiple latent student populations exist and that the Rasch model holds within each latent class with differing item difficulty parameters across classes. When compared with the LCA (Latent class analysis) model, by incorporating the Rasch model within a latent class, the MRM provides a model where items are allowed to be correlated within each class. Each student is characterized by two latent variables, a continuous quantitative variable that provides a measure of the trait of interest, and a categorical qualitative variable which differentiates among respondents who differ in their likelihood of correctly responding to items. Thus, a student performance on an item is determined by its discrete qualitative group membership and the continuous quantitative latent ability. (p. 387)

委託研究　進捗状況報告書：1．2.
＊21st Century Skills と Standard Setting＊

中村洋一(清泉女学院短期大学教授)

1. “21st Century Skills” と “Standard Setting”

21 世紀、さらには 22 世紀を生きる世代のために、今、英語教育研究は何ができて、何をすべきなのだろうか？日本における英語教育研究は、その歴史の中で、時代の要求を見極めながら、その時代のあり方を探ってきた。しかし、英語教育のみならず日本の教育全体に関する昨今のニュースは、ここに来て大きな変化の時代に差し掛かっていることを伝えている。

> 文部科学省の中等教育審議会は、現行のセンター試験を廃止し、2020 年度から新しい大学入試を導入することを答申している。その中で、「学力評価テストは、マークシート方式に記述式を加え、成績は一点刻みではなく段階で」、「将来的には、教科の枠を超えた設問のみでの評価を目指す」、ことを提言している。また、英語に関しては、「『読む』『書く』以外に『聞く』『話す』の評価を求め、英検や TOEFL (トーフル) など民間試験の活用の検討」も提言した。
>
> 主要六教科の基礎学力を測るため「高等学校基礎学力テスト」を一九年度から新設。原則マークシート方式で、高校二、三年次に年二回程度受けられる。最低限の学力を担保する狙いで、推薦入試などの参考資料に活用できる。各大学の個別入試は「一般」「推薦」、面接や小論文で個性を評価する「AO」の区分を撤廃。各大学は人物像や評価基準などの受け入れ方針を示した上で、人物を多面的に評価するよう提言した。
>
> (http://www.chunichi.co.jp/article/front/list/CK2014122302000072.html)

大学入試に関するこの答申は、日本の中等教育の将来に大きな影響を与える方向性を示している。しかし、従来までの日本における教育文化を鑑みれば、この答申の一つひとつの提言を実現するまでには、解決すべき多くの課題が未解決のまま山積している現状をどうしていくのか、という悲観的な見通しもぬぐいきれない。「一点刻みではなく段階で」成績を算出するとあるが、その具体的な統計的手法は研究され、見通しがついているのか?「教科の枠を越えた設問」の内容的妥当性をどのように担保していくのか?「民間試験」と学校教育の教育内容とのすりあわせはどのように処理して

いくのか?「年二回程度受けられる」テストの公平な結果処理について、具体的な対応策はあるのか?項目応答理論の適応による等化は、その可能性を持つ概念だと考えられるが、そのようなテスト理論の理解・啓蒙は、充分と言えるのだろうか?このようないくつかの“?”に、次のような反応も寄せられている。

新たな大学進学テストは、センター試験と異なり、従来の「教科型」に加えて数学と理科など複数の教科を合わせた「合教科型」や教科の枠を超えた「総合型」も出題。記述式問題を取り入れ英語の民間試験も利用する。年複数回実施し、将来はコンピューター出題方式の導入を目指す。各大学が個別入試に使う成績は、1 点刻みではなく段階別評価で示す。実現すれば共通 1 次試験導入 (1979 年) 以来の大改革だが、新テストの内容の具体化はこれからで、高校や大学側には不安や反発の声もある。

(http://www.yomiuri.co.jp/national/20141222-OYT1T50099.html)

中央教育審議会 (文部科学相の諮問機関、安西祐一郎会長) は 22 日、大学入試「改革」に関する答申を出しました。高校に「全国学力テスト」を新たに導入して入学者選抜に活用するなど、競争主義に拍車をかける危険性を抱えています。

答申は、「知識の暗記・再現に偏り、真の『学力』が育成・評価されていない」と指摘。「基礎学力を評価」するために“高校版全国学力テスト”ともいうべき「高等学校基礎学力テスト」の導入を打ち出しました。

現在のセンター入試は「大学教育に必要な能力を評価」として、「大学入学希望者学力評価テスト」に名称を変えて実施。その上に各大学が小論文や面接などで多面的な評価を行うとしています。

基礎学力テストは高校 2、3 年生が年 2 回程度実施。「希望参加」ながら選抜に活用され、半強制となる可能性が強いものです。

学力評価テストも複数回実施。試験内容はセンター試験に比べて難易度に幅を持たせ、複数教科を組み合わせた問題や記述回答を増やすとするにとどまりました。英語は「聞く」「話す」も評価するとして民間の資格・検定試験の活用を打ち出しました。

両テストとも結果は「一点刻み」でなく「段階別表示」にとどまっており、大学や学部学科ごとに入試を行うという世界に例のない競争的な制度の抜本的改革にはほど遠い内容です。

(http://www.jcp.or.jp/akahata/aik14/2014-12-23/2014122301_03_1.html)

「教科型」、「合教科型」、「総合型」の出題とあるが、どのように「合教科」あるいは「総合」の問題項目を構成していくのか、その下位項目の検討は実施予定の時までに「間に合うのか」?「コンピューター出題方式」のノウハウの研究やインフラの整備は進んでいるのか?また、「コンピューター出題方式」に関して、このテストに関わるテストユーザー全体での理解は浸透しているのか?「『高等学校基礎学力テスト』を “高校版全国学力テスト”」と位置づけるような印象論に対する、理論的根拠の提示は準備しているのか?まだまだ乗り越えなければならない、基礎的な課題は多々あると言わざるを得ない。さらに、この答申の検討過程で、英語教育に関してもっと根本的な問題が議論されたことも報じられている。

■日本の英語教育に根本的疑問も

文科省は 12 月 2 日、「英検」や「TOEFL (トーフル)」などの民間資格試験を、大学入試に活用できるかどうかを検討する有識者会合を立ち上げた。席上、活用の是非とは別に、有識者から日本の英語教育そのものへの根本的な疑問が相次いだ。

一部の教育関係者からは、「英語教育は必要」としながらも、差し迫った課題ではないとの意見も聞かれた。

■中高生の半数…「英語使うことない」

ベネッセ教育総合研究所が今年 3 月に全国の中高生約 6,200 人を対象にアンケートを行ったところ、中高生ともに 9 割以上が「仕事で英語を使うことがある」など社会生活での英語の必要性を感じていることが分かった。

一方で、「自分自身が英語を使うイメージがあるか」と尋ねたところ、中学生の 44%、高校生の 46%が「英語を使うことはほとんどない」と回答。調査を担当したベネッセ教育総研の加藤由美主任研究員は「日本の大部分の子供たちは教室の外に出れば、英語を使う環境にないのが現状。ただし、メディアなどにより『英語が必要』という意識はある」と説明する。

さらに学校での授業内容についても、中高の約 8〜9 割が「英文を日本語に訳す」「単語の意味や英文の仕組みについて先生の説明を聞く」と回答するなど、受け身的だ。一方で、授業で自分の考えなどを英語で話す機会は中学 2 年の 55%をピークに、学年が上がるごとに低下。高校 3 年の時点で 26%にとどまっており、「授業での学びと、英語を使うことにも大きなずれがある」 (加藤主任研究員) のが現状だ。

■財界は「企業が語学教育せざるを得ない」と嘆く

だが、教育界の英語教育熱は高まる一方だ。文科省が進める改革では、「読む」「書く」「聞く」「話す」−の 4 技能をバランス良く盛り込んだ実用的な学習環境づくりが喫緊の課題とされ、議論が進んでいる。

12 月 2 日の文科省の有識者会議では「(英語教育の) 必然性はない」と述べた委員らに対し、財界側から出席した日本経済団体連合会 (経団連) の教育問題委員会企画部会長の三宅龍哉委員が「ビジネスにおいては必然性は高い。社員を海外駐在などへ送り出す際、(企業が) 語学教育をせざるを得ない現状だ」と正反対の意見を述べた。

■専門家は「能力や希望に応じた多様な学習の場を」と指摘

立教大は、平成 28 年度の一般入試から「英検」などの民間資格試験の活用を他大学に先駆けて決めた。塚本伸一副総長は「卒業生にその力量を身につけさせるためにも高度な英語教育は欠かせない」と話す。

立教大では平成 20 年、より実戦的な英語を学べる「異文化コミュニケーション学部」を新設すると、教養英語中心の文学部英米文学科の志願者が激減し、新設学部に人気が集中した。塚本副総長は「学生が求めていたものが教養としての英語ではなく、ツール (道具) としての英語だということが分かった」と語る。英語を遠いものと感じる生徒らがいる一方、英語を積極的に身につけたいと考える層も薄くはない。

塚本副総長は「高校進学率がほぼ 100 パーセントとなる中、(高校などの英語教育に) 一律の基準を設けるのは無理があるのではないか」と疑問を呈する。

小野名誉教授は「外交官や通訳など高度な英語力が必要とされる人たちと、アジアへ向かうビジネスマンらとでは、求められる単語数や発音などは自ずと異なる。それぞれの能力や、将来の希望などに応じた多様な教育の枠組みを作っていくことが大切だ」と指摘している。

(http://headlines.yahoo.co.jp/hl?a=20141222-00000524-san-soci)

日本における英語教育の歴史は、振り子のように「行ったり来たり」を繰り返してきた。昨今の報道を見ると、また、そのあまり生産的ではない議論の轍に戻ってしまうのではないかという懸念も感じざるを得ない。そうしないためにも、今こそ、英語教育研究は何ができて、何をすべきなのだろうか、という問に、新たな方向性と可能性を持つ答えを追求していくことが必要である。

その検討のひとつとして、教育の global standard として提唱され、研究が始まっている 21st Century Skills との関連の中で、英語教育の目標に関する方向性を見極めていくことが重要であると考える。上記の長い引用の中には、「21 世紀型スキル」といった用語は出てこないが、中教審の答申の背景には、21 世紀型スキルの概念が大きく影響していることが伺われる。

「ICT 等を活用した評価について」検討する本研究の一部として、この 21 世紀型スキルと英語教育研究の関連を追及していきたい。中教審の答申に現場からの懸念や反発があることの原因のひとつは、答申の背景にある 21st Century Skills という概念に対

する理解不足ではないかと思われる。実際に、2014 年の教員免許状更新講習の機会に、参加された現場の先生方に「21 世紀型スキルという概念をご存じですか?」という質問をした時には、「初めて聞きました」という反応が大多数であった。

下の写真は、2014 年 12 月に、台北の街で偶然見つけた看板である。地元台湾の大学生に聞いたところ、学習塾の看板だという。台北では「21 世紀的生存能力」を「培養」することを掲げた「学習塾」が出現している。世界各地で提唱され、研究が進められているが、日本においては、広く深い認知が進んでいるとは言い難い 21st Century Skills に関して、本稿後半で取り上げていきたい。

また、本稿の関心事のもうひとつとして、Standard Setting の課題を取り上げていきたい。ここ数年の英語教育の課題のひとつに、英語教育の目標設定に関わる Can Do List の開発がある。Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) に端を発し、日本においても、投野編 (2013) で CEFR-J が提唱されている。2001 年に出版された CEFR の Prefatory note (p. ix) には “a process which has been actively pursued since 1971” とあり、実に、30 年という歳月をかけて開発が行われてきたことが分かる。CEFR-J に代表される日本の取り組みにおいても、この歳月の長さを肝に銘ずるべきである。ともすれば、「すぐに役立つ」「他力本願的な」 Can Do List を欲する声も聞こえるが、日本の英語学習において、妥当で実現可能性の高い Can Do List を作成していくためには、腰を落ち着けて、長い時間的なスパンも視野に入れ、慎重な検討を蓄積していく必要がある。特に CEFR “3 Common Reference Levels” に言及のある “a common framework scale should be *objectively determined* in that they are based on a theory of measurement (p. 21) という観点に留意しなければならない。「目標の内容、その設定方法を検討する際には、それと同時に、目標の到達と未到達はどうすれば判断できるのかという課題に触れる必要」があり、「それができなければ、単なる表面的な解決に終わってしまう」 (大友監修, 2009, p. 2) 懸念がある。本稿の第一の関心事である、21 世紀型スキルを考慮した英語教育の目標を検討する際も、その到達・未到達を判断するための科学的な方法論を研究していくことが必要である。

「今、英語教育研究は何ができて、何をすべきなのだろうか?」の問いに対する答えとして、本研究で、次世代を担う人材に必要な能力として、“21st Century Skills” の枠組みの中で英語教育の目標を捉え、「教科型」を超えた「総合型」の英語能力の教育と測定・評価における “Standard Setting” の方法論を研究し、その具現化を追求して

いくことを取り上げていきたい。

2. 先行研究の概観

"21st Century Skills"

「21 世紀型スキルの学びと評価プロジェクト (Assessment and Teaching of Twenty-First Century Skills Project: ATC21S) は 2009 年 1 月にロンドンで開催された「学習とテクノロジーの世界フォーラム」において立ち上げられた。シスコムシステムズ、インテル、マイクロソフトという世界的なテクノロジー企業がスポンサーとなり、2010 年にはオーストラリア、フィンランド、ポルトガル、シンガポール、イギリス、アメリカが参加し、メルボルン大学の評価研究センター内に研究開発本部を置いて活動が始まった (三宅監訳, 2014, p. 1)。」同書は、ATC21S は「ここ 2、3 年日本国内でも話題になっている」が、「今の日本の教育システムにおける「学び」と「評価」では必ずしも充分とは言えない」 (p. v) と指摘し、「21 世紀型スキルを育成する授業と評価を実現する上で「ラーニングプログレッションズ (Learning Progressions)」を考慮すべき」 (p. vi) と説明している。ATC21S のウェブサイト < http://atc21s.org > にも、"It's not what you know, it's how you learn" と、このプロジェクトの目指すところが端的に表現されている。

Griffin et al. (2012, pp. 18) は、 Chapter 2: Defining Twenty-First Century Skills で、21st Century Skills の構成を大きく 4 つに分類し、10 のスキルを以下のように定義している。

Way of Thinking

1. Creativity and innovation
2. Critical thinking, problem solving, decision making
3. Learning to learn, Metacognition

Way of working

4. Communication
5. Collaboration (teamwork)

Tools for working

6. Information literacy
7. ICT literacy

Living in the World

8. Citizenship – local and global
9. Life and career
10. Personal and social responsibility – including cultural awareness and competence

Griffin et al. の定義は、いわば「たたき台」のようなものであり、それが 21 世紀型スキルの構成概念を検討する出発点となっている。21 世紀型スキルの定義は、「つく

り変え続けているもの」であり、「私たちの生活に離れたどこか別世界から与えられたり、私たちの期待とは無関係に育成される (あるいは育成させられる)」ものではない、「21 世紀型スキルとは何か、どう育成したらよいかという 2 つの問いへの答えは私たち一人ひとりがつくり上げていくべき」ものである (三宅監訳, 2014, p. iii)。現在においても、上記の「抽象的」な 10 のスキルを、さらに細分化した下位構成要素の定義、あるいは、操作的定義の検討が進められている。文部科学省においても、「求められる資質・能力の枠組み思案」といった検討がなされ、「21 世紀型能力は『生きる力』としての知・徳・体を構成する資質・能力から、教科・領域横断的に学習することが求められる能力を資質・能力として抽出し、これまでの日本の学校教育が培ってきた資質・能力を踏まえつつ、それらを『基礎』『思考』『実践』の観点で再構成した日本型資質・能力の枠組みである」といった方向性を打ち出している (http://www.mext.go.jp/b_menu/shingi/chousa/shotou/095/shiryo/__icsFiles/afieldfile/2013/07/18/1336562_01_4.pdf)。

Fullan & Langworthy. (2014). は、Foreword に "… what they describe is a new pedagogy emerging, not in laboratories or universities, but at the front line, in classrooms as far apart as Denmark, Canada, England, Australia, Colombia, and California" と記しているように、"New Pedagogies" が "Deep Learning" を促進する具体的な実践例を示しながら、21st Century Skills を養成する教育方法の可能性を示し、21st Century Skills の教育で用いるべきタスクに関して、留意点を以下のようにまとめている (p. 22)。

> Learning Re-structured
>
> Deep learning tasks re-structure learning activities from a singular focus on content mastery to the explicit development of students' capacities to learn, create and proactively implement their learning. In their most effective instances, deep learning tasks are:
>
> 1. guided by clear and appropriately challenging learning goals, goals that ideally incorporate both curricular content and students' interests or aspirations.
> 2. include specific and precise success criteria that help both teacher and student know how well goals are being achieved.
> 3. incorporate feedback and formative evaluation cycles into the learning and doing processes, building students' self-confidence and 'proactive dispositions'.

同書 Foreword に "Of course much of what Fullan and Langworthy describe is not new at all." と述べているように、上記 1 ～ 3 の "clear and appropriately challenging goals"、"specific and precise success criteria"、"feedback and formative evaluation cycles" は、それほど目新しい概念ではない。しかし、従来の英語教育研究は、必ずしも、この 3 点を、クリアーできているとは言い難い。さらに、21st Century Skills の教育が求める、integrated performance に関して上記 3 点を検討することを考えれば、その作業は、か

なり複雑なもので、慎重な検討が必要となるものである。まさしく、“…, the structure and design of deep learning tasks are of critical importance (p. 28).” である。

Fullan & Langworthy. (2014). は、21st Century Skills 教育のアセスメントについて “We need to develop measures for both deep learning outcomes and for the new pedagogies and learning environments that support them. (p. 43)” とし、“The bottom line is that the issue of new assessments is at very early stage of development, and represents a major challenge in the immediate future (p. 46)” であることを指摘している。

Lai & Viering. (2012). は 21st Century Skills の構成概念に関する論文を調査し、Recommendations and Discussion (pp. 43 - 51) の章で、以下の 6 点を提言している。

> Assessment systems should provide multiple measures that support triangulation of inferences.
>
> Assessment tasks should be of sufficient complexity and/or offer sufficient challenge.
>
> Assessment should include open-ended and/or ill-structured tasks.
>
> Assessment should use tasks that establish meaningful and/or authentic, real world problem contexts.
>
> Assessment tasks should strive to make student reasoning and thinking visible.
>
> Assessment should explore innovative approaches to address scalability concerns.

最後の scalability concerns に関して “Such integrated tasks would, however, give rise to multidimensionality in student responses, which is not handled well in commonly-used psychometric models, such as unidimensional IRT approaches (Bennett et al., 2003). Thus, innovative approaches to modeling student responses that reflect the multidimensionality inherent in complex, integrated tasks should be considered, such as multidimensional scaling techniques (p. 51).” と指摘されている点は、本研究のテーマとも関連の深いものである。

また、*Standards for educational psychological testing*. が最新版への改訂の理由のひとつとして “The impact of technology was considered throughout the volume (p. 4)” とあげたように、昨今のコンピュータ技術は、飛躍的な発展を見ている。21st Century Skills の構成要素にも内在される *Tools for working:* 7. Information literacy, 8. ICT literacy もまた、考慮されなければならない。学習者達の literacy を高めるには、それを越え、さらに ‘proactive dispositions’ を促進するような方法論の研究が必要である。

“Theoretically, at the end of learning experiences with new pedagogies, students should breeze through standardized tests that measure mastery of curricular content. Of more importance would be measuring the full range of students’ deep learning competencies (Fullan & Langworthy, 2014, p. 40).” と指摘されているように、mastery of curricular のアセスメントに留まることなく、the full range of students’ deep learning competencies のアセスメントへの転換を検討する時が来ているのである。英語教育においても、言語知識のみのアセスメントに留まることなく、 integrated performance in English の観点

に焦点をシフトしていくための検討が重要である。

“Standard Setting”

Griffin et al. (2013-a, b) のような先駆的な研究も始まっているが、21st Century Skills におけるアセスメントの研究は、緒に着いたばかりの感がある。“It is anticipated that assessment designers will continue to make progress on leveraging technology to support inferences about a broader range of 21st century skills (Griffin et al., 2013-a, p. 71)” の指摘のように、今後の研究に負うところが大きい。

また、Kong et al. (2014) の、e-learning の 21st Century Skills への寄与を検討する研究では、“both formal and informal learning context”、“both individualized and collaborative learning approaches”、“supported by evidence of improvement and awareness of progress” といった e-learning の利点をあげ、 “Technology plays a crucial role in supporting schools on realizing the desirable learning goals, learning process and learning outcomes …” としており、興味深い指摘である。

本研究のもうひとつの関心事である、21st Century Skills の教育における Standard Setting に関する先行研究については、未だ、文献を検索中である。

“In the 2001 version of this chapter we wrote: ‘Clearly, there is a need for new ideas and more research. New methods, improved implementation of existing methods, and increased efforts to validate any performance standards are needed.’ … Eleven years have passed and those statements are still true (Hambleton et al., 2012, p. 69).” とあるように、新規開拓が必要な分野である。

しかし、Jiao et al. (2011). の研究は、次のように結論をまとめており、今後の可能性として、Mixed Rasch Model の研究は追求していく価値があると思われる。

> This paper proposes a method for establishing performance cut scores based on the Mixture Rasch Model under the assumption that students in different proficiency levels are distinctly different from each other in terms of qualitative characteristics represented by their item response patterns and quantitative characteristics represented by their latent ability along a continuous theta scale. The proposed procedure is based on fitting the Mixture Rasch Model to the data and using the intersecting point on the corresponding density functions between adjacent distributions to define cut scores for distinguishing reflect a large-scale reading test, the proposed Mixture Rasch Model based method results in a reasonably high level of classification accuracy (p. 514).

上記の 21st Century Skills 及び Standard Setting の先行研究を基にして、さらに文献研究を進め、本研究の課題に関する根本的な理解を深めていきたい。

参考文献

American Educational Research Association, American Psychological Association and National Council On Measurement in Education. (2014). *Standards for educational psychological testing.* American Educational Research Association.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. CUP.

Fullan, M. & M. Langworthy. (2014). *A rich seam: How new pedagogies find deep learning*. Pearson & Nesta. (http://www.michaelfullan.ca/wp-content/uploads/2014/01/3897.Rich_Seam_web.pdf#search='A+Rich+Seam%3A+How+new+pedagogies+find+deep+learning.') [小柳和喜雄 訳. (2014). 『豊かな鉱脈 – 新しい教育方法(学) は、どのように深い学びを見いだせるか?』. ピアソン・ジャパン & 日本マイクロソフト. (http://dl.pearson.co.jp/RichSeam_web-v1.1.pdf)].

Griffin, P., McGaw B. & E. Care. (eds.). (2012). *Assessment and teaching of 21st century skills*. Springer.

Griffin, P., Care, E., Bui, M., & Zoanetti, N. (2013-a). 'Development of the assessment design and delivery of collaborative problem solving in the Assessment and Teaching of 21st Century Skills project'. In E. McKay (Ed.), *ePedagogy in online learning: New developments in web-mediated human-computer interaction.* IGI Global.

Griffin, P., Care, E., Francis, M., Hutchinson, D., Arratia-Martinez, A. & McCabe, C. (2013-b). 'Assessment and learning partnerships: The influences of teaching practices on student achievement'. University of Melbourne. [report:http://education.unimelb.edu.au/__data/assets/pdf_file/0009/809253/LP00991123_Teaching_Practices_Report.pdf].

Hambleton, R., M. Pitoniak, and J. M. Copella. (2012). 'Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results'. In Cizek, G. J. (ed.) *Setting performance standards: Foundations, methods, and innovation.* Routledge.

Jiao, H., Lissitz, R. W., George Macready, G., Wang, S., & S. Liang. (2011). 'Exploring levels of performance using the mixture Rasch model for standard setting'. *Psychological test and assessment modeling, Volume 53, 2011 (4), 499-522.*

Lai, E. R. & M. Viering. (2012). *Assessing 21st century skills: Integrating research findings.* National Council on Measurement in Education, Vancouver, B. C. &

Pearson. (http://researchnetwork.pearson.com/ wp-content/uploads/assessing_21st_ century_skills_ncme.pdf#search='Assessing+21st+Century+Skills%3A+Integrating+ Research+Findings.').

Kong, S. C., Chan, T.-W., Griffin, P., Hoppe, U., Huang, R., Kinshuk, Looi, C. K., Milrad, M., Norris, C., Nussbaum, M., Sharples, M., So, W. M. W., Soloway, E., & Yu, S. (2014). 'E-learning in school education in the coming 10 years for developing 21st century skills: Critical research issues and policy implications'. *Educational technology & society, 17* (1), 70–78.

三宅なほみ 監訳. (2014). 『21 世紀型スキルの学びと評価』 (Griffin et al., 2012 の日本語訳). 北大路書房.

大友賢二監修 中村洋一・小泉理恵編集. (2009). 『言語テスト: 目標の到達と未到達』. 英語運用能力評価教会.

投野由紀夫 編. (2013). 『CAN-DO リスト作成・活用 英語到達度指標 CEFR-J ガイドブック』. 大修館書店.

ATC21S のウェブサイト http://atc21s.org (2014 年 7 月 22 日取得).

DISCO 人々の「学ぶ・働く」を考える キャリアリサーチ. 「21 世紀型スキル」は世界標準の力. (三宅なほみ氏 インタビュー) http://www.disc.co.jp/uploads/2012/03/ 2012.1.10miyakeshi_jinzai.pdf (2014 年 7 月 22 日取得).

文部科学省. 「21 世紀型能力」. http://www.mext.go.jp/b_menu/shingi/chousa/shotou/ 095/ shiryo/__icsFiles/afieldfile/2013/07/18/1336562_01_4.pdf (2014 年 7 月 22 日取得).

産経新聞. 2014 年 12 月 25 日. http://headlines.yahoo.co.jp/hl?a=20141222-00000524-san-soci (2014 年 12 月 25 日取得).

しんぶん赤旗. 2014 年 12 月 23 日. http://www.jcp.or.jp/akahata/aik14/2014-12-23/ 2014122301_03_1.html (2014 年 12 月 24 日取得).

中日新聞. 2014 年 12 月 23 日. http://www.chunichi.co.jp/article/front/list/CK2014 122302000072.html (2014 年 12 月 24 日取得).

YOMIURI ONLINE. 2014 年 12 月 22 日. http://www.yomiuri.co.jp/national/20141222-OYT1T50099.html (2014 年 12 月 24 日取得).

委託研究　進捗状況報告：1．3.
＊語彙能力分析から見たプレイスメントと診断的評価の諸相＊

法月　健(静岡産業大学教授)

1.　なぜ究明されなければならないのか？

1.1.　語彙能力分析とプレイスメント評価について

筆者は 2011～2013 年度、日本英語検定協会英語教育研究センター委託研究「言語テストの規準設定」の研究プロジェクト(大友賢二研究代表)の下で、受容語彙能力を測定するテストを使った能力別クラス編成テスト(以後、プレイスメント(テスト))における分割点法について研究を行ってきた。

受容語彙能力テストのプレイスメントテストへの応用については、その利点が後述する多くの文献で論じられているが、筆者が勤務校の現実的なニーズから開発・実施したプレイスメントテストが研究の契機になったと言える。

このプレイスメントテストが実施されるまで、筆者の勤務校では、毎年のように異なる外部テストが採択されて、プレイスメントが実施されていた。いずれも高い信頼性と妥当性を持つことで定評のあるテストではあったが、多くの受験者が高校までに経験した 50 分程度のテストに比べて倍以上の実施時間を要し、初級学習者を中心に、解答を途中でほとんど放棄するような受験者が多く見られた。

テスト実施後には、相応の診断的情報を得ることができたが、個別受験者の個別項目への正解･不正解の状況は開示されない場合が多く､その後の教育指導に活かして、議論されることもなかった。

いずれも優れたテストであるにもかかわらず、筆者は多くの受験者にとって入学後早々に受験させられるこれらのテストは、その品質とは関係なく、苦痛に過ぎず、むしろ英語嫌いを助長させているのではないかと危惧していた。

このような状況の中で、筆者は、30 分程度で全員の解答が終了する、後述の 80 問の受容語彙テスト、SCELP を開発した。SCELP は数年間実施された後、諸事情で中止となったが、毎年確認した信頼数の係数は、.90 を超え、テスト自体に大きな問題があったとは思えない。

法月(2013)では、プレイスメントとして実施されていた当時のデータの一部を使って、規準設定の観点から分割点を統計的に分析したが、この目的においても SCELP の有効性が示唆される結果が得られた。

SCELP は特定の英語熟達度テストに準拠した語彙テストではなかったことと、選択肢の数が問題の数と比べて、相対的に少なかったことから、新たな受容語彙力テスト(VKS)が開発された。VKS は英検の語彙集に基づき、各級の重要語彙を集めて、作成

された。できるだけ広域の能力層をカバーできるように、4～準 1 級の重要語で構成される Version 1 とより難しいと考えられる準 2～1 級の重要語を含む Version 2 のテストを仕上げた。VKS の問題数は 50 問で SCELP に比べて少ないが、選択肢に対する問題数が多く、各問の単語の理解度を 4 段階で受験者が解答直後に評価できるシステムを設けたことで、正解が自信のある正解なのか、当て推量に基づく所産なのか、あるいは不正解が、まったく自信がない無作為選択による必然の結果なのか、自信に反しての何らかの予期せぬ理由(誤解釈、過剰解釈等)による誤りなのか、大まかな想像を立てることができる。

法月(2014)では、SCELP 同様に VKS が規準設定において有効で、テスト項目の難易度と英検の級別の対照情報や理解度との比較など、SCELP の分析にはない新たな可能性をもたらしたことを確認することができた。一方で、分割点に対応する項目をどこに設定するかという問題や分割点付近の能力の特徴記述も今後の課題として残された。

最初の問題点については、特に選択式の問題においては、ラッシュモデルで一般的な 50%の正答確率よりも高い正答確率を持つ項目、つまり、より難度の低い項目を分割点の規準に対応させたほうが理解度が高くなり、より適切な分割点設定につながるのではないか今後検討する意義がある。2 番目の課題については、語彙能力の諸相を文脈における受容語彙能力や発表語彙能力の観点も含めて分析し、より具体的な能力の特徴記述につなげていくことで、簡易プレイスメントの役割だけでなく、診断的評価を構築することも考えられるだろう。

1.2　診断的評価における語彙能力－筆者の指導事例と経験から

筆者は英検等の資格試験を指導する際に、Excel を使って学習者に解答させ、正誤結果を瞬時に示す方法を使っている。

	解答	正答	正誤
R問1	1	1	1
R問2	3	2	0
R問3	1	1	1
R問4	2	2	1
R問5	1	1	1
R問6	3	2	0
R問7	1	1	1
R問8	3	2	0
R問9	1	1	1
R問10	3	2	0
R問11	1	1	1
R問12	3	2	0
R問13	1	1	1
R問14	3	2	0
R問15	1	1	1
R問16	3	2	0
R問17	1	1	1
R問18	3	2	0
R問19	1	1	1
R問20	3	2	0
R問21	1	1	1
R問22	3	2	0
R問23	1	1	1
R問24	3	2	0
R問25	1	1	1

	語彙	長文穴埋め	＊長文内容理解	リーディング	会話内容理解	説明内容理解	＊日常問題解決	リスニング	合計
	R大問1	R大問2	R大問3	合計(-作文)	L大問1	L大問2	L大問3	合計	
得点	14	5	14	33	8	7	8	23	56
得点／問題数(%)	56%	83%	70%	65%	67%	58%	80%	68%	66%

＊長文内容理解は各問2点(得点)
＊日常問題解決は各問2点(得点)
＊作文問題は14点満点(得点)

図 1　筆者が指導に使う英検の解答入力フォーム

図 1 はその一般的様式を一部提示した内容であるが、学習者は問題に沿って、解答の欄に答えを入力する。解答が終了すると、教師(筆者)は、あらかじめ入力しておいた(解答時には受験者に見えない)白文字の正答、正誤(正答だったら 1、誤答だったら 0 と標示するように数式が入れてある)の欄の文字を黒に変える。さらに分野得点・セクション合計・合計の数値も黒字に変えることで、解答後にすぐに個別の正誤、得点等の情報を学習者にフィードバックしている。

ゼミ活動等の個別指導においては、同じ問題を数か月後に再実施して 2 回の結果を比較しながら、誤答を中心に、学習者に面談しながら、解説していくことがある。

図 2 は、1 回目と 2 回目の解答を比較したフィードバック用の提示例であるが、① 2 回とも間違えてしまった解答にピンクの背景色を使い、② 1 回目は正解したが、2 回目は間違えてしまった解答は黄色、③ 1 回目は間違えたものの 2 回目は正解したものについてはクリーム色で塗り分け、④ 2 回とも正解したところは白の背景のままにしてある。一旦、2 回目と 1 回目の正誤の欄を軸にして、0−1 順で並べ替えることで、簡単な手作業で塗り分けが操作できる。

	1回目解答	2回目解答	正答	1回目正誤	2回目正誤
R問1	2	1	1	0	1
R問2	4	4	4	1	1
R問3	1	2	2	0	1
R問4	3	2	4	0	0
R問5	2	2	2	1	1
R問6	3	3	3	1	1
R問7	3	4	3	1	0
R問8	4	1	3	0	0
R問9	1	3	2	0	0
R問10	1	1	1	1	1
R問11	3	4	2	0	0
R問12	3	3	2	0	0
R問13	4	2	4	1	0
R問14	2	4	4	0	1
R問15	3	3	3	1	1
R問16	1	4	4	0	1
R問17	1	1	1	1	1
R問18	3	4	1	0	0
R問19	2	3	3	0	1
R問20	4	4	4	1	1
R問21	3	2	2	0	1
R問22	1	4	1	1	0
R問23	2	3	4	0	0
R問24	3	3	3	1	1
R問25	2	1	1	0	1

1回目	語彙	長文穴埋め	*長文内容理解	リーディング	会話内容理解	説明内容理解	*説明2内容理解	リスニング	合計
	R大問1	R大問2	R大問3	合計(-作文)	L大問1	L大問2	L大問3	合計	
正答数(*点)	11	4	14	29	10	3	6	19	48
正答率	44%	67%	70%	57%	83%	25%	60%	56%	56%

*R大問3とL大問3は1問2点

2回目	語彙	長文穴埋め	*長文内容理解	リーディング	会話内容理解	説明内容理解	*説明2内容理解	リスニング	合計
	R大問1	R大問2	R大問3	合計(-作文)	L大問1	L大問2	L大問3	合計	
正答数(*点)	15	6	14	35	9	6	6	21	56
正答率	60%	100%	70%	69%	75%	50%	60%	62%	66%

1回目×、2回目×
1回目○、2回目×
1回目×、2回目○

図 2　英検問題 1 回目解答と 2 回目解答の診断的フィードバック事例

筆者は現在、①と②のケースを中心に、(時間がある場合は、③や④の項目についても)質問をしながら、学習者に解答理由を説明させる方法を時折使用している。①の解答結果は全体的に内容がよく理解できていなかった場合に多く見られるが、大部分を理解できていても、文章・会話中や問題文中の一つの単語の意味がわからなかったために正解に至らなくなることも少なくない。②については、初回解答時よりも学習を通じて習熟度や理解度が高まった分、深く考えすぎて間違えてしまったと学習者が反省する場合もある。反対に、③、④の解答様式であっても確たる根拠がないまま正答を選んでいたり、特定の単語の誤った解釈から選択に至った答えが偶然に正解になる

こともある。

このような診断的評価に基づく指導を通じて筆者が痛感しているのは、ある受験者の個別問題への正解、不正解は、必ずしもその問題に対する理解の度合いを示しているわけではなく、いずれの結果も個別言語事項への不十分な理解に影響を受けていることが少なくなく、それらの多くが特定の語彙知識の欠如を示していることである。筆者がインフォーマルに実施した英検2級学習者10名のあるテストへの解答結果を分析したところ、文章穴埋め問題の部分点とその他の問題形式の合計得点との相関(重複部分を除去した相関)が他の部分のものよりも高く、 .7 を超える結果を示したが、文脈下で語彙を認識する力が英検の測定する英語熟達度と強く関連していると言えるかもしれない。今後、より大規模な調査で検証する価値があるだろう。

学習者が認識しなければならない問題点は、文法・構文、発音・イントネーション、文章構造への理解等、すべてを列挙することは不可能であるが、時間をかけて指導したり、学習自身の習得や習熟に委ねなければならない事項も多い。一方、語彙については、綴り、意味、発音、文脈的理解・使用等、体系的・系統的な指導がしやすく、関連する語彙を意識させたり、統合的な言語的、非言語的知識に結び付けていくことも可能である。

語彙能力が言語能力の中枢にあることは間違いなく、教授者兼研究者の立場から、語彙能力の診断的評価及び指導の役割について、今後、組織的な調査を行い、議論していく意義があると思われる。

1.3 プレイスメントと診断的評価における語彙能力

研究の発端は、筆者の勤務校で行った受容語彙能力プレイスメントテストであったが、規準設定の研究を進める中で、単に分割点を決定するだけでなく、分割点の意味を明確に定義することが重要であることを認識し、VKS のような簡易なプレイスメントテストであっても、有用な診断的情報を提供することが可能であり、その具体的な方法を模索する必要があると考えるに至った。

プレイスメントについては、診断的情報提供を考えると、より多角的な言語能力測定が望まれるかもしれないが、どのような方法であってもプレイスメントから供給できる診断的情報には曖昧性が伴う。具体的な診断的フィードバックは、授業開始後の診断的評価の役割であるが、VKS や VKS を改良した簡易プレイスメントの中での診断的機能を今後の研究の中で追究していきたい。

授業後の診断的評価については、まず「具体性」が必要であるが、理想的には、「徹底性」、「個別性」の側面も求められるだろう。たとえば、筆者は勤務校で担当している英検 2 級対策講座の中で、診断的観点から受講生の大半が間違えた項目を抽出して解説指導することがあるが、正解した学習者にとっては、半ば不要な解説になってしまうことも少なくない。反対に大半の受講生が正解している問題であっても、実際

には理解していない受講生が多かったり、不正解している受験生にとっては、多くの受講生が間違えた問題以上に優先して診断的情報を必要とする問題である可能性もある。個別指導のような密度では提供できないものの、一般授業についても情報の個別性、徹底性を高めていく努力は重要であると考えられる。

E-ラーニングのようなシステムの導入も将来的な研究テーマには掲げていきたいが、現実的な次への一歩として、学習者への理解度アンケート(選択式及び記述式)の実施、学習者の解答プロセスに関する口頭記述の録音、学習者への面談調査等を、今後の研究の具体的な方策として、検討したい。

具体的な研究計画の作成の中で、変更や修正の可能性はあるが、現在のところ、次の様な研究課題を、将来の研究調査の中で掲げていきたいと考えている。

将来の研究課題案

1)VKS のような簡易プレイスメントツールを使用することで、どのような分割点設定を適切に行うことが可能か。

2)分割点付近や編成された能力グループの特徴をどのような方法で、どのように記述することが可能か。

3)1)や 2)の手続きを通じて、各クラスや担当教員、受講生に対して、どのような診断的情報の提供を行うことが可能か。

4)授業開始後に、授業活動、テスト、他の評価手法を活用して、語彙能力を軸とした言語能力の観点から、どのような診断的指導を行うことが可能か。

プレイスメントと診断的評価は一見すると独立した研究テーマであり、一回の研究調査の中で扱うことは容易でないかもしれない。前者については、ともすると受験者はテスト受験を強要されている意識にとらわれることがあるため、彼らにとっても有用な診断的情報の提供は歓迎され、安定したプレイスメントテストを実施するうえでも重要な要素になるかもしれない。教授者は必ずしもプレイスメントから得られる細かな診断的情報を必要としないかもしれないが、その後の診断的評価の参考や動機づけとなるならば、プレイスメントと診断的評価を連動して議論することは、彼らにとっても意義のある問題提起となるだろう。

2. 関連する過去の研究と参考文献

2.1 関連する研究

効果的な英語の授業を実践するために、学習者に適した教材や教授法を使用することが重要であることは言うまでもないが、可能ならば、指導する学習者の習熟度ができるだけ同じようなレベルに集中している能力別編成クラスであることが望まれる。

能力別クラスを編成するために、市販や公的に利用できる標準化された英語熟達度テストが活用される場合、テストの内容や難易度が教育プログラムや学習者に適さないことも少なくない（Kokhan, 2013)。個別の教育機関・プログラムの目標を直接的に反映したテストを活用することが可能ならば理想であるが､様々な現実的な制約の中で、適切なテストの開発・分析、それに基づく合理的な判定を機関内のスタッフの力で賄うことは、決して容易でない（法月, 2014)。

近年、様々な形式の語彙テストを使った能力別クラス編成(プレイスメント)に関する研究が盛んに行われており、Laufer and Goldstein (2004) は、受容語彙能力を測定するテストは、受験者の将来のリーディング、ライティング、総合的言語能力、さらには学術的達成の成否を予測するのに適していて、プレイスメントや入学許可の目的で使用するのに優れているとしている。また、Beglar and Hunt (1999) や小泉・飯村(2010) のように、語彙テストが日本の高校や大学の英語プログラムにおけるプレイスメントに有効であることを検証した研究も見られるが、そのほとんどは外部テストを活用したものである。

法月(2013) は、大学(学科)独自のニーズを鑑みて、外部テストではなく、関連語のペアを選ぶ自家製（in-house）の受容語彙テスト（SCELP）を開発して、利用するに至った背景を以下の観点から説明している。

1) 対象受験者の能力水準(ほとんど英語を勉強したことがない留学生から相応の水準の学習者が受験できるテストが必要)
2) テストの所要時間(ペーパーテストの多くは項目数が多く、解答に相当時間がかかる。初級学習者は学習意欲の喪失につながる可能性もある)
3) 指示文・選択肢の使用言語(日本語訳を選択する問題形式にすると留学生の多くが､問題文(語)を理解できても､選択肢に対応できなくなることがある。英語のみにすると初級学習者の解答負担が大きくなる)
4) 正確かつ簡便な実施(語彙テストは統合型テストに比べて短時間で解答可能。作文や面接テストと異なり､採点の主観性の問題がなく一斉実施も可能である)
5) 意味と形の関係(Yes/No 形式のテストは、「理解している」と受験者が誤って解釈している可能性が高く、理解度も不明。英語定義のみや日本語訳のみの問題では、1)～4)のような問題も生じる)

法月(2013) の分析によって、SCELP が単一の教育機関の集団準拠的なプレイスメントのニーズを十分に満たしていたことが確認できたが、テストを構成する異なる水準の単語に関する知識を有するか否かが、教育的に何を意味しているのかを明確に規定することはできなかった。また、図 3 の例のように、各問題ブロックの問題が 4 問であるの対して、選択肢が 5 つしかないためか、分析のために解答及びアンケートや

面接に参加してもらった学習者が、やや難しい項目に対しては無作為に近い状態で解答し、かなりの割合で正答している状況も確認できた。

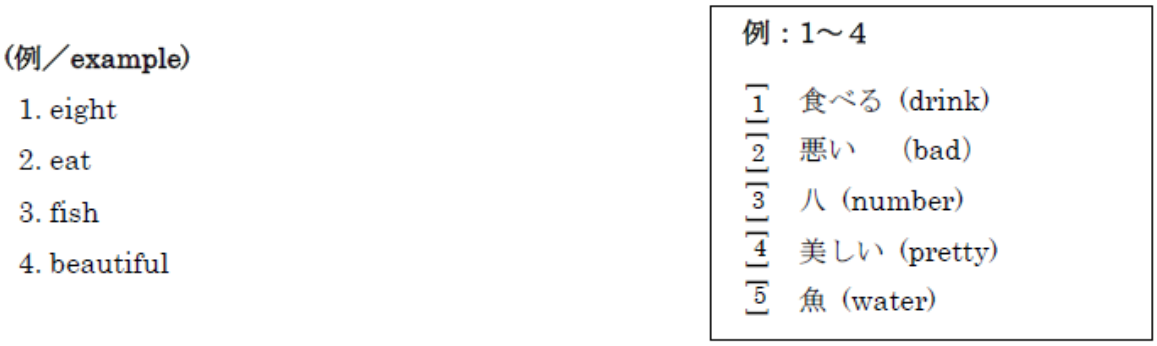


図 3　SCELP の問題形式　（例題）

Milton (2009)　は、近年の研究結果から、受容語彙力と IELTS の評定、Cambridge FCE の合否、CEFR との対応結果を明確に提示しているが、幅広い習熟度の英語学習者が受験していて、英語教師も学習者の習熟度をその合格級で判断することの多い実用英語検定の語彙を基に、テストを開発する価値はあると考えられる。法月 (2014) は、実用英語検定に出題されることの多い重要単語を級別に頻度順に分類している単語集を基に、一部問題が重複する 4～準 1 級 (VKS1)、準 2～1 級 (VKS2) の受容語彙力を測定する 2 つのバージョンのテスト (VKS) を開発した。

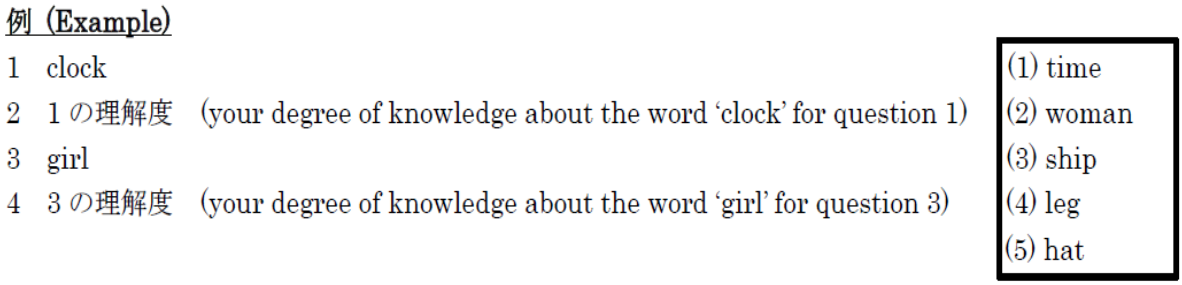


図 4　VKS の問題形式　（例題）

図 4 は VKS の問題形式を説明するために受験者に提示された例題であるが、問題数を 80 問から 50 問に減らして、各ブロック 2 問に対して選択肢 5 つを用意することで、各問の選択の幅を広げた。また、SCELP は 2 か国語提示であったが、英語情報を主要なヒントに頼る習熟度の高い留学生が日本語情報を有効に活用できていない状況が確認されたこともあって、項目難易度の公平性を維持するため、VKS では同級水準の英語の類義語のペアを選ぶ形式に変更した。

さらに図 4 の 2、4 番のように、偶数番号の項目に対しては、前の番号の問題の「単語の理解度」に 4 段階(5－かなり知っている単語　4－何となく意味がわかる単語　2

ー見たことはあるが意味は分からない単語　1ー見たこともないし､意味もわからない単語)で評価させることとした。

VKS は異なる水準の 2 つのバージョンを有するテストであり、単一の教育機関内の集団準拠型のプレイスメントにとどまらず、英検の 6 つの級(4～1 級)の語彙基準に照応させることで、基準準拠型の観点から規準設定 (standard setting) を追及する価値があると考えられる（規準設定については、Cizek　& Bunch, 2007 ; Cizek, 2012; Tannenbaum & Cho, 2014 等を参照)。

法月 (2013&2014) の分析を通じて､段階評価に基づく潜在ランク理論（植野･荘島、2010)は､規準設定の基盤となる分割点を決定するのに有用なランク関連指標を提供するのに対して、ラッシュモデルの分析は、様々な規準設定法の審査判断における客観性を高め、順序尺度と間隔尺度を融合した規準設定法へと発展させることも可能であることが明らかになった。

さらに法月(2014)の分析結果から、項目の難易度と英検の対象級は全体的には呼応しているものの、中学校、高校と必ずしも段階的に英語学習の習熟度を高めてきていない学習者がいるためか、個別項目においては、級区分とはかなり異なる結果となるケースも見られた。また、難度の高い項目が必ずしも理解度が低い項目でないケースも確認できた。

法月 (2013 & 2014)　は、分割点を受験者の正答率(ラッシュ能力値)と受験者の潜在ランクの切り替わる最初の地点に置き、その能力地点の受験者が 50%よりも高く、50%に近似する正答確率の難易度に相当する項目を境界項目 (borderline items) と位置付けた。VKS は SCELP に比べて 1 問当たりの選択肢の数が多いが、それでも当て推量のみで正答できる確率が各問題の 2 つ目の項目では 1/4 に達するため、ラッシュモデルで一般的に規準とされる 50%の正答確率が､VKS の形式の境界項目としては低すぎるのではないかという疑問が残った。

各能力グループ受験者の分割点以上の項目の理解度平均を調査してみると、理解度が高いとは言えない「3」前後の数値を示し、特に上位 2 番目のグループにおいては 2.53 と低く、自己評価で多少控えめなマーキングをしている可能性はあるものの、全体的に十分に理解されていない語彙が各能力群に対応する項目として位置付けられている状況が明らかになった。

法月 (2013&2014) の研究では、文脈と切り離した語彙知識に焦点を置き、受験者の能力値から分割点を決定しているが､項目難易度を規準にして分割点を設定する場合、Wang (2003) が指摘するように、項目の正答確率を意味する応答確率 (response probability) を変えることで、どのように分割点の設定に影響を与えるかについて、検討する価値がある。特に近年海外の公的試験における規準設定において広く用いられる Bookmark 法で一般的に採択されることの多い 2/3 や 67%のような応答確率に上げることで、より理解度が高いと考えらえる対象項目と照応して、分割点以上の知識や

技能の内容記述(performance level descriptors)の構築を進めていくことが望まれる(Bookmark 法については、Cizek & Bunch, 2007; Lewis, Mitzel, Mercado & Schultz, 2012; Hsieh, 2013; 大友、2013 & 2014 等を参照)。

プレイスメント簡易ツールとしての VKS の役割を今後も追究していく一方で、文脈下での語彙知識や技能の活用について、4 技能や複合的技能の言語活動の中で探っていく必要がある。文脈下での語彙使用を追究していくと、正解しながらも特定の語彙を誤って解釈した結果であったり、逆に不正解であっても内容をほぼ正確に理解していながら、選択肢の表現を拡大解釈したことが原因だったことも、受験者に解答過程や心理を口頭で報告させる手法 (verbal reports) を使用することで明白になることがある (Ericsson & Simon, 1993; Bowles, 2010; Norizuki, Ito & Shimatani, 2011; Cohen, 1998; 2012)。また、このような方法を実践することで、山田・法月(2014) が指摘するように、学習者が学習ストラテジーを使って問題を解決する能力を伸ばし、自律学習を推し進める力となっていくことも期待できるだろう。今後は、Jang (2012) が論ずるような、規準設定と診断的評価との関係の観点からも、解答過程に関する分析を活用していく価値があるだろう。

2.2 関連研究の参考文献

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level vocabulary tests. *Language Testing, 16* , 131-162.

Bowles, M. A. (2010). *The think-aloud controversy in second language research*. New York: Routledge.

Cizek, G.J. (2012). *Setting performance standards: Foundations, methods, and innovations.* (2nd ed.) New York: Routledge.

Cizek, G.J. and Bunch, M.B. (2007). *Standard Setting, A Guide to establishing and evaluating performance standards on tests.* 320. Sage

Cohen, A. (1998). *Strategies in learning and using a second language.* London: Longman.

Cohen, A. (2012). Test-taking strategies and task design. In G. Fulcher & F. Davidson (Eds.) *The Routledge Handbook of Language Testing*. (pp.262-277). New York: Routledge.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Hsieh, M. (2013). Comparing yes/no Angoff and bookmark standard setting methods in the context of English assessment. *Language Assessment Quarterly, 10, 331-350.*

Jang, E.E. (2012). Diagnostic assessment in language classrooms. In G. Fulcher & F. Davidson (Eds.) *The Routledge Handbook of Language Testing*. (pp.262-277). New York: Routledge.

Kokhan, K. (2013). An argument against using standardized test scores for placement of

international undergraduate students in English as a Second Language (ESL) courses. *Language Testing, 30, 467-489*

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54,* 469-523.

Lewis, D.M, Mitzel, H.C., Mercado, R.L., & Schultz, E.M. (2012). The bookmark standard setting procedure. In Cizek, G.J. (Ed.). *Setting performance standards: Foundations, methods, and innovations.* (2nd ed.) New York: Routledge.

Milton, J. (2009). *Measuring second language vocabulary acquisition.* Bristol, UK: Multilingual Matters.

Norizuki, K., Ito, A., & Shimatani, H.(2011)Exploring item-examinee response characteristics in search of diagnostic functions of TOEIC® tests for university students in Japan. *JLTA Journal*、*14*, pp. 1-20.

Tannenbaum, R.J. & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly*, *11*, 233-249.

小泉利恵・飯村英樹 (2010). 「ニューラルテスト理論の特徴：古典的テスト理論・ラッシュモデリングとの比較から」 『日本言語テスト学会紀要』, 13, 91−109.

法月 健 (2013). 「受容語彙力を測定するプレイスメントテストにおけるラッシュモデルと潜在ランク理論に基づく規準設定の試行」 『言語テストの規準設定 報告書第 2 号』 公益財団法人英語検定協会英語教育センター委託研究. (pp.81-103).

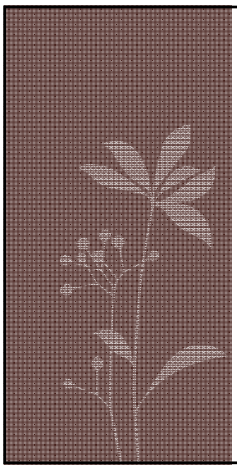
法月 健 (2014). 「実用英語検定の級別頻出単語に基づく英語受容語彙力テストの開発と規準設定」、『言語テストの規準設定 報告書第 3 号』、公益財団法人日本英語検定協会英語教育研究センター委託研究. (pp.77-101).

大友賢二 (2013). 「予備調査：CITO variation on the bookmark method」『言語テストの規準設定 報告書第 2 号』 公益財団法人日本英語検定協会英語教育研究センター委託研究. (pp.1-38).

大友賢二 (2014). 「CITO variation on the bookmark method の一考察」、『言語テストの規準設定報告書第 3 号』、 公益財団法人日本英語検定協会英語教育研究センター委託研究. (pp.2-26).

植野真臣・荘島宏二郎(2010). 『学習評価の新潮流』、東京：朝倉書店.

山田登・法月健 (2014). 「学習ストラテジーを活用しての効果的リスニング法」、『静岡産業大学情報学部研究紀要』, 16, 45-69.



ICT等を活用した評価についての調査・研究

進捗状況報告(1/16/2015)

池田　央、村木英治、
大友賢二
中村洋一、法月　健

ICT等を活用した評価についての調査・研究

これまでの研究活動

＜会議＞
日時:2014年9月13日(土)12:00−14:00
場所:日本英語検定協会　会議室
議題:研究課題の設定と今後の方針
＜進捗状況報告＞
締め切り:2015年1月9日
内容:設定課題が究明されなければならない理由と先行研究資料

2

進捗状況報告書

- 報告書(1)再考:言語テストの規準設定
- 報告書(2)語彙能力分析から見たプレイスメントと診断的評価の諸相
- 報告書(3)21st Century Skills と Standard Setting

- 関心のある課題
  Mixture Rasch Model
  Continuous Response Model

3

報告書(1)の概要

1. なぜ、規準設定の方法を究明する必要があるのか？
2. 主な先行研究資料として、どのようなものがあるか？
3. 年度末までの課題:Mixture Rasch Modelの考察と展望

4

1．基準設定を究明する必要

情報通信技術（ICT：Information and Communication Technology)の中での評価の課題

課題の中心を「規準設定」（standard setting)と結びつけて考える

Standard setting？
a process by which a standard or cut score is established

外国語習得水準 ABC
Cではなく、Bと判断するには、どんな条件が必要か？
どんな状態であれば、BではなくAと判断できるか？

5

1.1.CEFRとCAN-DO statements

- なぜ、「規準設定」が「ICT等を活用した評価」という領域に関連するのか？
- 第1の理由は、現在の英語教育の流れCEFR, CAN-DO statementsの持っている重要な課題と「規準設定」は密接に結びついているから。

向後（2015),中津原（2010),野口・大隈（2014), North(2000), Fulcher(2003), Weir(2005)

6

1.2. Standard Setting 研究結果の流れ

- なぜ、「規準設定」が「ICT等を活用した評価」という領域に関連するのか？

その第2の理由は、規準設定研究結果の流れが、じつに様々であり、可能であれば、一つの方向を求める必要がある。

- Hambleton & Pitoniak (2006).
- (1) 否定的見方、（2）中立的見方、（3）肯定的見方

7

1.3. 大学入試と段階別表記

なぜ、「規準設定」が「ICT等を活用した評価」という領域に関連するのか？
その第3の理由は、最近話題になっている大学入試の「段階別表記」と深い関わりを持っているから。

中央教育審議会高大接続特別部会（10/24/2014),
荘島（2010）
国際教養大学の「暫定入学」
南風原（2014）

8

1.4. Mixture Rasch Model の考察

これまでの「規準設定法」に向けられた多くの問題があるが、こうした課題を乗り越えた客観的・統計的解決法はないのだろうか？再考：規準設定の第4の理由はそこにある。

Mixture Rasch Model:
純粋な順序尺度に基づく「潜在ランク理論」に、連続尺度の精度向上を目指す「項目応答理論」Rasch Model の特性を補充することで、より明確な、より実用的な規準設定法を見出して行こう。

Kelderman & Marcredy (1990), Mislevy & Verhelst (1990), Rost (1990)

9

1.5. 異なるデータを用いた規準設定

規準設定に関しては、選抜等を目的とした場合と学習等を目的とした場合との2つの視点から、それぞれ異なった角度の考察が必要であろう。この視点を混同すれば、さらなる混乱をきたすことになる。ICT等を活用した評価の中で規準設定を検討しなければならない第5の理由は、こうした異なるデータを用いた規準設定に関する課題の解明である。

委員に対する最終課題の一つ：第3章「課題の考察と展望」
Mixture Rasch Model を有効に活用するために、どのような目的を達成するには、どのような手順を設定するのが最も適切か。

10

1.6. 参考文献

AERA, APA & NCME (1999)
Cizek, G.J. and Bunch, M.B. (2007)
Fulcher, G. (2003)
Hambleton, R.K. & Pitoniak, M.J. (2006)
Jaeger, R.M. and Mills, C.N. (2001)
Kaftandjieva, F. (2004)
Kelderman, H. & Marcready, G.B. (1999)
Peterson, C.H., Schulz, and Engelhard, Jr. G. (2011)
Rost, J. (1990)
Tannenbaum,R.J.,& Cho, Y. (2014)
Templin, J. & Jiao, H.(2012)
Weir, C.J. (2005)
Zieky, M.J., Perie, M. & Livingston, S.A. (2008)
朝倉拓也(2014)、南風原朝和(2014)、向後秀明(2015)、中央教育審議会高大接続特別会(2014)、中津原文代(2012)、野口裕之・大隈敦子(3014)、荘島宏二郎(2010)

11

2. 主な先行研究資料

2. 1. 言語テスト一般の先行研究
2. 2. 専門雑誌によるMixture Rasch Model関係の先行研究

2.2.1. *ETS Research Report Series*: 45
2.2.2. *Psychometrika*: 121
2.2.3. *Applied Psychological Measurement*: 84
2.2.4. *Journal of Educational Measurement*: 44
2.2.5. *Educational Measurement: Issues and Practice*: 6
2.2.6. *Language Testing*: 12
2.2.7. *Language Assessment Quarterly*: 14

12

2. 主な先行研究資料

2. 3. 先行研究:abstractと寸描

2.3.1. Cizek and Bunch. (2007).
2.3.2. Jiao, Lissitz, Macready, Wang and Liang. (2011).
2.3.3. Jiao, Macready, Liu, and Cho. (2012).
2.3.4. Mislevy and Verhest (1990).
2.3.5. Peterson, Schulz, and Engelhard, Jr. (2011).
2.3.6. Rost (1990).
2.3.7.Templin and Jiao (2012)

13

3. 年度末までの課題

委員に対する最終課題の一つは、さらに第3章「Mixture Rasch Model の考察と展望」を作成。

Mixture Rasch Modelを有効に活用するために、どのような目的を達成するには、どのような手順を設定するのが最も適切かを委員の間で検討し、それをまとめる。

＊＊＊＊＊

14

Mixture Rasch Model の考察と展望：3．1.
＊単純 Rasch モデルと混合 Rasch モデル＊

池田　央 (立教大学名誉教授)

１．単純 Rasch モデルの特徴

言語テスト結果の分析に項目応答理論(Item Response Theory、略称 IRT)のモデルが使われることも多くなってきた。代表的な IRT モデルにはその項目特性曲線(Item Characteristic Curve、ICC)として 1-、2-、3-パラメータモデルと呼ばれるものがある。

それは学力、能力、適性(以下能力という言葉で代表させる)といった、測ろうとする応答者の特性水準が高まるにつれて、与えられた質問に正答できる確率も次第に高まっていくという性質を利用して、いちばん可能性の高いその人の適正水準を推定しようとして生まれた数学的理論である。ただその高まり具合は直線的ではなく、はじめの低い段階ではなかなか上がらないが、やがて上昇率も高まって伸びのスピードは増す。しかし、ある段階を過ぎると伸びも次第に緩み、やがて止まって全体として S 字型カーブをとることが多い。IRT モデルの項目特性曲線はそのことを表現した形になっている。

そのうち最も包括的な形は 3-パラメータモデルで、各項目について識別力(傾き)a,難易度(難しさ)b、当て推量 c (択一式の場合)の違いを考慮し、項目ごとにそれらを推定しようとするものである。しかし求めるのは計算も複雑だし、安定した値を得るには受験者数も最低 1000 人以上は必要といわれている。

そこで、c の値は考えないで、a と b だけを使って推定しようというのが 2-パラメータモデル、さらに識別力 a もどの項目も同じ値を持つ(ここでは便宜上 a＝1.7)として分析を進めていくのが 1-パラメータモデルと呼ばれるものである。これは起源は異なるが、最初に提唱したデンマークの G.Rasch(1901-1980)のモデルと形が一致することから Rasch モデルと呼ばれることも多い。

この項目特性曲線をグラフに表わしたのが次ページの図 1 で、式にすると

$$P(\theta)= 1/[1+\exp\{-1.7(\theta-b)\}]$$

と表せる。θ は応答者の能力レベルを表し、b が項目の難しさを示す指標(パラメータ)である。そして能力 θ の人がその質問項目に正答できる確率が 0.5(正誤半々)に当たる θ の位置で該当項目の難しさを表す。そしてある能力値 θ_0 を持つ応答者がその質問に正答できる可能性(確率)はグラフの高さ $P(\theta_0)$ で表されることになる。

ここでは応答者の能力を表す尺度 θ の集団平均が便宜上 0、標準偏差は 1 になるように標準化されたものとして定義されている。したがってこれを 10 倍し、50 を加えたものを改めて

X=10 θ ＋50

とすれば、X は使い慣れた偏差値のような形(平均 50、標準偏差 10)にすることができる。

以上のように Rasch モデルでは、各質問項目の性質が項目の難しさを示すパラメータ b の値のみで決まる。そこで b の値が、b＝－1.5，－1.0，－0.5，0.0，0.5，1.0，1.5 であるような 7 つの項目(A～G で示す)の特性曲線を描いてみたのが図 1 である。b の値は項目によっていろいろ変わるが、それ以外の条件は同じであるから、グラフは同じ形のものが横に平行移動したような形になっている(この例は b が等間隔でひとつの理想的な項目配分になっている)。

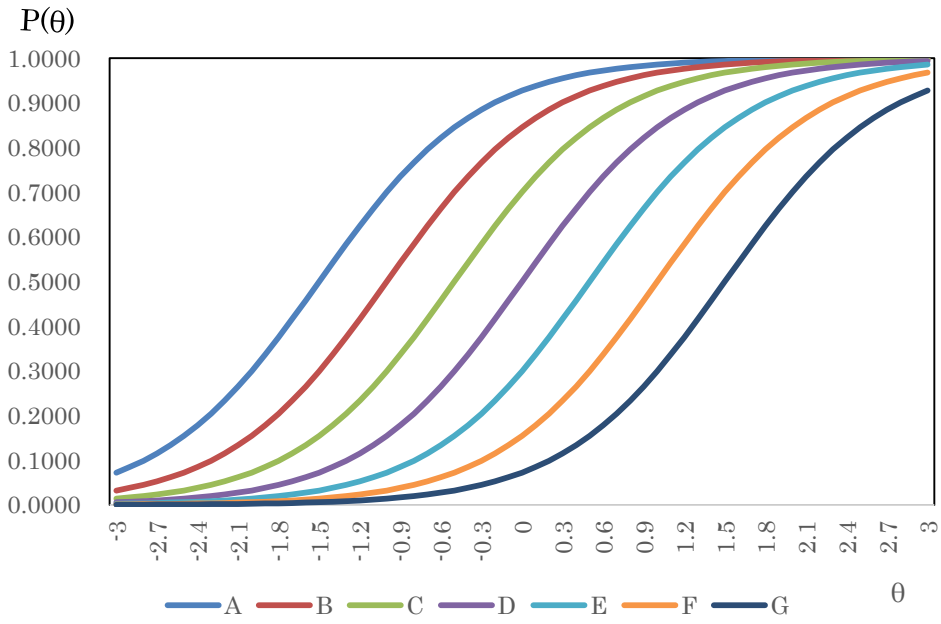


図1　7つの項目特性曲線(左からA~G)

2. 単純 Rasch モデルが持つ限界

しかし、こうして定義されたいわば単純なモデルが実際のテストデータで成り立っているか、確かめようとすると、話はそう簡単ではない。ここで θ は一本の直前上に大小が定義される能力と考えている。しかし知的能力とか学力や適性といった複雑な人間の特性が身長や体重と同じように一本の線上に定義されるようなもの(一次元尺度)であろうか。

図 1 のように等間隔ではないとしても、どの質問項目も同じ形をした特性曲線で表すことができるのだろうか。応答する集団が同質の集団であればまだしも、いろいろ違ったグループや思考性を持つ集団が交っていても同じように扱ってよいのであろうか。しかもそれは項目によって違いの有るものもあれば、無いものもあろう。

いろいろな民族や宗教、育った言語も違う人々からなる米国では、同じ質問を与えてもその捉え方は応答者の出処によって異なり、それを同一の尺度で表すのは不公平な偏り(バイアス)を生むとして、差異項目機能(DIF：Differential Item Functioning)という手法が生

まれ研究も進んだ。

しかし、そうした目に見える(manifest)形の違いならば、応答集団をそれぞれのグループに分けて同質の下位集団ごとに項目分析を行い、その下位集団ごとにパラメータを決めてもよいが、表面的に見ただけではわからないような応答者が持つ潜在的特性(latent trait)や内在的(internal)思考法の違いなどからくる偏りに対しては、新しい分析手法を考える必要がある。そうして生まれたのが混合 Rasch モデルである。

3. 混合 Rasch モデルの考え方

混合 Rasch モデル(Mixed あるいは Mixture Rasch Model：MRM)が適用できるのは、ある質問項目に答える際に応答者に複数の潜在的な応答パターンがあり(Latent class といってもよい)、考え方や応答方略のタイプが同じパターンに属する人の間では Rasch モデルが成立するが、別タイプの考え方や方略を取る人はパラメータが違う別の Rasch モデルに従う(Rasch モデルであることには変わりない)といった場合である。そこでは複数のタイプの潜在的 Rasch モデルが混じり合ってその質問項目が構成されていると考えられる場合である。各タイプに属する応答者の割合は同じでないので。それらを一つにまとめた単純 Rasch モデルで分析するとその特性曲線は人数の大きな割合を持つ潜在クラスに影響されて歪みを持った形となり、モデルが misfit したと解釈される。

それを図に示すと図 2 のようになり、研究対象にされる対象集団はいくつかの潜在下位集団に別れるというモデルである(この図では 2 つ)。各下位集団(その大きさはいろいろ異なる)はそれぞれパラメータ b の値は異なるがいずれも単純 Rasch モデルに従っていると考えられるわけである。そして各応答者の能力値 θ はその人が属する下位集団に準じて決められる。

ここでは詳しく説明する余裕はないが、その計算は複雑で、モデルの成り立ちもいろいろなケースが考えられて単純ではない。下位集団数も適合度を調べていちばん近いものが採用される。図 2 は一例で、対象とする集団全体は A,B 2 つの潜在的下位集団に分かれ、7 つの質問項目 X1～X7 の項目パラメータ b の値はそれぞれの下位集団ごとに異なるという場合である。従って応答者の能力推定値はその応答者が属すると判定される下位集団での θ によって表されるわけである。

しかし、場合によってはもっと複雑なケースも考えられる。図 3 は 7 つの質問項目すべてが関わるのではなく、X1～X5 までは一つの Rasch モデルが適用され、X2 と X4～X7 にはまた別の Rasch モデルが成り立つというような場合である。いずれにしても、こうした分析には複雑な手法が取られるわけで、コンピュータソフトなしには実行は難しい。Mixed Rasch モデルについても多くの文献があり、次ページにあげているのはその一例で手掛かりにされたい。

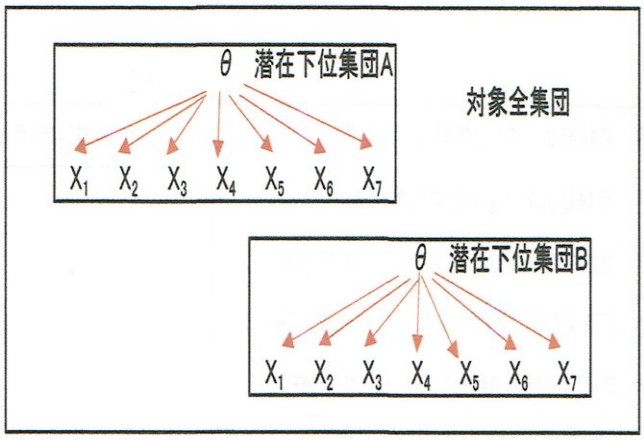


図 2：各項目が 2 つの潜在下位集団に関わる例

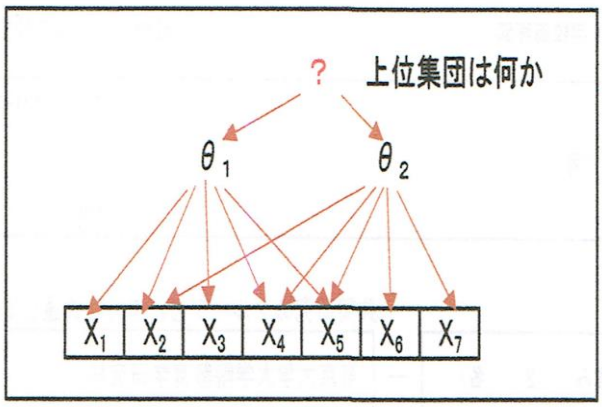


図 3：構成する潜在下位集団が項目で一致しない例

参考文献

Bolt, D.M., Cohen, A.S., & Wollack, J.A., (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics. 26*, 381-409.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Gilford Press. (当書の付録 E に説明あり)

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across on manifest and latent examinee groups. *Journal of Educational Measurement. 27,* 307-327.

Mislevy, R. J. (2006). Cognitive psychology and educational measurement. In R. L. Brennan, (Ed.), *Educational measurement*(4th ed.). West Port, CT: American Council on Education and Praeger Publishers.

Mislevy, R. J., & Verhelst, N., (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*,195-215.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271-282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology. 44*, 75-92.

Rost, J.,& von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer, & I. W. Molenaar (Eds.) *Rasch models -foundations, recent developments, and applications* (pp.371-379). New York: Springer-Verlag.

von Davier, M., & Rost, J. (1995). Polytomous mixed Rasch model. In G. H. Fischer, & I. W. Molenaar (Eds.) *Rasch models -foundations, recent developments, and applications* (pp.257-268). New York: Springer-Verlag.

von Davier, M., & Rost, J. (2007). Mixture distribution item response models. In C. R. Rao, & S. Sinharay (Eds.). *Handbook of statistics, vol. 26, Psychometrics.* (pp.643-661). North-Holland: Elsevier B. V.

【Mixed model 用分析ソフト】

WINBUGS　http://www.mrc-bsu.cam.ac.uk/bugs/

WIN-MIRA　(von Davier & Rost, 1995)

「潜在ランク理論」用として、日本では

Exametrika　荘島宏二郎氏の(http://www.rd.dnc.ac.jp/~shojima/exmk/index.htm)

EasyNTT　熊谷龍一氏作成の(http://irtanalysis.main.jp/)

Neutet　橋本貴充氏作成の(http://www.rd.dnc.ac.jp/~hashimot/neutet/index.html)

Mixture Rasch Model の考察と展望：3．2.
＊　Standard Setting の視点から ＊

大友賢二 (筑波大学名誉教授)

1. Standard Setting

これまでのテスト研究の文献には、standard setting や performance standards に関する多くの用語が見られる。例えば、passing scores, cut scores, cutoff scores, performance levels, achievement levels, mastery levels, proficiency levels, thresholds, standards, performance standards (Hambleton & Pitoniak (2006)) などである。この performance standards や standard setting に含まれる「規準を設定する」というのはどんなことを意味するのであろうか。 それは、ごく簡単に言えば、the process of establishing cut scores on examinations (Cizek (2006))ということである。

こうした中で、分割点(cut scores)の設定を求める方法は、これまで、数多く論ぜられてきている。例えば、その一つに、bookmark method (Lewis, Mitzel, & Green (1996)) がある。そこで用いられている項目応答理論(IRT)による分析結果、response probability of 0.67 等が話題になったことがある。しかし、これは、「古典的精神物理学」(psychophysicsal method)などの利用と見なされた経過もあるが、Angoff method よりも NAEP(National Assessment of Educational Progress)standard setting には適切である (Peterson, Schlulz & Engelhard Jr. (2011))という見方もあった。このように、規準の設定方法に関しては、これまで多くの研究がなされてきているが、その結果は、全てが一つの方向を向いているとは言い難い。否定的見方、中立的見方、そして肯定的見方(大友(2013))がある。さらに、合否判定などの規準は、単にテスト得点のみではなく、他の多くの要素も含めて決定される必要があると言う。まさに、それは考えなければならない重要な視点であり、それを否定するものではない。しかし、テスト得点の角度から見た場合はどうすべきかという課題を解決することは、教育測定に関する研究者にとっては、きわめて重大な責務である。

この「規準設定」が、なぜ、わが国の外国語教育の中で、とりあげられて検討されなければならないのであろうか？その理由のいくつかに関しては、大友(2015)で既に述べているが、その題目だけに触れれば、次のようなことがある。(1)CEFR と Can-do statements の持っている重要な課題と「規準設定」は密接に結びついている。(2)「規準設定研究結果」の流れが、じつに様々であり、可能であれば、一つの方向を求める必要があると考えられる。(3)わが国における大学入試の「段階別表記」などとも深い関わりをもっている。

2. MRM の考察

こうした課題解決のために、より高い信頼性、より優れた妥当性の方法が求められてきていた。そのひとつが、Mixture Rasch Model (MRM)である。これは、Kelderman and Macready (1990), Mislevy and Verhelst (1990), Rost (1990)らによって初めて提案されたと言われている。ここでは、純粋な順序尺度に基づく「潜在ランク理論」(Latent Rank Theory)に連続尺度の精度向上を目指す「項目応答理論」(Item Response Theory)である Rasch Model の特性を補充することで、より明確な、より実用的な規準設定法を見出していこうとするものである。

専門雑誌による Mixture Rasch Model 関係の先行研究論文の数は、極めて多い。たとえば、*ETS Research Report Series* の 45、*Psychometrika* の 121、*Applied Psychological Measurement* の 84、J*ournal of Educational Measurement* の 44、*Educational Measurement: Issues and Practice* の 6、*Language Testing* の 12、*Language Assessment Quarterly* の 14 などである。多くの先行研究の中の一つは、Jiao, Lissitz, Macready, Wang & Liang (2011) であるが、その summary and discussions の中で、Mixture Rasch Model の利点を要約すると、以下のように述べている。

(1)First, in the current practice of standard setting, there is no statistical validation of the pre-specified numbers of proficiency levels.

(2) Second, the mixture Rasch model analysis results can help identify the minimally competent or borderline examinees.

(3) Third, in the process of finding the cut scores using the model based approach, it is easy to cross-validate the cut scores using a subset of items and/or examinees.

(4) Fourth, the proposed method can be used as a yardstick for comparing multiple standard setting methods irrespective of whether the method is an examinee-centered or test-centered method because the procedure incorporates
both test-centered and examinee-centered information.

これを簡単に述べると、第 1 に、現在、ほかで用いられている規準設定法においては、事前に明記される能力水準(pre-specified number of proficiency levels) に関する統計的な妥当性検討はなされていない。第 2 の理由としては、この MRM で分析された結果を使えば、最低能力とか境界線上の受験者(minimally competent or borderline examinees)を確認することができる。第 3 の理由としては、このモデルに基づいた方法で分割点を設定する過程では、テスト項目やテスト受験者の部分集合(subset of items and/or examinees)を用いて、分割点の相互の妥当性検討は容易に行うことができる。第 4 としては、その手順はテスト中心の情報にも受験者中心の情報にも使えるので、テスト中心の設定方法であれ受験者中心の設定方法であれ、いずれの場合も、多様な設定方法を比較検討するための基準として、この MRM は使用することができる、ということである。

3. MRM の展望

MRM は、Rasch model と LCA (Latent Class Analysis) model を用いて、複数の潜在的母集団(latent population)の要素を持ったテストデータを分析するものである。そこでは、受験者は 2 つの潜在的な変数：つまり、関連する特性を測定できる継続的で量的な変数、と分類的で質的な変数をもっている。したがって、テスト項目における受験者の行動は、「部分的な質的なグループ集団と継続的量的潜在能力」(discrete qualitative group membership and continuous quantitative latent ability)で決定されていると考えているのである。

MRM が規準設定のための開発に到達するまでの過程では、この MRM は Kelderman & Macready (1990)などによって DIF(differential Item Functioning)の開発の目的で、また、Mislevy & Verhelst (1990) などによって、項目反応パターンの差(differences in item response patterns)の検討などのために開発されていた。DIF に関する多くの研究は、Zieky (2006) などにも見られるが、非常に簡単に理解できる説明は、“A test is labeled with “DIF” when examinees with equal ability, but from different groups, have an unequal probability of item success.” American Board of Internal Medicine: Item Response Theory Course)である。Templin, Cohen & Henson (2008) や Jiao, Lissitz, Macready, Wang, & Liang (2011) などによって、はじめて規準設定のための MRM 開発が盛んに行われるようになったと言われている (Templin & Jiao (2012))。したがって、まだ、研究の歴史は十分長いとは言えないので、Thus, a large sample is recommended to reduce variability. (Jiao et al (2011:516))という summary and discussions での発言なども見られる状態である。

我が国における MRM 関連の研究は、極めて少ないが、注目すべき事項も少なくない。その中の「潜在ランク理論」(LRT : latent rank theory)(植野・荘島(2010))の研究、また、法月(2012)(2013)などの実践も貴重である。ただ、外国における研究で論じられている LCA (latent class analysis) model と、わが国で述べている LRT (latent rank theory) の比較検討、それが規準設定のための MRM にどう結びつくかの検討は必要であろう。MRM のコンピュータ・ソフトは、WINMIRA 2001 (von Davier, M. (2001))、mdltm (von Davier, M. (2005))などに示されている(Jiao, Lissitz, Macready, Wang & Liang (2011))。そのソフトを用いて、目的とする規準設定が十分可能であるかどうかの検討、さらには、Exametrika Ver.5.3 (荘島)との比較研究も重要な課題の一つであろう。

最近の話題にのぼっている 2 つの事項にこの MRM による規準設定がどう関係しているかを検討する必要もあろう。その 1 は、*Educational Measurement: Issues and Practice:* Vol.33, Issue 4, 2014 でも盛んに言及している *Standards for Educational and Psychological Testing (5^{h} Revision)*, 2014 である。そのなかに新たに加えられた項目の一つ<Score, Scales, Norms, Score Linking and Cut Score>のなかの<Cut Score>の内容の徹底的検討はぜひ必要である。その 2 は、JMOOC (Japan Massive Open Online Course) の出現である。2012 年にアメリカで複数立ち上がった「オンラインで公開された無料

の講座を受講し、終了条件を満たすと修了証が取得できる」MOOC という教育サービスは、世界中で急速に広がりを見せており、今後の高等教育のあり方を大きく変える可能性をもっている。ここに、外国語教育を入れた場合に MRM による規準設定をどう生かすことができるか、というのも大きなそして重要な課題の一つであろう。

****

参考文献

AERA, APA, & NCME. (2014). *Standards for educational and psychology testing*. (Fifth Revision) AERA.

Cizek, G.I. (2006). Standard setting. In T. Haladyna & S. Downing (Eds.),*Handbook of test development* . (p.225). Lawrence Erlbaum.

Hambleton, R.K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (p.435). American Council on Education / Praeger.

Jiao, H., Lissitz, B., Macready, G., Wang, S., & Liang, S. (2011). Exploring levels of performance using the mixture Rasch model for standard setting. *Psychological Test and Assessment Modeling*, Vol. 53, 499-522.

JMOOC: *Japan Massive Open Online Course*: http://www.jmooc.jp

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, 27, No. 4, 307-327.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996). *Standard Setting: A Bookmark Approach.* In D.R. Green (Chair), IRT-Based Standard-Setting Procedures Utilizing Behavioral Anchoring. Symposium presented at the 1996 Council of Chief State School Offices 1996 National Conference on Large Scale Assessment, Phoenix, AZ.

Mislevy, R.J., & Verhelst, N.D. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, 55 (2), 195-215.

Peterson, C.H., Schlulz, E.M., & Engelhard, Jr.G. (2011). Reliability and Validity of Bookmark-based Methods for Standard Setting, *Educational Measurement: Issues and Practice.* 30 (2). 3-14.

Rost, J. (1990). Rasch model in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement,* 14, 271-282.

Templin, J., Cohen, A., & Henson, R. (2008). *Constructing tests for optimal classification in standard setting*. Paper presented at an annual meeting of the National Council on Measurement in Education, New York

Templin, J. & Jiao, H. (2012). Applying model-based approaches to identify performance categories. In G.J. Cizek(Ed.) *Setting performance standards.* (Second Edition). 379-397. Routledge.

von Davier, M. (2001). *WINMIRA 2001*. Retrieved from http://www.ipn.uni-kiel.de/abt_ppm/tt0506/winmiramanualmvd.pdf

von Davier, M. (2005). *mdltm*: Software for the general diagnostic model and for estimating mixtures of multidimentional discrete latent traits models <Computer software>. Princeton, NJ: ETS

Zieky, M.(2006). Fairness Reviews in Assessment. In S.M. Downing & T.M. Haladyna (Eds.) *Handbook of Test Development.* (pp.359-376). Lawrence Erlbaum Associates, Publishers.

法月　健(2012)「規準設定における潜在ランク理論の有用性」(pp.127-136).
『英検：英語教育研究センター委託研究：言語テストの規準設定報告書　第 1 号』
法月　健(2013)「Rasch Model と　LRT を併用した分割点設定法」(pp.37-46).
(第 7 回日本テスト学会賞記念講演研究協力：成蹊大学)
大友賢二(2013)「英語教育とテスト：第二言語習得における基準設定をめぐって」
(pp. 20-24)(第 7 回日本テスト学会賞記念講演：成蹊大学)
大友賢二(2015)「ICT 等を活用した評価についての調査研究：再考：言語テストの規準設定」(pp.1-4).(英検：英語教育研究センター委託研究進捗状況報告書)
植野真臣・荘島宏二郎(2010)　『学習評価の新潮流』　(pp.81-111).朝倉書店.

＊＊＊＊＊

Mixture Rasch Model の考察と展望：3．3.
＊21st Century Skills と MRM＊

中村洋一 (清泉女学院短期大学教授)

1. Standard setting における Mixture Rasch Model の可能性

Standards for educational psychological testing は Cut Scores に関する standard の規定について‘clearly’、‘reasonable’、‘distinct’といったキーワードをあげ、明確に standard を設定する重要性を指摘している (2014, pp. 107 - 109)。

Standard 5.21

When proposed score interpretations involve one or more cut scores, the rational and procedures used for establishing cut scores should be documented clearly.

Standard 5.22

When cut scores defining pass-fail or proficiency levels are based on direct judgments about the adequacy of item or test performances, the judgmental process should be designed so that the participants providing the judgments can bring their knowledge and experience to bear in a reasonable way.

Standard 5.23

When feasible and appropriate, cut scores defining categories with distinct substantive interpretations should be informed by sound empirical data concerning the relation of test performance to the relevant criteria.

21st Century Skills プロジェクトの研究では、Griffin & Care (eds., 2012, p. 25) が、‘Principles for Twenty-First Century Standards and Assessments’の項で 21st Century Skills の assessment には、new psychometric approaches が必要だと指摘している。

・Be technically sound. … In the absence of reasonable measurement precision, inferences from results, and decisions based on them may well be faulty. The requirement for precision relative to intended purposes means both that intended uses and users must be clearly specified and evidence of technical quality must be established for each intended purpose. Establishing evidence of quality for innovative approaches to assessing twenty-first century skills may well require new psychometric approaches.

大友・渡部 (2012, 2013, 2014) は Standard setting に関する方法論の幅広い研究をまとめ、様々な方法論の比較や可能性を検討している。その中で法月は、(報告書 (2)「受容語彙力を測定するプレイスメントテストにおけるラッシュモデルと潜在ランク理論に基づく規準設定の試行」(2013), p. 84) において、Mixture Rasch Model (MRM) が、将来的な統計的解決方法のひとつとして提唱されていることを紹介している。MRM の先駆的研究を進めている Jiao et al. (2011) の研究は、“… the proposed mixture

Rasch model based method results in a reasonably high level of classification accuracy (p. 514)”とまとめ、Mixture Rasch model の今後の可能性を示唆している。日本の英語教育における Standard setting の作業に関する研究はやや遅れている感があるが、今後の研究において、適用可能な統計的解決方法のひとつとして MRM を取り上げ、検討を深めていきたい。

2. Mixture Rasch Model 検討のポイント

Rost (1996, pp. 449 - 463) は、“Logistic Mixture Models”あるいは、“Discrete mixture distribution models”といった用語を使用しているが、MRM の基本的な概念として、“Discrete mixture distribution models (MDM) assume that observed data do not stem from a homogeneous population of individuals but are a mixture of data from two or more latent populations”(p. 449) と述べている。また、Templin & Jiao (2012, p. 387) は、“The MRM assumes that multiple latent student populations exist and that the Rasch model holds within each latent class with differing item difficulty parameters across classes. When compared with the LCA model, by incorporating the Rasch model within a latent class, the MRM provides a model where items are allowed to be correlated within each class.”とこのモデルを概観している。21st Century Skills は複合的に構成される performances として定義されているが、その assessment において、‘two or more latent populations’‘multiple latent student populations’が想定できることは、大きな利点になり得る。さらに“Further, it is possible to extend the procedure to tests consisting only of polytomous items or tests containing both dichotomous and polytomous items.”(Jiao et al., 2011, p. 516) という指摘から鑑みるに、open ended responses を要求する task も大幅に採用していくことが考えられる 21st Century Skills の assessment において、その採点データの処理方法に MRM が寄与する可能性は高い。今後の研究では、日本の英語教育の視点から、 21st Century Skills を構成概念として含むテストの実施データを用いて、 Standard setting における MRM による処理の可能性を明らかにしていきたい。

Templin & Jiao (2012, p. 388) は、“There are multiple estimation methods for the MRM.”とし、“the marginal maximum likelihood estimation (MMLE) method with the expectation-maximization (E-M) algorithm (used in mdltm)”, “M-Plus”, “the conditional maximum likelihood estimation method (used in Winmira)”, “the Markov Chain Monte Carlo estimation method (used in WINBUGS)”をリストアップしている。テストデータの推定方法として、どの方法が適切なのかを検討することもポイントのひとつである。さらに、それぞれ特徴のある推定方法を採用しているプログラムのうち、どれが有用なのかについても検討が必要である。Templin & Jiao (2012) は the Mulitdimensional Discrete Latent Trait Model (mdltm) software を使用して levels of performance の研究をしている (p. 388)。Baghaei & Carstensen (2013) は、WINMIRA を使用して a reading

comprehension test composed of 20 multiple-choice items の分析を行い、“Item fit for each class”といった観点からの分析も行っている。WINMIRA は Rost (1996, p. 459) のような比較的初期の文献にも紹介されており、また、ウェブサイトでその内容に関する情報や、マニュアルが入手可能で、Kagi online store から購入も可能であり、検討のとりかかりとしては、魅力的である。いくつかのモデルを比較したシンポジウムの overview (Rupp, 2009) や Rasch Measurement Analysis Software Directory: http://www.rasch.org/software.htm といったウェブサイトもあり、プログラムについては、そのような情報を糸口に検討を深めていきたい。

3. 英語教育 + 21st Century Skills の Standard setting

英語教育 (の assessment) でカバーすべき構成概念と、以下に示す Griffin et al. (2012, p. 18) の 21st Century Skills の構成要素との関係も検討しなければならない。

1. Creativity and innovation
2. Critical thinking, problem solving, decision making
3. Learning to learn, Metacognition
4. Communication
5. Collaboration (teamwork)
6. Information literacy
7. ICT literacy
8. Citizenship – local and global
9. Life and career
10. Personal and social responsibility – including cultural awareness and competence

実際に使用されている英語の 4 技能の assessment tasks は、トピックに依存する特徴、task に含まれる複合的な情報の処理、ICT 利用の言語学習、 Computer Based/Adaptive Testing、異文化理解といった側面を含んでおり、本来的には「複合的な構成概念」を持っていると言うこともできる。そのような観点に立って整理しつつ、英語教育 (の assessment) でカバーすべき構成概念と、21st Century Skills の構成要素との関係の検討を深めたい。このような検討のためには、Templin & Jiao (2012, pp. 515 - 516) “The estimated ability parameters will be a constrained to a limited number of values. The impact of sample size may be important and requires further study.”との指摘があるように、大きなサイズのテストデータが必要である。すでに大きなサイズのテスト実施データがある、たとえば、英検のインタビューテストのように、面接者の評価が複数のスケールとして提示されるデータを用いて MRM の分析を行い、その構成概念の再検討と併せて、結果を累積していけば、英語教育 + 21st Century Skills の Standard setting の研究に大きく貢献できるのではないかと考えるものである。

参考文献

American Educational Research Association, American Psychological Association and National Council On Measurement in Education. (2014). *Standards for educational psychological testing.* American Educational Research Association.

Baghaei, P. and Carstensen, C. H. (2013). ‘Fitting the mixed Rasch model to a reading comprehension test: Identifying reader types’ in *Practical Assessment, Research & Evaluation., Vol. 18, No. 5.*

Davier, M. (1997). ‘WINMIRA – program description and recent enhancement’ in *Method of Psychological Research Online, 1997, Vol. 2, No. 2.* Pabst Science Publishers.

Griffin, P., McGaw B. & E. Care. (eds.). (2012). *Assessment and teaching of 21st century skills*. Springer.

Jiao, H., Lissitz, R.W., Macready, G., Wang, S., and Liang, S. (2011). ‘Exploring levels of performance using the mixture Rasch model for standard setting’ in *Psychological Test and Assessment Modeling,* Vol. 53, 2011 (4). (pp. 499 - 522).

Templin J. & H. Jiao. (2012). ‘Applying model-based approaches to identify performance categories’. in Cizek, G. J. (ed). (2012). *Setting performance standards, second edition.* (pp. 379 –379).

Rost, J. (1996). ‘Logistic Mixture Models’ in Linden, J. & R. K. Hambleton. (1996). *Handbook of modern item response theory*. Springer. (pp. 449 – 463).

大友賢二 研究代表・渡部良典 研究副代表. (2012). 『言語テストの規準設定 報告書第 1 号』 公益財団法人日本英語検定協会英語教育研究センター委託研究.

大友賢二 研究代表・渡部良典 研究副代表. (2013). 『言語テストの規準設定 報告書第 2 号』 公益財団法人日本英語検定協会英語教育研究センター委託研究.

大友賢二 研究代表・渡部良典 研究副代表. (2014). 『言語テストの規準設定 報告書第 3 号』 公益財団法人日本英語検定協会英語教育研究センター委託研究.

Kagi on line store: http://order.kagi.com/cgi-bin/store.cgi?storeID=1HN&&

Rupp, A. A. (2009). “Software for Calibrating Diagnostic Classification Models: An Overview of the Current State-of-the-Art” *retrieved at* http://www.education.umd.edu/EDMS/fac/Rupp/AERA-SIG%20Software%20Symposium%20(Handout%20Package).pdf, on Feb. 6, 2015.

Rasch Measurement Analysis Software Directory: http://www.rasch.org/software.htm

WINMIRA manual: http://208.76.84.140/~svfklumu/wmira/winmiramanual.pdf

WINMIRA 2001: http://208.76.84.140/~svfklumu/wmira/index.html

Mixture Rasch Model の考察と展望:3．4.
＊統計的解決に基づく分割点設定は可能か？＊

法月　健　(静岡産業大学教授)

1．統計的解決に基づく分割点設定法：議論の背景

テスト結果を基に分割点を決定する規準設定の方法は数多く存在するが、テスト中心のモデルと受験者中心のモデルのいずれかに分類されることが多い (大友, 2008)。受験者中心のモデルについては、個々の受験者の能力を熟知している規準設定の評定者がいない場合は不適当であり、テスト中心のモデルについては、境界水準の受験者の個別のテスト項目の正答確率や項目群への正答数等を正確に予測して評定することが要求される(Pitoniak and Morgan, 2012)。

いずれにしても従前の規準設定法は、綿密な計画な下に実施されても、人間の判定に基盤を置く恣意的なものだと Lissitz (2013) は述べ、将来の発展が期待される規準設定法として、混合ラッシュモデル(Mixture Rasch Model: MRM)等に代表される統計的解決法を提唱している。MRM は、「複数の潜在的な母集団を含んだテストデータを分析するため、ラッシュモデルと潜在クラス分析 (Latent Class Analysis: LCA) モデルを統合した」モデルであり、近年、様々な研究が行われている (Rost & Langeheine, 1994; Kreiner, 2007; Choi, 2010; Jiao, Lissitz, Macready, Wang & Liang, 2011; Templin & Jiao, 2012 等)。

MRM は数千人以上の大規模試験の分析には活用が期待できるが、Jiao et al. (2011)等が示唆するように、数百名以下のデータ分析では大きな測定誤差を生じ、十分に適応しないものと考えられる。法月 (2013 & 2014)、大友 (2013)は、比較的小規模なサンプルサイズにおいても、ラッシュモデル (Rasch Model: RM) と潜在ランク理論 (Latent Rank Theory: LRT) の分析(植野・荘島, 2010)を併用することで、同一の間隔尺度上でテスト項目の難易度と受験者能力を直接比較しつつ、統計的に付与された潜在ランクを考慮に加える分割点設定法(以下、RM-LR 法)を提唱している。

英検の級別重要頻出語彙を基に「受容語彙力テスト (VKS1 & VKS2)」を開発した法月 (2014)は、RM の能力推定値と項目難易度を LRT のランク・メンバーシップ・プロファイル(RMP)の表に、図 1 の要領で位置づけて、RM-LR 法の分割点設定を試行している。

①算出された RMP の Excel 表内に RM 能力推定値(CHIPs 値)が入った列を挿入し、数値の高いほうがリストの上に来るように並べ替える。この表に「対応する項目」の難易度(CHIPs) や番号等の情報を追加する。対応する項目とは、(A) その難易度がある受験者能力値と同じかそれよりも低く(正答確率.50 以上)、(B) 次に高い受験者

能力値よりもその難易度が上回っているものを意味するが、(A)の条件を満たし、(B)の条件を満たす項目が無い場合は、その能力推定値に最も難易度が近い項目を「対応する項目」として、扱う。

②①の並べ替えの際に CHIPs 値が同じでランクが異なる場合は、ランクの高いほうがリストの上に来るように設定する。

③各ランクに所属する確率も提示する。①の並べ替えの際に、②の条件に加えて、CHIPs とランクがともに同じ場合は、隣接する境界ランクの「より高い」ランクに所属する確率(例、境界ランクが 5 と 4 の場合は、5 の確率)が高い方がリストの上に来るように設定する。

	A	B	C	D	E	F	G	H	I	J	K
1			潜在ランク	能力	対応する項目	Item Number	級	ランク・メンバーシップ・プロファイル			
2	正答数	正答率	推定値	CHIPs	CHIPs	δ(≤θ)	δ(≤θ)	Rank 1	Rank 2	Rank 3	Rank 4
14	36	0.720	4	54.55	54.19 & 54.00	問2057. 2069. 2073 & 2087	準1 級 & 1 級	0.000	0.001	0.118	0.881
15	36	0.720	4	54.55	54.19 & 54.00	問2057. 2069. 2073 & 2087	準1 級 & 1 級	0.000	0.003	0.127	0.870
16	36	0.720	4	54.55	54.19 & 54.00	問2057. 2069. 2073 & 2087	準1 級 & 1 級	0.000	0.006	0.187	0.807
17	36	0.720	4	54.55	54.19 & 54.00	問2057. 2069. 2073 & 2087	準1 級 & 1 級	0.000	0.005	0.304	0.691
18	36	0.720	3	54.55	54.19 & 54.00	問2057. 2069. 2073 & 2087	準1 級 & 1 級	0.001	0.057	0.626	0.316
19	35	0.700	4	53.913	53.5035	問2067	準1 級	0.000	0.002	0.145	0.853
20	35	0.700	4	53.913	53.5035	問2067	準1 級	0.000	0.003	0.147	0.850
21	35	0.700	4	53.913	53.5035	問2067	準1 級	0.000	0.011	0.176	0.813
22	35	0.700	4	53.913	53.5035	問2067	準1 級	0.000	0.005	0.248	0.747
23	34	0.680	4	53.276	52.821	問2065	準1 級	0.000	0.022	0.476	0.502
24	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.005	0.112	0.883
25	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.012	0.148	0.840
26	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.002	0.204	0.793
27	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.020	0.189	0.790
28	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.006	0.282	0.713
29	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.005	0.292	0.703
30	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.017	0.308	0.675
31	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.013	0.362	0.624
32	33	0.660	4	52.639	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.007	0.396	0.597
33	33	0.660	3	52.639	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.004	0.156	0.611	0.229
34	32	0.640	4	52.0475	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.016	0.468	0.516
35	32	0.640	3	52.0475	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.024	0.618	0.358
36	32	0.640	3	52.0475	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.001	0.058	0.677	0.264
37	32	0.640	3	52.0475	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.003	0.153	0.758	0.086
38	32	0.640	3	52.0475	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.007	0.220	0.691	0.082
39	32	0.640	3	52.0475	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.036	0.177	0.709	0.078
40	31	0.620	4	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.016	0.296	0.687
41	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.001	0.060	0.581	0.358
42	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.002	0.083	0.568	0.347
43	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.001	0.040	0.638	0.321
44	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.004	0.074	0.607	0.316
45	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.000	0.035	0.668	0.297
46	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.003	0.086	0.664	0.247
47	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.001	0.049	0.719	0.231
48	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.004	0.120	0.662	0.213
49	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.003	0.091	0.772	0.134
50	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.002	0.111	0.756	0.131
51	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.005	0.109	0.757	0.129
52	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.022	0.283	0.590	0.106
53	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.010	0.299	0.590	0.102
54	31	0.620	3	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.005	0.217	0.682	0.096
55	31	0.620	2	51.4105	51.41 & 51.09	問2063 & 2035	準1 級 & 2 級	0.249	0.615	0.131	0.004
56	30	0.600	3	50.819	50.77 & 50.46	問2049 & 2053	2 級 & 準1 級	0.011	0.082	0.526	0.382
57	30	0.600	3	50.819	50.77 & 50.46	問2049 & 2053	2 級 & 準1 級	0.002	0.053	0.683	0.262
58	30	0.600	3	50.819	50.77 & 50.46	問2049 & 2053	2 級 & 準1 級	0.006	0.214	0.635	0.145

図 1　VKS2 の最上位グループ決定の分割点候補　(法月, 2014)

2. MRM を活用した分割点設定法

MRM では、同一クラス(階層) の受験者能力推定値と項目難易度は、RM-LR 分析の RM 値と同様に、単一の間隔尺度上に表されるが、RM-分析と異なり、難易度パラメター値や、難易度順位がクラス間で異なることもある (Rost & Langeheine, 1994)。

Kreiner (2007) は認知症の判別検査 (1,147 名)の分割点を MRM と関連の統計手法を使って、分析している。まず、赤池情報量規準 (Akaike's information criterion: AIC) に沿ってクラス (階層) 数を 2 つに分け、次に、①尤度 (likelihood) 、②ベイズ情報量規準 (Bayesian information criterion: BIC)、③局所的均質性 (local homogeneity)の分析を基に、分割点の候補を 2 つに絞り込む。最終的に、診断結果が陰性で実際に発症していない「真の陰性 (true negative)」の確率 (specificity) が高い地点は、両候補の数値が高く、大差がないため、診断結果が陽性で実際に発症している「真の陽性 (true positive)」の確率 (sensitivity) が顕著に高い値を示している地点を分割点に設定している。「真

の陰性」、「真の陽性」の折衷法は、Brown (1996: 261-262) が論じる教育測定における習得者と未習得者を分割する対照グループ法の理念にも通じる。

Jiao et al. (2011) は、10,000 人の読解テスト (シミュレーション) データを使って、分割点設定を行っている。AIC と逸脱度 (deviance) は 5 つの階層、BIC は 3 つのクラス(階層)に分割することを支持する結果を示したが、3 階層モデルの最も低い第 1 階層、中間の第 2 階層、最も高い階層の各平均は、5 階層モデルの第 1・2 階層をまとめた平均、中間の第 3 階層の平均、第 5・4 階層をまとめた平均の値に近似するため、階層内の均質性を高め、受験者の階層をより細かく分類できる 5 階層モデルを選択している。クラス(階層)数が決まると、各階層と隣接する階層の確率密度関数 (density functions) が交差する地点を算出し、分割点を設定することが可能となる。

Templin & Jiao (2012) は、LCA や MRM 等の有限混合分布モデル(finite mixture model) の体系について説明し、規準設定の評定者による審査結果を基にモデルを構築する適応型 MRM は、モデルがデータによく適合している場合は、探索的に分割点の数を検証する純粋 MRM よりも正確な規準設定が期待できるとしている。

Templin & Jiao (2012) の議論は、MRM のような単一次元 (unidimensional) モデルにとどまらず、複数技能の多次元的測定を仮定した診断的分類モデル(Diagnostic Classification Models: DCM) を活用した新しい規準設定の概念に及んでいるが、Choi (2010) は、MRM と DCM を融合した診断的分類混合ラッシュモデル (Diagnostic Classification Mixture Rasch Model: DCMixRM) を提唱している。DCMixRM は、サンプルの大きさや細目化された構成技能・知識の推定精度等において、様々な制約や困難点があるが、このようなモデルが受験者の習得状況を点検し、評価システムの改善を導く新たな方法になるかもしれない。

3. 期待される今後の研究

RM-LR 法も MRM も、統計的解決に基づく分割点設定法として、その有用性が今後も議論されることが期待されるが、最終的な規準判定の裁量や、学習者や教授者等の教育プログラム関係者に提供される診断的フィードバックの内容の吟味は、今後も人間的な柔軟な判断に委ねられるべきものであると言えよう。また、Templin & Jiao (2012)が主張するように、評定者による審査結果を基に構築する適応型モデルや評定者を伴う従来の分割点設定法との多角的な比較検証も行っていく価値があるだろう。法月 (2013 & 2014)の試行した RM-LR 法では、いずれも分割点は、受験者の正答率(RM 能力値)と受験者の潜在ランクの切り替わる最初の地点に置き、その能力地点の受験者が 50%よりも高く、50%に近似する正答確率の難易度に相当する項目を境界項目 (borderline items) と位置付けた。しかしながら、RM で一般的規準とされる 50%の正答確率は、当て推量のみで正答する確率が 1/4 に達する選択式応答の境界項目としては低すぎないか、疑問が残った。同様に MRM においても、Kreiner (2007) の分割点規

準として真の陰性の確率、真の陽性の確率の議論や、Jiao et al. (2011) の隣接クラスの密度関数の交差点から分割点を設定する方法の観点から、分割点設定の妥当性を検証する意義があるだろう。

特定領域の習得が遅れている学習者への診断的フィードバックやモデルに適合しない項目の改善は RM-LR 法の分析でも課題になったが、DCMixRM のような理論や MRM や DCM 分析の弾力的運用が、今後、英検のような大規模試験に応用できないだろうか。

参考文献

Brown, J.D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

Choi, H.J. (2010). *A model that combines diagnostic classification assessment with mixture item response theory models.* Unpublished doctoral dissertation, University of Georgia, Athens, GA.

Jiao, H., Lissitz, B., Macready, G., Wang, S., & Liang, S. (2011). Exploring using the mixture Rasch model for standard setting. *Psychological Test and Assessment Modeling, 53,* 499-522.

Kreiner, S. (2007). Determination of diagnostic cut-points using stochastically ordered mixed Rasch models. In von Davier, M., & Carstensen, C.H., (Eds.). Multivariate and mixture distribution Rasch models: Extensions and applications. (pp.131-146). New York: NY: Springer.

Lissitz, R.W. (2013). Standard setting: past, present, and perhaps future. In M. Simon, K. Ercikan & M. Rousseau (Eds.) Improving large-scale assessment in education: Theory, issues, and practice. (pp.154-174). New York: Routledge.

Pitoniak, M.J., & Morgan, D.L. (2012). Setting and validating cut scores for tests. In C. Secolsky & D.B. Denison (Eds.) *Handbook on measurement, assessment, and evaluation in higher education.* (pp. 343-366). New York: Routledge.

Rost, J., & Langeheine, R. (1997). A Guide through latent structure models for categorical data. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp.13-37). Munster, Germany: Waxmann.

Templin, J., & Jiao, H. (2012). Applying model-based approaches to identify performance categories. In G.J. Cizek (Ed.), *Setting performance standards. (Second Edition)* (pp.379-397). New York, NY: Routledge.

法月　健 (2013).「受容語彙力を測定するプレイスメントテストにおけるラッシュモデルと潜在ランク理論に基づく規準設定の試行」.『言語テストの規準設定報告書第 2 号』.公益財団法人日本英語検定協会英語教育研究センター委託研究. (pp.81-103).

法月　健 (2014). 「実用英語検定の級別頻出単語に基づく英語受容語彙力テストの開発と規準設定」.『言語テストの規準設定　報告書第 3 号』.公益財団法人日本英語検定協会英語教育研究センター委託研究. (pp.77-101).

大友賢二(監修)(2008).『言語テスト：目標の到達と未到達 vol. 2』英語運用能力評価協会.

大友賢二 (2013b,12 月). 「英語教育とテスト：第二言語習得における規準設定をめぐって」（第 7 回日本テスト学会賞記念講演） 成蹊大学.

植野真臣・荘島宏二郎 (2010). 『学習評価の新潮流』、東京：朝倉書店.