サイナビ！ブックレット Vol. 18

テスト理論から見た大学入試改革論

南風原 朝和

2017 年 10 月 27 日発行（Ver. 1.0）● 発行元：ちとせプレス

1990 年度入試から始まった大学入試センター試験が，2021 年度入試からは「大学入学共通テスト」に変わります。報道されているように，国語と数学に数問程度，記述式問題が導入され，英語は民間の試験が導入されるようです。この案に至るまで，複数回受験とか，段階評価など，いろいろな話が出ては消え，と紆余曲折がありました。こうした最近の大学入試改革論は，テスト理論の観点からはどう評価されるのか，テスト理論を含む心理統計学が専門で，文部科学省の「高大接続システム改革会議」の委員も務められた東京大学の南風原朝和教授に寄稿していただきました。

第 **1** 回

高大接続システム改革会議とその前と後

私が委員として参加した文部科学省の高大接続システム改革会議は，2015 年 3 月から翌 2016 年の 3 月まで 14 回開催され，2016 年 3 月 31 日に最終報告が出されました [(1)]。

この委員会は，2014 年 12 月に出された中央教育審議会の答申 [(2)] をふまえて設置されたものです。そして，その中央教育審議会の答申は，2013 年 10 月の教育再生実行会議の提言 [(3)] を受けて出されたものです。

さらに，高大接続システム改革会議の後は，文部科学省のもとで小規模の検討・準備グループをつくって具体的な検討が続けられ，2017 年 7 月 13 日に，これまでの大学入試センター試験に代わる「大学入学共通テスト」の実施方針が公表されました [(4)]。全体として，検討にそれなりに時間はかかっています。

その過程で，教育再生実行会議で示された，テストの結果をいわゆる「1 点刻み」ではなく「段階評価」とし，「一発勝負」ではなく「複数回受験」を可能とするという方針は，中央教育審議会を経て，高大接続システム改革会議でトーンダウンして実質上，消滅しました。一方，記述式問題の導入については，教育再生実行会議では何もなく，中央教育審議会の途中から浮上して，高大接続システム改革会議でかなり議論された後も保持されました。

そして，現時点で最も大きな注目を集めている英語の民間試験の活用は，高大接続システム改革会議の最終報告では「民間の資格・検定試験の知見の積極的な活用の在り方なども含め検討する必要がある」[(5)] という程度の記述だったものが，その後の実施方針では大きく踏み出し，民間試験に全面移行する勢いとなっています。これには私も正直，驚いています。

このように議論の内容は大きく変容してきたのにもかかわらず，中央教育審議会で打ち出された「新しい共通テストは 2021 年度入試から」というタイムスケジュールだけは変わることがありませんでした。そのことを含め，いろいろと不思議に思うことや，納得のできないことなどもあり，その都度，専門の立場から発言・発信をしてきました [(6)]。現時点でも不透明な部分や，とても無理だと思うことなど多々ありますが，議論の中で，国の方針がどんなふうに決まってしまうのか等，いろいろと学ぶことはありました。

本稿の内容

本稿では，一連の大学入試改革論の中で出てきたキーワード，すなわち「記述式問題」「段階評価」「複数回受験」「英語民間試験」「障害のある受験者への配慮」について，それらをテスト理論の用語・概念と結びつける形で紹介・解説し，テスト理論の観点から評価することを試みたいと思います。なお，テスト理論というのは，テストによる能力推定やテストの質の評価のために用いられる，テスト得点に関する統計的理論のことです。本稿が，大学入試改革論とテスト理論

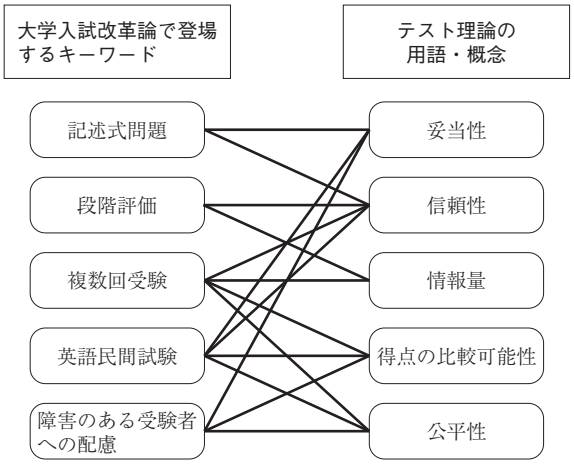

図 1　キーワード関連づけ

の両方の理解を深めるのに役立てば幸いです。

入試改革論のキーワードとテスト理論の用語・概念

図 1 は，左側に大学入試改革論のキーワードを並べ，右側にテスト理論の用語・概念を並べて，関連するものを線で結んだものです。この図 1 を適宜参照しながら，読んでいただければと思います。

最初に，テスト理論の用語・概念と，その相互関係について，簡単に説明しておきたいと思います。

妥当性・信頼性・情報量

まず，妥当性と信頼性です。

図 2 に示したように，テストの得点は，測りたい能力の高低だけで決まるのではなく，それに必ず誤差が加わります。誤差の中には，たとえば，タブレット端末を使ってテストをする場合の，タブレット端末使用の慣れ・不慣れの個人差のように，もう一度テストしても同じように得点を左右する安定的，系統的な誤差と，採点者の判断のブレのような，不安定で偶然的な誤差とがあります。

測りたい能力と，それとは相関しない系統的な誤差，さらにそれらとは相関しない偶然的な誤差の合計でテスト得点が構成されると考えると，テスト得点の分散（集団における散らばり）も，図 2 に示したように，それら 3 つの構成要素それぞれの分散の和になります。そのテスト得点の分散のうち，測りたい能力を反映する成分の分散の割合が「妥当性」です。つまり，妥当性は，テストの得点が測りたい能力をどれだけ反映しているかという程度を表す重要な指標です。

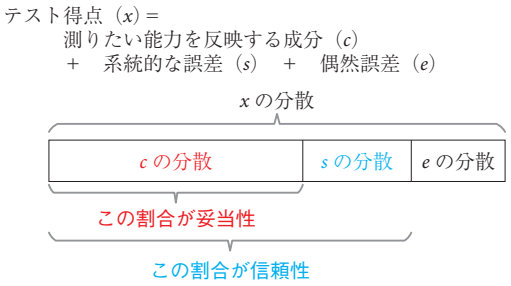

図 2　妥当性と信頼性

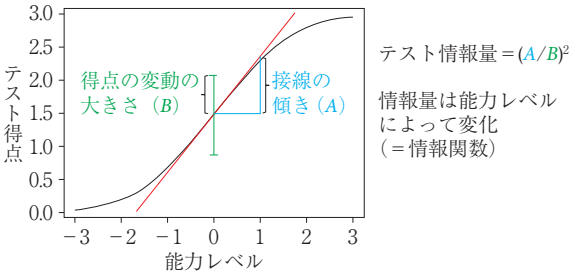

図 3　情報量

一方，テスト得点の分散のうち，測りたい能力を反映する成分と系統的な誤差の分散が占める割合は「信頼性」となります。つまり，信頼性は，テストの得点が偶然誤差によって左右されない程度，その意味で安定している程度を表す指標です。

妥当性と信頼性の間には密接な関係があります。図 2 からわかるように，妥当性が高ければ，誤差の分散は相対的に小さくなりますから，信頼性も高くなります。それに対し，信頼性が高くても，系統的な誤差の分散が大きい場合には，得点が安定はしても妥当性は低いことになります。安定して，的外れのもの（先の例で挙げたタブレット端末使用の慣れ・不慣れの個人差など）を測っている可能性があるのです。

次に，情報量についてです。

図 3 の黒い曲線は，能力レベルに応じて，テスト得点の期待値が高くなっていく様子を示すもので，「テスト特性曲線」と呼ばれるものです。その曲線の傾きが大きいところは，能力の差異が鋭敏にテスト得点に反映されますので，テスト得点を知ることで能力の個人差がかなりわかることになります。その意味で，テスト得点は多くの情報をもっていることになります。ただし，能力レベルが一定でも偶然誤差によってテスト得点が大きくばらつくとしたら，そのぶん，情報は減少します。情報量はその双方を勘案する形で，図 3 の右に示したような式で定義されます。

テスト特性曲線の傾きも，また偶然誤差の大きさも能力レベルごとに異なるので，情報量は能力レベルによって変化する関数となります。これは重要な性質で

す。同じく偶然誤差の大きさを査定する信頼性は，受験者集団に対して 1 つに決まる指標であり，その点で情報量との違いがあります。たとえば，全体として信頼性の高いテストであっても，ある能力レベルでは情報量が低く，そのレベルでの測定には適していない可能性があります。たとえば，難度の高いテストは，全体としての信頼性は高くても，能力レベルの低いところではテスト特性曲線がほとんどフラットになり，能力レベルに差があってもテスト得点の差に反映されないため，能力の個人差を見分ける機能を発揮できないことになります。

得点の比較可能性・公平性

次の「得点の比較可能性」は，文字通り，異なるテストを受験したり，異なる採点者が採点したりした場合に，互いに得点の比較が可能か，という問いに関するものです。

最後の「公平性」は，テストがある特定の属性をもつ受験者集団に対して不当に有利，または不利になっていないかを表す概念です。公平性は妥当性と深い関係があります。なぜなら，公平でないテストは，テストが本来測るべきでない属性によって得点が左右される，つまり系統的な誤差の大きなテストになるからです。

それでは，以上の準備をふまえて，次回から本論に入っていきたいと思います。

文献・注

(1) 高大接続システム改革会議 (2016).「最終報告」
http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2016/06/02/1369232_01_2.pdf

(2) 中央教育審議会 (2014).「新しい時代にふさわしい高大接続の実現に向けた高等学校教育，大学教育，大学入学者選抜の一体的改革について～すべての若者が夢や目標を芽吹かせ，未来に花開かせるために～（答申）」
http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/__icsFiles/afieldfile/2015/01/14/1354191.pdf

(3) 教育再生実行会議 (2013).「高等学校教育と大学教育との接続・大学入学者選抜の在り方について（第四次提言）」
http://www.kantei.go.jp/jp/singi/kyouikusaisei/pdf/dai4_1.pdf

(4) 文部科学省 (2017).「高大接続改革の実施方針等の策定について（平成 29 年 7 月 13 日）」
http://www.mext.go.jp/b_menu/houdou/29/07/1388131.htm

(5) 文献（1）の p. 58。

(6) 以下に，本稿執筆時にウェブで公開されているものの一部を掲載します。
「入試選抜の測定問題」（講演）『大学入試センターシンポジウム 2014 ── 大学入試の日本的風土は変えられるか』pp. 61-74.
http://www.dnc.ac.jp/albums/abm.php?f=abm00004972.pdf&n=ｼﾝﾎﾟｼﾞｳﾑ 2014 報告書 Web.pdf
「共通試験と個別試験に求められるもの ── 測定論の観点から」（講演）『第 24 回東北大学高等教育フォーラム　新時代の大学教育を考える［13］報告書　大学入試における共通試験の役割 ── センター試験の評価と新制度の課題』pp. 7-23.
http://www.adrec.ihe.tohoku.ac.jp/wp/wp-content/uploads/2016/11/IEHE-TOHOKU-Report-68.pdf
「大学入試新テスト記述式案 ── 高校国語ゆがめる恐れ」『日本経済新聞』2016 年 11 月 28 日朝刊
http://www.nikkei.com/article/DGKKZO09986080W6A121C-1CK8000/
「高校の国語教育ゆがむ恐れ ── 開始時期こだわらず検証を」（インタビュー）『AERA』2016 年 12 月 19 日号, pp. 23-24.
https://dot.asahi.com/aera/2016121400209.html
「大学新テスト方針案公表 ── 記述式・英語委託熟考を」『日本経済新聞』2017 年 5 月 22 日朝刊
https://www.nikkei.com/article/DGKKZO16614060Z-10C17A5CK8000/
「真価問われる高・大・新テスト改革」（インタビュー）『産経新聞』2017 年 7 月 19 日朝刊
http://www.sankei.com/life/news/170719/lif1707190002-n1.html
「大学入学共通テストの課題」『NHK 視点・論点』2017 年 9 月 1 日放送
http://www.nhk.or.jp/kaisetsu-blog/400/278834.html

第 2 回

記述式問題と妥当性・信頼性

第 1 回で書いたように，新しい「大学入学共通テスト」では，国語と数学に数問程度，記述式問題が導入される予定です。

このうち国語の記述式問題では，「多様な文章とともに，図表などを含めて，複数の情報を統合し構造化して考えをまとめたり，その過程や結果について，相手が正確に理解できるよう根拠に基づいて論述したりする思考力・判断力・表現力等を評価する」とされています (1)。

現在，大学入試センター試験の国語の受験者数は 50 万人を超えており，その人数が記述式問題に解答するとしたら，答案の採点は膨大な作業になります。その大量の記述式の答案を限られた期間内に正確に採点することができるかどうかが心配されていますが，特に採点の信頼性の観点からは，「採点者が異なっても，一貫した採点ができるか」ということが問われます。つまり，たまたまある採点者にあたったから得点が高くなったとか低くなったというような偶然誤差が十分に抑えられるかという問題です。

受験者数が限られている個別大学の試験では，たとえば 1 つの問題については，全受験者の答案を 1 人の採点者が通して採点するというような方法で対応したりしているようです。このようにすれば，どの答案がどの採点者によって採点されたかという偶然性による得点の変動は抑えることができ，その意味での信頼性は確保できます。ただし，誰がその全体を採点する人になるのかによって結果が変わる可能性があるという意味では，問題が残ります。その問題を解消するには，複数の採点者が全答案を採点し，結果が一致しない場合は合議によって得点を決めるという，より慎重な方法がとられます。

受験者数が 50 万人を超えるような試験ではこのような対応はできません。そのため，「大学入学共通テスト」では，採点者間での採点結果の変動を抑えるために，たとえば，「全体を二文でまとめ，一文目には……を，二文目には……」というように，設問において一定の条件を設定して解答させる「条件付記述式」の問題が採用される予定です [2]。そして，採点においては，正答の条件（形式面，内容面）への適合性を判定するとされています [3]。そうすることで，誰が採点しても結果が同じになるように，できるだけそれに近くなるように，ということを狙っているわけです。

この条件付記述式によって，採点者間の採点のブレが実際に解消されるかどうかは，これから計画されているプレテストによる検証を待つところですが，仮にそれが実現するとしたら，その記述式問題は，限りなく客観式問題に近いものであることになります。実際，これまで提出されたモデル問題例には，本文中から特定の語句を抜き書きする問題などがあり，たしかにそういう客観式に近い問題であれば採点のブレは少なくなると考えられます。

しかし，そのような問題で，先に引用した記述式問題の狙い，すなわち「多様な文章とともに，図表などを含めて，複数の情報を統合し構造化して考えをまとめたり，その過程や結果について，相手が正確に理解できるよう根拠に基づいて論述したりする思考力・判断力・表現力等を評価する」ことが可能でしょうか。これは，測りたい能力が実際にどれだけ測れるかという妥当性についての疑問です。

このように，記述式問題については，採点の信頼性を高めようとすると，その内容や形式が制限され，その結果として，より重要な妥当性についての疑義が生じてくる，というジレンマがあります。大規模試験において，採点のための莫大なコストをかけて，この種の記述式問題を導入する必要があるかについて，私は疑問に思っているところです。

解答形式に関する思い込み

入試改革に関する一連の議論の中で私が強く感じたのは，「記述式であれば深い思考や表現力が測れるが，客観式（具体的にはマークシート式）ではそれは無理」という考えが，広く，かつ根強くもたれていることです。実際には，記述式問題で深い思考や表現力を測ることができるためには，問題文が適切であるだけでなく，採点基準およびそれに基づく実際の採点が適切であることが必要です。記述式であれば良いということではけっしてありません。

一方，マークシート式問題でも，工夫次第で，深い思考や表現力を測ることが可能です。たとえば，アメリカの大学が採用している共通テストである SAT には，“Writing and Language Test” という下位テストがあります [4]。これは，マークシート式問題で文章推敲力を評価し，それによって文章表現力を測るテストであり，注目に値するテストです。また，現在の大学入試センター試験のマークシート式問題にも，深い思考や表現力を問う良問が多く含まれていると私は評価しています。

段階評価と情報量

2014 年 12 月の中央教育審議会の答申では，「各大学はその教育方針に照らし，どのような評価方法を組み合わせて選抜を行うかを，応募条件として求める「大学入学希望者学力評価テスト（仮称）」の成績の具体的提示等を含め，アドミッション・ポリシーにおいて明確に示すことが求められる」と述べられています [5]。そして，その成績については，「「1 点刻み」の客観性にとらわれた評価から脱し，……段階別表示による成績提供を行う」とされていました [6]。

つまり，当時「大学入学希望者学力評価テスト」と仮称されていた新しい共通テストでは，成績をおおくくりの段階別表示とし，各大学は，応募条件として受験者にどの段階を求めるかを示すこととされていたのです。大学はその求める段階によってランクづけられ，受験者は自分が得た段階によって応募できる大学が限定されるということですから，もしこの案の通りに進んでいたら，これまでの大学入試のあり方が大幅に変わることになり，良くも悪くも，当時言われた「明治以来の改革」となったことでしょう。しかし，成績を段階評価することについてはさまざまな問題点が指摘

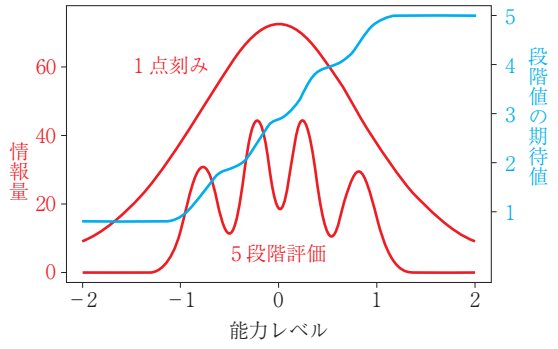

図 1　段階評価と情報量

され，実質上，とりやめになりました[7]。

段階評価の最大の問題点は，選抜に必要な個人差の識別が十分にできないこと，言い換えれば受験者の能力レベルに関する情報量が少なくなることです[8]。

図 1 は，段階評価にすることによって，テストの情報量がどれくらい減少するかを具体的に見ていただくために用意したものです。ここでは，100 項目からなるテストを考え，それを 1 点刻みの 100 点満点で採点した場合と，20 点以下は段階値が 1，そして，21 点から 40 点までは段階値が 2，という具合に段階値 5 までの 5 段階評価をした場合を考えます。計算の簡単のために，100 個の項目は困難度などの統計的特性をすべて同じにしてあります。

赤い曲線のうち，上のほうが 1 点刻みの場合の情報量で，下のほうの波打っている曲線が 5 段階評価の場合の情報量です。これを見ると，能力レベルの特に高いところと特に低いところでは，5 段階評価の情報量がほとんど 0 になっていること，そして，その他の能力レベルでも 1 点刻みの情報量に比べてかなり低くなっていることがわかります。

たとえば，段階評価にすることによってテストの情報量が 3 分の 1 に減少するとしたら，テストの項目数を 3 倍にすることでやっと 1 点刻みの情報量になるということです。言い換えれば，テストの項目のうち 3 分の 2 にあたる部分を捨てるのと同等だということです。せっかく入試改革の方針に沿って良い項目をつくってテストを構成しても，それで得られる情報を最後に捨ててしまうのでは意味がありません。

ところで，図 1 で，5 段階評価の情報量の曲線が波打っていることについてですが，それは，能力レベルと 5 つの段階値との関係によって生じる現象です。図 1 の左下から右上に向かって上昇している青い曲線は，各能力レベルでの段階値の期待値です。期待値というのは，たとえば，ある能力レベルにおいて段階値が 1 になる確率が 0.8 で，段階値が 2 になる確率が 0.2 で，段階値が 3 以上になる確率が 0 だとすると，1 × 0.8 ＋ 2 × 0.2 ＝ 1.2 と計算されます。その能力レベルにおいて，段階値が 1 になる場合も 2 になる場合もあるけど，平均すれば 1.2 というのが期待値の意味です。

この曲線は，ある能力レベルのあたりで傾きが急になり，その後は緩やかで，またある能力レベルのあたりで急になるという形になっています。傾きが緩やかなところは，そのあたりでは能力レベルがある程度異なっても，段階値はほとんど変化しないことを意味します。結果として，その能力レベルのあたりでは情報量が低くなります。一方，傾きが急なところは，そのあたりでは能力レベルの少しの違いで上の段階値になったり下の段階値になったりするので，そこでは情報量が高くなります。これが，5 段階評価の情報量の曲線が波打って上下している仕組みです。

段階評価については，テストの情報量が減少すること以外にもいくつか問題点があります。これについては，第 3 回で取り上げます。

文献・注

(1) 文部科学省 (2017).「大学入学共通テスト実施方針」p. 17。
http://www.mext.go.jp/b_menu/houdou/29/07/__icsFiles/afieldfile/2017/07/18/1388089_002_1.pdf
(2) 文献（1）の p. 3, p. 30。
(3) 大学入試センター (2017).「第 1 回，第 2 回モニター調査実施結果について」p. 23.
http://www.dnc.ac.jp/corporation/daigakunyugakukibousyagaku-uryokuhyoka_test/model.html
(4) SAT の “Writing and Language Test” のサイト参照。
https://collegereadiness.collegeboard.org/sat/inside-the-test/writing-language
(5) 中央教育審議会 (2014).「新しい時代にふさわしい高大接続の実現に向けた高等学校教育，大学教育，大学入学者選抜の一体的改革について～すべての若者が夢や目標を芽吹かせ，未来に花開かせるために～（答申）」p. 12.
http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/__icsFiles/afieldfile/2015/01/14/1354191.pdf
(6) 文献（5）の p. 15。
(7) ただし，第 4 回で述べるように，英語の民間試験の活用の文脈で，段階評価案が再び浮上しています。
(8) 中央教育審議会の答申は，そのことは承知のうえで，わざと共通テストの情報量を少なくし，後は各大学の個別試験でという案だと思います。ですから，ここでの論点は，情報量が多いか少ないかではなく，情報量が少ないとどのような問題があるかということです。

さらに段階評価について

テストの結果をおおくくりの段階評価とすることについては，前回述べた，テストの情報量が減少すること以外に以下のような問題点があります。

まず，いくつの段階に分けるか，またどこを段階の境界とするかという「段階への分け方」が恣意的にならざるをえないことです。資格認定や課程修了認定での「基準設定」（standard setting）であれば，段階数は合・否の2段階に決まりますが，どこを分割点（cut-off score）とするかは完全に客観的に決めることはできず，最終的には実施機関による判断となります[(1)]。

入学者選抜のための共通テストの場合は，段階数からして特に根拠のある値があるわけではなく，段階間の境界に至ってはなおさらそうです。そのような根拠薄弱な段階評価によって，受験者がある大学への応募条件を満たしたり満たさなかったりするのは，不合理でしょう。

次に，段階評価がおおくくりのものであればあるほど，それぞれの段階内には大きな個人差があるのにそれが無視され，その一方，段階の境界付近では，実際にはわずかな能力差であるにもかかわらず異なる段階に振り分けられて，能力差が誇張されるという問題があります。その結果，ある段階の中の下位に位置する者は，かなりの能力向上があっても同じ段階にとどまる可能性が高くなり，努力が反映されにくくなります。それでは学習意欲も高まらないでしょう。その一方で，段階の境界付近では，誤差変動だけでも段階が変化し，決定的な影響をもってしまいます。

1 点刻みの 1 点差で合否が決まることに対しては批判もありますが，段階評価でも段階の境界付近では同じことです。むしろ，1 点刻みの 1 点には実質的な意味がないことが広く了解されているのに対し，同様に実質的な意味のない境界付近のわずかな差が，異なる段階となることによって実質的に意味があるかのように誤解されることのほうが問題でしょう。

段階評価には，さらに，選抜における柔軟で多様な活用を阻害する側面があります。たとえば，東京大学の推薦入試では，ほとんどの学部で，大学入試センター試験で「概ね 8 割以上の得点」を求めており，一部の学部学科ではその基準点を「780 点程度」としています[(2)]。このように各大学あるいは各学部が柔軟な基準設定ができるのは，1 点刻みで成績が提供されているからです。また，テストの得点に重みをつけて，他の多様な情報と総合することも，段階評価では難しくなります。

複数回受験について

共通テストの段階評価の提案は，ある段階以上を大学への応募条件とするなど，いわば資格試験的なテスト利用を想定してなされたものでした。そして，一般的な資格試験は合格できるまで何度でも受験できるのと同様に，共通テストも「一発勝負」ではなく，複数回受験を可能にするという提案がなされていました。

複数回受験は，入学者選抜の目的に使える情報が増えることになりますから，潜在的には選抜の信頼性の向上に資する可能性があります。ただし，複数回受験が可能になるためには，どの回に受験しても，その結果が互いに比較可能であることが必要です。そのためには，すべての回の結果が，共通の尺度上に位置づけられるか，あるいは換算表のようなものが用意されることが求められます。

一連の議論の中では，それを可能にする方法として，項目反応理論（Item Response Theory, 略して IRT）を適用することが話題になっていました。ここで簡単に，IRT の概要と，それを用いて能力の推定をすることで，異なるテストの結果が共通の尺度上で表現される仕組みについて説明しましょう。

IRT の仕組み

IRT の基本は，テストの個々の項目に正答する確率が，能力レベルの上昇に伴ってどのように高くなっていくかを表す項目特性曲線です。図 1 には，難度の低い（易しい）項目 1 から難度の高い（難しい）項目 3 まで 3 つの項目の特性曲線が例示されています。項目 1

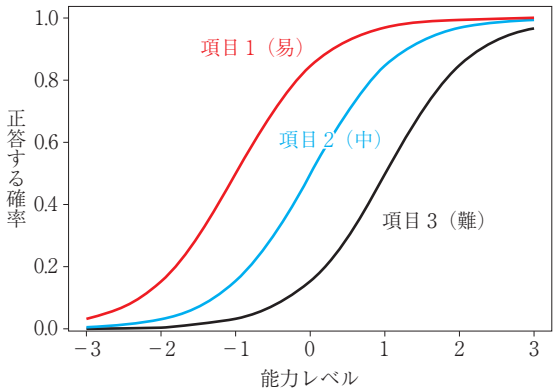

図 1　項目特性曲線

表 1　さまざまな正誤パターン

	パターン 1	パターン 2	パターン 3
項目 1（易）	○	○	○
項目 2（中）	×	○	○
項目 3（難）	×	×	○

の難度が低いことは，どの能力レベルで見ても，他の項目より正答確率が高いことでわかります。テストに含まれる項目すべてについてこの関数を合計すると，能力レベルの関数としてテスト得点（正答数）の期待値を表すテスト特性曲線となります（第 1 回参照）。

いま仮に，その 3 つの項目でテストが構成されているとします。すると，項目それぞれに正・誤の可能性がありますから，全体では $2^3 = 8$ 通りの正誤パターンがあります。表 1 に示したのは，そのうちの 3 つのパターンです。

次が重要なポイントですが，それぞれの項目は，その正答確率が能力レベルの関数としてわかっている（実際には，データから推定されている）ので，それぞれの正誤パターンとなる確率も，能力レベルの関数として表すことができます。たとえば，パターン 2 の○○×となる確率は，項目 1 に正答する確率×項目 2 に正答する確率×（1－項目 3に正答する確率）です [3]。

図 2 は，3 通りの正誤パターンのそれぞれについて，そのパターンが生じる確率を能力レベルの関数としてグラフ化したものです。これを見ると，たとえばパターン 2 の○○×となる確率は，能力レベルの上昇とともに高くなっていき，能力レベルが 0 と 1 の間くらいで最大となって，その後は低くなっていくことがわかります。易しいほうの 2 問に正答して，最も難しい 1 問に誤答するというパターンが生じるには，ある程度の能力レベルが必要ですが，それ以上に高い能力レベルだと，最も難しい 1 問にも正答する可能性が高くなるので，パターン 2 の確率は低くなるということです。

では，このパターン 2 のように解答した受験者の能力レベルはどの程度だと推定できるでしょうか。最尤（ゆう）推定法とよばれる推定法では，このパターンを生じる確率が最大となる能力レベルをもって，その受験者の能力レベルの推定値とします。パターン 2 の場合，計算すると能力レベルが 0.59 のとき確率が最大になりますので，能力レベルは 0.59 と推定されます。同様にパターン 1 の○××の場合は，－ 0.59 という推定値になります [4]。

このように，項目の特性曲線さえ与えられていれば，どの項目の組み合わせでテストを構成しても，そのテストにおける正誤パターンから，能力レベルの推定が

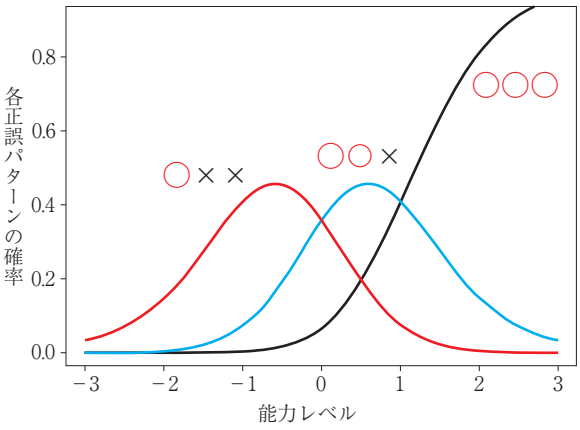

図 2　最尤推定法の原理

可能になります。つまり，複数回受験で，回ごとにテストを構成する項目が異なっても，問題なく，公平な能力推定ができ，互いに比較可能になるということです。

共通テストでの IRT 適用の可能性

このように，理論的には IRT を適用することで，複数回受験を行っても得点の比較可能性が確保できることがわかります。しかし，その前提として，各項目の特性曲線が推定されていなければならず，そのためには，プレテストをして実際に解答データを得る必要があります。IRT を活用しているテストの実際の運用では，本番のテストにおいて，一部，本番の採点には使用しないプレテスト項目を混ぜておいて，そこで項目特性曲線の推定のための解答データを収集する方法などが採用されています。

しかし，この方法を採用するためには，非常に多くの項目が用意されていて，プレテスト項目として混ぜた項目が，その後，本番のテストで使用されてもほとんど影響がないような状況が必要です。このような条件は，現在の大学入試センター試験でも，また今後の共通テストでも実現することが難しいといわざるをえません。

ところで，一連の議論の中では，IRT を利用して能力推定をし，結果は段階評価で，という案が出たことがあります。しかし，IRT は 1 点刻みよりももっと細かく，そして高い精度で能力推定をするための理論ですので，おおくくりの段階評価とは目指す方向が違います。

文献・注

（1）アメリカにおけるミニマム・コンピテンシーテスト（課程

修了のための最低基準をクリアしているか否かを判断するテスト）における基準設定については，専門家が「基準設定に関する文献が何か決定的な点をもっているとしたら，それは，コンピテンシーテストで，擁護しうる基準を設定することの困難さについてである」と述べています（Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: Macmillan. pp. 485-514.（井上俊哉訳，1992「学生のコンピテンスの証明」池田央・藤田恵璽・柳井晴夫・繁桝算男監訳『教育測定学』下巻，みくに出版，pp. 215-257.）の翻訳版の p. 226）。

(2)「平成 30 年度東京大学推薦入試学生募集要項」p. 5，p. 35。
http://www.u-tokyo.ac.jp/content/400065221.pdf

(3) このように掛け算で確率が求められるためには，能力レベルを所与としたとき，ある項目に正答するかどうかが，他の項目に正答するかどうかと独立であるという「局所独立性」を仮定する必要があります。これは IRT の適用において非常に重要な仮定で，テストの内容構成によっては無理な仮定にもなりますが，本稿では詳細は割愛します。関心のある方は，やや専門的な議論になりますが，以下の説明などを参照してください。

南風原朝和 (2000).「個人正答確率に基づく局所独立性の概念の明確化 ── 実験的独立性および一次元性との関係を中心に」
http://www.p.u-tokyo.ac.jp/~haebara/local_ind/

(4) パターン 3 の○○○の場合，つまり全問正答の場合は，能力レベルが高ければ高いほどそのパターンとなる確率が高くなりますので，最尤推定値は無限大になってしまいますが，実際の推定プログラムでは，有限の値になるように調整がなされます。

第 4 回

英語の民間試験と妥当性・信頼性

現在の大学入試センター試験では，英語を話す力と書く力を直接に評価することができません。間接的には，発音・アクセントの問題や，語句を並べ替えて文章を完成させる問題などで，これらを評価する工夫はされていますが，直接的な評価はしていません。このうち書く力については，大学によっては個別試験で評価しているところも少なくありませんが，話す力については個別試験でも評価をしていないのが現状です。

一方，高等学校の学習指導要領では，読む・聞く・話す・書くの 4 技能を総合的に育成することとしています[(1)]。学習指導要領と現在の大学入試の英語のギャップを埋めるためには，現在の大学入試センター試験の後継となる大学入学共通テストの英語で，話す力と書く力も評価できるようにすることが考えられます。しかし，50 万人規模の共通テストでそれを実行することは困難であるという判断から，すでに 4 技能の評価をしている民間試験を活用する案が出てきました。そして，少なくとも 2024 年度入試までは，各大学の判断で，共通テストの英語と民間試験のいずれか，または双方を選択利用することとされています。

しかし，民間試験は，当然ながら日本の大学への入学者選抜を目的に開発されたものではありませんから，その目的に照らして，はたして妥当な試験かどうかを検証する必要があります。第 1 回で述べたように，妥当性は，テストの得点が測りたい能力をどれだけ反映しているかを表すものです。そのため，それぞれの民間試験が，それら本来の目的のためには妥当な試験であったとしても，それを別目的で使用する場合には，あらためて妥当性を調べる必要があるのです。

また，入学者選抜に使うのであれば，民間試験の信頼性についても確認しておく必要があります。第 2 回で，記述式問題の採点の信頼性について述べましたが，英語を話す力の採点は，発音や応答の速さなど，記述式問題以上に多くの要因が反映されますので，信頼性を確保するのはより困難であることが想定されます[(2)]。

多くの民間試験のうちのどれを採用するのかについて，文部科学省は最近，「「大学入学共通テスト実施方針」（平成 29 年 7 月 13 日）では，資格・検定試験をセンターが「認定」するとしているが，これは法的根拠に基づく認定制度ではない。本要件は，あくまでも成績提供システムに参加するための要件として定めるものである。（資格・検定試験そのものの質や内容を評価するものではない。）」としており[(3)]，試験の質や内容は評価しないと明言しています。だとすると，国立大学協会など，これらの試験を使用する可能性のある大学側が妥当性と信頼性を評価して，入学者選抜の要請に応えるものであるかどうか判断しなくてはならないでしょう。

複数の民間試験の得点の比較可能性

いま仮に，そのような観点から，入学者選抜にふさわしい民間試験が複数選ばれたとします。その場合，その次の問題は，それぞれ内容や難易度の異なる別々の民間試験を受けた受験者の成績をどのように互いに比較するかということです。

この成績比較の問題について文部科学省は，「ヨーロッパ言語共通参照枠」（Common European Framework of Reference for Languages），略して CEFR と呼ばれる参照枠を利用することを想定しています。CEFR は欧州評議会が，ヨーロッパの諸言語に共通して使えるように作成した言語教育と評価のためのガイドラインで，ヨー

表 1　各試験団体のデータによる CEFR との対照表

CEFR	Cambridge English	英検	GTEC CBT	GTEC for STUDENTS	IELTS	TEAP	TEAP CBT	TOEFL iBT	TOEIC / TOEIC S&W
C2	CPE（200+）				8.5-9.0				
C1	CAE（180-199）	1 級（2630-3400）	1400		7.0-8.0	400	800	95-120	1305-1390 L&R 945 ～ S&W 360 ～
B2	FCE（160-179）	準 1 級（2304-3000）	1250-1399	980 L&R&W 810	5.5-6.5	334-399	600-795	72-94	1095-1300 L&R 785 ～ S&W 310 ～
B1	PET（140-159）	2 級（1980-2600）	1000-1249	815-979 L&R&W 675-809	4.0-5.0	226-333	420-595	42-71	790-1090 L&R 550 ～ S&W 240 ～
A2	KET（120-139）	準 2 級（1284-1800）	700-999	565-814 L&R&W 485-674	3.0	150-225	235-415		385-785 L&R 225 ～ S&W 160 ～
A1		3 級 -5 級（419-1650）	-699	-564 L&R&W -484	2.0				200-380 L&R 120 ～ S&W 80 ～

（注）各試験団体の公表資料より文部科学省において作成。平成 28 年 3 月現在。

ロッパ以外でも広く利用されているものです。CEFR では，表 1 に示したように，言語能力が A1 レベルから C2 レベルまでの 6 段階に分けられており，A1，A2 は基礎段階の言語使用者，B1，B2 は自立した言語使用者，そして，C1，C2 は熟達した言語使用者とされています。

この表の上のほうに並んでいるのが民間試験の主なものです。そして，表の中には，それぞれの民間試験でどのような成績をとれば，CEFR の各段階に相当するかを示しています。この内容は，それぞれの試験団体自身が公表したものを文部科学省でまとめたものです [4]。たとえば，英語圏の大学への入学には B2 レベルが必要とされることが多いようですが，それが英検ですと準 1 級に合格すること，そして TOEFL iBT ですと，72 点以上をとることがそれに相当するということです。このように，CEFR という共通の枠組みに対応づけることを通して，結果的に，異なる試験の間で成績を比較できるというのが文部科学省の説明です。

しかし，この方法には 3 つの問題点があります。

1 つは，それぞれの試験の成績と CEFR の段階との対応づけが，客観的に決まるものではなく，最終的にはそれぞれの試験団体の判断によるということです [5]。したがって，一度示された対応づけが，その後の判断によって大きく変化することもあります。

表 2 はその一例ですが，TOEFL iBT の成績と CEFR の段階との対応づけが，2008 年と 2014 年というわずか 6 年間の間でかなり変更されていることがわかります [6]。たとえば，先ほど述べた B2 レベルに必要な成績が，2008 年には 87 点であったものが，2014 年には 72 点で良いというふうに変わっています。

表 2　TOFLE iBT と CEFR との対応づけの変化

CEFR	2008 年	2014 年
C2		
C1	110 以上	95 以上
B2	87 以上	72 以上
B1	57 以上	42 以上
A2		
A1		

このように，それぞれの試験の成績と CEFR の段階との対応づけには曖昧さや不安定さがあるので，それをもとに，異なる試験の成績を互いに比較するのは，大学入学のための共通テストという重要さを考えると，問題があるように思います。

もう 1 つの問題は，仮に各試験の成績が CEFR の段階にしっかりと対応づけられたとしても残る問題です。

表 3 は，2015 年に文部科学省が国公立約 500 校，約 9 万人の高校 3 年生を対象に英語力の調査を行ったうちの，公立高校の生徒の結果です。CEFR の段階別にそこに何％の生徒が含まれるかという割合を示しています [7]。これを見ると，読む・聞く・話す・書くのどの技能においても，CEFR の B1 レベル以上は，3

表 3　文部科学省の平成 27 年度英語力調査結果（高校 3 年生）

CEFR	読む	聞く	話す	書く
B1 以上	2.1%	2.3%	1.2%	0.7%
A2	29.9%	24.2%	9.8%	17.2%
A1	68.0%	73.6%	89.0%	82.1%

％未満であること，言い換えれば，下の 2 つの段階，A1 と A2 に 97％以上が入ってしまうことがわかります。このように 2 つの段階にほとんどの生徒が入ってしまうような段階分けでは個人差を十分に見ることができず，とても共通テストとして選抜の目的に使うことはできません。

さらに，いまの問題と関連して，第 2 回，第 3 回で述べた段階評価の問題があります。たとえば同じ A2 レベルでも，その下の A1 に近いレベルと上の B1 に近いレベルでは相当な個人差がありますが，この評価法では同じ扱いになります。一方，たとえば A2 と B1 の境界では，誤差変動くらいでも段階が変化し，大きな影響をもってしまいます。

この最後の問題については，CEFR の 6 段階よりも細かい段階評価をする案も出ていますが，上記の問題点や素点がもっている情報量が失われるなどの問題点は，本質的には改善されません。したがって，もし民間の試験を活用するとしたら，素点や標準得点を使用するほうが良いということになりますが，その場合，異なる試験間の成績比較は相変わらず難しい問題として残ります [8]。

文部科学省は，できるだけ多くの民間試験を採用するように呼び掛けていますが，本来は，試験の質や内容を評価して，大学の入学者選抜の目的に最もふさわしいもの 1 つを選ぶのが筋でしょう。そして，多面的・総合的に評価した結果，もしもどの民間試験よりも，現在の大学入試センター試験の後継となる大学入学共通テストの英語を採用するのがより良いと判断されることがあれば，あるいはそのように判断する大学があれば，それを採用するのも当然，あってよいことでしょう。

障害のある受験者への配慮

2016 年 4 月に，いわゆる「障害者差別解消法」が施行され，障害者に対して合理的な範囲で配慮をすることが法律で求められることになりました。もちろん，障害のある受験者への配慮は，この法律の施行を待つまでもなく，これまで大学入試センター試験や各大学の個別試験で行われてきています。

たとえば，手の運動機能に障害があると，マークシートをうまく塗りつぶすことができません。しかし，テストで測りたいのは，そのような運動機能ではなく，それぞれの教科・科目の学力ですから，マークシートを塗りつぶすことができない場合は，選択肢に直接チェックするのでもよい，というような措置をして，その受験者の学力が十分に発揮できるようにしています。また，視覚障害のある受験者に点字で出題することもなされています。また，その場合を含め，障害のために解答に余分の時間がかかると判断される場合には，試験時間を延長することもあります。

このような対応は行われてきたものの，今後，大学入学共通テストに記述式問題を導入することや，民間試験を使って英語を話す力を評価するなどの新しい方向性により，障害のある受験者への対応について，これまで以上の工夫が求められることになります。

このシリーズの最後のテーマとして，「障害のある受験者に対してある配慮を行うことが公平である」とはどういうことか，という重要な問題を考えてみたいと思います。

たとえば，先に例示した「試験時間の延長」という措置を考えます。このとき，「障害のある受験者に時間延長を認めることが公平であるためには，障害のない受験者に時間延長をしても成績に影響しないことが必要である」という主張を見ることがあります。また，実際，障害のない受験者を対象に，ある措置をとることで成績が（ほとんど）変化しないことを示すことで，その措置を障害のある受験者に適用することを正当化しようとする研究も少なくありません。

しかし，実際に示す必要のあることは，「障害のある受験者にその措置を適用する」，そして「障害のない受験者にはその措置を適用しない」という異なる条件で，もしその両者の受験者が，「テストにおいて測りたい能力が等しいならば，テスト得点の期待値も等しくなる」ということです。もっと一般的にいえば，テストが公平であるということは，「測りたい能力が同じであれば，テスト得点の期待値が，障害の有無など測りたい能力以外の属性によらず等しくなること」ということになります [9]。

これはいわば理論的な定式化で，実際には「測りたい能力が同じ」ということの証明は難しいです。しかし，それでも，このような定式化をし，それを参照することで，先に述べた誤解から導かれる「障害のない受験者に適用すると成績が向上するような措置を，障害のある受験者に適用するのは不公平だ」という主張

に根拠がないことを示すことができ，合理的配慮の範囲を拡大できる可能性が出てきます。

ただし，この問題については，理論的に定式化される「公平性」のほかに，人々が主観的に抱く「公平感」「不公平感」も無視するわけにはいきません。この点は，このシリーズで扱った他の問題についても同様で，テスト理論のような理論的な観点のみから，今後の入試のあるべき方向性を決めることはできません。しかし，実証性・専門性を欠いたまま議論が進んでいくのは，何としても食い止めなければなりません[10]。

大学入試改革に関する今後の議論が，これまで以上に実証的・専門的な根拠をもって進められることを願って，このシリーズを閉じたいと思います。4 回にわたり読んでいただき，ありがとうございました。

文献・注

(1)「高等学校学習指導要領解説　外国語編・英語編」
http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2010/01/29/1282000_9.pdf

(2) 以下の報告には，話す力の評価において，「採点者内信頼性（intrarater reliability）を確保することの難しさに，採点者一同が打ちのめされたことが浮彫になった」(p. 37) との記述があります。
羽藤由美・神澤克徳 (2016).「CBT 英語スピーキングテストの開発と実施 ── 入試への導入に向けた試みの検証」『京都工芸繊維大学情報科学センター広報誌』**34**, 30-48.
https://kitspeakee.files.wordpress.com/2016/01/e5ba83e5a0b-1no-34e68a9ce3818de588b7e3828a.pdf

(3)「英語 4 技能大学入試成績提供システム（仮称）への参加要件について（案）」
http://www.mext.go.jp/b_menu/shingi/chousa/shotou/134/shiryo/__icsFiles/afieldfile/2017/09/13/1395611_6.pdf

(4)「大学入学者選抜改革について」の p. 37。
http://www.mext.go.jp/b_menu/houdou/29/07/__icsFiles/afieldfile/2017/07/18/1388089_002_1.pdf

(5) CEFR の段階への対応づけについては，以下のような詳細なマニュアルがあり，それに沿って判断やその集計等を行うこととされています。
Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR), A manual*. Strasbourg, France: Council of Europe.
http://communicationskillsworkshop.pbworks.com/f/CEFR+Manual+Final.pdf

(6) Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum ETS RM-15-06). の p. 8 の表と記述をもとに作成。
https://www.ets.org/Media/Research/pdf/RM-15-06.pdf

(7)「平成 27 年度英語力調査結果（高校 3 年生）の速報（概要）」の p. 5 をもとに作成。
http://www.mext.go.jp/b_menu/shingi/chousa/shotou/117/shiryo/__icsFiles/afieldfile/2016/05/24/1368985_7_1.pdf

(8) 異なる試験間の成績の対応づけは linking と呼ばれ，一連の方法が用意されていますが，その結果については，対象者の属性に依存する問題点が指摘されています（下記の第 10 章参照）。たとえば，テスト A とテスト B の対応づけをする際，テスト A の測定内容に慣れている受験者を対象者とする場合，テスト B の測定内容に慣れている受験者を対象とする場合に比べて，テスト A のより高い点がテスト B のそれぞれの点に対応するという結果が生じやすくなります。
Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer.

(9) この議論についてくわしくは，下記シンポジウムの指定討論を参照。
高橋知音・佐藤克敏・立脇洋介・近藤武夫・南風原朝和 (2016).「研究委員会企画シンポジウム 4　障害のあるテスト受験者への合理的配慮とエビデンス」『教育心理学年報』**55**, 304-312.
https://www.jstage.jst.go.jp/article/arepj/55/0/55_304/_article/-char/ja/

(10) 下記は，最近の大学入試改革論も視野に入れたテスト理論入門の良書です。
光永悠彦 (2017).『テストは何を測るのか ── 項目反応理論の考え方』ナカニシヤ出版

著　者

南風原 朝和（はえばら・ともかず）：東京大学高大接続研究開発センター長／大学院教育学研究科教授。主著に『心理統計学の基礎 ── 統合的理解のために』（有斐閣，2002 年），『続・心理統計学の基礎 ── 統合的理解を広げ深める』（有斐閣，2014 年），『量的研究法』（臨床心理学をまなぶ 7，東京大学出版会，2011 年）など。