

# 1. 「IRTとCBTの光と影－高大接続改革の夢か現か幻か－」

名古屋大学名誉教授・名古屋大学アジア共創教育研究機構客員教授

野口 裕之

## はじめに

私は、高大接続改革に公的な関わりは一切ありません。フリーの立場です。長年、大学の一研究者としてIRTやCBTに関する研究、そして最近では言語テストに関する研究を行ってきました。CEFR（Common European Framework of Reference：ヨーロッパ言語共通参照枠）は、日本では“セファール”と呼ばれることが多いのですが、英語のケンブリッジ・イングリッシュ・ランゲージ・アセスメント（Cambridge English Language Assessment：ケンブリッジ英語検定機構）がYouTubeに流しているものを聴いてみると、C-E-F-R（スィー、イー、エフ、アール）とはっきりと発音しています。ですので、私はここではC-E-F-R（スィー、イー、エフ、アール）と言いたいと思います。

これから私がお話しすることは、私個人の考え方や見解が大きく反映しています。他の考え方や見解があることも重々承知しておりますので、うまく建設的な議論が出来ると思います。高大接続改革を否定的にとらえているわけではありません。より良い方向で、より現実味のある話で進めなければならないと思っていますので、ここでは建設的な議論ができればと思っています。

本日の流れですが、1では「高大接続改革とテスト」、要するに大学入学共通テストのことで、それについてお話しします。2では「IRTとは何か」、3では「CBTとは何か」、この2と3で基本的なことをお話ししたいと思います。4では「英語4技能外部試験をどう位置付けるか」についていろいろと考えていきたいと思っています。そして最後に「まとめ」をお話しします。

## 1. 高大接続改革とテスト

「大学入学共通テスト」は、大学入試センター試験に代わるテストで、“共通テスト”と略されています。大学入学希望者を対象に、高等学校段階における基礎的な学習達成度を判定し、大学教育を受けるために必要な能力について把握することを目的としています。このため、各教科・科目の特質に応じ、知識・技能を十分に有しているかの評価も行いつつ、思考力・判断力・表現力を中心に評価を行うものとしています。〈図1～3参照〉

<図1>

### 1. 高大接続改革とテスト

大学入学共通テスト

1. 大学入試センター試験に代わるテスト:「共通テスト」と略される
2. 大学入学希望者を対象に、高等学校段階における基礎的な学習の達成の程度を判定し、大学教育を受けるために必要な能力について把握することを目的とする。

このため、各教科・科目の特質に応じ、知識・技能を十分に有しているかの評価も行ないつつ、思考力・判断力・表現力を中心に評価を行なうものとする。

<図2>

3. 大学入試センターが問題の作成、採点その他一括して処理することが適当な業務等を行なう。
4. 【英語】大学入学者選抜においても、「読む」「聞く」「話す」「書く」の4技能を適切に評価するため、共通テストの枠組みにおいて、既に民間事業者等により広く実施され、一定の評価が定着している資格・検定試験を活用する。
5. 大学入試センターが必要な水準・要件を満たしているものを認定し、その試験結果およびCEFRの段階別成績表示を要請のあった大学に提供する。

<図3>

6. 国は、CEFRの段階別成績表示による対照表を提示する。
7. 各大学は、認定試験の活用や、個別試験により英語4技能を総合的に評価するように努める。
8. CBTの導入については、引き続きセンターにおいて、導入に向けた調査・検討を行なう。

「大学入学共通テスト」実施方針(案) (2017.05.16.文科省公表)より抜粋  
文言は適宜変更した。

私は高大接続改革について特に詳しいわけではないので、<図3>の一番下に入れましたが、2017年5月16日に文部科学省から公表された『「大学入学共通テスト」実施方針(案)』から抜粋し、文言を適宜変更して、使用しています。

3に、「大学入試センターが問題の作成、採点その他一括して処理することが適当な業務等を行う」とありますが、ここで大学入試センターの存続が決定しているということが分かります。以前、「改組する」という話を聞いたことがありましたが、この実施方針(案)にはそのように書かれています。

4では英語について、『大学入学者選抜においても「読む」「聞く」「話す」「書く」の4技能を適切に評価するため、共通テストの枠組みにおいて、既に民間業者等により広く実施され、一定の評価が定着している資格・検定試験を活用する』とあります。CEFRでは「話す」が、今、私がお話ししているような「話す」と、「ディスカッションする」の2つに分かれていますので、実は5技能なのですが、その2つを「話す」で1つにまとめ、4技能としてお話を進めます。

5では「大学入試センターが必要な水準・要件を満たしているものを認定し、その試験結果およびCEFRの段階別成績表示を要請のあった大学に提供する」とあります。つまり、大学入試センターが主体であることがわかります。

6に「国は、CEFRの段階別成績表示による対照表を提示する」とあります。この表は新聞等

に掲載されることもありますので、皆さんもご覧になったことがあるかもしれません。

7に「各大学は、認定試験の活用や、個別試験により英語4技能を総合的に評価するように努める」、8では「CBTの導入については引き続きセンターにおいて、導入に向けた調査・検討を行う」とあります。つまり、高校生を対象とした、学力の基礎となる知識や技能を評価するテスト「学びの基礎診断」でも、調査・検討をするということです。後でもお話ししますが、こういうことは性急にすすめてはだめです。きちんと研究し、試行し、そして具合の悪いところを修正していくといった、ある程度時間のかかることなのです。それにも関わらず、導入を開始する目標年度を先に設定してしまうと、良いものは生まれてきません。

少なくとも、『「大学入学共通テスト」実施方針（案）』にはIRTの3文字は出てきません。いつの間にかどこかに飛んでしまいました。そしてCBTの3文字は一箇所だけ出てきます。私の個人的な感じ方かもしれませんが、英語の外部試験・検定試験の英語の4技能については頻繁に取り上げていますが、例えばTOEFLは「Speaking（話す）」と「Writing（書く）」はIRTベースのテストではありませんが、「Reading（読む）」と「Listening（聞く）」はIRTがベースとなっています。つまり英語の外部資格試験・検定試験には、すでにIRTとCBTが反映しているのです。

## 2- 1. IRTとは何か

IRTとは“Item Response Theory”の略で、「項目応答理論」ともいいます。一般的には“アイ・アール・ティー”とっていますが、私は「項目応答理論」というほうが好きです。また心理学の分野では“項目反応理論”ということが多いようです。これはどのようなものであるかということ、“テスト項目や受験者集団に依存せずに、テストの受験者の能力値を算出することができるテスト理論”のことです。“古典的テスト理論（Classical Test Theory：CTT）”に対して、“現代テスト理論”として位置付けられています。現代テスト理論には、IRTだけではなく、G-theory、つまり“Generalizability theory（一般化可能性理論）”を入れることもあります。G-theoryは古典的テスト理論の延長線上にあります。IRTは延長線上にありません。全く違うモデルなのです。

## 2- 2. IRTと古典的テスト理論

古典的テスト理論では、項目の困難度を表すのに正答率（通過率）というものが使われます。これはどれくらいの比率の受験者が正答しているかであり、例えば100人の受験者のうち60人が正答していれば、正答率または通過率は0.6ということになります。つまり、受験者の能力を表すのは、基本的に正答数得点・正答した項目の数、つまり正答数です。しかしながら、同じ問題でも受験者集団が変われば得点・正答数が違ってきますので、“受験者集団に依存”して、正答率は変わります。また、同じ受験者が項目の異なるテストを受けた時、例えば困難度の高い項目から構成されたテストを受ける時と、困難度の低い項目から構成されたテストを受ける時では得点が違ってきます。困難度の低い項目から構成されたテストを受けるほうが得点・正答数は高くなります。困難度の高い項目から構成されたテストは得点・正答数が低くなります。つまり、その人の能力が急に変わった訳ではないので、解答したテスト問題の“項目の困難度に依存”して得点・正答数が変わり、また正答率も変わるということになります。

これに対してIRTは、“受験者に依存しないで項目の特性を表わすことができる”、そして“解答

した項目に依存しないで受験者の能力・特性を表わすことができる”のです。受験者に依存しないで項目の特性を表わすことができるのも大切なことなのですが、解答した項目に依存しないで受験者の能力を表わすことができるので、例えばTOEFLでは受験する時期が異なっても、いつ受験しても、同じスケール上のスコアとして受験者の能力が表わされるのです。出題される問題は同じではありませんが、解答した問題の項目は異なっている、同じスケール上のスコアとして、受験者の能力が表わされます。IRTは“解答した項目に依存しないで能力を表わすことができる”テスト理論なのです。

**【司会から質問】**

TOEFLなどで、お父さんが何年も前に受けた時の点数と、自分が今年受けた時の点数が比較できるというのは、このことによるということですか。

**【野口先生回答】**

基本的にはそうです。ただし、今はIBT (Internet-based test) であり、お父さんの時代はPBT (Paper based test) です。PBTの時代は「Listening (聞く)」と「Reading (読む)」だけでしたが、今は4技能ですので、PBTとIBTの換算表というものが一応ありますが、そのまま比較することはできないと思います。

ではIRTは夢のような方法なのかということ、実はそうではありません。いいことばかりではありません。なぜなら、いつ受験しても同じスケール上に乗るようになるのは、相当な努力が必要なことなのです。後でもお話ししますが、それを開発していくのに人的資源、開発するための時間、それと当然のことながら資金がかかります。TOEFLを行っているETS (Educational Testing Service)、IELTS (International English Language Testing System) やケンブリッジ英語能力検定試験 (Cambridge English Language Examination) などは、開発に携わるヒトが多数います。研究者もいますし、問題を作成する実務家を多数抱えています。そして研究開発に非常に多くの時間をかけ、努力をしているのです。つまり、数年で簡単に出来るという話ではないということです。

IRTはフレデリックM. ロード (Frederic M Lord)<sup>30</sup>によって、今から65年前の1952年に基礎が確立<sup>31</sup>され、ロードとメルビンR. ノヴィック (Melvin R. Novick) の両氏により、1968年には既に数理的に体系化<sup>32</sup>されていました。1964年の東京オリンピックの頃には、理論的にはある程度体系化されていたということです。TOEFLのようなETSで開発実施される試験を中心に、すでに実用水準で用いられていました。日本で注目され、公的試験などで用いられるようになったのは最近のことなのです。と言いましても、20年ほどは経っていると思います。テストといえば、アメリカという印象が強いですが、アメリカのみならずヨーロッパやオーストラリアなどでも言語テストを中心に、広く用いられています。例えばPISA調査<sup>33</sup> (学習到達度調査) は、オーストラリア・メルボルン郊外にあるACER (Australian Council for Educational Research) が中心と

<sup>30</sup> 米国のETS (Educational Testing Service) でテスト研究の中心的役割を果たした。

<sup>31</sup> E. M. Lord, A Theory of Test Scores, Psychometric Monograph, vol.7, Psychometric Society, 1952]

<sup>32</sup> E. M. Lord and M. R. Novick, Statistical Theories of Mental Test Scores, Addison-Wesley, 1968

<sup>33</sup> OECD (経済協力開発機構 Organization for Economic Co-operation and Development) が進めているPISA (Programme for International Student Assessment) と呼ばれる国際的な学習到達度に関する調査。15歳児を対象に読解力、数学的リテラシー、科学的リテラシーの三分野について、3年ごとに本調査を実施している。

なって、3回前まで行なっていました。事情は分かりませんが、その調査主体はその後、アメリカのETSに移ったそうです。

### 2-3. ラッシュ・モデル

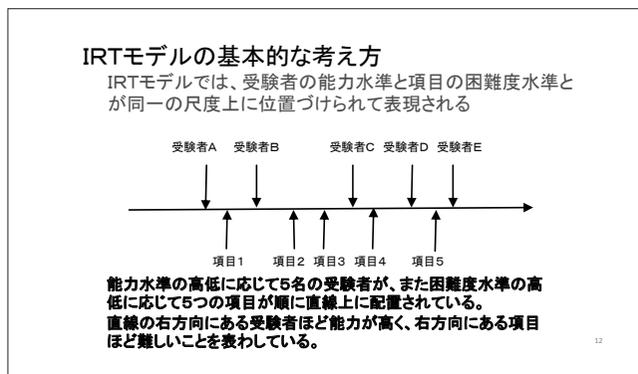
言語学系や英語教育の専門の方はよく耳にされると思いますが、「ラッシュ・モデル」というものがあります。数理的にはIRTと同じモデル群に含まれますが、別の文脈から導かれたモデルです。本講演ではラッシュ・モデルも合わせて、IRTモデルと呼ぶことにしますが、「ラッシュ・モデルとIRTは似て非なるもの」と、英語教育の静哲人先生は本<sup>34</sup>を出版されています。確かにその通りです。受験者の能力を表わす値と項目の困難度を表わす値を分離して独立に表現したいということがあり、デンマークの数学者ゲオルク・ラッシュ（Georg Rasch）がこのモデルを提案したのです。この数学者の名前を採り、「ラッシュ・モデル」としています。

先ほども申し上げましたが、テスト項目に依存しないスコア・能力値、それから受験者集団に依存しない項目の困難度などが必要でした。実はこれはコペンハーゲン大学、それからデンマークの軍隊の中で能力を図ることと関係しながら発展してきたらしいのですが、これについて詳しくありませんので“らしい”としか言えません。デンマークで生まれたこのモデルを、シカゴ大学のベンジャミン・ライト（Benjamin Wright）が中心となって育て、研究および普及活動が進められました。ヨーロッパやオーストラリアでは言語テストの標準的な分析モデルです。ケンブリッジ英語能力検定試験は、「ラッシュ・モデルを使っている」ということを、どこかにはっきりと書いてあったと思います。ですが、アメリカでは標準的なモデルではないようです。むしろ「ラッシュ・モデルでよいのか」といった言い方をされることもあります。

とにかく、何にせよ、いわゆる先進国ではIRTやラッシュ・モデルはすでに実用水準で使われてきていました。決してもの珍しいものではないということです。日本では、最近になって様々な資格認定試験などで使われるようになって来ました。

### 2-4. IRTモデルの基本的な考え方

<図4>



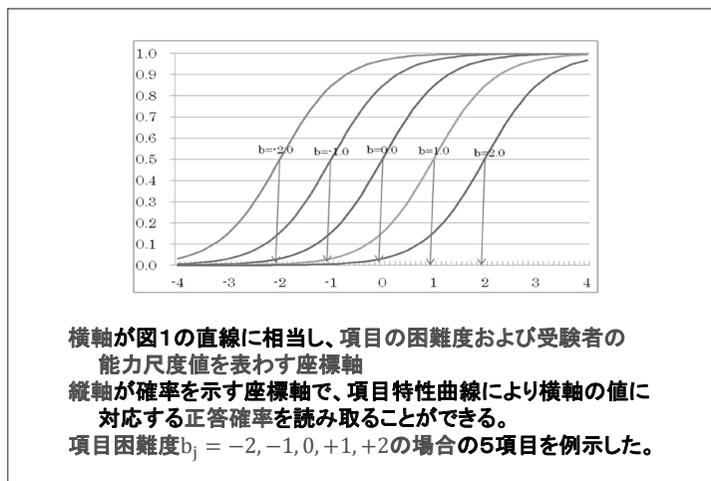
<sup>34</sup>「基礎から深く理解するラッシュモデリング 項目応答理論とは似て非なる測定のパラダイム」静 哲人、関西大学出版部 2007年

IRTモデルの基本的な考え方についてですが、IRTモデルでは“受験者の能力水準と項目の困難度水準が、同一の尺度上に位置付けられて表現”されます。〈図4参照〉つまり一本の横矢印線を尺度して考えると、受験者A、B、C、D、Eの能力値と、項目1、2、3、4、5の困難度が、同一の尺度上に位置付けられて表現されるということです。そして矢印線の右側に位置付けられた受験者ほど能力が高く、右側に位置付けられた項目ほど困難度が高いということになります。また、左側に位置付けられた受験者の能力は低くなり、左側に位置付けられた項目の困難度が低くなるということを、この図は表わしています。このように一本の同じ尺度上に、能力値と困難度を表しましたが、ではこれをどうするのかというと、例えば受験者Bが項目1に正答する確率、つまりどの程度で正答出来るのか、あるいは項目5にどの程度で正答することができるのかということを見るには、モデルを示す必要があります。

〈図4〉の直線上の位置で、『「受験者<sup>ひく</sup>-項目」の距離が0.0より大きいほど、当該受験者がその項目に容易に正答することができ、0.1より小さいほどその項目に正答するのは困難である』といいます。どういうことかということ、例えば受験者Cは項目1までの距離が長く、項目3については距離が短い、つまり項目1は非常に容易に正答することができるが、項目3には、容易ではないですが、恐らく正答できるであろうということになります。一方、受験者Cの能力値と項目5の困難度の値は「受験者<sup>ひく</sup>-項目」でいうとマイナスになりますので、項目5については正答することが難しいということになります。

## 2-5. 項目特性曲線

〈図5〉



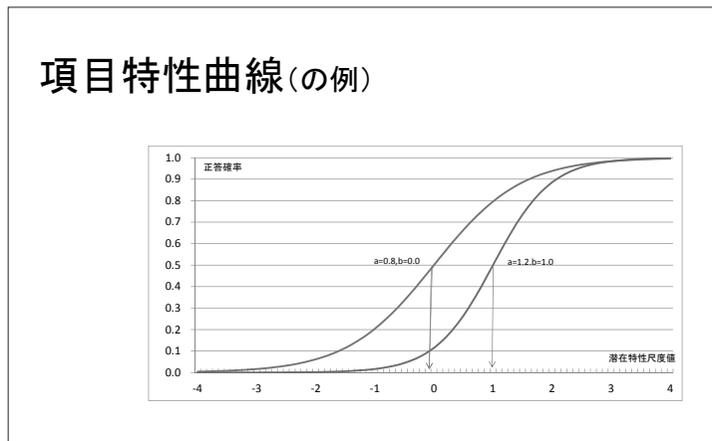
実際には正答のしやすさを「正答確率」というもので表現します。ここで確率を導入します。〈図5参照〉横軸の値（潜在特性尺度値）に対して、縦軸は正答確率を表わす座標軸で、0.1～1.0の値が設定されます。そしてここに色をつけた5本の曲線があります。これらは「項目特性曲線」と言います。例えば、赤色（真ん中の曲線）の項目に対して受験者の能力値が横軸の値・潜在特性尺度値が1.0の位置にとあったとすると、この赤色（真ん中の曲線）の項目に正答できる確率

は約0.15ですが、潜在特性尺度値が0.0の位置にあった場合には正答できる確率が0.5と、相対的に高い確率で正答できるというように、能力値の変化に対して正答確率が変化する、つまり能力が高ければ正答確率が高くなる、能力が低ければ正答確率は低くなるということが示されます。また、紫色（一番右側の曲線）の項目の場合は、能力値が0.0の位置にある人はほとんど正答できませんが、2.0の位置にある人は確率0.5で正答できます。赤色（真ん中の曲線）の項目の場合は、能力値が0.0の位置にある人でも正答確率が0.5であり、能力値が2.0の位置にある人はほとんど1に近くなります。ということは、この図は右にあるものが難しい項目であり、左にある項目のほうがやさしい項目ということを表しています。つまり、この曲線の位置が項目の“困難度”を明示しているのです。

曲線の傾きがすべての項目で同一なのでラッシュ・モデルになるのですが、IRTモデルとしては次のような言い方ができます。

「項目応答理論 (IRT) では、項目の特性 (困難度および識別力) が項目特性曲線を用いて表される。これは項目ごとに潜在特性尺度値を横軸に取り、その特性尺度値を持つ受験者が正答する確率を縦軸方向にプロットして得られる曲線である。項目特性関数にどのような関数を用いるかに応じて、具体的なモデルが決まる。」

<図6>



<図6>は、先ほどの5本の曲線があるもの<図5>と異なります。先ほどは、すべて傾き具合が等しかったのですが、この2本の曲線は傾き具合が違います。これは「2パラメタ・ロジスティック・モデル」といわれるもので、2つの曲線の横にそれぞれ項目パラメタと呼ばれるものが記入してあります。この2つの数値で曲線が唯一に決まります。この曲線が項目特性曲線で、赤色(右側の曲線)の項目は青色(左側の曲線)の項目よりも難しいということが分かります。また「スロープが急な方が、識別力が高い」という言い方をします。「識別力パラメタというのは適切ではない、スロープ・パラメタということが望ましい」と言う人もいますが、とにかくこれで表わされます。実際のテストにIRTを適用する前にはこの曲線のパラメタを推定する必要があります。

IRTモデルは具体的には色々あります。「2値型応答モデル」は0か1、すなわち項目に対する応答が正答か誤答か、あるいはパーソナリティ検査のような“はい”か“いいえ”かのよう、2

つの段階で表わされます。「多値型応答モデル」は、部分得点が与えられるなど、段階づけられたカテゴリで表わされる場合に適用されます。例えば、0, 1, 2, 3の場合、完全な正答なら3、完全な誤答なら0、少しは見るべきところがあるというなら1、ほぼ正解だが少し誤りがあるというなら2になります。この0, 1, 2, 3の4段階のように、部分得点が与えられるなど段階づけられたカテゴリ (Graded Response) に適用されます。

## 2-6. IRTモデルの代表的なもの

先ほどもお話ししましたが、よく用いられる2値型応答モデルの代表的なものに、「2パラメータ・ロジスティック・モデル」があります。〈図6〉のモデルは次のような関数で表されます。添え字のjは項目番号を表します。

$$P_j(\theta) = \frac{1}{1 + \exp\{-1.7a_j(\theta - b_j)\}}$$

この式では位置が $b_j$ 、スロープつまり傾きが $a_j$ であり、この2つが決まれば関数形が決まるということを表わしています。 $b_j$ は困難度パラメータで項目特性曲線の位置を表し、パラメータ値が大きい方が、項目特性曲線が右寄りになり困難度が高いことを表しています。また $a_j$ は識別力パラメータで項目特性曲線の立ち上がりの程度(勾配)を表し、パラメータ値が大きい方が項目特性曲線の立ち上がり急になり識別力が大きいことを表しています。

もう一度、〈図6〉を参照します。赤い曲線(右側の曲線)の横に小さく $a=1.2$ ,  $b=1.0$ 、青い曲線(左側の曲線)の横には $a=0.8$ ,  $b=0.0$ と書いてあります。これは、この曲線にはパラメータ $a$ 、パラメータ $b$ の2つがあり、その数によって、この曲線が決まるということを示しています。この図で青い曲線(左側の曲線)の $b=0.0$ が赤い曲線(右側の曲線)の $b=1.0$ よりも左側にあるということは、同じ能力値だと赤い方が正答しにくく、青い方が正答しやすいということを示しているので、青い方がやさしい項目、つまり $b$ の値が小さい方が困難度は低く、 $b$ の値が大きい方が困難度は高いというふうに、このパラメータを解釈することができます。モデルとしてはこのように説明することができますが、実際のテストは、このパラメータの値を推定しなければならないのです。

このほかに、識別力は全ての項目で等しく困難度のみをパラメータとする「1パラメータ・ロジスティック・モデル」、多枝選択形式の項目における“あて推量”をパラメータ $c_j$ として取り込んだ「3パラメータ・ロジスティック・モデル」などがあります。

3パラメータ・ロジスティック・モデルですが、多枝選択形式の場合にはrandom guessing(あて推量)で正答することがありますので、能力値が低いところでも正答確率は0.0ではなく、もう少し上がるのではないかと推定されます。その部分についてもパラメータを入れているのがこのモデルで、アメリカではよく使われています。なお、ラッシュ・モデルは数理的にはこのようにIRTモデルの中に位置付けることもできるということです。日本での英語教育界ではラッシュ・モデルのほうがよく知られているのではないかと思います。

## 2-7. 実際のテストにIRTを適用するには

では実際のテストにIRTを適用するにはどうしたらよいかということですが、先ほどパラメタaは識別力、パラメタbは困難度を表わすと言いましたが、これらを推定しておく必要があります。これらをどうやって推定するのかという話をすると、半期分の講義ができるくらい延々と続く内容になってしまうので、ここでは「推定しておく必要がある」としておきます。受験者の測定結果は、受験者が各項目に対して正答したか誤答したかをまとめて、正誤パターンで表わします。0, 1, 2, 3の4段階のようなGraded Responseの場合はどうするかというと、その4段階の項目応答パターンを使いますが、ややこしくなりますので、今は正答か誤答かの2値型で考えます。正答したら1、誤答したら0といったパターンで表わします。そのパターンとすでに推定されている項目のパラメタ値を用いて、能力値を推定します。つまり項目については、識別力・困難度が推定されたモデルが与えられおり、それらのパラメタ値と、受験者がそれぞれの項目に解答したパターン(1.0.1・・・)を使って受験者の能力値、つまり同一尺度上の位置を推定するということです。

## 2-8. IRTの特徴のまとめ

いろいろとお話ししましたが、IRTの特徴をまとめると1)～6)のようになります。

- 1) 項目の困難度が受験者集団とは独立に特性尺度上の一点として定義される。
- 2) 受験者の特性(能力)尺度値が解答した項目群とは独立に特性尺度上の一点として定義される。
- 3) 項目の困難度と受験者の特性(能力)尺度値とが同一の尺度上に位置づけて表わされる。
- 4) 項目の特性は特性尺度値と正答確率の関係を表す「項目特性曲線」で全て記述される。

2)に「解答した項目群とは独立に特性尺度上の一点として定義される」とあります。これはCAT(Computerized Adaptive Testing コンピュータ適応型テスト)があるので可能になるのです。コンピュータに向かってテストを受ける場合、自分が解いている問題と隣の人が解いている問題が異なっても、つまり違う項目を解答しても最終的には、同じ尺度上で能力値が推定できます。ここがIRTの“光の部分”だと思います。ただ、パラメタ値を推定しておくなど、いろいろと大変なことがあります。精緻なモデルほど、それを実際に使うのは大変です。そのモデルが適合するような条件が必要です。テストそのものも、その条件を満たすように作っておかなければなりません。これは実は次につながります。

- 5) ある項目に正答するか誤答するかは、他の項目に正答したか誤答したかの影響をうけず、相互に独立である。(局所独立の仮定)

能力が高ければたくさんの項目に正答することができ、能力が低ければたくさんの項目に正答できないということは誰もが分かることですが、そうではなくてある能力値、つまり「横軸の値を固定したときに、複数の項目に対する解答は他の項目に“正答した”か“誤答した”かの影響を受けない」という、この仮定が成り立っている必要があります。これについては、成り立たない状況を考えたほうが分かりやすいです。例えば大問1の小問(1)に対して解答します。その結果を使って(2)に解答します。そうすると(1)に正答できていないと(2)には正答できないのです。(1)に誤答したら(2)はアウトです。(1)に正答していれば(2)に正答できる。(1)に誤答したら(2)は誤答するしかない。(1)の結果によって(2)が決まる、つまり影響を受けているということになります。そういう場合は使えません。

つまり、項目をそうではない作りにはしておかなければならないということです。先ほどのように、(1)の結果が(2)に影響するような作りはだめだということです。例えば英語の問題で、1つのパラグラフを読んで複数の問題に解答する場合がよく問題になるのですが、よほど関係が深い事柄について問わない限りは、同じパラグラフ中に2つ設問があっても、その答えが影響しあうということはありません(局所独立の仮定)。

今、申し上げたようにIRTはモデルとしては精緻な条件があります。今は横軸一本の1次元の話ですが、多次元IRTモデルもあります。いろいろな仮定を満たすかどうかという時に、厳密に満しているのかというと、そうともいえない状況が出てきます。つまり現実のテストに適用する時、仮定を厳密に満たさないと使わないとするのがよいのか、それとも数理的な厳密さから若干外れるところもあるかもしれないが、実用レベルでは問題ない場合に“実用的に大丈夫”と考えて使ってよいのかどうかという判断、これが非常に大事です。もちろん厳密な理論展開がモデルでは重要です。しかし一方において適用する時には、例えば、1つのパラグラフを理解しているかどうかで両方に影響するだろうということがあっても、(1)の結果が(2)に影響することがないなら、“実用的に大丈夫”と判断してよいのではないのでしょうか。もちろん、そういったことを確かめる必要はあると思います。確認をしてエビデンス(証拠・根拠)を付ける必要はあると思いますが、現実のテストでは厳密なだけが全てではないということです。

6) 測定精度が特性(能力)尺度値の関数として表され、尺度値毎にきめの細かい測定精度の評価が可能になる。

古典的テスト理論では、テストの精度を表すのに「信頼性係数」を用います。信頼性係数は1つのテストに対して1つです。それに対してIRTの場合はそうではなくて、測定精度を尺度値の関数として表わします。〈図6参照〉尺度値とは、図の横軸の「潜在特性尺度値」で、この尺度値の関数です。ここには出しませんでした。テスト情報曲線を描けると、潜在特性尺度上でどの部分の人たちに対して精度の良い測定になっているか、どこの部分の人たちに対して精度の良い測定になっていないかということを明らかにすることができます。

5)、6)もIRTの良いところです。テスト全体ではなく、受験者一人ひとりについて、どのくらいの精度で測定されているということがきちんと分かります。ただ“局所独立の仮定”というものがあり、それを満たすようにテストが作られていなければなりません。テストに限らず、どんなものでも条件を満たすように作らなければ使いようがありません。

IRTは様々な場面で適用の可能性ががあります。特にテストの“等化(equating)”とテスト項目の“特異項目機能(Differential Item Functioning: DIF)”を検出する場面で有用です。“差異項目機能”ということもあります。

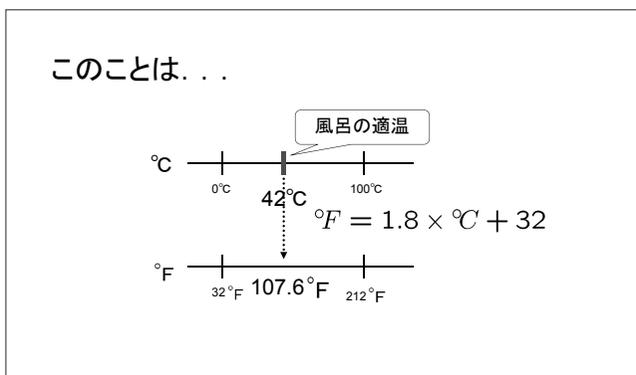
## 2-9. IRT尺度の等化

IRTと等化についてですが、「等化」とは英語ではequatingです。これを実施することにより、複数のテストの測定結果を相互に比較可能にします(テストの等化)。より厳密にいうと、同一の能力特性を測定する目的を持ち、同一の仕様(specification)で開発されているが、異なる問題項目から構成される複数のテストの得点を、相互に比較可能な共通尺度上の得点に変換して表わす操作のことを言います。例えば、TOEFL iBTは同じ内容のテストを全員が解くのではなく、

隣の人と異なる問題を1セット解きます。問題のセットが異なるので、問題項目も異なります。これは、異なる問題項目から構成される複数のテストの得点を相互に比較可能な共通尺度上の得点に変換する、つまり等化が行われているから可能なのです。

IRT尺度では、項目パラメタ値や特性尺度値を表現する目盛の原点と単位とは、線形変換の範囲内で自由に決定することができます。“線形変換”と難しい言い方をしていますが、何も難しいことではありません。「 $y=ax+b$ 」というような一次関数で表わされ、異なる原点と単位を持つ複数のIRT尺度は、適切な線形変換を行うことによって、すべて共通の原点と単位を持つ共通尺度に合わせることができます。これを「IRT尺度の等化」といいます。

<図7>

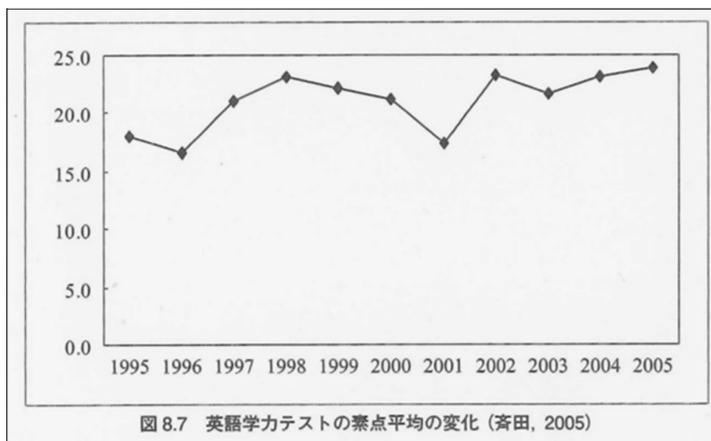


IRTでは能力を表すのに、よく $\theta$ を使います。すなわち「 $\theta^* = k \times \theta + 1$ 」という線形変換を行っても、対応する尺度 $\theta^*$ 上で正答確率は不変であることが示されます( $k \neq 0$ )。これについては「お風呂のお湯の温度」を例に挙げます。<図7参照>

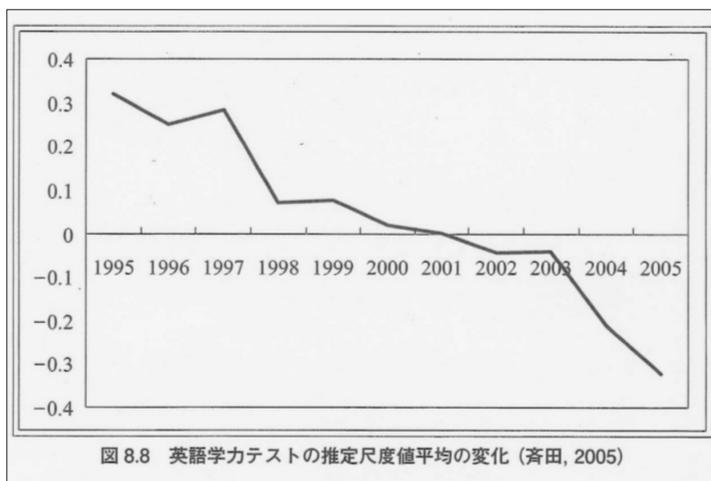
お風呂の適温が42°Cとします。これは摂氏で42°Cです。それを華氏にすると107.6°Fです。その変換は「 $^{\circ}\text{F} = 1.8 \times ^{\circ}\text{C} + 32$ 」という式で表されます。この式により“42”というスコアを“107.6”というスコアに変換することができます。また、例えば日本で測定した結果が“°C”で表わされ、米国で測定した結果が“°F”で表わされている場合、2つのデータを直接比較することが出来ません。比較するために何とか同じ尺度上に乗せたいという場合は、この式で変換すればよいのです。“42”と“107.6”は数値(スコア)は異なりますが、実際のお風呂のお湯の熱さは同じです。つまり、“能力は変わらない”ということです。表わす尺度が異なっている時、IRT尺度の場合には線形変換で変換することによって相互に比較することが可能になるのです。ただし、この場合は温度の変換なので、係数の“1.8”や“32”はすでに分かっている数字ですが、具体的なテスト、特に新しく開発したテストなどは、この数字を推定しなければなりません。それを「等化係数の推定」といいます。この推定は非常に技術的な問題です。それを推定することができたならば、変換して同じスケール上に乗せることが可能になります。

配布資料には載せていませんが、IRT尺度化・等化について、ある例をお話しします。齊田智里先生の研究で、学力の変化を明らかにする試みがありました。

\*<図8>



\*<図9>



\*<図8、9>

野口裕之・大隅敦子 (2014) . テスティングの基礎理論. 研究社 122頁 図8.7 と 123頁 図8.8より

ある県の高校の入学時の英語力をみるテストの話です。毎年問題が変わりますが、テストの素点で見ると、このように変化しています。<図8参照>実は2002年度からの新学習指導要領の実施に伴い、中学から高校へ入る生徒の英語の学力が落ちてきているという、現場の感覚があったそうです。でも、得点の平均値は下がっていません。もちろん、各年度で問題が異なるので得点を相互に比較することはできません。そこで年度毎の尺度を等化して共通尺度化したそうです。すなわち、各年度で問題項目が異なりますので、年度によって難易度に違いがあると思いますが、そこは考慮せず素点の平均点をプロットした場合にはグラフの折れ線に下落傾向は見られません。これをIRT尺度化し、等化するとこのようなグラフ、すなわち、下落傾向が明らかになっ

たそうです。〈図9参照〉要するに英語の学力が確かに落ちているということを示しました。どうしてこのようなことができるかという、IRT尺度化・等化したグラフは、相互比較可能な共通尺度で表されているからです。素点では相互比較することはできません。

### 3- 1. CBTとは何か

CBTとはコンピュータを利用したテストのことです。CBT (Computer Based Test) のうち、Adaptive (適応型) なものをCAT (Computerized Adaptive Test) といいます。CBTはコンピュータを用いて実施するテスト方式ですので、コンピュータの可能性の拡大に対応して、テストの可能性を拡げています。例えば、紙に印刷された項目だけでなく、画像・音声・動画を使用することは、ある意味では言語テストで大事だと言われている真正性 (authenticity) が向上します。そして実技試験にも用いられています。これは米国の方が進んでいます、日本でも例えば医学など進んでいる分野があるようです。テストの進行に関しては、反応時間 (応答時間) の記録や、次に実施する項目の最適化が可能です。

ただ気をつけなければいけないのは、こういった新しいものがでてくると、一気に飛びついてしまいがちですし、新しいもので実験してみると新規効果と言いますか、それで点数がぐっと上がるということがあります。だからといって否定すべきものでもありません。テストの可能性を拡げるものです。またPBTに完全に取って代わるものでもありません。PBTも必要です。日本はすべてIBTですが、TOEFLの数年前のデータでは、全世界で96%がIBTで実施され、4%はPBTで実施されています。なぜPBTで4%実施されているかということ、インフラが整っていない地域や国の人たちもテストを受けられるように対応を行なっているからです。必要に応じて使い分けなければいけません。

そして、コンピュータを使えば何でも出来るというものではありません。技術面の信頼性をきちんと評価することが大切です。また、テストは媒体も大切ですが、測定する内容が最も大事であることを忘れてはいけません。すなわち“測定の妥当性検証”です。日本ではこれについて、あまり語られることがありません。Computerized Testがだめだということではありません。コンピュータそのものも、どんどん進化していきますので、それに応じて適用範囲も増えていくでしょう。ただPBTをすべて捨てるというのはいかがなものかということです。

### 3- 2. CATについて

CATは、受験者ごとに最適な項目を選択・編集して実施するテスト方式です。すべての受験者に対して一定水準の精度での測定が可能です。しかし、これには項目プールが必要です。たくさん項目を貯めておくことが大切なのです。基本的にCATは、直前に実施した項目に対して正答した場合には次により難しい項目を、誤答した場合にはよりやさしい項目を実施するという手続きを繰り返します。ということは項目プール、アイテムバンク、アイテムプールともいいますが、大きなプールを用意しておかなければなりません。つまり、“項目を作り続ける必要がある”ということです。そして“実際にそれができるのか”ということなのです。できなかつたら意味がありません。例えば毎年、もしくは半年に1回実施するテストで、一度使用した項目を再使用することなく捨ててしまうとすると、それだけの項目を次々と開発する力、つまり十分な人的資源

や資金がないとCATを維持することが難しくなります。

### 3-3. CATの利点

<図10>

**CATの利点**

1. すべての受験者に対して高い精度の測定が実施可能
2. 精度を落とすことなく、受験者一人あたりに実施する項目数を減らすことができる
  - ➡ テスト実施時間を節約できる
3. 難しい項目が続いて、受験者にフラストレーションや不安を起こさせたり、易しい項目が続いて飽きさせたりすることがない
4. 受験者の予定に最大限配慮して実施できる
  - ➡ ただし、膨大な数の項目から構成される項目プールが必要

CATでは受験者ごとに解答する項目やその数が異なりますが、測定結果を同一の尺度上に表現して測定精度を確認できます。それにはIRTの適用が基本的な条件になります。CATの利点は、1番目として“全ての受験者に対して高い精度の測定が実施可能である”ということ、2番目として“精度を落とすことなく受験者1人あたりに実施する項目数を減らすことができる”ということ、つまり受験者の解答に合わせて最適な項目を選んで出題していくので、能力の高い人にやさしい問題を解かせたり、能力の低い人に難しい問題を解かせたりする必要がありません。能力推定値に近い困難度の問題を次々と出題していくので、短い時間で精度の高い測定を実施することが可能です。3番目として“難しい項目が続いて、受験者にフラストレーションや不安を起こさせたり、やさしい項目が続いて飽きさせたりすることがない”ということです。

また、4番目として“受験者の予定に最大限に配慮して実施できる”としました。いつでも受験できるということなのですが、それには膨大な数の項目から構成される項目プールが必要であり、しかもこれには推定された項目パラメタが付いていないといけません。だから利点というものはありますが、表裏なのです。こういった利点を出そうと思ったら、必要とされるものをきちんと整理しておかなければならないのです。光と影というわけではありませんが、良い点があればそれを支えるために努力が必要です。そう簡単にできるものではないということです。

### 3-4. CBT化にあたって

High stakes test の場合、すなわち、そのテストの結果が個人の処遇に大きな影響を与えるようなテストの場合には、信頼性の高い技術が要求されますので、CBTを実施する施設の整備が十分である必要があります。以前、TOEFLを受験した人から聞いた話です。たまたまキーボードがうまく反応しないものにあたってしまったらしいのですが、そのことを申し出なかったそうです。その理由を尋ねると、「言い出しにくかった」ということです。もちろん申し出れば無料でもう一回受験できると思います。しかしこれはお金の問題ではありません。大切なのは、何のトラブルも無く、気持ちよく受験できる施設であるか、つまり信頼性の高い技術によってきちん

と整備された施設および環境を整えているかということなのです。

IRTが必須というわけではないですが、適用可能なテスト仕様になっているかどうかも大切です。それから膨大な数の項目が蓄えられた項目プールの整備が可能か、テストの実施状況を管理する人材は確保できるのか、そして、これはとても大事なことなのですが、特別な配慮を必要とする受験者に対応できるのかということも大切です。

それから、技術の進歩に対応して、テスト開発をする人的資源、時間、資金は十分用意できるのかということなのです。これを抜きにして「あれが出来ます。これが出来ます。こうすればいいんです」といっても、出来るはずはありません。どこかで必ず破綻します。CBTが悪いといっている訳ではありません。CBT化にあたっては、それなりに手当てをしなければならないということをお願いしたいのです。

#### 4- 1. 英語4技能外部試験をどう位置付けるか

英語4技能外部試験をどう位置付けるかについてですが、当初はIRT、CBTという3文字が結構躍っていましたが、最近はトーンダウンしています。しかし、英語の試験に関する改革は変わらず大きな議論を巻き起こしています。外部試験のテスト開発機関の考え方で微妙に揺れますが、現在名前が挙がっている外部試験はIRTベースかつCBTのものが多いです。そういう意味では、IRTとCBTが活用されているといえます。しかしながら、“共通テスト”と大学が独自に実施する“個別試験”との総合的な議論の影がどうも薄くなっている気がして仕方がありません。“共通テスト”の議論は活発にやっていますが、“個別試験”をどのようにしていくのかはあまり議論されていません。これは双方を総合的に考えていかないと、絶対いけないはずですが。例えば4技能試験を個別試験で行うこともありだと思っています。

教科「外国語」の科目「英語」に関して気をつけていただきたいのですが、「英語、英語」といいますが、教科は「外国語」です。科目が「英語」です。フランス語、ドイツ語、中国語等も今のセンター試験、それから各大学の試験でも行なっているところもあります。そのことは横に置いて、「英語4技能」が強調されています。

2020年度から2023年度までの4年間は、共通テストとTOEFLやCambridge英検などの大規模英語試験の中から大学入試センターが認定する“認定試験”を併存させることになりました。しかし、複数の認定された英語試験を各大学はどのように他の教科・科目と合わせて総合的に入学者選抜に用いればよいのかについては、何も触れられていません。では各大学が自由に考えていいのでしょうか。これはありがたいことです。他から色々といわれるよりは各大学が独自に考えて、「ウチの大学にとってはこうするのがよい」と自由に設定できるのは良いことだと思います。しかし実際はどうなるのでしょうか。

#### 4- 2. CEFRによる「認定試験」の対応付け

大規模英語試験は、その目的も異なれば、テストの仕様も異なっています。例えば、IELTSとTOEFLは両方とも大学入学のためのテストです。TOEFLは主として北米地域、IELTSはイギリスやオーストラリア地域で使用されていますが、この2つの大規模英語試験は、特に「Speaking」の仕様が異なります。TOEFLは「Speaking」を含め、その他のすべてをCBTで行っています。一方、

IELTSはそうではありません。「Speaking」に関しては、人が行っています。これはどういうことかという、CBTのようにコンピュータが相手だと、Interaction（相互作用）の能力がどれだけなのかを測ることができません。以下の例はもしかすると適切ではないかもしれませんが、もし人が相手だったら、「今、おっしゃったことの意味が分かりません」と質問したりするやり取り、つまり何か少しでもInteractionをすることができます。その能力を持っているかどうかを査定したいなら、CBTではなくIELTSのように人が行なうほうがよいのです。そこがCBTにはありません。

<図11>

・ それにもかかわらず、大学入学共通テストでは英語以外の外国語科目をどのようにするのかについての議論が見えて来ない。

表1 各種外国語能力試験とCEFRレベルの関係

CEFR	Cambridge English	IELTS	TOEFL	英検	TOEIC/ TOEIC&W	GTEC CBT	DELTA/ DALF	漢語水平考試
C2	CPE	8.5-9.0					DALF C2	筆記6級
C1	CAE	7.0-8.0	95-120	1級	1305-1390	1400-	DALF C1	筆記5級
B2	FCE	5.5-6.5	72-94	準1級	1095-1300	1250-1399	DELTA B2	筆記4級
B1	PET	4.0-5.0	42-71	2級	790-1090	1000-1249	DELTA B1	筆記3級
A2	KET	3.0		準2級	385-785	700-999	DELTA A2	筆記2級
A1		2.0		3級-5級	200-380	-699	DELTA A1	筆記1級

大規模テストの測定結果をCEFRの能力レベルに対応付けた表が文部科学省から資料として公開されています。こちら<図11>は文部科学省が公開した表そのものではありません。この表の中のA 1、A 2はいわゆる初心者(Basic User)です。B 1、B 2は独立した言語使用者(Independent User)です。C 1、C 2はProficient User、つまり熟達者です。ケンブリッジ大学、オックスフォード大学はC 1以上のレベルでないと入学することはできません。CEFRはCommon European Framework of Reference for Languages（欧州言語共通参照枠組）の略称です。欧州域内のテスト開発機構などを含む外国語教育専門家が、言語学習や教授法、評価法に関して、国や言語の違いを超えて、相互理解およびコミュニケーションを促進するための共通基盤となる“言語参照枠組み”を提示した260頁に及ぶ文書で、2001年に欧州評議会（Council of Europe: COE）によって出版されました。“共通参照枠組み”であって規準（基準）ではありません。

この表をみるとTOEFL IBTはB 1から上はありますが、A 1、A 2は空白です。測定範囲に入っていない。あるいはCambridge Englishでは、A 1が測定範囲に入っていない。これについては後でお話します。

実はCEFRは異なるテスト間の測定結果を対応付けるものではありません。表を横に見ると、テスト間が対応付けられているように見えるのですが、そうではなく、“テストとCEFRのレベルとが対応付けられている”という、それだけの話です。したがって、フランス語、ドイツ語、中国語などの大規模試験も1つの表に入れることができます。IELTSとTOEFLはETSが換算表を作成し公表していますが、その他はテスト間では得点を換算することはできませんが、CEFR

では「対応付け」は可能です。英語に限らず他の言語も共通に、学習者のその時点での言語能力を6段階のレベルで表わすものであるので、フランス語、ドイツ語、中国語などの大規模試験も1つの表に入れることが可能なのです。なぜそのようなことができるのかというと、これが大事なことなのですが、CEFRは個別言語で設定される能力基準ではないからです。スタンダードではないのです。言語能力のレベルのイメージが共有できるように、そのレベルの学習者がその言語を使ってどのようなことができるかを、Can-do statements、つまり言語行動の能力記述文で表した“参照枠組み”です。フワッとしたレベル・イメージなのです。決してテスト間の得点を対応付けるものではありません。つまり、各言語テストの測定結果をCEFRの6段階のレベルに対応させることはできますが、言語テスト間の得点の対応関係を保証するものではないのです。

そうすると、大学入学共通テストでは、認定試験の結果は得点ではなくCEFRのレベルで表わすしかありません。しかしながら、換算表を作成しているIELTSとTOEFLは、等化よりも少しゆるいもので対応付けているのだと思いますが、CEFRでは異なる英語のテストによる測定結果を、比較可能な共通尺度上に表わす等化という得点の変換をすることができないのです。言い換えると、得点の換算表ではないのです。

#### 4- 3. 高校3年生の英語能力とCEFRの導入

平成26年度に全国から無作為抽出した国公立高校の3年生約70万人に対し、文部科学省が行った「英語教育改善のための英語力調査<sup>35</sup>」の結果についてお話しします。もちろん、これが大学入学共通テストの受験者層と完全に一致するわけではありませんが、CEFRでは「Reading」がB 2-0.2%、B 1-2.0%、A 2-25.1%、A 1-72.7%、A 1とA 2を合計すると97.8%です。「Listening」はB 2-0.0%（5人）、B 1-0.7%、A 2-12.8%、A 1-86.5%であり、A 1とA 2を合計すると99.3%です。つまり、TOEFLではA 1、A 2レベルは測定範囲外ですから、高校3年生の英語力は、ほとんどがTOEFLの測定範囲外のレベルにあります。なお、「Speaking」「Writing」はそれよりも低い水準にとどまっていることが明らかになっています。このような結果が出ているのに、英語4技能を評価するために外部試験を本当に使うのかということなのです。このあたりのレベルの高校生が大学進学を希望しない可能性があるといっても、確かに数値は少し変わってくるでしょうが、この英語力調査結果の数値を見るとそれは成り立たないように思います。大学入学共通テストの受験者層と同質というわけではありませんが、大部分の受験生がA 1、A 2レベルにあることは確かであり、実力差が見えず、CEFRをそのまま導入しても、入学試験で可否を決定する情報にはなり得ないのではないのでしょうか。

#### 5. まとめ

まず、大学入学共通テストに関して、「IRTベースの試験の導入」「CBTによる全国一斉実施」という議論は現段階では実現せず、「外部4技能英語試験の導入」が最も現実味を帯びているということは事実だと思います。またIRTベースの試験を開発するには、IRTが仮定する条件を満たす仕様・スペックをしっかりと組み立てることが大切であり、本当にIRTベースにすることがで

<sup>35</sup> [http://www.mext.go.jp/component/a\\_menu/education/detail/\\_icsFiles/afieldfile/2015/07/03/1358071\\_01.pdf](http://www.mext.go.jp/component/a_menu/education/detail/_icsFiles/afieldfile/2015/07/03/1358071_01.pdf)

きるかどうか、各教科・科目について基礎的な研究を進める必要があります。英語の4技能は比較的やりやすいほうなので、実際にすでに行われているテストがありますが、教科によってはかなり厳しいものもあると思います。

それからCBTに関しては、テクノロジーの進歩を大規模試験に取り入れるという姿勢、柔軟さは必要ですが、全国規模の共通テストに導入する前に、大学・学部・学科など、小さな単位での実施を繰り返して問題点を洗い出し、それを解決しておく必要があります。というのは、何でも“初期故障”があります。その故障を1つずつ潰していく必要があるのです。例えば、新幹線300系のぞみ号は本格走行をする前に、乗客を乗せないで一年間も東京—新大阪間を往復しました。その間に様々なことをチェックし、あらゆる問題点を洗い出し、それを解決していきました。そういった“慎重さ”が大事なのです。時間に追われ「いつまでに何とかしろ」といって、勢いで進めていません。

大学入学共通テストは、英語以外の教科・科目も合わせて、全体として統一的な理念・構成・仕様のもとで開発・実施されることが望ましいのではないのでしょうか。そして、英語の4技能を測定することが必要であるならば、民間の知恵や技術も導入しながら大学入試センターで新しい英語テストを開発することも考えていいのではないのでしょうか。人、つまり専門家と時間と予算をつけて、きちんと開発するということが必要です。TOEFLにしる、ケンブリッジ英検にしる、それらは日本の学習指導要領のことを考えて開発していません。あるいは日本に限らず、ある特定の国の教育課程に合わせて導入されることを考えてはいないのです。自分たちが考える英語能力、そして英語能力の発達ということをもとに作っています。それを、大きな大学入学共通テストの枠組みの中に組み込む方がいいのか悪いのかについては、いろんな判断があると思いますが、やはり統一した理念の枠組みの中で実施されるのが望ましいのではないかと思います。外部試験は意味が無いということではありません。共通試験と個別入試とを総合的に検討することが必要であるということで、例えば個別入試で4技能試験をしっかりと導入してくださいということです。

どの試験を活用するかについては、アカデミック・イングリッシュに重きをおくならTOEFLやIELTS、そうでなく、あるところにポイントを押さえ、CEFRのレベルでそれ以上かどうかを見るならば、ケンブリッジ英検というものもあります。TOEICは国際ビジネスコミュニケーション協会のテストです。日本で原案を作ったテストで、どちらかというビジネスを念頭においた内容です。それがいいという考え方もありますし、アカデミック・イングリッシュがいいという考え方もあります。それからSpeakingではInteractionの能力を測るべきだというならばそれなりの対応が必要であり、そうではなくコンピュータで基本的な発話能力をみることができればよいという考え方もあります。このようにいろんな考え方があります。各大学がそれに応じたものを選び、4技能を測ることでよいのではないかと思います。

私の個人的な感想ですが、共通一次試験導入の時よりも性急な印象があります。共通一次試験導入の際には、一次試験と二次試験の役割について、しっかり議論されていたように思います。外部英語試験の導入は“大学入学共通テスト”ではなく、各大学が教育理念や大学・学部・学科の特徴に応じて“個別試験”の中で適切なものを選択して利用することが望ましいと思います。例えば、外国語大学や外国語学部は、それ以外の学部とは違った考え方があっていいでしょう。それを反映

したテストを行なうことが相応しいということです。

高大接続改革の中で大学入学試験をどうするかは大きな問題ではありますが、もはや全ての大学について一律に議論できる状況ではありません。希望者全員が入学できる大学もありますし、留学生を海外に集めに行く大学もあるのです。こういった現実を踏まえないで、大学がみんな同じように大学教育について議論しても、テストをうまく使いこなすことはできません。実態に見合ったところで議論しなければならないのです。大学の個性に応じて利用しやすい“共通テスト”であっていいのではないのでしょうか。

最後にもう一度繰り返しますが、何よりもテストの研究開発には、“人的資源”“時間”“資金”が必要であることを肝に銘じておかなければなりません。これは絶対にそうなのです。これらをケチって研究開発しても、良いものは出来ないのです。しかしその中で、テスト開発に携わる研究者や実務家は皆頑張っているのです。