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

平成 27 年（2015 年）3 月

独立行政法人大学入試センター入学者選抜研究に関する調査室

独立行政法人大学入試センター
入学者選抜研究に関する調査室報告書２

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

目　　次

独立行政法人大学入試センター入学者選抜研究に関する調査室主催
後援：文部科学省・日本テスト学会

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

プログラム

◎開催日時　：　2014 年 11 月 29 日（日）　開場 12 時半・開始 13 時

◎会　　場　：　東京工業大学大学院社会理工学研究科デジタル多目的ホール

◇開会挨拶　山本廣基　大学入試センター理事長

◇趣旨説明　荒井克弘　大学入試センター試験・研究統括官
入学者選抜研究に関する調査室室長

◇基調講演　演題「PISA：抜本的改革への挑戦 ―何をどう測定し　どう評価していくのか―　Major innovations in PISA 2015 and beyond」
山本兼太朗　Deputy Director of the Center of Global Assessment, Principal Research Scientist of ETS（Educational Testing Service, USA.）

◇報告１　演題「試験の日本的風土」
前川眞一　東京工業大学大学院社会理工学研究科教授

◇報告２　演題「入試選抜の測定問題」
南風原朝和　東京大学大学院教育学研究科長

◇報告３　演題「センター試験で何が測られるのか？」
大津起夫　大学入試センター研究開発部長

◇報告４　演題「高大接続のこれから」
村上　隆　中京大学現代社会学部長

◇ CBT サンプルのデモ
大久保智哉　大学入試センター研究開発部助教

◇パネルディスカッション　　コンピュータ・教育測定・大学入試
Kentaro Yamamoto 氏　・　前川眞一　氏　・　南風原朝和　氏　・
大津起夫 氏　・　村上　隆 氏
総合司会　大塚雄作　大学入試センター試験・研究副統括官
入学者選抜研究に関する調査室室長補佐

■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』概要

大学入試の抜本的な変革をめざして、教育再生実行会議、中央教育審議会高大接続特別部会等を中心に議論が積み重ねられてきています。そこでは、複数回受験、知識活用力等を重視した多面的評価、合教科・科目型試験の開発、100 点満点の素点に代わる段階別評定の利用、コンピュータ利用（CBT：Computer Based Testing）等々の新たな入試の提案がなされてきています。いずれも重要な提案、論点ですが、そこには測定理論的に多くの課題が付随しています。また、とりわけ、わが国においては、それらの改変が必ずしも円滑に進まないと思われるテスト文化が根強く潜んでいます。そこで、教育測定の専門的視点から、大学入試のよりよい改革のためにどのような問題が具体的に残されているかを整理すると共に、それらの克服のためにどのような対処が必要とされるのかについて、このシンポジウムを通じて浮き彫りにしていきます。

■ 講演者プロフィール

基調講演：山本兼太朗　（Kentaro Yamamoto）

Deputy Director of the Center of Global Assessment, Principal Research Scientist of ETS（Educational Testing Service）。1987 年イリノイ大学大学院教育心理学研究科で Ph.D.取得。同年、ETS の心理測定学・統計学部門に研究員として赴任、2004 年より現職。2006 年より、OECD の PISA、PIAAC 等にその専門的立場から関わり、2015 年、2018 年の PISA の心理測定学分野からの担当責任者を務めている。

報告１：前川眞一　（まえかわ・しんいち）

東京工業大学大学院社会理工学研究科教授。1984 年東京大学大学院人文科学研究科 心理学専門課程博士課程単位取得満期退学、アイオワ大学博士課程修了、 Ph.D. 1985 年アイオワ大学統計学科客員助教授 1986 年より大学入試センター研究開発部の助教授、教授を経て、2001 年より現職。専門は統計科学、テスト理論。日本テスト学会理事。

報告２：南風原朝和　（はえばら・ともかず）

東京大学大学院教育学研究科教授。2013 年度より教育学研究科長・教育学部長。アイオワ大学大学院博士課程修了、Ph.D. 1982 年より新潟大学教育学部講師、助教授、1993 年東京大学教育学部助教授、2002 年より同教授。専門は心理統計学、心理測定学。日本テスト学会副理事長。

報告３：大津起夫　（おおつ・たつお）

大学入試センター研究開発部長。1983 年北海道大学文学研究科修士課程（心理学専攻）修了。民間企業勤務、北海道大学文学研究科教授をへて 2003 年度より大学入試センター研究開発部教授。2014 年度より現職。専門は心理統計学。

報告４：村上　隆　（むらかみ・たかし）

中京大学現代社会学部長。東京教育大学大学院教育学研究科博士課程中退。博士（心理学・筑波大学）。名古屋大学教育学部教育心理学科助手・同助教授・同大学院教育発達科学研究科教授を経て、2007 年度より中京大学教授。2014 年度より現職。その間、2002 ～ 06 年度全国大学入学者選抜研究連絡協議会会長、日本留学試験実施委員などを歴任。専門は計量心理学。

開会挨拶

独立行政法人大学入試センター理事長
山本廣基

○司会（大塚）　大変お待たせいたしました。ただいまより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」を開会させていただきます。本日はあいにく雨模様になりまして、お足元の悪い中、多くの皆様方にお集まりいただきまして本当にありがとうございます。5時半までの長丁場でありますけれども、どうぞよろしくお付き合いのほどお願いいたします。私は、本日進行を担当いたします大学入試センター試験・研究副統括官の大塚雄作と申します。未熟ではありますが、進行を務めさせていただきますので、最後まで何とぞよろしくご協力のほどお願いいたします。

それでは早速、大学入試センター理事長、山本廣基よりご挨拶させていただきます。

○山本理事長　本日は、師走を目前にした何かとご多忙のなか、土曜日の午後に全国より多くの方々が、大学入試センターシンポジウムにご参加下さいましたことを、心より御礼申し上げます。

先週、11月20日の中央教育審議会の総会におきまして、「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」という長い標題の答申案が承認され、来月下旬、文部科学大臣に答申される予定となっています。皆さん方は十分ご存じのこととは思いますが、平成24年8月に文部科学大臣から「大学入学者選抜の改善をはじめとする高等学校教育と大学教育の円滑な接続と連携の強化のための方策について」が諮問され、総会直属の高大接続特別部会が設置されました。この間、昨年の10月には政府の教育再生実行会議の第4次提言があり、合計21回の特別部会が開催されて、この答申案に至ったところです。高等学校教育、大学教育のあり方と並んで、大学入学者選抜のあり方を一体的に改革することが求められています。特別部会におきましては、とりわけ入試改革に議論が集中したようで､本日もこうして多くの方々にお集まりいただくなど、入試改革は当事者のみならず、社会的にもきわめて関心の高いことが伺えます。

答申案に盛り込まれている入試改革案は多岐にわたっています。新テストとして、高等学校基礎学力テスト、大学入学希望者学力評価テストという名前の二つの大規模なテストの実施、そして入学者選抜に関わる支援やその調査研究・人材養成などがあげられ、これらを全て、大学入試センターを抜本的に改組した新センターで行うこととされています。それだけに私どもといたしましても、身の引き締まる思いでこの入試改革の動向を見守ってきているところです。

言うまでもなく、入学者選抜は、大学教育への適応力であるとか、それに必要な学力などについて測定し、その測定結果に基づいて選抜が行われるということになりますが、入

試改革を議論するにあたっては、教育測定論の専門的見地からの十分な議論も必要であると考えています。そこで、改めて、本日は、教育測定を専門としておられる先生方にお集まりいただき、今回の入試改革に関わる測定論的課題について取り上げていただくことにいたしました。

とりわけ、基調講演をお願いしています山本兼太朗先生は、アメリカのプリンストンにありますETS（Educational Testing Service）よりはるばるお越しいただきました。現在中心となって担当されておられるPISAの動向について貴重なお話を伺う機会を得ることができますことを心より感謝いたします。PISAも3年ごとの実施となっているということで、いろいろな改革に取り組まれているようでして、それを実際に先導されている先生のお話はとても参考になることと楽しみにしております。

答申に盛り込まれている入試改革の提案の方向性は多数の共感を得る内容ではあろうかと思いますが、専門的見地から測定論的な課題が突きつけられますと、問題点も多々浮き彫りにされてまいります。我が国の入試の在り方についてのより具体的な議論に結び付けていくためにも、そういった課題を共有することが必要です。本日、報告をお願いいたしました前川眞一先生、南風原朝和先生、村上隆先生は、それぞれ測定論の第一線の先生方です。測定理論の最先端の視点からの課題をお伝えいただけるものと思います。特に、答申案ではCBT（Computer-Based Testing）の導入なども求められていますが、それを実際に行うためには、今までとは全く異なる試験に対する見方を理解し、社会全体で共有することも求められます。すなわち、これまで選抜試験に対して私たちが持っている風土を変えていくことが、入試改革の一つの要件になるということを、シンポジウムを通じて皆さま方と共有していければと思います。

答申で謳われていることを実現に向けて進めていくためには、今日のシンポジウムで議論される以外にも多くの困難な課題があろうかと思います。ただ、私たちの叡智をもって、これらを一つひとつ乗り越え、新たな時代を見据えた教育改革、入試改革を進めていかなめればならないと強く考えているところです。

午後いっぱいの長いシンポジウムになりますが、ご報告の先生方共々、実りあるシンポジウムになりますよう、皆さまのご協力をお願いして挨拶といたします。

○司会　山本理事長、どうもありがとうございました。

趣旨説明

独立行政法人大学入試センター試験・研究統括官／
入学者選抜研究に関する調査室室長
荒井克弘

○司会　続きまして、シンポジウムを始めるに当たりまして、大学入試センター試験・研究統括官、荒井克弘より趣旨説明をいたします。

○荒井試験・研究統括官　今、ご紹介いただきました大学入試センターの荒井でございます。

私から趣旨説明をさせていただきます。

手短にということで、4 枚スライドを用意いたしましたが、3 枚目と 4 枚目を中心にお話しさせていただきたいと思います。

まず「大学とは？―教育選抜の意味」と書いてあるスライドでございます。現在の大学入学者選抜の課題とは何か、ここで皆さまと認識を共有しておきたいと考えました。

ご存じのように大学進学者は増えております。大学入学者選抜の実態は、学力エリートの選抜というよりは、これからの知識社会を支える対象層を選抜し、教育し、社会的に配分していくという役割に変わってきました。そう考えております。そういたしますと、高校から大学への進学も、そのプロセスは「選抜」というよりも教育プロセスであり、教育課程の積み上げを意味する。そのために、つなぎ目として、大学入試は可能な限り妥当なそして信頼性の高い教育測定に重点が置かれなければならない。これが私の理解するところでございます。

ここで「大学教育に必要な学力の確保」「教育・学習の支援」「教育診断の信頼性と妥当性の向上」と書きましたが、第 2 象限から第 4 象限に向かって大学入学者選抜の目的は変わってきた、というのが、この図の意図でございます。

4 枚目のスライドですが、大学入試改革といってマスコミや世間が大騒ぎするが、いつも同じことを繰り返し議論しているだけではないか、そのようにお考えになる方もおられます。しかし、大学入試の歴史をレビューしてみますと、一度たりとも入試改革が同じポジションに位置していたことはありません。戦後の 70 年間を見ましても、そのことにいささかの疑いも生じないと確信いたします。

日本では、戦後に 4 つの入学共通テストが入れ替わり導入されてきました。進学適性検査、能研テスト、共通 1 次試験、センター試験です。さらに今後、また新たな共通テストの議論が浮上しております。少なくとも、これまでの 4 つの試験についてはそれぞれの時代の課題に応じて改革努力が進められてきたという経緯があります。

進学適性検査については、敗戦直後の教育混乱期にいかにして大学入学者を選抜するかが課題でした。能研テストは、高度経済成長期のマンパワー計画の一環として構想が練られました。共通 1 次試験は学歴社会の弊害や受験競争の激化の解消をめざし、教育の正常

化に寄与するという課題がありました。それを目的に実施されましたが、結果として、共通１次試験は大学の序列化、志願者の輪切りを顕在化させたという批判を浴びて廃止になりました。センター試験は、沸騰した国民的な入試批判をどのようにかわし鎮静化するか、その課題を解決するために登場したという経緯があります。

それぞれの時代、入試改革をどのように進めていくか、その解決を主導する専門家の方々も大きく変わってきました。例えば、進学適性検査の時代には心理測定の専門家が中心的な役割に担っていました。それから、能研テスト、共通１次試験になりますと、学歴社会、選抜社会、そして教育計画が問題の中心ですから、教育社会学、教育経済学、社会学の専門家が前面に出てくる時代になりました。

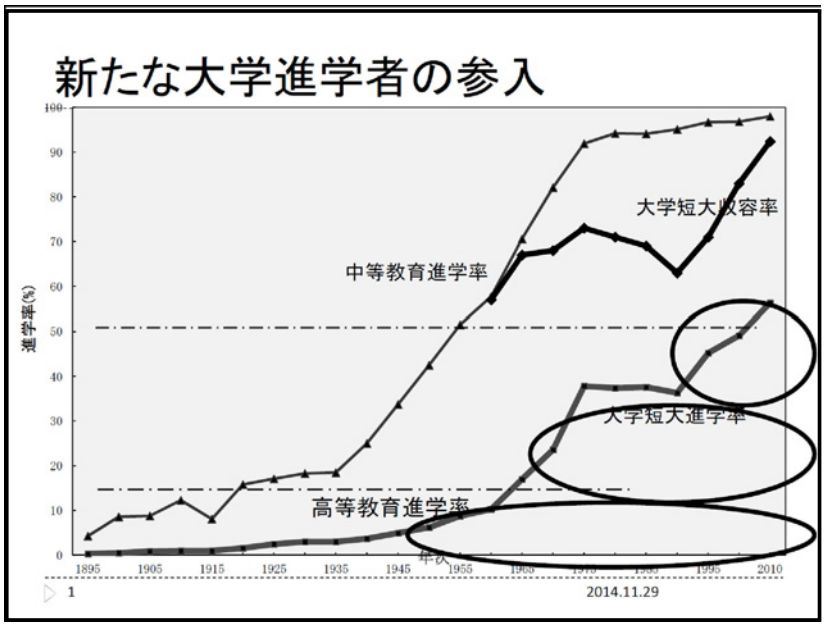
センター試験の次世代のことを考えますと、教育測定の技術論が以前にまして重要になる、と先ほど申し上げました。教育測定の適切さ、的確さです。テスト用語で言えば信頼性、妥当性の高い試験の実現が不可欠だろうと思います。今後は、この教育測定論の専門家の方がたを中心に、さまざまなことを検討し、実施しなければならない時代がくる、と思います。

そこで、先ほど理事長からご紹介がありましたように、今回のシンポジウムでは基調講演をお願いしました山本兼太朗先生を初めとして、各報告をご担当いただくパネリストの方がたは、前川眞一先生、南風原朝和先生、大津起夫先生、村上隆先生の５人ともに、すべて測定論の第一線に立っておられる研究者の方々です。さらにコンピューターテストのデモをやってくれる大久保智哉先生、それから今日の総合司会と進行、パネルディスカッションのコーディネーターを務めてくださる大塚雄作先生も測定論の専門家です。いよいよもう私などが付け入る隙もないほどに、教育測定の専門的な知見をベースにしなければ議論が成り立たない、そういう時代になってきたという実感がいたします。

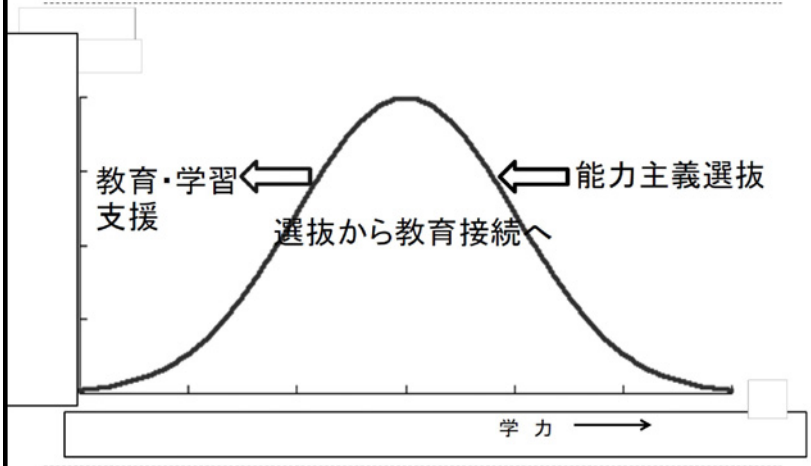
以上が今回、このシンポジウムを企画した趣旨でございます。最後までよろしくお願いいたします。

ありがとうございました。（拍手）

○司会　荒井先生、どうもありがとうございました。

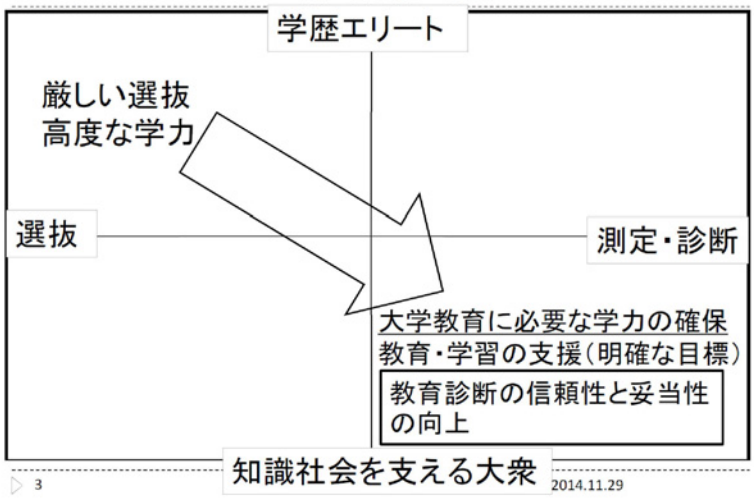


大学進学の普及と入試の変容



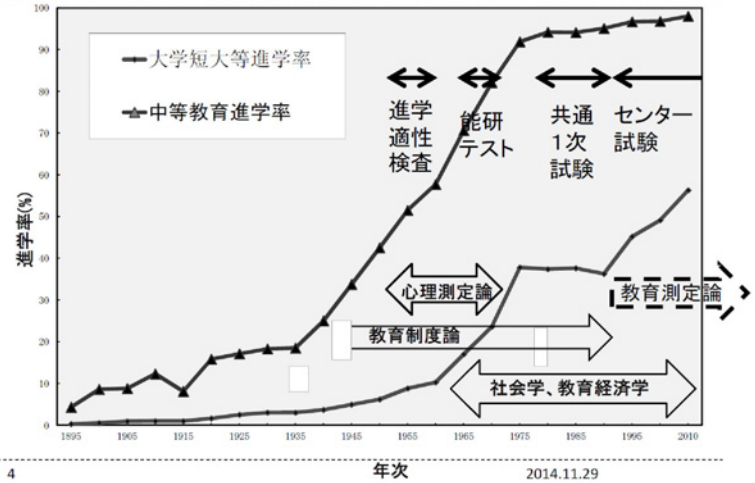
2 2014.11.29

大学とは？—教育選抜の意味



3 2014.11.29

入試改革を主導する専門家たち



4 2014.11.29

【基調講演】

PISA：抜本的改革への挑戦
―何をどう測定しどう評価していくのか―
Major innovations in PISA 2015 and beyond

Deputy Director of the Center of Global Assessment, Principal Research Scientist of ETS
山本兼太朗

○司会　続きまして、アメリカの入試センターとお考えいただければよいかと思いますが、ETS（Educational Testing Service）の研究部門を牽引しておられます山本兼太朗先生に基調講演をお願いいたします。

標題にあります PISA につきましては、特に教育測定の専門的な立場からの責任者として、今月も世界中を飛び回っておられ、そのお忙しい中、充実したスライドをご準備いただきました。山本先生はアメリカから来られるときに、東海岸は雪で飛行機が遅れて一日遅れて昨日夜のご到着になったり、また、便が変更になったことでシカゴのトランジッションの際の手違いで荷物が日本に届かなかったり、いろいろと大変だったようです。そんなことでお疲れのことと思いますが、今朝はお元気に会場に来て下さいました。それでは、山本先生、よろしくお願いいたします。

○山本氏　初めまして、山本兼太朗です。

もう 40 年近くアメリカに住んでいて、日本語で講演することはまずないので、すごく緊張していますよろしくお願いします。今回お話ししようと思っているのは、ラージスケールアセスメント（large-scale assessment）とは何か、なぜ PISA の 2000 年から 2012 年までをもう一度ふり返って分析しなければならなかったのか、それと、PISA2015 のフィールドテストを今年行いましたのでその結果を少し話します。特にこれは、先々週まで 4 週間かけて PISA の会議がありましたが、PISA のフィールドテストに関しては、PISA に直接関与している方以外は誰も知らないことでしたので、今回初めてお話しすることです。それから、PISA2015 のデザインの大まかな概要、それと PISA のコンピューター・ベースド・アセスメント（computer based assessment）のサンプルを見せます。

日本の大学入試に関しては、私はこれまで関わってきておりませんので、PISA の新たな開発過程から入試に示唆を及ぼすものは何なのかと考えてみましたが、まず、私たちがやっている PISA に関しては、何を能力として測ろうとしているのかを明らかにすること、そして、コンピューター・ベースド・アセスメントによって全て測定しようとしているということ、それと同時に、受験者集団（population）のトレンド（trend）をモニターしようとしていること、そのために、2000 年から 2015 年まで同じ尺度を用いてきているということなどの点で、大学入試にも参考になるものがあるのではないかと思います。

ラージスケールアセスメント（large-scale assessment：LSA）とは

ラージスケールアセスメントというのは、大規模な能力調査による評価とお考えいただければと思いますが、それを実施する主な理由は、政策決定のためのデータに基づいた解釈を提供することです。政策研究の適切な役割は政策を定義することではなくて、むしろそれはデータに基づいた、判断が可能となる証拠を確立することです。そして前提にしなければならないことは、ラージスケールアセスメントは個人の能力の測定が目的ではないということです。

したがって、ラージスケールアセスメントというときに、まず最初に考えることは、個人的、また社会一般の成功を支えるコンピテンシーといったものに対応するアウトカムを識別し、計測するということです。この図（スライド 3）のアウトカムには、測定された値が入ることになります。そして、何がコンピテンシーの向上に貢献しているのか、また妨げているのか、コンピテンシーはどのように社会的または経済的な傾向としてあらわれるかを調査することになります。それによって、コンピテンシーをより高度に成立させるべき政策の手段を識別する手助けとなるわけです。

大学入試に関して考えれば、この図のリテラシーのところに、代わりに大学入試で測定しようとしている能力が入りますが、入試の場合はそれだけをアセスメントすることになります。それ以外のバックグラウンドとなる変数についてはほとんどアセスメントすることはありません。それはある意味で当然なことで、個人の能力をアセスメントすることが入試の場合には一番重要なことですから。

しかし、政策に関して考えるときには、どんなバックグラウンドがあるのか、例えばどんな社会経済的地位（SES：socioeconomic status）にあるのか、どんな教育を受けたか、第一言語（Native language）は何か、どんな 障害（disability）をもっているのか、それから年をとることによってどんな影響があるのかといったことが、通常考えられている変数と言えます。それがリテラシーにどのような影響を持ち、そして、その能力によって、社会的、経済的な結果として、例えば、就職にどう結び付くのか、健康とか家族構成、言語学習（Language Learning）、それから、どのように地域の状況に貢献するかといったことが考えられています。ですから、ラージスケールアセスメントでは、能力とそれに関わると考えられるさまざまなバラエティのすべてを計測することが試みられることになります。

PISA とは

PISA は、国際的に一番よく知られている大規模な能力調査です。PISA は 15 歳の学校に通っている生徒を対象にして、数学的リテラシー、読解力、科学的リテラシー及び属性に関しての設問などで構成されています。2000 年から 3 年ごとに 2012 年まで行われており、これからも 3 年ごとに、次は 2015 年、そして 2018 年と続きます。

PISA は、15 歳の持っている知識や技能を実生活のいろいろな場面でどれだけ活用できるかを見るものであって、特定の学校カリキュラムをどれだけ習得しているかを見るものではありません。各国間に共通しているものは、そういう生活に実用されている部分になります。思考プロセスの習得、概念の理解、及び、各分野のさまざまな状況の中でそれらを生かす力を重視しています。社会は今、持っている知識以上に、持っている知識や技能を使って何ができるかを評価している、そういう考え方に基づいて PISA は作られています。

PISA は概ねこのスライド（スライド 6）にありますようなデザインで構成されています。クラスター（cluster）というのは 30 分ほどで回答できる程度の質問のセットです。そのセットは科学リテラシーとか読解力とか、その 1 分野のドメインの中の 1 つのものだけです。クラスター 1、クラスター 2 と 60 分かけて回答し、休憩がありまして、その後また、60 分かけてクラスター 3、クラスター 4 と進みます。そして、バックグラウンド、属性の質問（Background Question：BQ）がありまして、その後休んで、他のドメインのアセスメントが行われます。他のドメインとは、例えばファイナンシャルリテラシー（financial literacy）といった特別なリテラシーです。

2015 年の PISA が 2000 年から 2012 年までと比べて随分変わるという点は、全面的にコンピュータ化するということです。それから、信頼性のある尺度を使って前回のサイクルからの差を測るということです。トレンド（trend）の測定ですね。また、新しい科学リテラシーが定義され、それに加えて新しいテスト分野が利用されることになります。そのなかで、一番大切な点は、安定した経年推移（stable trend）を見ていくということです。今まで PISA で行ってきた 2000 年から 2012 年までのデザインとアナリシスの方法では、トレンドについては誤差（error）が混入してきているということがあるからです。

過去の PISA の課題

まず、このグラフを見てみます（スライド 9）。これは、2000 年から 2006 年にかけて読解力のスコアがどのくらい変わったかということを国別に示したものですが、国によっては 30 点以上変化していることがわかります。PISA の点数は平均が 500 点で標準偏差が 100 点という単位ですから、6 年間で 30 点変わるというのは、ものすごく大きな変わりようということが言えます。

このグラフ（スライド 10）は、もう発表されている結果ですけれども、2000 年と 2009 年の読解力のトレンドを見てみたものです。スレッシュホールド（threshold）というのは、各項目に正答する条件確率が 0.62 となる PISA の点数に当たるものです。このグラフは、2000 年と 2009 年に共通した項目についてスレッシュホールドをプロットしたものです。2000 年から 2009 年の尺度を決めるにはこれらのアイテムだけが関与していることになります。このスレッシュホールドがかなり変化している項目がありますが、これはかなり大きいことなんです。なぜ変わったかということについては、後ほど触れたいと思います。

次のグラフを見ていただきます。習熟度のレベルというのがありますが、それは 407 点、480 点、553 点、626 点で区切られていて、それ以上、以下ということで、習熟度のレベルは、1 以下、1、2、3 とか 4 といった具合に決められています。スレッシュホールドの値によってアイテムのセットを作って、そのアイテムのセットを使って習熟度のレベルがどういうものかのディスクリプションを作っています。それでこのグラフから窺えることは、25 %のアイテムが、2000 年から 2009 年にかけて、習熟度のレベルが変わっているということですね。それでもなお、2000 年から 2009 年にかけて、習熟度のレベルのディスクリプションは変わっていなかったんです。ですけれども、レベルのディスクリプション自体はそれを使って解釈しているわけですから、そのアイテムのセットは大切なものです。

このような変化の理由は、2000 年から 2012 年までは、等化（equating）のモデルを使っているわけですが、サイクルごとに全く別のスケーリングをしまして、それでイクエイティングはそのアイテムのパラメータのディストリビューションをマッチして作っていま

す。そのために、アイテムの値が 2000 年から 2009 年にかけて変わることになります。等化の方法としては、基本的には、受検者の能力の分布に基づくか、それともアイテムパラメータに基づくか、どちらかに基づいて行われることになりますが、特にラッシュモデルを使うときには、アイテムパラメータを使う方法が比較的容易と言うこともあり、それを使ってやっていました。そういうことでアイテムパラメータの値が変わっているんですね。

それをもうちょっと別な方法で見てみます。

これ（スライド 12）はドメインごとに、例えば数学（Math）とか科学（Science）とか読解力（Reading）に関して、アイテムの数がどのくらい変遷しているかを見たものです。PISA では、ドメインがメインとマイナーに分けられていて、2000 年から 2012 年、グリーンのバックグラウンドが入っているのはメインのアセスメントを示したものです。2000 年は読解力がメインで、数学と科学はマイナーのドメインでした。ですから、読解力は 129 問ありました。次の 2003 年にはマイナーに回って、そのうちの 28 問が残り、2006 年も 28 問で、それらをもとにしてトレンドが検討されることになります。そして 2009 年には 131 問です。2000 年から 2009 年までで、28 問以上、30 問ぐらい残ったと思うんですけれども、それに加えて新しいアイテムが入れられて 131 問になっています。2012 年にはその 131 問のうちの 44 問が残って、それがマイナードメインとして使われています。

そうしますと、例えば 129 問から 28 問に移ったとき、すべての読解力の中のアイテム、サブスケール（例えば、テキストの環境、形式、タイプ、側面、用途、状況）が必ずしも同一のスケールではありませんから、それが多次元という形で少し入っているわけですね。そのアイテムのセレクションのされ方によって、国によってはすごくインパクトが大きくなるわけです。ということで、例えば偶然誤差（random error）だけではなくて、体系的誤差（systematic error）がトレンドの中に入ってきてしまうことになるわけです。

これ（スライド 13）は 2015 年のフィールドテストに使ったアイテムの数です。83 問、98 問、85 問というのは、数学と読解力と科学のトレンドを見るためのアイテム数ですけれども、新しい科学の問題数は 213 問でした。これをグラフで見てみます。このくらい変わっているんですね。3 分の 1 から 4 分の 1 ぐらいのアイテムしかマイナードメインのときには残っていないわけです。

そういうことで、トレンドを検討することが、アセスメントの重要な趣旨にもかかわらず、体系的な誤差（systematic error）が PISA の中には入っていると思います。それをまず変えるということを 2015 年のプロポーザルに私たちは書きました。

このような考え方なんですけれども、以前は、例えば 2006 年から 2012 年にかけて、グラフのバーの高さは構成概念（construct）のカバレージ、すなわち、測定の対象となる構成概念がドメインにおいてどれだけ網羅されているかということを表しておりまして、アイテムの数に直接関係させて描いてあります。それが 3 分の 1 から 4 分の 1 ぐらいになっていますね。2009 年と 2012 年。これは数学ですけれども。

新しいデザインは、2015 年、例えばメジャードメイン、これは科学ですけれども、2018 年と 2021 年にはほとんどアイテムを全部使って、だけれども項目ごとの被験者の数を減らすわけです。ということは、データの数は同じだけれども被験者の数を減らして、何をしなければならないかというと、2015 年からデータを借りなければならないんですね。つまり、2015 年のデータと 2018 年、2021 年のデータを同時に解析するということです。

アイテムパラメータは、2000 年から 2012 年の PISA、また TIMMS とか PIAAC とか、私たちはリテラシーサーベイ（literacy survey）をやっていたんですけれども、その際の経験からすると、アイテムパラメータというのはすごくステーブルです。

そういうことで、全く別々、以前のデータをまるで使わずに新しいサイクルごとにアナリシスするのは、非常にデータを無駄遣いすることになります。そういうことで、トレンドに関しては、丸きり以前のそのような考え方を変えたわけですね。すなわち、過去のすべてのデータの情報を共有するということです。それによって、すべてのドメインごとに 1 つの尺度に乗せることができ、各項目のスレッシュホールドの値はサイクルごとに変わらなくなるということになります。また、過去のデータを補正することもできますので、長期間にわたってトレンドをモニターすることができることになります。

また、各項目質問と各国の関連性をモニターすることもできます。2012 年まではインターナショナルなアイテムパラメータのセットがありまして、各国間との相関性をまるで無視して同じアイテムパラメータを使って全部のアナリシスをしていたんですね。例えばアイテムパラメータがたまたまその国のデータと合っていなくても、そのままインターナショナルなアイテムパラメータを使ってやっていました。そうしてみますと、数サイクル同じ間違ったパラメータの項目で実施していて、今度、新しいサイクルで改正されたということがデータで見ると読めるわけです。

それと同時に、このデザインを見てわかると思いますが、2015 年のマイナードメインというのは、データの数は足りないわけですね。ですから、2015 年のマイナードメインに関しては、ある国においては項目の数に対して被験者の数が少な過ぎることになりますから、アイテムパラメータを推測することができないということになります。そこでどうするかというと、2000 年から 2012 年までに使ったデータをもとに解析して、新しいパラメータを作って分析を進めなければならないということです。

過去の PISA データの再分析

そういうことで、2000 年から 2012 年までのデータをもう一度分析することにしました。私は古いデータを使うことが頻繁にあるんですけれども、データベースを作るときには、一つ一つの小さな問題を解決していかなければならないと言うことがあって、なかなか大変です。2000 年から 2012 年までの古いデータを扱うことになりますが、アイテム、スコアリングに関しての個々の決定事項は記録として残っていませんので、いろいろと苦労があるわけです。すべて合わせますと 193 万人のデータが残っていますから、それをもとにして前のデータをもう一度持ってきて、今度は全く別に全部網羅しているデータベースを作って分析しました。

これ（スライド 17）は数学の例ですけれども、以前発表された PISA の結果とほぼ同じ結果が、ラッシュモデルを使って再現できています。2000 年の結果は特異（anomaly）であったと見られているのですが、ここでもその特徴が出てきています。このような結果が得られたということは、データベースが正確に再生できたということかと思います。

それをもとにして、スライド 18 にあるように、いろいろなモデルで分析してみました。

ラッシュモデルというのは、もともとモデルがフィットしていないときにはよく見えませんから、2 パラメータモデル（2PL）、それから 2 パラメータモデルとラッシュモデルをハイブリッドしたモデル、また、項目と国とのインタラクション（IBCI：item-by-country

interaction）を考慮したラッシュモデルと 2PL をハイブリッドしたモデルとの適合度を検討してみて、最終的には、IBCI のモデルを採用しました。

ラッシュモデルというのは同じ傾きをもつ項目特性を仮定していて、項目の特性の違いは傾きのパラメータを与える特性曲線の位置の違い、しばしば b と表される難易度パラメータのみということになります。2PL は傾きのパラメータと位置のパラメータの二つのパラメータを含むモデルということになります。

IBCI の例を、スライド 20 に示しました。これは PISA の例ではありませんけれども、たまたまフレンチカナダのデータが随分違って出てきていまして、もし以前のようにインタラクションを見ないで同じパラメータを使っていますと、フレンチカナダの結果はバイアスをもつことになります。

モデルフィットの指標としては、基本的には AIC（赤池情報量規準）と BIC（ベイズ情報量規準）を使いまして、主に BIC を利用しました。それと同時に平均偏差（MD：Mean Deviation）、これは、平均からの偏差と分布の確率密度（density）を掛けて積分したものです。RMSD（Root Mean Squared deviation）は、それを二乗したものを積分して、その平方根を求めたものです。

これらを見ます（スライド 22）と、モデルフィットは明らかに、ラッシュモデルと 2PL モデルのハイブリッドしたものでアイテム・バイ・カントリー・インタラクションをとりいれたモデルで最も適合度が高く出ています。例えば、無答案がどんな無答案なのかといったことを解釈する違いによっても別々に出てきます。

スライド 24 は RMSD を使って、それが 0.15 以上のアイテム・バイ・カントリー・バイ・サイクルの組み合わせで、数学を見ますと 1,500 以上の組み合わせの可能性があるんですけれども、それを見ると 549、0.15 以上あったケースが 4 に下がっておりまして、明らかにモデルフィットだけではなくて、ディビエーションの大きさ自体もかなり減っています。

スライド 25 を見ると、2000 年から 2012 年まではラッシュモデルを使っていましたが、どれだけラッシュモデルによる項目が残っているかを見ますと、43 %、19 %、14 %ぐらいということですね。

最終的には、PISA の 5 回の調査をもとに総括的なデータベースを作りました。ラッシュモデルと比べて 2PL とのハイブリッドでアイテム・バイ・カントリー・インタラクションモデルが一番よくフィットしました。3 つの基本リテラシーの累積データは、すべて項目反応理論（IRT）に基づいたパラメータで推定しました。

フィールドテストの分析結果

次に、2015 年のフィールドテストのアナリシスの結果を見てみます。もう一度言いますね。このデザインを実現するために、また PBA（paper-based assessment）から CBA（computer-based assessment）とモードを変えるために、フィールドテストが、小規模な PISA のメインサーベイではなくて、リンキングスタディ（linking study）になっているんですね。だから、今までの PISA のフィールドテストの性格とは丸きり別な性格を持っています。

フィールドテスト・アセスメント・デザインはだいたいこのようになっているんですけれども、無作為に 3 つのグループに分けて、グループ 1 というのは PBA のトレンドアイテムだけを使って、最初のフォームは、1 から 6 までのフォームは科学と数学を 1 時間ご

とに使っています。それと同じ問題がフォーム 31 から 36 まで、コンピュータベースでグループ 2 は使っています。グループ 3 はすべて新しいアイテムです。

結果として見えてきたのは、数学リテラシーでは、クラスターが 4 番目に出てきたときには 16 分ぐらいでできたものが、最初に出てきたポジションの場合には 19.76 分、20 分ぐらいかかっているんですね。これは既に ICT のトレーニングを 5 分ぐらいした後、またこれ位かかっているんですけれども、結局これは効率（efficiency）が上がっているのではないかと考えています。

しかし、レスポンスが早くなっても正答率が下がったら困るわけですね。そういうことで正答率を見ますと、実際少し下がっているわけです。2015 年のフィールドテストを見ます（スライド 31）と数学リテラシーの場合には 0.44 から 0.43 まで、1 %とか 3 %ぐらい正答率が下がっています。でも、これは 2015 年だけですから、前の 2009 年にどれだけ下がっていたかを見ないと比較できないんですね。それを見ると、PBA なんですけれども、2009 年には 4 %から 8 %ぐらいまで下がっているわけですね。だからコンピュータベースにしたときに、読解力について、2015 年で 3 %落ちていたのが 2009 年では 8 %落ちています。要するに、CBA では位置効果というのが減っているということです。

レスポンスタイムに関して、能力別に見てみました。能力別に見ますと、これは能力を一番下と中間と上 3 分の 1 の 3 つに分けているんですけれども、そうしてみますと、能力がある人のほうが時間をかけてやっていますね。能力のない人のほうが早くやっている。アイテムによって、このグラフは科学のトレンドの質問の国別のレスポンスタイムの平均値なんですけれども、ここではっきり出てくるのは、すべての国で質問自体の特性として見られること。黒いラインは標準偏差なんですけれども、それも質問の平均値と同様に動いています。質問のクラスター内での位置とは余り関係ありません。スピードテストではないということですね。

他のテスト、特に記憶力が多大に影響しているテストは、能力の違いがスピードにてきめんに出てくることがよくあります。能力はスピードでもあるということですね。しかし、PIAAC のデータもそうなんですけれども、PISA のデータは能力とスピードと反面しています。

これ（スライド 34）は数学のトレンドです。質問のレスポンスタイムの平均値を能力の上の 3 分の 1 と下の 3 分の 1 と比べたものです。時間のかかる質問は特に難しい質問が多いんですけれども、能力のない人は問題に関係なく、ほとんど同じような時間がレスポンスして、能力のある人は、時間のかかる問題に関しては長く時間をかけてやっています。

無作為の 3 グループはほぼうまくできましたが、小さい学校の場合、国によっては同等化は完全ではありませんでした。

採点者信頼性（corder　reliability）の点ですが、PISA の 3 分の 1 ぐらいはオープン・クエスチョンですから、その回答については採点者によって採点されます。その信頼性についてもモニターしていますけれども、採点者信頼性はかなりよかったですね。平均でも 94 ～ 95 %ぐらいと高いんです。しかし、国によってはずっと低い国もありますし、項目によってはかなり低いものもあります。だから、メインサーベイのときには信頼性の低い問題はもう一度トレーニングし直すことになります。

それから無回答の率ですが、これは CBA の方が PBA よりも低くなっていました。後に

続くクラスターの方が正答率が低下しますので、レスポンスタイムがデータの質の向上に貢献しているということかと思います。レスポンスタイムは項目の位置や国の違いではなく、項目の特殊性と学生の能力に強く影響されていました。

CBT と PBT の差異

次に、PBA と CBA のモードエフェクトなどに関してご紹介してみます。

PISA は評価尺度の各サイクルの互換性が重要な位置を占めます。これは数字的な等化だけではなくて、評価尺度の解釈の互換性ということも含みます。アイテムパラメータが同じになるということは、評価尺度の解釈に互換性があるということを意味します。それは等化というよりも、共通のスケールを作るという意味合いがあります。すなわち、サイクル間の項目パラメータを等化することです。共通項目に関して、PISA2000 から 2012 のパラメータと 2015 年のトレンドを見る項目のパラメータが同じであることがまず最初で、その後、各国間のパラメータが同じであるということ、その後、PBA と CBA の共通項目が同じであるということなどを確認していくことになります。それらに関して IRT を使ってやってみました。

まず最初（スライド 38）は、数学リテラシーについて、表側はアイテム、表頭は国をとったものです。緑が MD（mean deviation）が 0.2 より小さいところです。赤は MD が 0.2 より大きいところ。科学リテラシーは 95.1 %グリーンでした。いずれも尺度の互換性がはっきりと示されました。

スライド 39 は、2000 年から 2012 年までのアイテムパラメータを使って、2015 年の PISA のフィールドテストのデータを見てみたものです。科学の新しいアイテムについてはアイテムパラメータが既知ではありませんから、この新しいアイテムだけを使ってアイテムパラメータを推定して、それをすべての国でどれだけ MD があるかを見たもので、98.1 %が 0.2 以下の MD となっていました。

PBA と CBA の互換性はもうちょっと複雑でして、3 つのグループが無作為に抽出されていることが比較の基本ですけれども、これ（スライド 41）を見るとわかると思いますが、一番上の◆の CMC（complex multiple choice）は、多肢選択の項目ですが、2 段階の選択が含まれることから、通常の多肢選択項目と違い、問題のタイプにかかわらず、傾きパラメータ a に関しては、かなり高い相関性があります。

これ（スライド 42）は、位置パラメータ b に関するものです。トレンドを見るためのアイテムですけれども、PBA と CBA、数学リテラシー、読解力、科学リテラシー、すべて入れてプロットしたものです。これもかなり高い相関性をもっています。高い相関性は、傾きパラメータ a、位置パラメータ b の両パラメータで見られるということです。この結果を見ますと、モードエフェクトは、ほとんどないと考えてよいと思います。ですから、75 ～ 85 %ぐらいのアイテムは、PBA と CBA で、同じアイテムパラメータを使って差し障りないということになります。

PISA2015 の調査設計

PISA2000 年から 2012 年にかけてクラスターの数は、まず、マイナードメインのときには 2 か 3 ぐらいになるんですね。今度、新しいデザインでは 6 ですから、そのうち 4 から 5 ぐらいのクラスターの分は同じアイテムパラメータを使える。これはもう 2 倍の数のアイテムで、同一のパラメータを使えるわけですね。

新しいデザインに関しては、読解力と数学は、両方マイナードメインですけれども、それぞれ 6 クラスターを使っています。ファイナンシャルリテラシーは特別なリテラシーでクラスターが 2 つです。協調的問題解決（CPS：collaborative problem solving）は 3 クラスターを使います。科学は 6 ＋ 6 で、トレンドの 6 クラスターと新しい 6 クラスターです。これが新しいデザインです。

基本的には 6,300 人、150 の学校で 42 人ずつ 66 フォームを使って、以前は 13 フォームでしたから、ペーパーベースのときにはフォームの数が増えるととても管理できなかったんですが、これはコンピュータベースで実施されますから、すべてのアドミストレーションのデザインはコンピュータ内部でやっていますから、フォームの数が増えても丸きり問題ないですね。

緑は科学でグレーが読解力です。科学は 12 クラスターあります。その 12 クラスターのコンビネーションはすごく難しいんですよ。それを完全にポジションでバランスして、それと他のドメインとの兼ね合いもバランスしなければならないですから、これを使って、このデザインをもとにして、そうすると新しいクラスターは 8 回ずつ、古いクラスターは 4 回ずつ出します。

これは同じことですけれども、これを乱数を利用して合わせますと、66 × 6 の 396 フォームを作ることになるんですね。それをもとにして、もともと科学と数学と読解力の 2 ウェイの共分散（covariance）と、それと全部使っている 3 ウェイの共分散を使って結びつけています。

科学領域のサンプル項目

科学のサンプルアイテム（スライド 52 ～ 59）は、このようなものを使っています。これは PISA で作られた新しいアイテムですけれども、このアイテムは実際には使いません。2015 年のために作られましたけれども、これは結局、悪いアイテムではないんですよ。いいアイテムなんですけれども、これと同じようなコンストラクトの他のアイテムがありましたから、それを使って、これは使わないことに決めました。でも、これはとてもいい例だと思います。シミュレーションのアイテムで、スライダーを変えることができたり、シミュレーションを実行することができたりします。

多肢選択式の問題も含まれていますが、右の 2 つのスライダーを使って実際にデータを作るといった問題もあります。また、多肢選択ではなく、理由を記述するオープンエンドの問題も含まれていて、これらは実際にはコーダーが採点します。これはコンピュータから抽出して、それをもとに完全にオーガナイズしてコーダーに振り分けます。コーダーのリライアビリティも同時に見ますし、コーダーのレコードとかもすべて記録します。

以上です。（拍手）

質疑応答

○司会　どうもありがとうございました。

ちょっと時間がありますので、何か事実確認的なご質問などありましたらどうぞお手をお挙げいただければと思います。マイクをお持ちします。

ないでしょうか。では、私から 1 つ。

スレッシュホールドのグラフがありましたけれども、スレッシュホールドという考え方は日本では余り使われていないと思いますが、項目の正答率が 0.62 となる点数ということですよね。

○山本氏　そうです。だから本来、マスタリングということを考えると 0.8 というのをよく使っていました。長くアダルトリテラシーを作っていましたけれども、そのときには 0.8 を使っていました。それは概念をマスターしたときには 0.8 がいいんじゃないかということですね。0.62 というのは、私としては低過ぎると思うんですけれども、それは OECD で決めたことですから。

○司会　0.62 というのは我々からするとちょっと中途半端な気がするので、どうやって決まったのかなと。

○山本　0.62 というのは、一番下のレンジがありますね。そこのところで 0.5 になるアイテムを探しているんですよ。レベルの中のすべてのアイテムを、テスト特性曲線をつくったときに 0.5 になるように。それとレンジ自体も変えていますから。

そのときに、それを 2000 年につくったので、それがそのままつながっているんですね。だから本来は、もし習得レベルのアイテムのセットが変わったならば習得レベルのディスクリプション自体も変えるべきなんですけれども、今回、2000 年から 2012 年までのリスケーリングをしてアイテムパラメーターが決まりましたから、今回はその習得レベルのディスクリプションは変えなくて済むんです。

○司会　細かい部分は私もよくキャッチアップできていないんですけれども、新しいテスト開発をするときに、非常に用意周到といいますか、フィールドテストも含めてクラスターのポジションによってどう違うかという分析までされているというのは、これから新しいテストを開発するときに見習わなくてはいけない点だなと思いました。

いかがでしょうか。皆さんから何かございますでしょうか。

ご所属とお名前をお願いします。

○安野氏　国立教育政策研究所の安野と申します。

ちょっと教えていただきたいのは、今年のフィールドテストのときにペーパーベースのものとコンピュータのものを比較しましたが、そのときに 1 つ、実施しているときにちょっとどうかなと思いましたのは、コンピュータ型のほうで画面で見てやるのとペーパーで割と差異があるのは、記憶ではなくて思考力に関わる問題でメモが必要であったりとか、そういった問題は上位層で差が出るのではないかという懸念がちょっとあったんですけれども、実際に統計的には、そのようなアイテムは出てきましたでしょうか。

○山本氏　出てきますよ。例えば数学のアイテムで特に出てきているのは、例えば方程式を書くときに、知っている人は方程式を手で書くのは比較的できるんですけれども、方程式をエディターでつくるのは難しかったりしますね。そんなところがよくできていなかったということはあったかと思います。

○安野氏　今回やはり日本は日本語と数式の入力が非常に難しかったんですね。そこら辺も国によって、言語と入力による違いというのは出てきていますか。
○山本氏　出てきていますね。特にレスポンスのスピードとか、とにかく国によってはスピードが随分違っています。平均のスピードとしてだいたい 40 %ぐらい、30 %くらいかな、早い国があるんですね。
　だけれども、日本の PISA のフィールドテストはアダプテーションに関してちょっと問題があったんです。他の国と比べて。要するに、半角と全角の……
○安野氏　数字に全角と半角があったりとかですね。
○山本氏　そうそう。それがちょっと問題があったりして、それも全部改良されましたけれども、そんなことも多分エラーの中に入っていると思います。
○安野氏　思考力を要する問題が、コンピュータベースに移したときに本当にペーパーより正確に測れるのかなと、ちょっと疑問に感じたんですね。
　ありがとうございます。
○南風原氏　東京大学の南風原です。
　先ほど項目のパラメータが国によって違う部分があると。ということは項目の働き方が、どの国で測定するかによって違いがあるということだと思うんですが、それを認めた上で同じスケールで国の比較ができるというのは、どういうロジックになるんでしょうか。
○山本氏　例えば、今まで、アダルトリテラシー部分なども含めて、すべて同じようにやってきましたけれども、例えば、大木があって、1 つの大きな幹があると想定していただいて、それで、アイテムによっては枝が出てくるという感じでしょうか。国によって違う枝が出てくるということですね。今までアダルトリテラシーに関して見たときに、95 %ぐらいはインターナショナルなアイテムパラメータでよくフィットしていて、5 %とかそのくらいユニークなアイテムパラメータがありました。ですから比較できる部分は大きな幹のところであって、それでユニークな点として出てくるのは 5 %ぐらいずつ出てきているということです。比較するのはそこの 95 %、インターナショナルのリファレンスに関して、95 %のアイテムに関して比較できているということです。
　残りの 5 %の部分は、必ずしも国の違いということだけが理由ではなくて、アイテムのトランスレーションが悪かったり、回答方式へのアダプテーションが悪かったり、そういうこともあると思います。例えば、前にあったんですけれども、オランダで作られたアイテムで、自転車のサイズを選ぶ問題がありました。ナンバー 5 がすごく大切なナンバーだったんで、実際アイテムを作ったオランダでは、ナンバー 5 を数字の 5 ではなくて書き出したんですよ。そうすると、オランダではそのアイテムが極端に難しくなってしまったということがあります。
○南風原氏　悪い項目は、使うんですか、使わないんですか。
○山本氏　使います。使いますが、別のアイテムパラメータをアセスメントすることで。アイテム・バイ・カントリーのインタラクションがあるということは、必ずしも悪いアイテムとは限らないんですね。他の別な理由で、例えばもっとファミリアな、もっと頻繁に使われている語彙が出ていたとか、そういうことに関してもアイテムパラメータが少し変わる場合がありますから。
○司会　まだまだ他にあると思いますけれども、またパネルディスカッションのときにご

質問、ご意見等お寄せいただければと思います。

山本先生、どうもありがとうございました。（拍手）

それでは、コンピュータの入れ換え等もありますので、報告に入る前に休憩とさせていただきます。

〈午後 2 時 15 分　休憩〉

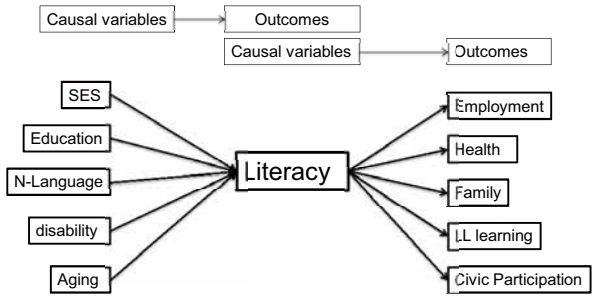
Major Innovations in PISA 2015 and Beyond

Kentaro Yamamoto
ETS
11/29/2014

Purposes of LSA

- The primary reason for developing and conducting large-scale assessment is to provide empirically grounded interpretations upon which to inform policy decisions. The appropriate role for policy research is not to define policy; rather it is to establish a body of evidence from which informed judgments can be made. (Messick)

Rationale of LSA



What is PISA

- The OECD's Program for International Student Assessment (PISA) is one of the most widely known large scale surveys.
- PISA assesses sampled in-school 15-year-old students in three core domains of Reading, Mathematics and Science, and 30 min. of background questions (BQ).
- Assessment every three years since 2000.
- Participation has grown from 228,784 students from 43 countries in 2000 to 514,531 students from 32 countries in 2012, and more in 2015

What PISA measures

- PISA measures what students have learned and been taught and how well they can extend what they have learned and apply that knowledge in new or unfamiliar settings, both in and outside of school.
- The modern societies tend to reward individuals not for what they know, but for what they can do with what they know.
- This focus on applications of skills instead of measuring the acquired knowledge has been implemented and gaining strength as seen in National Adult Liteacy Survey, NALS (1992), International Adult Literacy Surveys, IALS (1994), Adult Literacy and Life Skills Survey, ALL (2004), and most recently in Programme for the International Assessment of Adult Competences, PIAAC (2012).

General PISA Design

- A cluster is a set of items from one domain and expected to take less than 30 minutes to respond by most students. A set of items in a cluster is unique.

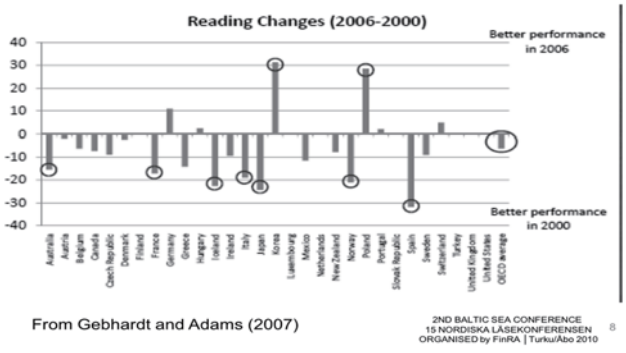
Cluster 1	Cluster 2		Cluster 3	Cluster 4	BQ		Other Domain
30分	30分	休	30分	30分	30分	休	60分

Goals of PISA 2015

- Fully computer delivered
- Reliable trend information on Reading, Mathematics, and Science
- Expanded construct of Science
- New domain: Collaborative Problem Solving
- Stable trend

Major Issues of Past PISA

Other Significant Changes in PISA Reading Literacy Scores

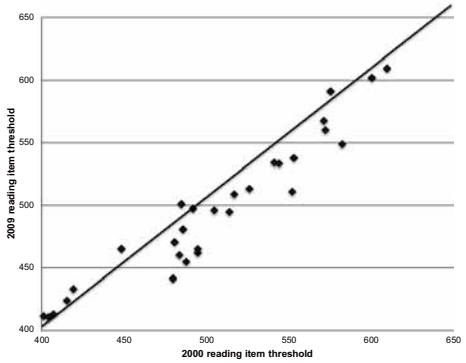


From Gebhardt and Adams (2007)

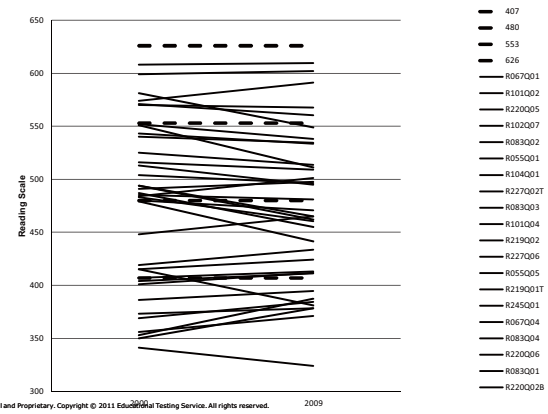
2ND BALTIC SEA CONFERENCE
15 NORDISKA LÄSEKONFERENSEN
ORGANISED by FinRA | Turku/Åbo 2010

8

Thresholds of Trend Reading Items Between 2000 and 2009 PISA



Thresholds of PISA Reading Trend Items

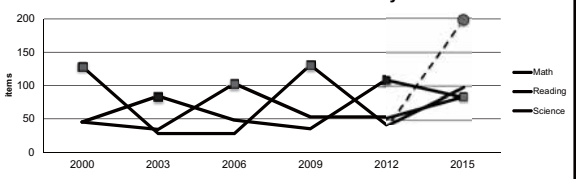


Number of Items Used Before

	2000	2003	2006	2009	2012
Math	45	84	48	35	109
Reading	129	28	28	131	44
Science	45	34	103	53	53

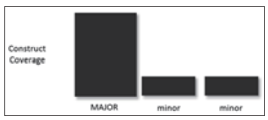
Note: Major domains in green

Number of Items Used in Past PISA by Domain



Conceptual Designs of Trend

Old PISA Design Concept



2006 2009 2012

New Design Concept

2015 2018 2021

- Width of bars represents the number of respondents per item.

What do we gain?

- large linked databases
  - Math, Reading, Science with 1.9 million students
  - Financial Literacy with a set of 2012 countries
- Integrated (5-cycle) evidence on:
  - For an item: Item functioning over time
  - For a country: Item by country interactions over time
  - Study improvement of model-data fit with all data
  - Non-Rasch items are not (necessarily) bad items

What do we gain, again?

- Linking of cycles 2015, 2018,…
  - 'historical' databases can be amended
  - Gives raise to a tool for studying longer term trends
  - Allows country by item (rather than by cycle) treatment
  - Detect changes in item difficulty, by curriculum effects of PISA, or 'leakage' of items into instruction
- Forward looking:
  - Basis for comparisons of PBA and CBA assessment
  - Stable parameters for concurrent linking of 2015 cycle

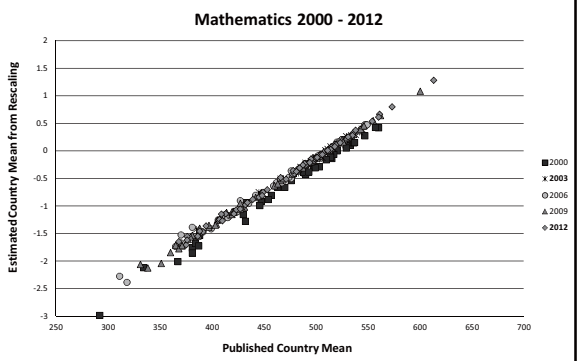
Creating the Database

- The analysis used all available (published data)
- Funded under ETS research allocation
- Used existing public documentation & data

Year	2000	2003	2006	2009	2012
Student	228,784	276,165	398,750	515,958	514,531
Country	43	41	57	74	68

Total N: 1,934,188

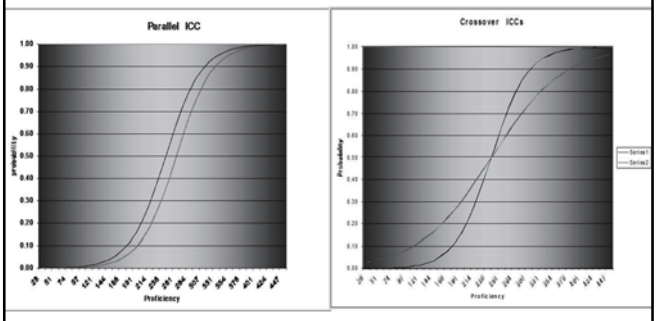
Comparison to Published Data



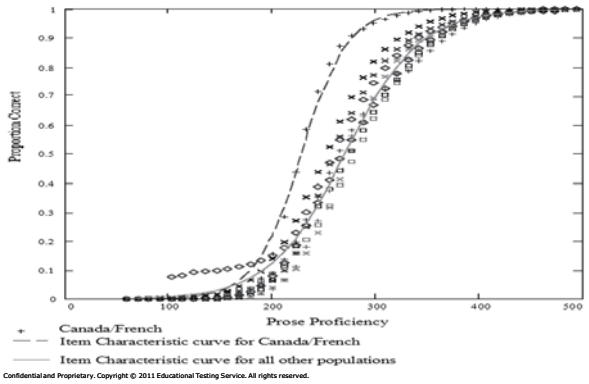
Creating the Database: More General IRT Models

- The data for math, reading, science, and financial literacy were concurrently calibrated using more general IRT models
  - Rasch/PCM Model (baseline), (2000-2012)
  - 2PL/GPCM Model
  - Rasch/2PL (PCM/GPCM) Hybrid
    - (item slopes released for some items)
  - Rasch/2PL (PCM/GPCM) IBCI
    - (accounting for item-by-country interactions)

Rasch Model | 2PL Model



Example of Item by Country Interaction



Model Fit and Item Fit Evaluation

- Overall Model fit was evaluated using AIC and BIC
- Item fit was evaluated using the Mean Deviation (MD) and the Root Mean Squared Deviation (RMSD)
- MD and RMSD were calculated for all items in each country-language group

$$MD = \int \left(P_o(\theta) - P_e(\theta)\right) f(\theta) d\theta$$

- RMSD values greater than 0.2, and MD values greater than 0.2 or smaller than -0.2, respecting the small sample size per country, were considered as deviations

$$RMSD = \sqrt{\int \left(P_o(\theta) - P_e(\theta)\right)^2 f(\theta) d\theta}$$

Model Fit

Scoring	Model	Math	Reading	Science	Fin Lit
Omit is Wrong	Rasch	26400730	30968125	29908518	3857879
	2PL	26118134	30675531	29585732	3814562
	Hybrid	26175012	30691983	29591677	3818645
	Hybrid IBCI	**25946516**	**30472304**	**29302806**	3804268
Omit is not-reached	Rasch	24170443	28198612	27918199	3619855
	2PL	23952026	27966585	27676158	3584239
	Hybrid	23999181	27984532	27681378	3590496
	Hybrid IBCI	**23787968**	**27720962**	**27372668**	3564699

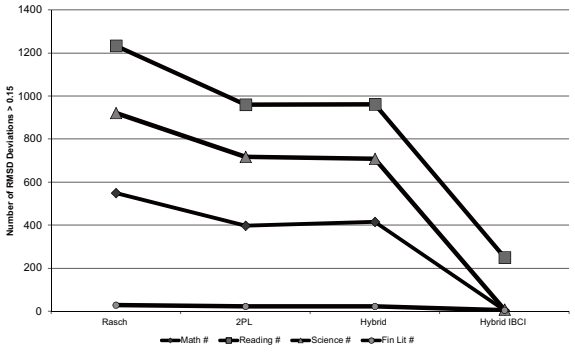
BIC (Schwarz, 1978) reported in the table

Improved Model Fit

	Large Deviations	Rasch	2PL	Hybrid	Hybrid IBCI
Math (N=15,795)	**N**	**549**	**397**	**415**	**4**
	Percent	3.48%	2.51%	2.63%	0.03%
Reading (N=18,603)	**N**	**1,233**	**960**	**962**	**250**
	Percent	6.63%	5.16%	5.17%	1.34%
Science (N=16,223)	**N**	**921**	**717**	**708**	**8**
	Percent	5.68%	4.42%	4.36%	0.05%
Fin Lit (N=718)	N	29	23	23	5
	Percent	4.04%	3.20%	3.20%	0.70%

Note: Deviations defined as item*country*cycle RMSD greater than 0.15

Improved Model Fit



Retained Rasch Items

	Total Number of Items	Rasch # Retained	Rasch % Retained
Math	179	77	43%
Reading	223	42	19%
Science	133	19	14%
Financial Lit.	40	15	38%

Note: Rasch items (slope=1.0) retained under the Hybrid and Hybrid IBCI models

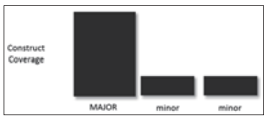
Summary

- Created a comprehensive database using all available PISA items across the 5 cycles
- Compared to the Rasch model Hybrid IBCI model fits the data best for all content domains
  - A number of items were still retained as Rasch items
  - Most items exhibit some item-by-country-by-cycle interactions that can be accounted for by the Hybrid IBCI model
- Created a database using Hybrid IBCI model containing common item parameters across each of the three domains that can be used to estimate trends with PISA 2015

Field Test Analysis

Conceptual Designs of Trend

Old PISA Design Concept



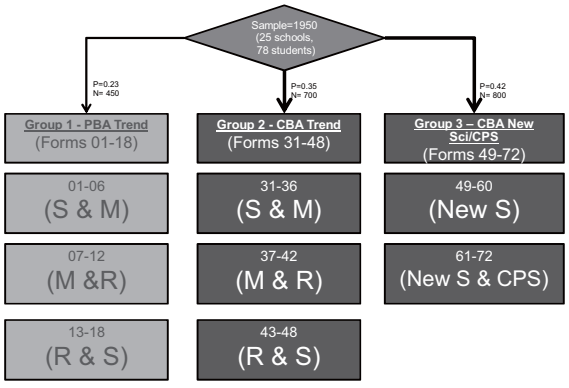
2006 2009 2012

New Design Concept

2015 2018 2021

- Width of bars represents the number of respondents per item.

FT Assessment Design (CBA)



Response Time Decreases by Cluster Position (min)

- Students spent fair amount of time.
- Considering response accuracy was not affected by cluster position, reduced response time may indicate response efficiency.

	Position 1	*Position 2*	*Position 3*	*Position 4*	*Position 4- Position 1**
Mathematics	**19.76**	**18.84**	**17.26**	**16.03**	**-3.72**
Reading	**24.11**	**21.47**	**20.75**	**18.44**	**-5.67**
Science	**22.32**	**20.61**	**19.21**	**17.11**	**-5.21**
CPS	**22.67**	**21.26**	**20.96**	**18.79**	**-3.88**
FL	**19.49**	**14.49**			**-5.00**

Note: Cluster response time is a sum of time spent on responding to every item. The two FL clusters alternated positions.

CBA Reduces Cluster Position Effect

P+: Proportion correct averaged across countries

2015

	Position 1	Position 2	Position 3	Position 4	Position 4-Position 1
Mathematics	0.439	0.454	0.432	0.427	**-0.012**
Reading	0.575	0.580	0.566	0.544	**-0.031**
Science	0.409	0.418	0.401	0.385	**-0.024**

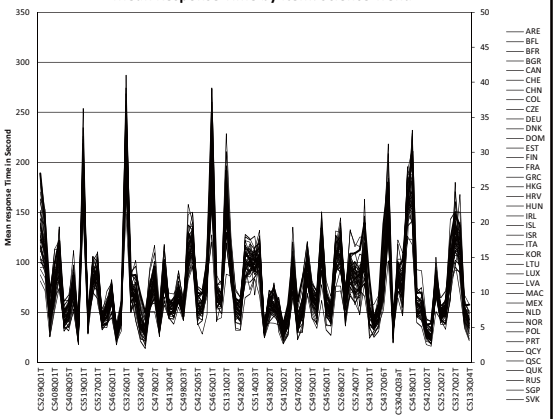
2009

	Position 1	Position 2	Position 3	Position 4	Position 4-Position 1
Mathematics	0.411	0.401	0.384	0.371	**-0.040**
Reading	0.581	0.557	0.532	0.499	**-0.083**
Science	0.490	0.478	0.457	0.435	**-0.055**

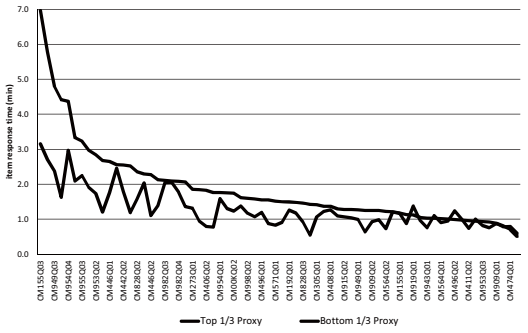
Cluster Level Response Time by Proxy Skill Level (min)

	Low Ability	Moderate Ability	High Ability	Difference (High-Low)
Mathematics	15.17	18.97	20.92	**5.65**
Reading	18.97	22.60	23.64	**4.67**
Science	17.45	21.31	22.86	**5.41**
CPS	19.69	22.23	22.77	**3.08**
Financial Literacy	15.76	19.97	22.11	**6.35**

Mean Response Time by Item: Science Trend



Average Response Time of Mathematics Items by Proxy Level



Data Quality: Summary

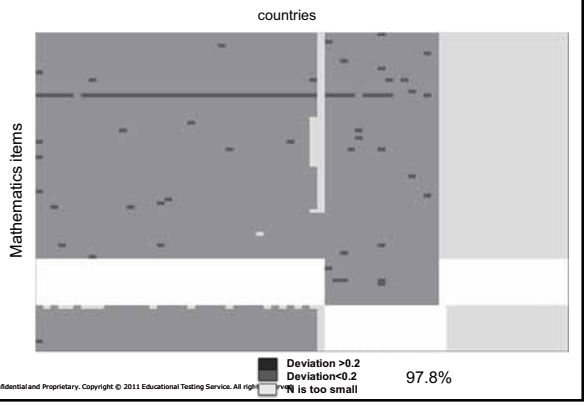
- Random student assignment worked well, but…
- High coder reliability overall, but…
- Advantage of CBA over PBA
  - Fewer omitted responses
  - Reduced cluster position effect on P+ decline
  - Response time improves data quality
- Item response time is driven by item demands and students' ability not by country nor by position

Establishing Comparability Through IRT Scaling

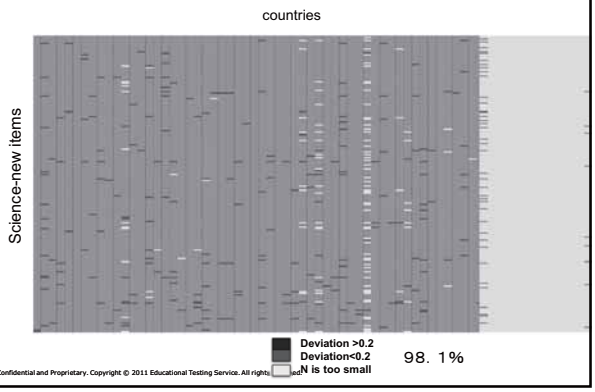
Establishing Comparability Through IRT Scaling - 1

- PISA 2015 needs to establish comparability across cycles, assessment modes, and countries for trend items in the domains Science, Math, Reading, and Financial Literacy; and comparability across countries for new Science items and CPS
- The best way to achieve this is through building common scales
- This was done through a multi-step process requiring:
  1) establishing comparability between historical PISA data (cycles from 2000-2012) and the PISA 2015 Field Trial in each domain for PBA
  2) establishing comparability across countries in each domain for CBA
  3) establishing comparability between modes of assessment (PBA, CBA) in each domain

Math Scale – PBA Deviation Against 2000-12 Parameters



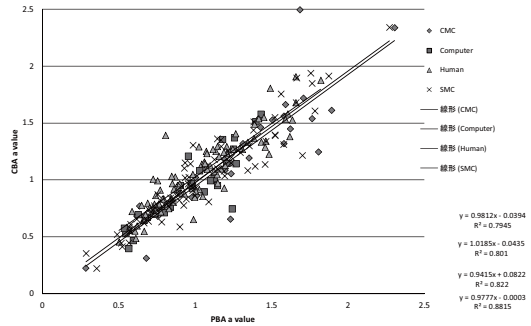
Science-new Scale – CBA Deviation Across Countries



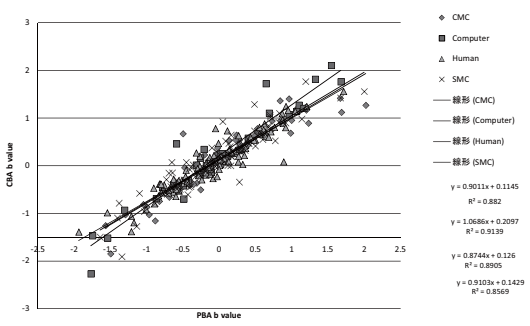
Examining CBA/PBA consistency with Field Test data

- The 3 student groups in the FT establish a basis for comparing CBA and PBA using ***randomly equivalent samples from the same student population(s)***
- Initial IRT modeling targeted the equivalency of item functions in PBA and CBA modes
- Evaluation of item parameters based on international comparability + graphical model checks

Comparison of Slope Parameter Estimates Across PBA and CBA



Comparison of Difficulty Parameter Estimates Across PBA and CBA



Summary: IRT Modeling for linking PBA and CBA scales

- A series of extended IRT models was estimated to examine the high level of agreement seen in graphical model checks
- The majority of items show strong measurement invariance, while the remainder of items still shows a weaker form of invariance
- Between PBA and CBA, we expect to see 75%-85% common parameters across domains
- This corresponds to roughly 4-5 of the 6 trend clusters, a stronger link than in past cycles

Cognitive Assessment Design 2015 PISA Main Study

Benefits and Features of Design

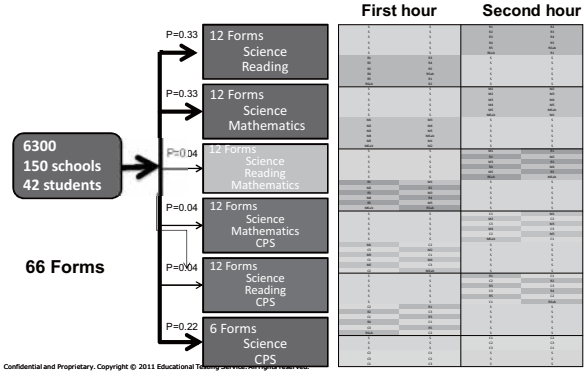
- Reduced systematic errors due to incomplete coverage of the constructs by expanded construct coverage across M-m-m cycles
- Three assessments (M-m-m) over 9 years can be thought as a package instead of 3 independent trend points
- Allow to introduce sizeable changes through renewed measurement constructs every 9 years

Many Clusters for Each Domain

A cluster is a set of items from one domain and expected to take less than 30 minutes to respond by most students. A set of items in a cluster is unique.

- Reading has 6 clusters of trend items,
- Mathematics has 6 clusters of trend items,
- Financial Literacy has 2 clusters of trend items,
- CPS has 3 clusters of new items,
- Science has 6 clusters of trend items and 6 clusters of new items.

MS CBA Design Detail



MS CBA: 2 Domain Forms

Forms 31-42
2 Domains
Science
Reading

P=0.33
N=2080

Form	Cluster 1	Cluster 2	Cluster 3	Cluster 4
31	S	S	R1	R2
32	S	S	R2	R3
33	S	S	R3	R4
34	S	S	R4	R5
35	S	S	R5	R6ab
36	S	S	R6ab	R1
37	R1	R3	S	S
38	R2	R4	S	S
39	R3	R5	S	S
40	R4	R6ab	S	S
41	R5	R1	S	S
42	R6ab	R2	S	S

Science Clusters Design1

SCIENCE CBA cluster pairs

		Second Cluster											
		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
first cluster	S1							1			2		
	S2				34				3				
	S3									4			5
	S4							6			7		
	S5	35										8	
	S6			36									9
	S7						10		19	20		21	
	S8	11				12					22		23
	S9		13				14		24			25	
	S10			15				26		27			28
	S11		16		17				29		30		
	S12					18		31		32		33	
1st cluster		2	2	2	2	2	2	4	4	4	4	4	4
2nd cluster		2	2	2	2	2	2	4	4	4	4	4	4
total		4	4	4	4	4	4	8	8	8	8	8	8

Trend Clusters: S1-S6

New Clusters: S7-S12

Science Clusters Design2

Trend (S1-S6) & New (S7-S12)

Pairs	1st	2nd
1	S1	S7
2	S1	S10
3	S2	S8
4	S3	S9
5	S3	S12
6	S4	S7
7	S4	S10
8	S5	S11
9	S6	S12
10	S7	S6
11	S8	S1
12	S8	S5
13	S9	S2
14	S9	S6
15	S10	S3
16	S11	S2
17	S11	S4
18	S12	S5

Both New (S7-S12)

Pairs	1st	2nd
19	S7	S8
20	S7	S9
21	S7	S11
22	S8	S10
23	S8	S12
24	S9	S8
25	S9	S11
26	S10	S7
27	S10	S9
28	S10	S12
29	S11	S8
30	S11	S10
31	S12	S7
32	S12	S9
33	S12	S11

Both Trend (S1-S6)

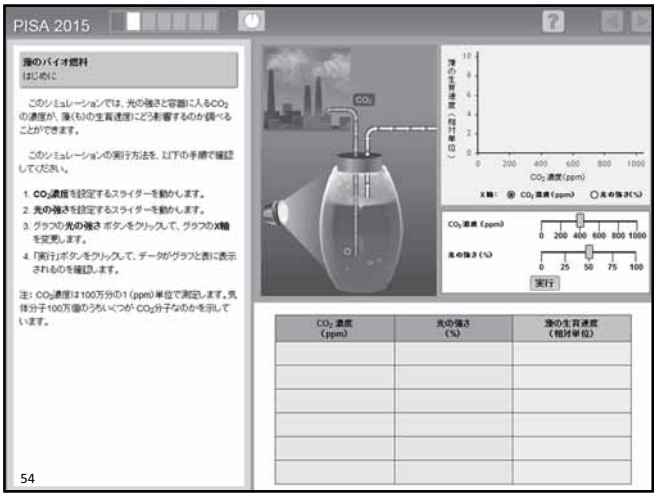
Pairs	1st	2nd
34	S2	S4
35	S5	S1
36	S6	S3

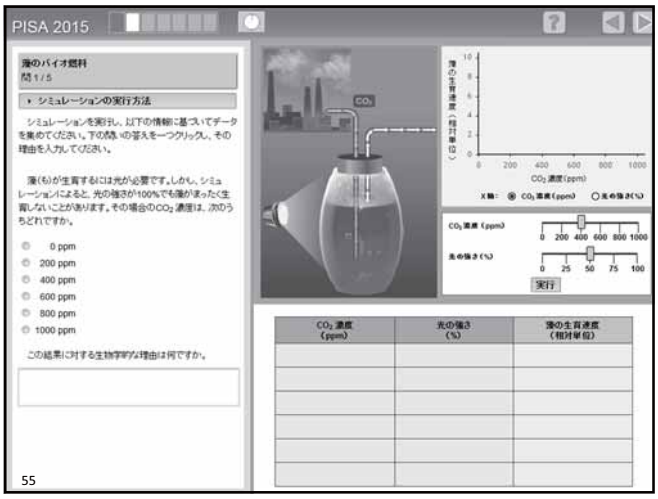
Science Clusters Selection

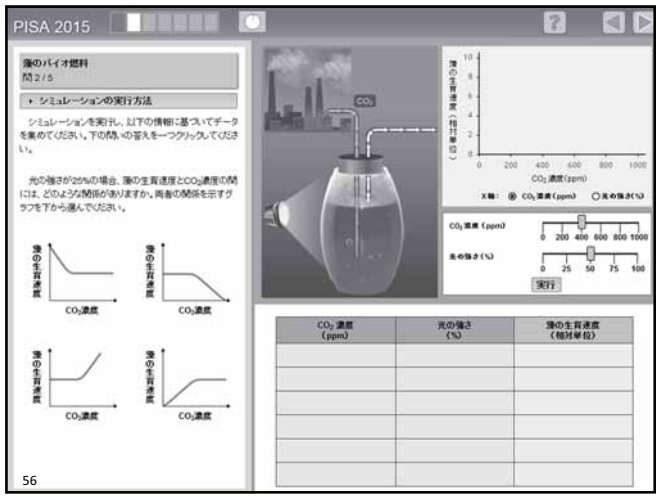
Form	Random number					
	1	2	3	4	5	6
31	1	13	6	9	22	25
32	2	16	12	10	31	32
33	11	5	17	14	26	29
34	35	4	7	19	23	30
35	34	15	8	20	24	28
36	36	3	18	21	27	33
37	35	4	7	19	23	30
38	34	15	8	20	24	28
39	36	3	18	21	27	33
40	1	13	6	9	22	25
41	2	16	12	10	31	32
42	11	5	17	14	26	29

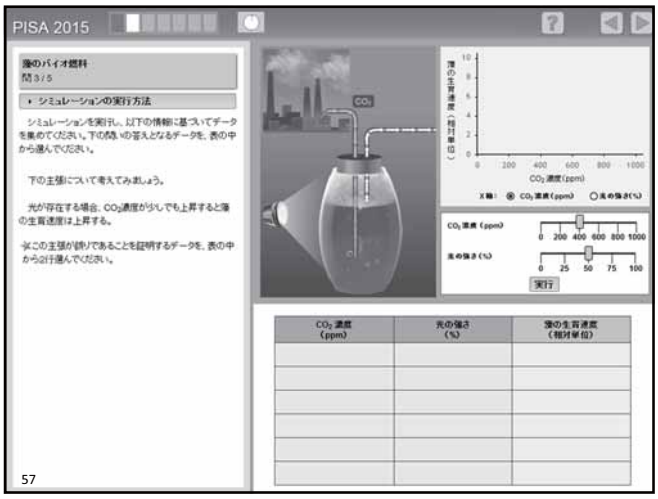
1-18: Trend & New Clusters
19-33: Both New Clusters
34-36: Both Trend Clusters

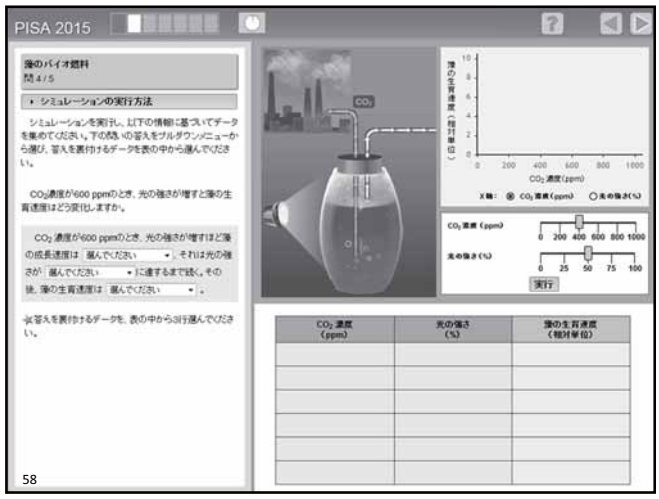
Sample Science Items




PISA 2015
藻のバイオ燃料
問 1 / 5
▸ シミュレーションの実行方法
シミュレーションを実行し、以下の情報に基づいてデータを集めてください。下の問いの答えを一つクリックし、その理由を入力してください。
藻(も)が生育するには光が必要です。しかし、シミュレーションによると、光の強さが100%でも藻がまったく生育しないことがあります。その場合のCO2 濃度は、次のうちどれですか。
0 ppm
200 ppm
400 ppm
600 ppm
800 ppm
1000 ppm
この結果に対する生物学的な理由は何ですか。
CO2 濃度 (ppm)
光の強さ (%)
藻の生育速度 (相対単位)
実行
55


PISA 2015
藻のバイオ燃料
問 2 / 5
▸ シミュレーションの実行方法
シミュレーションを実行し、以下の情報に基づいてデータを集めてください。下の問いの答えを一つクリックしてください。
光の強さが25%の場合、藻の生育速度とCO2濃度の間には、どのような関係がありますか。両者の関係を示すグラフを下から選んでください。
藻の生育速度
CO2濃度
CO2 濃度 (ppm)
光の強さ (%)
藻の生育速度 (相対単位)
実行
56


PISA 2015
藻のバイオ燃料
問 3 / 5
▸ シミュレーションの実行方法
シミュレーションを実行し、以下の情報に基づいてデータを集めてください。下の問いの答えとなるデータを、表の中から選んでください。
下の主張について考えてみましょう。
光が存在する場合、CO2濃度が少しでも上昇すると藻の生育速度は上昇する。
※この主張が誤りであることを証明するデータを、表の中から2行選んでください。
CO2 濃度 (ppm)
光の強さ (%)
藻の生育速度 (相対単位)
実行
57


PISA 2015
藻のバイオ燃料
問 4 / 5
▸ シミュレーションの実行方法
シミュレーションを実行し、以下の情報に基づいてデータを集めてください。下の問いの答えをプルダウンメニューから選び、答えを裏付けるデータを表の中から選んでください。
CO2濃度が600 ppmのとき、光の強さが増すと藻の生育速度はどう変化しますか。
CO2 濃度が600 ppmのとき、光の強さが増すほど藻の成長速度は 選んでください ↕ 。それは光の強さが 選んでください ↕ に達するまで続く。その後、藻の生育速度は 選んでください ↕ 。
※答えを裏付けるデータを、表の中から3行選んでください。
CO2 濃度 (ppm)
光の強さ (%)
藻の生育速度 (相対単位)
実行
58


PISA 2015
藻のバイオ燃料
問 5 / 5
この問いではシミュレーションを使う必要はありません。下の問いの答えを入力してください。
技術者が、藻(も)を育てる実用的なシステムをつくりました。容器の設計は右のシミュレーションと同じです。技術者は、そのシステムにおける藻の生育速度が、シミュレーションで予測した藻の生育速度と異なることに気がつきました。
この違いの原因になったと考えられる理由を、2つあげてください。
CO2 濃度 (ppm)
光の強さ (%)
藻の生育速度 (相対単位)
実行
59


Questions and Discussion
60
Confidential and Proprietary. Copyright © 2011 Educational Testing Service. All rights reserved.

Major innovations in PISA 2015 and beyond

Kentaro Yamamoto
ETS
November, 2014

Introduction

The OECD Program for International Student Assessment (PISA) is one of the most widely known large scale surveys on three of the main academic subject domain skills and background questions of secondary school students across many countries in the world. PISA assesses statistically sampled 15-year-old students who attend schools in participating countries in three core domains of Reading, Mathematics and Science every three years since 2000. Participation has grown from 228,784 students from 43 countries in 2000 to 514,531 students from 68 countries in 2012.

Over the years, important innovations in terms of both assessment domains and mode of delivery have been introduced. Through the inclusion of innovative assessment domains, PISA has sought to broaden the range of knowledge and skills measured in this international assessment of 15-year-old students. These innovative domains have included assessments of problem-solving competences in 2003 and 2012, and of non-cognitive dispositions such as self-assessments of learning strategies in 2000 and attitudes toward science in 2006. Beginning in 2006, in recognition of the expanding role of technology in educational systems, the workplace, and everyday life, innovative computer-based modules have been included in each cycle of PISA. These included the Computer-Based Assessment of Student Skills in Science in 2006 and the PISA 2009 Electronic Reading Assessment. In 2012, computer-based assessments were administered for Reading, Mathematics, and Problem Solving. In each cycle, computer-based modules were optional, but the number of countries participating in these options increased steadily over the years.

PISA examines not only what they have learned and been taught but also examines how well they can extrapolate from what they have learned and apply that knowledge in new or unfamiliar settings, both in and outside of school. This approach reflects the fact that modern societies tend to reward individuals not for what they know, but for what they can do with what they know. This notion of assessment of successful applications of skills instead of measuring the acquired knowledge has been implemented and gaining strength as seen in National Adult Liteacy Survey (NALS) in 1992, International Adult Literacy Surveys (IALS) in 1994, Adult Literacy and Life Skills Survey (ALL) in 2004, and most recently in Programme for the International Assessment of Adult Competences (PIAAC) in 2012.

Historical PISA (2000-2012)

In each cycle, one domain is identified as the major domain while the other two are treated as minor domains. This distinction is important because the amount of time allotted for the measurement of each domain is standardized for each participating school and student, as is the number of students who are sampled. This means that the total volume of data that is collected in each cycle is determined by the number of assessment items and the number of students who respond to each item.

PISA emphasized the major domain in each cycle at the expense of the volume of data collected in each of the minor domains. This was done primarily by reducing the number of items or the representation of constructs in each of the minor domains. Often, the number of items for the minor domains was reduced by a factor of 3 or 4, while maintaining a sufficient sample size per item. This choice was, among other things, based on the operational consideration of limiting the number of booklets in a paper□based assessment. In addition, it enabled the estimation of item parameters without

the need of borrowing or carrying over item parameter information from previous cycles.

The reduction of the number of items from nearly 130 for a major domain to about 30 or 40 for minor domains brings with it the potential for introducing systematic errors because items are neither deleted nor selected at random and the presence of item□by□country interactions is more pronounced in smaller item samples. In addition, one can argue that even if the "best possible" items are selected, a reduction in the number of items used in a given cycle by a factor of 3 or 4 will reduce the content coverage of practically every assessment of student skills. This item reduction approach may have introduced unintended bias due to poorer representation of the construct compared to the major domain assessment. While such effects can be ignored if there are no item□by□country interactions within a domain in terms of bias, a recent article by Gebhardt & Adams (2007) as well as recent chapters by Urbach (2012) and Carstensen (2012), point out that these effects are more salient in some countries.

PISA, like many large scale surveys, used a BIB spiral design in order to reduce respondent's burden while using relatively large number of items to represent the measurement construct as thoroughly as possible, i.e., a student received a subset of items instead of all items in the item pool. Following table below shows the booklet design for 2003 PISA when Mathematics was the major domain, Reading and science were the minor domains and problem solving was the innovative domain. Every cluster was balanced in terms of position distributed among 4 out of 13 unique booklets.

Table 1: 2003 PISA Booklet Design

Booklet	Cluster1	Cluster2	Cluster3	Cluster4
1	M1	M2	M4	R1
2	M2	M3	M5	R2
3	M3	M4	M6	PS1
4	M4	M5	M7	PS2
5	M5	M6	S1	M1
6	M6	M7	S2	M2
7	M7	S1	R1	M3
8	S1	S2	R2	M4
9	S2	R1	PS1	M5
10	R1	R2	PS2	M6
11	R2	PS1	M1	M7
12	PS1	PS2	M2	S1
13	PS2	M1	M3	S2

IRT model has been used for all past PISA surveys using random samples of 500 students from each of participating countries, i.e., about 150 students per item per country, thus it is too small to detect item by country interaction. Item by country interaction was assumed non-existent and a single model item calibration was carried out.

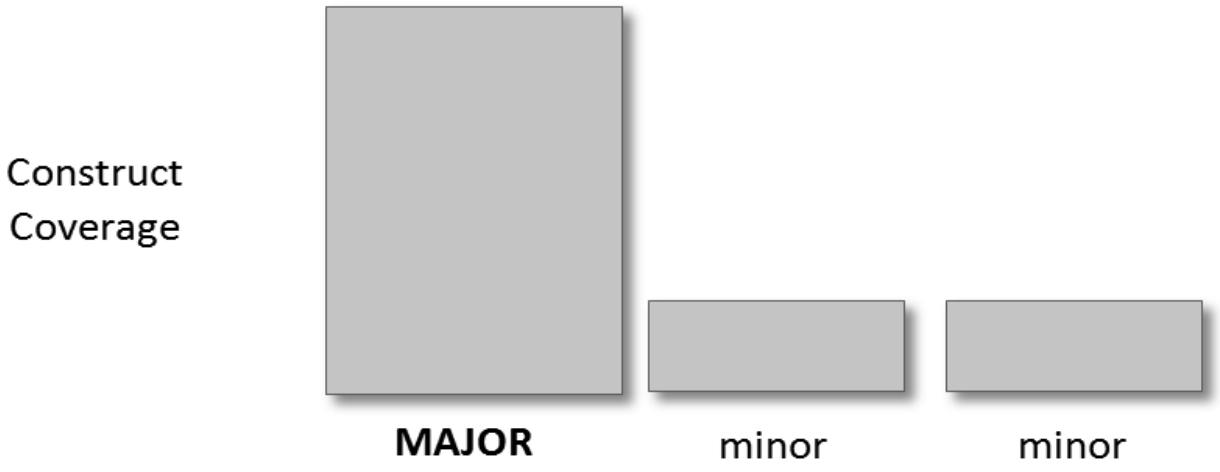
Every cycle, similar item calibration was carried out based on the data collected in one cycle without using data from other previous cycles. In order to evaluate trend on the comparable scale, scale equating was carried out by matching the distributions of item parameters. Since Rasch models was used in the past, i.e., an additive constant was used to match means of location parameters from multiple cycles. This resulted in the trend items administered in multiple cycles having different item parameters for each cycle.

2015 PISA conceptual design: Major minor distinction reduced

Any assessment must contend with two types of errors — random and systematic. Random errors do not result in bias but do increase uncertainty and, therefore, the precision of results. Systematic errors, on the other hand, introduce bias especially in the measurement of trends, and are less desirable because their direction is unknown and not easily quantified or controlled for by statistical means. All large-scale surveys, including PISA, struggle with these two sources of error and aim to control them by optimising the assessment design, as well as sample size, sampling methods, and other contributing factors. An increase in random errors will reduce the ability to detect differences among groups of interest and can typically be offset by increasing sample size. However, an increase in systematic errors not only reduces the ability to detect differences, but also may lead to the attribution of false differences; i.e., differences that are considered significant, even though the true differences are negligible, or even zero. Because of the possibility of introducing bias, a reduction in systematic errors is generally preferable over a reduction of random error components.

Figure 1 below provides a graphic representation of the relative difference in construct coverage between the major and minor domains as implemented from 2000–2012. The vertical height of each bar represents the proportion of items measured in each assessment cycle by domain, while the width conveys the relative number of students who respond to each item within each domain. The reduced height of the bars for the minor domains is intended to represent the reduction of items in that domain and therefore the degree to which construct coverage has been reduced.

Figure 1. Comparison of Construct Coverage in the 2000–2012 PISA Design by Major and Minor Domains



ETS proposed a new approach to measuring trends for PISA 2015 and believes it should be continued for 2018 and beyond. The design stabilizes the trend through reducing potential bias by including all (or almost all) of the items in each minor domain while reducing the number of students responding to each item. This strategy also keeps the volume of data the same for each cycle and increases the construct coverage for the minor domains, while reducing the number of students responding to each item. The result is that the construct representation for each minor domain is at a level comparable to the major domain cycle. More importantly, this approach reduces the potential for bias introduced due to item□by□country interactions and the switch from major to minor domains in the current design. The result both stabilizes and improves the measurement of trend.

The approach adopted in 2015 is represented graphically in Figure 2 below. As represented by the vertical bars, the construct coverage for the minor domain is comparable to the major domain design, at the expense of reducing the number of students who respond to each of the minor domain trend items.

This reduction of student responses per minor trend item is represented in the figure by the narrowing of the bars for the two minor domains.

Figure 2. Approach Used to Balance Major/Minor Domains in 2015 and Beyond

Under this approach for measuring trends, each domain goes through a "domain rotation" or a nine□year period that begins with a new or revised framework and continues with the two subsequent cycles in which it is a minor domain. As an example, for Reading Literacy, one domain rotation would be 2009, 2012, and 2015 and another will become 2018, 2021, and 2024 moving forward. Thinking about the assessment design in terms of this domain rotation clarifies the specific function of each cycle within that nine□year period and the importance of the construct. Over a domain rotation, each major and minor cycle serves a specific function in terms of its contribution to the measurement of trend. Information about item functioning is carried across each domain rotation, with the choice of which items to carry forward being based on the most accurate item parameter estimation (occurring when a construct is measured as a major domain). The set of items that are carried forward in the rotation represents the construct. In this way, the notion of trend is defined both by the coverage of the construct and by the statistical methodology employed.

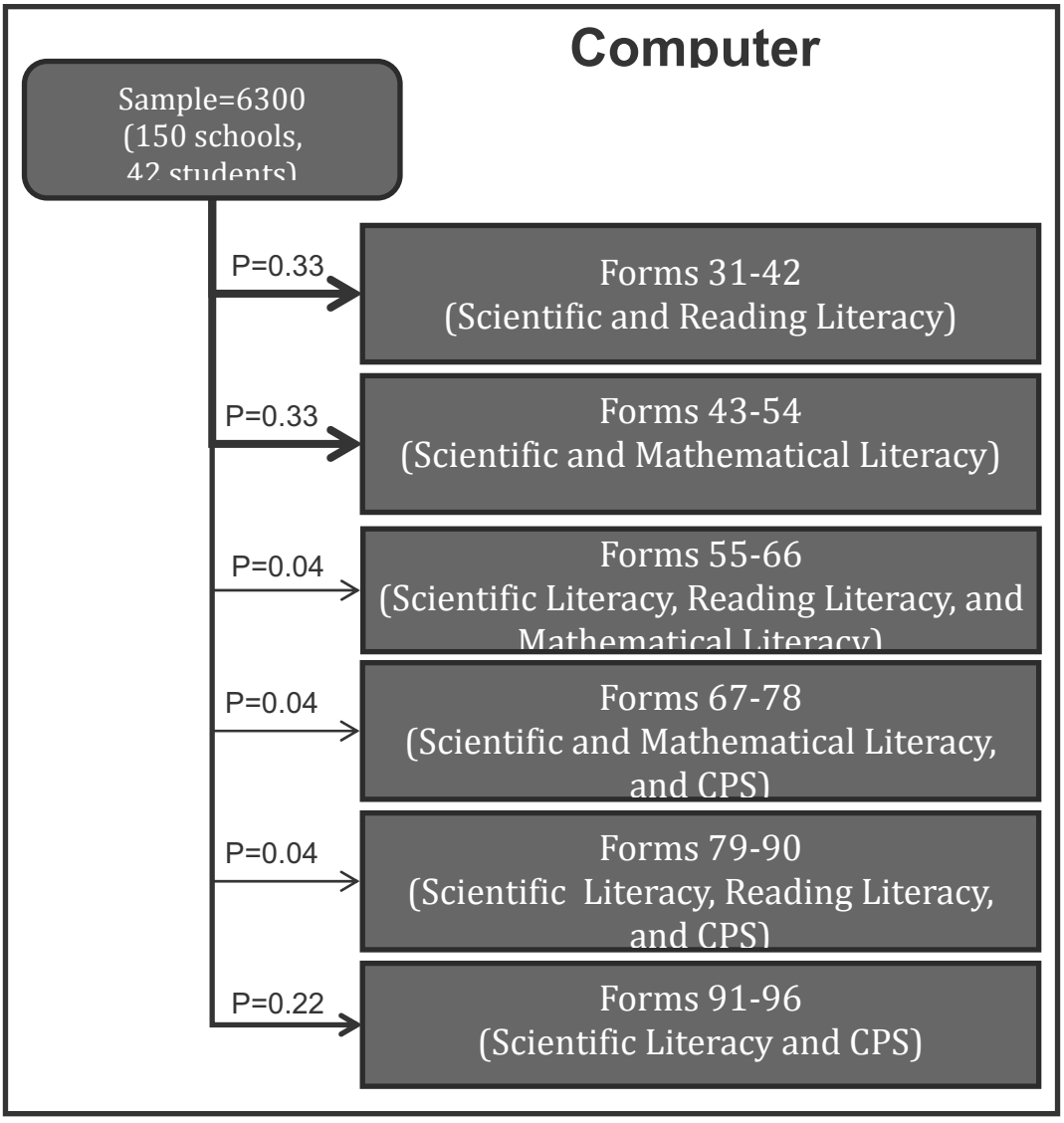
Main Study Assessment Design

The Main Study assessment design for PISA 2015 covers the domains of reading, mathematical and scientific literacy as computer-based and paper-based designs with the computer-based design adding the fourth domain of collaborative problem solving. A computer-based design for countries opting out of the assessment of Collaborative Problem Solving is described as part of the next section. These designs require participating countries to sample a minimum of 150 schools representing their national population of 15-year-old students. Countries taking the computer-based assessment will need to sample 42 students from each of 150 schools for a total sample of 6,300 students while countries taking the paper-based assessment will need to sample 35 students from each of 150 schools for a total sample of 5,250. It is important to understand that 88% to 92% of students will receive a form that consists of four 30-minute clusters (or sets of tasks) assembled from two domains, resulting in one hour of assessment time per domain, with a total of two hours of testing time per student. An additional 8% to 12% of students will receive forms that consist of four 30-minute clusters covering three of the four core domains. Scientific Literacy is included in each of these forms.

Main Study CBA Assessment Design

This design includes six intact clusters from each of the trend domains of Scientific, Reading and Mathematical Literacy based to the extent possible on the assessment cycle when each was the major domain – 2006 for Scientific Literacy, 2009 for Reading Literacy and 2012 for Mathematical Literacy. Additionally, it also includes six clusters of new Scientific Literacy and 3 clusters of Collaborative Problem Solving materials based on the new 2015 Frameworks. The six intact clusters will provide trend information for Mathematical and Reading Literacy. The six intact clusters of Scientific Literacy tasks will carry not only the trend information but also link to the new items developed to reflect the 2015 framework – this was done as part of the Field Trial analysis, including the mode study. In addition, three clusters of CPS items will be assembled for the Main Study. These materials will be organized according to the scheme shown in Figure 3 below (Forms 31-96).

Figure 3. Main Study Computer-Based Assessment Design



As reflected in Figure 3, there are 30 different test forms that combine two of the four domains – 88 percent of students receive one of these forms. These combinations include: *i)* Scientific and Reading Literacy (Forms 31-42), *ii)* Scientific and Mathematical Literacy (Forms 43-54), and *iii)* Scientific Literacy and Collaborative Problem Solving (Forms 91-96). In these test forms, students take one hour of Scientific Literacy (one cluster each of trend and new science) plus one hour of a another domain – Reading, Mathematical Literacy or Collaborative Problem Solving. These 30 test forms provide strong pairwise covariance information between Scientific Literacy and each of the three other domains.

In addition, there are 36 additional forms providing covariance information about three of the four domains. 12 percent of students receive one of these forms. In these forms, students receive one hour of Scientific Literacy plus two 30-minute clusters of items from each of the other three domains. These combinations are: *iv)* Scientific, Reading and Mathematical Literacy (Forms 55-66); *v)* Scientific and Mathematical Literacy and Collaborative Problem Solving (Forms 67-78), *vi)* Scientific and Reading Literacy and Collaborative Problem Solving (Forms 79-90). It is important to note that these three-domain test forms will ensure that all covariance estimates among the four domains are indeed based on the joint assessment of the domains.

As Scientific Literacy is the major domain for 2015, it is paired with one or two of the other three domains, and each of the different combinations of domains is balanced in terms of position to provide important covariance information. The design also reflects the fact that the random assignment of a form within a school follows a specific probability. According to this design, 33% of students within each school will be assigned to one of 12 Scientific and Reading Literacy test forms. Another 33% will be assigned to one of 12 Scientific and Mathematical Literacy test forms. In addition, some 22% of the sampled students within each school will be assigned to one of the 12 Scientific Literacy and CPS test forms. To provide additional covariance information, 4% percent of students will be assigned to one of 12 Scientific Literacy, Mathematical Literacy and CPS test forms; 4% to one of 12 Scientific Literacy, Reading Literacy and CPS test forms; and 4% to one of 12 Reading, Mathematical, and Scientific Literacy test forms[1].

The rotation of clusters – which identifies the form to be received by the respondent – will occur in a multi-step process that will take place when students are sampled.

STEP 1: Assignment of the base test form

The first step will be the assignment of base test forms. This assignment will be based on the 2-digit random number identified as "CC". This number will range from 31-96 and is directly linked to a specific base test form that is shown in Figure 4. These base test forms are useful in identifying the actual location and clusters for Math and Reading but only identify the location of Science clusters (i.e., which Science clusters are not yet assigned; these are only identified as "S"). The probability of assignment of each form type is also shown in column "Probability" and varies from 33% to 4% according to the combination of domains.

[1] These percentages are based on random assignment of test forms to students across schools. Each student in each classroom has a real probability of receiving any of the forms.

Figure 4. Main Study Computer-Based Assessment Base Test Forms

Probability of assignment	Base Test Forms	Cluster 1	Cluster 2	Cluster 3	Cluster 4
33%	31	S	S	R1	R2
	32	S	S	R2	R3
	33	S	S	R3	R4
	34	S	S	R4	R5
	35	S	S	R5	R6ab
	36	S	S	R6ab	R1
	37	R1	R3	S	S
	38	R2	R4	S	S
	39	R3	R5	S	S
	40	R4	R6ab	S	S
	41	R5	R1	S	S
	42	R6ab	R2	S	S
33%	43	S	S	M1	M2
	44	S	S	M2	M3
	45	S	S	M3	M4
	46	S	S	M4	M5
	47	S	S	M5	M6ab
	48	S	S	M6ab	M1
	49	M1	M3	S	S
	50	M2	M4	S	S
	51	M3	M5	S	S
	52	M4	M6ab	S	S
	53	M5	M1	S	S
	54	M6ab	M2	S	S
4%	55	S	S	M1	R1
	56	S	S	R2	M2
	57	S	S	M3	R3
	58	S	S	R4	M4
	59	S	S	M5	R5
	60	S	S	R6ab	M6ab
	61	R1	M1	S	S
	62	M2	R2	S	S
	63	R3	M3	S	S
	64	M4	R4	S	S
	65	R5	M5	S	S
	66	M6ab	R6ab	S	S

Probability of assignment	Base Test Forms	Cluster 1	Cluster 2	Cluster 3	Cluster 4
4%	67	S	S	C1	M1
	68	S	S	M2	C2
	69	S	S	C3	M3
	70	S	S	M4	C3
	71	S	S	C2	M5
	72	S	S	M6ab	C1
	73	M1	C2	S	S
	74	C3	M2	S	S
	75	M3	C1	S	S
	76	C1	M4	S	S
	77	M5	C3	S	S
	78	C2	M6ab	S	S
4%	79	S	S	R1	C1
	80	S	S	C2	R2
	81	S	S	R3	C3
	82	S	S	C3	R4
	83	S	S	R5	C2
	84	S	S	C1	R6ab
	85	C2	R1	S	S
	86	R2	C3	S	S
	87	C1	R3	S	S
	88	R4	C1	S	S
	89	C3	R5	S	S
	90	R6ab	C2	S	S
22%	91	S	S	C1	C2
	92	S	S	C2	C3
	93	S	S	C3	C1
	94	C2	C1	S	S
	95	C3	C2	S	S
	96	C1	C3	S	S

Where:

- *R1-R6* represent Reading clusters in computer (Trend)
- *M1-M6* represent Math clusters in computer (Trend)
- *S* represents Science clusters in computer (Trend and New)
- *C1-C3* represent Collaborative Problem Solving clusters in computer (New)
- *a* represents standard clusters and *b* represents easier clusters

STEP 2: Assignment of science cluster pairs

The second step will be the assignment of science clusters. This assignment will be based on the 1-digit random number , "S". This number will range from 1-6, will be uniformly distributed, and will be used in combination with the base form (e.g., selected by the first 2-digit random number) to identify which combination of Science clusters a student will receive. Figure 5 shows the lookup table where the 31-96 base forms are identified by the rows and the 1-6 lookup numbers are identified by the columns. The combination of these two numbers will identify which of the 36 possible combinations of science clusters will be inserted into the assigned base test form.

Figure 5. Lookup Table for Random Number "S"

Base Form (CC)	Random number (S)					
	1	2	3	4	5	6
31	1	13	6	9	22	25
32	2	16	12	10	31	32
33	11	5	17	14	26	29
34	35	4	7	19	23	30
35	34	15	8	20	24	28
36	3	36	18	21	27	33
37	35	4	7	19	23	30
38	34	15	8	20	24	28
39	3	36	18	21	27	33
40	1	13	6	9	22	25
41	2	16	12	10	31	32
42	11	5	17	14	26	29
43	1	13	6	9	22	25
44	2	16	12	10	31	32
45	11	5	17	14	26	29
46	35	4	7	19	23	30
47	34	15	8	20	24	28
48	3	36	18	21	27	33
49	35	4	7	19	23	30
50	34	15	8	20	24	28
51	3	36	18	21	27	33
52	1	13	6	9	22	25
53	2	16	12	10	31	32
54	11	5	17	14	26	29
55	1	13	6	9	22	25
56	2	16	12	10	31	32
57	11	5	17	14	26	29
58	35	4	7	19	23	30
59	34	15	8	20	24	28
60	3	36	18	21	27	33
61	35	4	7	19	23	30
62	34	15	8	20	24	28
63	3	36	18	21	27	33

Base Form (CC)	Random number (S)					
	1	2	3	4	5	6
64	1	13	6	9	22	25
65	2	16	12	10	31	32
66	11	5	17	14	26	29
67	1	13	6	9	22	25
68	2	16	12	10	31	32
69	11	5	17	14	26	29
70	35	4	7	19	23	30
71	34	15	8	20	24	28
72	3	36	18	21	27	33
73	35	4	7	19	23	30
74	34	15	8	20	24	28
75	3	36	18	21	27	33
76	1	13	6	9	22	25
77	2	16	12	10	31	32
78	11	5	17	14	26	29
79	1	13	6	9	22	25
80	2	16	12	10	31	32
81	11	5	17	14	26	29
82	35	4	7	19	23	30
83	34	15	8	20	24	28
84	3	36	18	21	27	33
85	35	4	7	19	23	30
86	34	15	8	20	24	28
87	3	36	18	21	27	33
88	1	13	6	9	22	25
89	2	16	12	10	31	32
90	11	5	17	14	26	29
91	1	13	6	9	22	25
92	2	16	12	10	31	32
93	11	5	17	14	26	29
94	35	4	7	19	23	30
95	34	15	8	20	24	28
96	3	36	18	21	27	33

Using the two random numbers, base test form identified by the 2-digit number as rows and the 1-digit random number as columns, Figure 6 shows the 36 possible Science Clusters combination (e.g., rotation of Science clusters S1 to S12) that will be inserted into the selected base test form.

Figure 6. Main Study Computer-Based Assessment Assignment of Science Cluster Pairs

Science Cluster Combination		
N	S	S
1	S1	S7
2	S1	S10
3	S2	S8
4	S3	S9
5	S3	S12
6	S4	S7
7	S4	S10
8	S5	S11
9	S6	S12
10	S7	S6
11	S8	S1
12	S8	S5
13	S9	S2
14	S9	S6
15	S10	S3
16	S11	S2
17	S11	S4
18	S12	S5

Science Cluster Combination		
N	S	S
19	S7	S8
20	S7	S9
21	S7	S11
22	S8	S10
23	S8	S12
24	S9	S8
25	S9	S11
26	S10	S7
27	S10	S9
28	S10	S12
29	S11	S8
30	S11	S10
31	S12	S7
32	S12	S9
33	S12	S11
34	S2	S4
35	S5	S1
36	S6	S3

Coding of open ended questions

Throughout all large-scale assessments, the essential activities associated with maintaining scoring consistency are the same. Typically, the required procedure to monitor scoring reliability is to have a portion of the items double scored (i.e., scored independently by two different scorers), and then compare the resulting scores to measure agreement.

The goals of such procedures are to ensure the accuracy and reliability of scoring within countries and identify scoring inconsistencies or problems early in the process so that interventions can be applied and problems resolved as soon as possible. In general, inconsistencies or problems arise because scorers misunderstand general scoring guidelines and/or a specific rule relating to a particular item. Double scoring yields a reliability measure that represents the extent to which two scorers agree on how a particular response should be scored, and thus demonstrates how comparably the scoring guideline is being interpreted and applied. The goal in PISA is to reach a within-country inter-rater reliability of 0.92 (92% agreement) across all items, with at least 85% agreement for each item.

When discrepancies among scorers occur, experience has shown that they fall into two distinct classes.

- _The first type of discrepancy reveals a consistent bias on the part of one scorer — for example, Scorer 1 may be consistently scoring more leniently than other scorers. To detect and address such discrepancies, countries are required to monitor reliability at key points during the scoring process so that problematic scorers can be retrained or, if necessary, dismissed.
- _The second type of scoring problem that can be revealed through analysis of the rescoring process is more challenging to address. This occurs when the scoring results reveal general inconsistencies between the scorers, with no pattern that can be attributed to one scorer or the other. This is relatively rare, but when it occurs it is generally the result of a problem with an item or an error in the scoring guides. One procedure for addressing this situation involves conducting a review of all inconsistently scored responses to determine if there is a systematic pattern and, if one is found, having those items rescored. Additionally, the scoring guides for such items can be

revised to clarify any issue identified as causing inconsistent scoring. When a specific problem cannot be determined, unique item parameters may be required for one or more countries to reflect this ambiguity in scoring.

For PISA 2015, the Open-Ended Coding System (OECS) tool was used to support the scoring of open-ended computer-based items. Once scoring is complete for each item, data can be imported into the OECS to be integrated across scorers and the OECS will produce summary reports that countries use to examine scoring reliability.

Consistent scoring within countries does not guarantee comparable scoring across countries. Therefore, in addition to examining the accuracy and reliability of scoring within countries, it is also important to check that scorers across countries are consistently applying the same criteria. This aspect is particularly important with the increased number of countries and languages participating in international surveys such as PISA. The approach used in PISA 2015, based on that used in the Adult Literacy and Lifeskills Survey (ALL) and PIAAC, involves the use of anchor responses. Anchor responses are a common set of items and associated responses prepared by test developers. For PISA 2015, anchor responses were provided for New Science items. Since it is the open-ended items that are the focus of this process, anchor responses for PISA are created using only those item types and it is these responses that are double scored. Because anchor responses are provided in English, scoring teams in each country are asked to designate two bilingual scorers responsible for the double-scoring process.

As was the case in 2015, the scoring design will dictate when in the scoring process the anchor responses should be scored as well as which of the two scorers should be the first scorer and which should be the second. The OECS simplifies this process for countries by organising responses according to this scoring design. Scorers receive a PDF file for each item

A part of the PISA 2015 CBA open-ended responses required human coding, a process that took place with the aid of the Open-Ended Coding System (OECS) – a computer tool that supported coders in coding the computer-captured responses according. During coding, the responses were organized and distributed to coders following a pre-determined coding design and the specific coding guides. Coding reliability of all computer based open-ended questions that required human coding was evaluated on a subset of responses. This tool reduced the coding administrator's burden of managing coding activities and ensured that the coding design requirements were met. Coders received a PDF file that contained all the responses assigned to the coder, including anchor responses (in English), if appropriate, with one response per page. Each page displayed a part of the question, the individual responses, and the acceptable coding categories for each question. The coder clicked the circle next to the selected code, which was then saved within the file.

During coding activities, OECS generate three types of reliability reports: i) proportion agreement, ii) coding category distributions, and iii) deferred and missing responses. The "proportion agreement" report shows the average agreement per item across all coders as well as average agreement per coder. The "coding category distributions" report is used to compare coding distributions across coders for each item. That is, it compares tallies or counts of the response categories for each item across all coders and identifies whether the distribution of response categories given by any single coder differs from the average distribution across coders. The "deferred and uncoded responses" report provides counts of responses that have not been coded during the coding process, being flagged as either deferred or uncoded to be scored later.

The human coding was conducted in the time frame allocated, and rare technical problems were addressed immediately so that a complete set of human coded responses was available for the analyses described below. A report was provided separately to each country summarising the overall level of agreement reached for each scale. Countries also are able to review level of agreement reached on each item.

Psychometric modeling

The reanalysis of PISA data from prior cycles (2000-2012) aimed to stabilize the trend measure and to ensure its quality. With PISA 2015 introducing CBA as the main mode of assessment, the concern was that it might influence the item parameter estimates for the linking items. Moreover, some linking items might not work equally well for all of the populations assessed in PISA 2015. Utilizing linking items that do not work equally well across subgroups (i.e., are not measuring the same construct in all populations) reduces the comparability of the trend measure. These linking items needed to be identified and excluded from the Main Study item pool. However, given the new scaling approach for PISA 2015 (the combined Rasch and 2PL model), it might still be possible to retain a larger share of these items.

Results from prior analyses (PISA 2000-2012) were replicated and then reexamined using the combined Rasch and 2PL model. The reanalysis produced a common parameter for each of the previously used items that is contained in the databases from PISA 2000 to 2012 and that can be considered as a starting point of scaling the PISA 2015 data. This ensures a solid database of item characteristics on a common international scale based on the past frameworks of each domain.

PISA has collected data in representative samples of 15-year old-students around the world every three years since 2000. In each of these five cycles (2000, 2003, 2006, 2009, 2012), both OECD- and partner- countries participated, resulting in almost 300 cohorts defined by assessment year and country. Many of the OECD countries as well as a substantial number of partner countries participated in each of the five PISA cycles so far.

In an effort to utilize the complete evidence on item functioning and scale coverage of the task material used in PISA, we compiled a database that merged all five cycles and all countries. This yielded a file that contains roughly 2 million student records assigned to each of the cycles by country/jurisdiction combinations. We utilized state-of-the-art multiple group IRT (Bock & Zimowski, 1997, Yamamoto & Mazzeo, 1992, von Davier & Yamamoto, 2004; von Davier & von Davier, 2007, Mazzeo & von Davier, 2008, 2013, Weeks, von Davier & Yamamoto, 2013) to specify a model that allows linking all items across all PISA cycles by country combinations.

Following four distinct models were considered; Rasch/PCM, 2PL/GPCM, Rasch/PCM and 2PL/GPCM , RaschPCM and 2PL/GPCM with item by cohort interactions.

Table 2 summarizes the improvement in item fit for the domains of mathematics, reading, and science. The table shows the results for the Rasch/PCM model, the 2PL/GPCM and the “hybrid” Rasch/2PL/GPCM model, with one set of item parameters for all countries, and a model that accounts for item-by-country (IBCI) interactions by releasing some country-specific parameters. These results are based on all cycles from 2000-2012 combined for the three main domains. In each domain, the IBCI model fits best (as characterized by the BIC), followed by the 2PL/GPCM, hybrid, and Rasch/PCM models. This can also be seen in the concomitant decrease in the number of items-by-country-by-cycle with RMSD values greater than 0.15. Approximately 3% of the items in math, 7% of the items in reading, and 6% of the items in science did not fit the Rasch model in one or more countries. On the other hand, around 1% of the items exhibit misfit in reading for the IBCI model and less than 0.1% of the items exhibit misfit in math and science under the IBCI model. For all subsequent analyses, the item parameter estimates from the IBCI model were used.

Table 2: Changes in Model Fit Summary

		Rasch/PCM	2PL/GPCM	Hybrid	IBCI
Math	# of item-country-cycle deviations	549	397	415	4
	Bayesian Information Criterion (BIC)	26400730	26118134	26175012	25946516
Reading	# of item-country-cycle deviations	1233	960	962	250
	Bayesian Information Criterion (BIC)	30968125	30675531	30691983	30472304
Science	# of item-country-cycle deviations	921	717	708	8
	Bayesian Information Criterion (BIC)	29908518	29585732	29591677	29302806

Total item-country-cycle values: Math = 15,795, Reading = 18,603, Science = 16,223
Deviations defined as RMSE values > 0.15

IRT Scaling of FT data

The new science items developed for 2015 are based on a revised science assessment framework. These new items exist in the CBA mode only because PISA 2015 represents a shift from a paper-based to a computer-based survey. The IRT scaling of the new science items was straightforward, but some changes to the scoring of two of the CPS units were necessary before the data could be used for IRT scaling. The CPS scale consists of seven units, which in turn comprise 165 items that can be used for the IRT scaling. The CPS units are based on simulated conversations with one or more computer-based agents that are designed to provide a virtual collaborative conversation. Test takers have to choose an optimal sentence from a multiple-choice list to go through the conversation with agents, or choose one or more actions programmed in the unit. It was found that data from two units had dependencies in the responses due to different paths that could be taken by students through the simulated chat. Therefore, the CPS chat items that showed this kind of dependency were combined into “composite items” by summing the responses for the different paths respondents could take. With this approach it was determined that each path-based response string could be scored to provide valid data and introduced into the IRT analysis. Table 3 gives an overview of the combination for composite items. The table shows how chat-based items in two CPS units (101 and 105) were combined into composite items in order to generate polytomous items for the purpose of reducing issues with local dependencies. Details about the items going into this rescoring can be found in the databases containing country-specific data as well as variable and value labels.

Table 3: Combination of CPS Items to Achieve Fair Scoring

New Item ID for Composite Items	*Combinations of CPS Items*
CC101201C	CC101201+CC101202
CC101203C	CC101203+CC101204+CC101205
CC101206C	CC101206+CC101207
CC101301C	CC101301+CC101302+CC101303
CC101304C	CC101304+CC101305
CC101307C	CC101307+CC101308+CC101309A+CC101309B+ CC101310+CC101311+ CC101312A
CC101312BC	CC101312B+CC101313
CC101317C	CC101317+CC101318+CC101319
CC105103C	CC105103+CC105104
CC105105C	CC105105+CC105106+CC105107
CC105201C	CC105201+CC105202
CC105208C	CC105208+CC105209+CC105210
CC105212C	CC105212+CC105213
CC105304C	CC105304+CC105305

Across the seven units and 165 chat/selection items, 10 had to be dropped during the analysis due either to no response variance or the presence of technical issues. Table 4 provides a list of these items.

Table 4: CPS Items Excluded from the IRT Analyses based on Classical Item Analyses

Domain	*Item*	*Mode of Administration*
CPS (10 items)	CC100403	CBA
	CC102202	CBA
	CC102208	CBA
	CC102212	CBA
	CC104303	CBA
	CC105108	CBA
	CC105303	CBA
	CC105403	CBA
	CC105405	CBA
	CC106306	CBA

The new science items were scaled together with the trend science items. In addition to 91 trend items, the science domain consists of 213 new items resulting in a total of 304 items. For science, five new items had to be excluded from the IRT analyses due to either a lack of response variance, or very low or even negative item total correlations. Table 5 gives an overview of these items. Most items showed no obvious defects, and the initial IRT scaling was conducted on 208 new science items.

Table 5: New Science Items Excluded from the IRT Analyses Based on Classical Item Analyses

Domain	*Item*	*Mode of Administration*
Science (5 items)	CS644Q02S	CBA
	CS638Q03S	CBA
	CS656Q06S	CBA
	CS661Q07S	CBA
	CS650Q03S	CBA

For science, data from 53 countries were received in time to be included in the IRT analyses, and for CPS data from 42 countries were received. For the IRT analyses, the sample was divided by country and language, resulting in 70 country/language groups for science, and 55 country/language groups for CPS.

While the item parameters of trend items were fixed to those obtained from the reanalyses of previous PISA cycles (historical data), the new science and CPS items had to be scaled based solely on the Field Trial data. For the CPS scale, both a multigroup Rasch model/PCM was estimated as well as a multigroup 2PL model/GPCM. For the new science items and CPS items a concurrent calibration was used to evaluate whether items are working equally across country/language groups or whether there are item-by-country/language interactions. Both model approaches were compared to each other. Item parameters for new science and CPS items that are provided for the countries and used to identify items for the Main Study are based on the 2PL model due to the improved model-data fit and because more information (with regard to slope parameters) about each single item is provided. These item parameters were also used for generating a standardized proxy (EAP) estimate standardized within countries that is available in the data delivery to countries).

Table 6: Comparison of Rasch model/PCM and 2PL model/GPCM for new items

	Likelihood:	*A-penalty*	*AIC*	*B-penalty*	*BIC*
CPS					
RM/PCM	-985477.57	686	1971641.15	3877.09	1974832.24
2PL/GPCM	-971208.69	994	1943411.38	5617.83	1948035.21
Science					
RM/PCM	-2215483.30	1266	4432232.60	7406.46	4438373.06
2PL/GPCM	-2192778.99	1698	4387255.97	9933.78	4395491.75

The item fit of the new science items and the CPS items was evaluated with regard to the concurrent calibration. Table 7 gives an overview of the percentage of RMSD and MD that was considered to be deviant using a rather strict criterion of RMSD > 0.20 and, MD > 0.20 and <−0.20.

Table 7: RMSD and MD Deviations of Trend Items and New Items overall Countries/Languages

	Science-new	*CPS*
	new items	new items
Percent of	CBA	CBA
RMSD > 0.20	0.93%	0.94%
MD > 0.20 and < −0.20	0.49%	0.51%

Table 7 shows that item deviations for new science and CPS items are generally small. The deviation frequencies were not found to be substantially higher for any one particular country or language

group. The results illustrate that the items show a good fit when using the same item parameters across different countries and languages. Moreover, both scales show sufficient IRT-based (marginal) reliabilities (Sireci, Thissen, & Wainer, 1991; Wainer, Bradlow, & Wang, 2007) with 0.80 for science (based on trend and new items) and 0.88 for CPS.

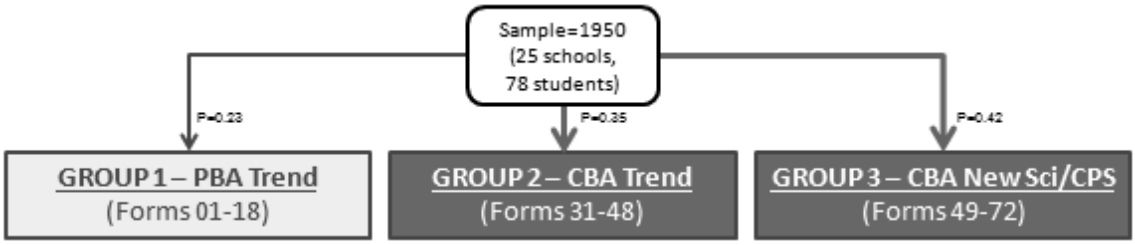
What we found from 2015 PISA FT and from 2012 PIAAC on CBA

The Field Trial analyses are based on data from 53 countries received by 31 July 2014. Of these 53 countries, 45 were considered computer-based assessment (CBA) countries, meaning that they implemented the survey using both paper and computer, while eight countries implemented the survey as a paper-based assessment (PBA). Altogether 120,303 cases including all multiple languages were included.

Almost all countries met the sample size requirements for the major assessment language. More specifically, 50 out of 53 countries produced more than 95% of required sample sizes: 1,950 for CBA countries, 900 for PBA countries, and 1,750 for CBA without CPS countries.

The CBA design specified that 1,950 students from multiple schools were randomly assigned to one of three groups according to prescribed probabilities: 1) 23% of the sampled students would be administered trend items through paper administration, referred to as Group 1-PBA Trend, 2) 35% would be administered trend items in computer administration, referred to as Group 2-CBA Trend, and 3) 42% would be administered new items through computer administration, referred to as Group 3-CBA and New Science/CPS. The successful implementation of the design above rested on strict random assignment of students to one of the three groups identified in Figure below.

In order to examine the transition to computer delivery, a set of 18 paper-based forms covering the domains of reading, mathematics, and science were constructed for Group 1 *PBA-Trend* from items administered in the past PISA surveys. A set of tasks "identical" to those contained in the 18 paper-based forms were then adapted and authored for computer administration, yielding 18 equivalent computer-based test forms. In addition, there were 12 computer-based test forms consisting of the new 2015 science tasks and 12 new test forms combining those 2015 science tasks with the new CPS tasks. The schematic for the CBA design is shown in Figure below.



Information yield by the CBA platform

Two key features of the computer delivery platform, not available in paper-based assessments, are the availability of detailed timing and process sequence information. Response times are recorded for each item in milliseconds; hence, they allow for precise, timing-related analyses. For instance, these data can be used to identify rapid responders (Wise & DeMars, 2005) and/or potential administration issues (e.g., groups of examinees who take substantively longer to complete the test than expected). Timing information can also be used to address issues of speededness and fatigue, between-country differences in allocated time, position effects, and interaction effects with variables such as examinee performance. Sequence information, on the other hand, can provide insights into how examinees progress through a set of items, including the number of times that an item is revisited, item sets that are skipped, and items that

are truly not reached. This information can be used in conjunction with the timing data to identify potentially problematic items, units, and/or clusters.

The Field Trial data indicate that timing data as well as process data have been successfully recorded for all data collections in the CBA countries. The available timing data was instrumental in evaluating the level of engagement and effort students spent over the course of the four 30-minute clusters. Analyses were conducted and results of these indicate that the CBA assessment provides valid information that can be used to evaluate response processes within and across countries.

Reduced position effects and improved omission rates in the CBA

Item position effects are a common issue in large assessment programs because substantial position effects can increase measurement error and introduce bias. The PISA 2015 Field Trial design balanced cluster position in order to monitor its impact on various measures. We examined the cluster position effects in terms of: 1) proportions correct, 2) response time, and 3) interactions by domain as well as by country.

In order to have a reference point to examine the magnitude of position effects, PISA 2009 data was examined in terms of proportions correct (P+) at the cluster level (PISA 2009 was the last time that science was the major domain). The values for PISA 2009 are shown in Table 8 below. In all content domains there is a decrease of 4 to 8 percentage points in the P+ values between positions 1 and 4. For the PISA 2015 Field Trial data (see Table 9), a smaller decrease of 1 to 3 percentage points in P+ values, compared to the 2009 values, is seen between positions 1 and 4.

Table 8: PISA 2009 PBA Proportions Correct Across Clusters and Across Countries

	Position 1	*Position 2*	*Position 3*	*Position 4*	*Position 4- Position 1*
Mathematics	0.411	0.401	0.384	0.371	-0.040
Reading	0.581	0.557	0.532	0.499	-0.083
Science	0.490	0.478	0.457	0.435	-0.055

Table 9: PISA 2015 CBA/PBA Proportions Correct Across Clusters and Across Countries

	Position 1	*Position 2*	*Position 3*	*Position 4*	*Position 4- Position 1*
Mathematics	0.439	0.454	0.432	0.427	-0.012
Reading	0.575	0.580	0.566	0.544	-0.031
Science	0.409	0.418	0.401	0.385	-0.024

Table 10: PISA 2015 CBA Cluster Timing Averaged Across Countries (in Minutes)

	Position 1	*Position 2*	*Position 3*	*Position 4*	*Position 4- Position 1**
Mathematics	19.76	18.84	17.26	16.03	-3.72
Reading	24.11	21.47	20.75	18.44	-5.67
Science	22.32	20.61	19.21	17.11	-5.21
CPS	22.67	21.26	20.96	18.79	-3.88
FL	19.49	14.49	NA	NA	-5.00

Note: * For financial literacy, the difference is taken between positions 1 and 2 because these instruments only had two clusters.

While the P+ values remain quite stable across positions for the 2015 Field Trial data, there is a notable decrease in the mean response times (around four to six minutes, i.e., nearly 20% reduction) for clusters administered in each of the four positions (see Table 10). These finding suggest that the decrease in response times over positions has little association with P+ values.

To further address the relationship between response time and examinee performance, we considered mean times grouped using a preliminary ability estimate or what we are calling a proxy. This ability estimate was derived based on the IRT scaling by domain that was carried out to evaluate the fit of trend item parameters. The IRT scaling in the three domains of math, reading, and science was used to generate an expected a posteriori (EAP) score that was averaged across domains to provide this preliminary ability estimate. This preliminary estimate is referred to as the ***proxy score***. This ***proxy score*** cannot be used for country comparisons since it is based on FT instrumentation which is preliminary, and sample sizes used in the FT do not allow sufficiently reliable estimates of country-level statistics. Table 11 below reports mean response times at three proxy score levels (low, moderate and high ability) based on an equal split into three groups (33.3% each). It is evident that low-ability students take several minutes less on average to complete a cluster. In some cases, the average differences between the low- and high-ability groups exceed five minutes.

Table 11: PISA 2015 Cluster Level Response Time by Proxy Level (in Minutes)

	Low Ability	*Moderate Ability*	*High Ability*
Mathematics	15.17	18.97	20.92
Reading	18.97	22.60	23.64
Science	17.45	21.31	22.86
CPS	19.69	22.23	22.77
Financial Literacy	15.76	19.97	22.11

The analysis of missing responses looks at the Field Trial data comparing the omission (student nonresponse) rate differences between CBA and PBA with the purpose to further examine the quality of data in the two modes. Items that were presented only in PBA or only in CBA mode as well as a very small number of items that exhibited technical issues were not included in this analysis. In total, 318 items were compared in terms of omission rates across CBA and PBA modes.

We defined the omission rate as the number of omitted responses divided by the total number of test takers who received the item in each country. A high omission rate is defined by omission rates larger than 10%. The effect of omissions was investigated by checking the differences between two approaches of calculating percent correct (P+) item score. These two approaches are referred to as the W rule (omitted responses treated as "wrong") and X rule (omitted responses excluded), which in effect differed by the inclusion or exclusion of omissions into the sample size incorporated in the calculation of P+ values. An omission effect is present when the P+ calculated under the W rule and the X rule, respectively, differs by more than 5%.

The correlation of omission rate and omission effect between CBA and PBA was 0.88 and 0.80, respectively. It was found that 13 countries had high omission effects in PBA, while only three countries were detected to have a high omission effect in CBA. Further, 23 countries were found to have more than 50 items that were labeled as high-omission-effect items in PBA, that is, the difference between *W* rule and *X* rule was over 5% and across at least 20 countries, while only nine countries were found under the same conditions in CBA. On average across countries, high omission effect for the CBA mode was 32.5 items, which is lower than the 44.4 items with high omission effect in the PBA mode.

Open response items accounted for the largest proportion of items with a high omission rate and impacted the *P+* value calculation the most. When counting the number of items that have omission rates higher than 10% across at least 20 countries, there were 47 CBA items, among which 98% were open-response items. In comparison, 54 items were labeled as having a high omission rate in PBA, all of which were open-ended items. The omission rate in all content domains was higher for PBA than CBA, except financial literacy, which had a 1.1% higher omission rate in CBA than PBA. Mathematics items took the highest proportion (approximate 35%) in both of the modes and were almost the same in CBA and PBA. Reading items were slightly higher in omission, by 2%, for PBA. Bigger differences were found for items in the domains of science (5% increase) and financial literacy (6% decrease) in PBA.

In conclusion, reductions in omission rate and omission effect were observed in CBA over PBA. In addition, the findings of higher omission rate and omission effect in PBA items compared to CBA, especially in open-response items, provides motivation to study the processing behavior of the students when transitioning to CBA. Process data may provide new insights into test-taker behavior during the test. It would be helpful to analyze process data stored in log files to further understand which student behaviors are related to correct, incorrect, and omitted responses and explore the patterns between omissions and item types as well.

Future directions of PISA

The future design and functionality of PISA is solely in the hands of OECD and PISA Governing Board and proposals consortium have made, so it is not certain about specific features or their timeline. However, as a contractor for both 2015 and 2018 contractors, I have a vantage point of view. It is clear that CBA is here to stay and measurement constructs in any domain will expand beyond PBA constructs. The future constructs may incorporate how students use skills to learn deeper and further in the domain. Such activities will include measuring what they have learned already as well as how they might be able to use in more realistic settings. Although, depth of knowledge being measured and realism of questions would introduce contextual factors that may interfere generalizability of measured skills. PISA items tend not to require special factual knowledge but higher organizational knowledge to utilize necessary and extra information to solve problems. CBA may require awareness of what he/she knows and does not know in order to pursue which direction to solve problems as information explorations take place.

Historical item construction criteria of strict adherence to over one third to be hand scored is somewhat arbitrary and ignores the advanced capability of computer scoring. It is expected that proportion of computer based scoring would increase as more computer based scoring algorithm mature. It has been believed among those who have been involved in the large scale assessment with emphasis on trend assessment that “to measure trend don’t change the measure”, meaning keep the intact clusters as large as possible across cycles. However, since 1990 we have been seeing the data in adult literacy surveys, cluster size can be fairly small to maintain consistent item parameters across cycles, certainly not as large as 30 minutes worth of sequence of items. If intact clusters can be as small as a set of items based on a single stem, more flexible item combinations can be possible and open a door to many different kind of adaptive testing.

Regularity of response time, number of actions and sequence of actions among students, language and countries are remarkable. Future analysis model would include additional information unique to CBA for estimating skill distribution of students population.

References

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of modern item response theory (pp. 433-448). New York, NY: Springer-Verlag.

Carstensen, C. H. (2012). Linking PISA competencies over three cycles – Results from Germany? In:

Gebhardt, E. & Adams, R.J. (2007). The influence of equating methodology on reported trends in PISA. Journal of Applied Measurement 8(3), 305‐322.

Mazzeo, J., & Davier, M. von (2008). Review of the Programme for International Student Assessment (PISA) Test Design: Recommendations for Fostering Stability in Assessment Results. doc.ref. EDU/PISA/GB(2008)28; Retrieved 12/12/2008 from http://www.oecd.org/dataoecd/44/49/41731967.pdf

Mazzeo, J. & von Davier, M. (2013). Linking Scales in International Large-Scale Assessments. In Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). Handbook International Large-scale Assessment: Background, Technical Issues, and Methods of Data Analysis. CRC Press (Chapman & Hall).

Prenzel, M., Kobarg, M., Schöps, K., & Rönnebeck, S. (Eds.) Research in the context of the Programme for International Student Assessment. Berlin: Springer.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28, 237-247.

Urbach, D. (2012). An investigation of Australian OECD Pisa Trend results. In: Prenzel, M., Kobarg, M.,Schöps, K., & Rönnebeck, S. (Eds.) Research in the context of the Programme for InternationalStudent Assessment. Berlin: Springer.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: an extension of the generalized partial credit model. Applied Psychological Measurement, 28(6), 389-406.

von Davier, M. & von Davier, A. (2007). A Unified Approach to IRT Scale Linkage and Scale Transformations. Methodology, 3, 3, 115-124.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). Testlet response theory and its applications. New York, NY: Cambridge University Press.

Weeks, J., Yamamoto, K., & von Davier, M. (2013). Design considerations for the program for international student assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis. Boca Raton, FL: CRC Press.

Wise, S. L. & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. Educational Assessment, 10, 1-17.

Yamamoto, K., & Mazzeo, J. (1992). Item response theory scale linkage in NAEP. Journal of Educational Statistics, 17(2), 155-173.

【報告１】

試験の日本的風土

東京工業大学大学院社会理工学研究科教授
前川眞一

○司会　そろそろ再開したいと思います。今日はシンポジウムのテーマにかかわりまして、4 つの報告を準備させていただいております。

まず東京工業大学の前川眞一先生より、「試験の日本的風土」ということで講演をお願いします。

前川先生、よろしくお願いします。

○前川氏　初めまして。東京工業大学の前川です。本日は「試験の日本的風土」という題でお話しさせていただきます。

この話、ここ 10 年ぐらいいろいろな学会で話したりしているので、お聞きになった方も何人かおられると思いますけれども、もう一度お話しさせていただきます。

日本の試験的風土

内容ですけれども、日本の試験的風土と呼ばれているもの、日本で行われている試験、テストの特徴、日本的試験文化と呼んだり日本的テスト文化と呼んだりしているものですが、それがどういうものなのか、もう一回改めて書き出して認識しよう、そういう話です。

「試験」と言われているものは日本語では「検査」「試験」「考査」「テスト」それから統計的検定とか漢字検定とかウルトラマン検定とか言われている、あれはみんな同じものです。英語で言うと「クイズ」「テスト」「イグザミネーション」の 3 つがそれに当たるということで、例えばインターネットで「テスト」と「イグザミネーション」とどう違うのかを引いてみると、ある大学のホームページが出てきて「うちの大学ではこう使っています」というような説明が見つかると思います。イグザミネーションがどちらかというと難しい方、テストが簡単な方、クイズというのはもう一つ簡単なもの、そういう感じです。

日本語はどう使い分けているか、ちょっとよくわかりませんけれども、学力検査、学力試験、学力考査、学力テスト、は使いますよね。しかし性格テスト、性格検査とは言うけれども、性格試験とは余り言わない。何か微妙な使い方があるのだと思いますけれども、いずれにしても、テスト、試験というものです。

日本的テスト文化にはどういうものがあるのか、箇条書きにしてみました。

ここで言っている日本のテストというのは入試を含む大規模なテストの話です。司法試験とか大学の入学試験とか、そういうもののことを考えています。

年に 1 度、同一問題で試験を一斉に実施する。

新しい問題が毎年出てくる。先ほど山本先生からフィールドテストという話がありましたけれども、日本のテストはフィールドテストを一切やりません。それが特徴の 1 つです。

それから、試験問題を公開します。さっきの PISA の問題とか、皆さんご存じの TOEIC とか TOEFL とかあの類のテスト、試験問題は公開されませんよね。日本は逆に試験問題を公開します。

それから、小さな問題がいっぱい寄せ集まって広い分野を測るという感じではなくて、問題の一まとまり、センター試験では「大問」と呼んでいるんですが、大問形式の問題が幾つか出るという形が多く使われている。

それから、ここはちょっと何のことわからないかもしれませんけれども、問題を作る人と問題の素材を集めてテストの冊子を編集する人が同じである。

それから、先ほどから心理測定とかサイコメトリシャンとかいう話か出ていますけれども、そういういわゆる統計的なテスト・心理測定をやっている人は問題作成に参加しないという特徴があります。

最後に出てくる点数なんですけれども、例えば 20 題の問題から成る試験を受けました。点数は 1 題の配点を 5 として 100 点満点にします。何題正答出来たかに配点を掛けた形で点数が出ます。それをここでは素点と呼んでいます。配点を利用した素点を使って、だから 20 題で配点が 5 だったら 100 点満点。0 点と満点が決まっているような、そういう尺度で成績が出てくるということです。だから、同じ 100 点満点のテストでも難しいテストの 100 点とやさしい問題の 100 点は本当は違うはずなのに、日本のテストは余りそこのところを考えていない、そういう特徴があります。

なぜこんなふうになっているのかということですが、赤いところが、なぜこうなったかという理由です。

同じ試験問題を年に 1 回、一斉に実施するということに関しては、問題が違ったら不公平ではないか、条件が違ったら不公平ではないかという感覚があるのだと思います。

それから、新作問題のみでの試験の実施にかんしてみ、プリテストとかフィールドテストとか言われているもの、前もって試験問題をどこかで試して良い問題だけを本番で使うという作業は、不公平だと思われている。

試験問題を公開することに関しては、1 つは教育的見地から、こういう問題を出したんだよというのを見せるということですが、他にも、非公開にしたい大規模試験はいっぱいあるんですけれども、日本の情報公開法とか規制緩和の影響で、それが今、出来ない状況にあるということ。

あと、大問形式の利用というのは、小さい問題で沢山幅広く測っていくためには、試験問題作成のブループリンドというかスペックというか、そういうものが必要なんですけれども、それを厳密に作って細かい問題をたくさん出すような形で出来ているかどうかという疑問があります。あと、大問形式だと多肢選択形式でも思考力を測る問題がつくれるのではないか、そういうことを考えているところがあります。

試験問題、特に入試問題に関しては大学の先生が作りますから、「おれたちが作った問題はいい問題だ。他のやつが文句を言う筋合いはないんだよ」みたいな形でつくっているというのが権威主義というところ。

あと、正答数に重みをかけて満点にするというのは、もうちょっと別のやり方、もっといいやり方があるんだよという心理測定学が主張することに対する無関心とか､それから、そうやって心理測定学者がつくった尺度得点に対する不信感みたいなものがある。こうい

うものが大体理由だと思われます。

日本の試験的風土の逆＝世界標準

では、その逆を考えてみましょうということですが、1 年間に 1 度、同一問題で試験を一斉に実施することの反対は、年に複数回、異なる問題で分散的に実施する。例えば IRT を使うとか CBT を使うとか、そういう話ですね。それから、新作問題のみで試験を実施することの反対は、統計的性質、例えば難易度みたいなもののわかっている問題のみで、良い問題だけでテスト問題を構成するという話。

それから公開に関しては、問題は非公開。ただし、個人には成績を送りますよという話。

大問形式の利用の反対は、細かいスペックに基づく広い分野を測定する、独立項目をたくさん利用する。別に独立項目でなくても、大問形式で沢山やっても問題はないんですけれども、問題数という点から独立項目を多用する。

それから、問題作成とテスト編集の融合の反対は、テスト理論を用いた難易度のコントロール。

素点の利用の反対は、テスト理論を用いた比較可能な尺度得点の利用ということになります。

この日本的試験文化の逆とここに書きましたものが、実は、まあ世界標準であろうと思われています。

「世界標準」とここに書きましたけれども、日本以外の国が全部こうやっているかというと、必ずしもそうではなくて、今はやりの海洋国家と言われている国がありますよね。アメリカとかオランダとかイギリスとか。あのあたりは明らかにこの日本文化の逆をやっている。海洋国家というのは戦争に負けたことがない国らしいので、どうも世界標準ということらしいんですが、日本的試験文化の逆が世界標準ということで、では、どう考えるかということですが、年に複数回、異なる問題で分散実施をする。年に複数回、受験機会が与えられることは、受験生にとっては非常にありがたいことではないかと私などは思います。ただ、作る方にとっては事務は煩雑化するし、問題が違ったら今と同じような形で作っていたのでは公平性とかセキュリティの問題が起きるのではないか。

コンピュータを使って、それから先ほどオープンエンディッドという言葉を山本先生はお使いになりましたけれども、自由形式記述の形でコンピュータに答えを書くような、ちょっと言葉は古いですけれどもいわゆるハイテクノロジーを使ったテストなんですが、これは山本先生の話にもありましたけれども、果たして紙と鉛筆のテストと同じ能力が測れているのかという疑問がありますし、それ以前に、本当にコンピュータを使って試験ができるのかという素朴な疑問があるかもしれません。しかし、それが出来るのは間違いないんです。

次に、独立小項目を多用する。広い範囲をカバーできるのでこれはいいことなんですけれども、大問形式にしないと思考力や活用力を測れないんだと考えたり、大問形式のほうが恐らく作るのは、1 つのテーマのもとに幾つかの項目を配置するわけですから、簡単に出来るという事情があるだろうし、小さい項目で広い範囲をカバーしようとすると、すごく沢山の項目を作らなければいかんのですが、なかなかそういう作題体制が今まではないということ。

それよりも、日本のこういう大規模試験の項目を作っている先生方は、「独立小項目は

品がない」、「こんなテストは作りたくない」、とおっしゃる先生がすごく多いんですね。ここには「作題者の感性」と書きましたけれども、これは本当に、そういうふうに思っておられる方が多いということがあります。

最後、尺度得点を受験生へ報告することに関しては、素点ではなくて難易度を調節したような点数ですから、複数回実施する、年をまたいだ成績が出てくるというときにその成績が比較可能になるので、受験生にとっても利用する側にとっても良いということなんですけれども、尺度得点は信用できないとか、何問正答したかという素点を実は知りたいとか、そういう要望があって、日本的試験文化が日本では使われているということだと思います。

試験の日本的風土への経緯

何故こんなふうになったのか。これは本当にこうなのかよくわかりませんけれども、入試に関して言いますと、個別大学の入学試験に代わるものとして共通第 1 次学力試験やセンター試験が導入された。1 年単位でその年の受験生のランキングができればいいという形で、別に年度ごとに難易度が変わっても 1 年ごとにランキングができれば、上から順に採っていけばいいやという感じで、余り尺度得点みたいなことにはこだわっていないのではないかという気がします。

それから、将来の展望の欠如というのは、例えば学力低下問題なんていうのが話題になりましたけれども、センター試験とか共通一次試験をデザインしたときに、日本人の学力が低下するなんて誰も考えていなかったということで、毎年の学力変化、学力の経年変化を測るような装置をテストの中に組み込んでいなかったということですね。今、思い出すんですけれども、90 年代に山本先生が日本の通産省にアダルトリテラシーのことで話に行かれましたよね。それで通産省のお役人から冷たくあしらわれて非常に困ったという話がありましたよね。日本人は勤勉なものですから、学力低下とか、どれだけ字が読めるかとか、そういうことは調べなくても間違いなく優れているんだよという意識があったんだと思います。

あと「心理測定への無理解、挫折」と書きましたけれども、この挫折というのは、先ほど荒井先生からお話がありました。進学適性テスト、それから能研テストというものがあったんですが、あれが余り受け入れられなかった。あのときに心理測定の専門家は作成に参加したんですけれども、どうもそれがセンター試験・共通一次の時は、過去にうまく受けられなかったことを反省し萎縮していたのではないでしょうか。

最後に、やはり作っているのがお役所ですから、日本人は昔はお役所がつくるものは信頼していましたから､少々不備があっても有り難がって使っていたようなところがあるし、逆に役所のほうは、少々欠点があっても説明責任は大して取らなくても良いみたいな、そういう形でテストが作られてきたということがあると思います。

試験の日本的風土の問題点と測定の基本モデル

では、いろいろ挙げましたけれども、一番の問題点は何かというと、これはやはり私の個人的な感じなんですけれども､最後の配点を利用した素点の利用､ここのところが一番。素点を使うテスト問題が出て、そのうち何題正答した、それを使って成績の尺度としているところが一番気になるところです。

ここからはちょっと教育測定という話になるんですが、観測されるテストの得点、例え

ば私が何かテストを受けて 65 点取りましたよというのが X です。私の点数 X は、実は私の真の得点、60 点に私の誤差、5 点が加わったものかもしれないし、私の真の得点 70 点に誤差マイナス 5 が加わったものかもしれないといった形で、観測されたテスト得点そのものが真の値をあらわしているのではなくて、このように、真の得点プラス誤差みたいな形で分解して考えるのが教育測定の基本です。

日本の試験文化では、あたかもこの部分、誤差の部分は存在しないよ、というように考えているのではないかと思っています。特に素点を利用するということに関しては。

この X というのは T と E の和だよというのを認めると、例えば X の中に T の部分がどれだけあるのかというような量でテストの信頼性みたいなものを定義することができる。だから、テストを作るときにはなるべく E の部分を小さくして、T の部分が大きくなるように作りましょうという考え方が出てくると思います。

心理測定とかテスト理論とかでテストをどう考えているかと言いますと、テストというのは能力や学力、性格や適性、その他いろいろなもの、ここでは「人間の精神的特性」と書きましたけれども、それを測定する道具であると考えています。だから長さを測る物差しとか温度を測る温度計とか体重を測る体重計とか電圧計とか、そういうものと全く同じものです。

物差しというのは、良い物差し、悪い物差しがありますよね。伸び縮みする物差しは余り良い物差しではないですよね。それと同じように、道具ですから善し悪しが存在する。先ほどの X ＝ T ＋ E という形で善し悪しを表現すると、E の部分が大きいテストは悪いテストということになります。

あと、温度計や何かもそうですけれども、室内の温度を測る温度計と何か実験をして 200 ℃とか 500 ℃とか 1,000 ℃ぐらいを測る温度計と、当然違う温度計を使わなければいかんわけですが、テストも同じで、下のほうの能力の人を測るテスト、上のほうの能力の人を測るテスト、当然ですけれども変えないと精度がよくないみたいな、そういうことも考えられるわけです。

テストを扱う理論として、古典的テスト理論とか項目反応理論とかいう話が出てきます。先ほどの山本先生の話には IRT（Item Response Theory）が出ていましたけれども、それはいわゆる項目反応理論で、項目単位ごとにそのテストの特性をあらわすパラメタみたいなものを推定する、そういう理論になります。

日本の大規模試験の現状

これはちょっと別の話になりますけれども、今、日本で行われている大規模試験はどんな感じかといいますと、上に書きました司法試験、公務員採用試験、それから医師の国家試験、この辺は余り、いわゆる測定論的にはそれほど厳密ではなく、日本的試験文化的につくられているという感じです。下に書きました情報処理技術者試験とか医学部の共用試験とかビジネス日本語試験とか、日本留学試験、日本語能力試験までは、一応テスト理論を利用して作ったりコンピューターで実施したりしている大規模試験です。情報処理技術者試験というのは国家試験なんですけれども、既にコンピュータを使って、テスト理論を使ってつくられています。

日本留学試験というのが下のほうにありますけれども、これは留学生のための入学試験でその受験生は日本人ではありません。外国人です。だからこれはある意味、世界標準に

基づいてつくられた試験ということで、コンピュータでは実施していませんけれども、テスト理論を使って実施しています。年間 30,000 人ぐらいかな、受験生がいます。年に 2 回やっています。

あと、一番下に書きましたけれども、日本の民間の会社でいろいろなテストがつくられていますけれども、それには IRT を使っている語学テストが結構たくさんある状況です。

そういうことで、日本的試験文化というものに関して言及した論文を、幾つかここに参考文献として挙げていますし、下から 2 番目、最近、新しい本が出ていますということで、とりあえず「日本的試験文化」というのは何かという復習的なお話をさせていただきました。

以上です。（拍手）

○司会　前川先生、どうもありがとうございました。

何か事実確認的なご質問がありましたらお受けいたしますけれども、いかがでしょうか。

大問と小問で古典的テスト理論と IRT と分けて対応づけされていましたけれども、山本先生がご紹介下さった最後のサンプル問題などは、どちらかといえば大問ですよね。あれも IRT でパラメータを求めていらっしゃるんですか。

○前川氏　項目反応理論― IRT というのは、今までは正答か誤答かで採点される項目を主に取り扱ってきたんですけれども、0 点、1 点、2 点、3 点、4 点、5 点みたいな形で大問的な項目は、２値項目に対して多値項目と呼び方をするんですが、それも通常の項目反応理論の形で取り扱うことが出来ます。

ちなみに、多値項目の IRT を最初に提案されたのは日本人の先生で、1960 年代だと思いますけれども、鮫島史子先生です。

○司会　大問などは、そこに含まれる項目はある意味で独立でないとみなせるのではないかと思うんですけれども、そういった局所独立の仮定が多少ずれていても余り問題ないということですか。

○前川氏　5 段階の値で採点される 1 つの項目という形で、大問ごとに独立と考えれば良いのでは。

○司会　大問の中にシミュレーションしたり理由を書かせたりという問題があるではないですか。それはどのように扱われることになりますか。

○前川氏　だから大問の合計で何題ということですよね。

○山本氏　前川さんの言っているのは、最後のオープンエンドの問題があったでしょう。あれはあれ自体が大問になるということですね。だから全部のシミュレーションのアイテム自体は正答、誤答に関して、例えば全部で 5 問ぐらいありましたけれども、4 問は正答、誤答でできますけれども、最後のアイテムは多値項目の問題で、ポリトゥマスアイテム（Polytomous　item）と英語で言いますけれども、そのような項目に適した IRT モデルを使います。

○司会　その 5 つの項目は独立として、パラメータをそれぞれ求めていくことになりますね。

○山本氏　そうです。局所的な従属性（local　dependency：ローカルディペンデンシー）に関して質問されているんですけれども、その点については、ETS でもすごくよく研究され

ているんですね。ローカルディペンデンシーによってどれだけバイアスが出てくるかといった観点になりますが、それを探しても見つけられたことがないんですよ。ハワード・ウェイナー（Howard　Weiner）というリサーチャーが ETS にしばらくいましたけれども、彼はすごく努力して探していましたけれども、見つけられないんですね。結局、結果としてあるのは、ローカルディペンデンシーがあるとアイテムパラメータが、スロープがもっと急になって、それで結局、情報量（information）を過大に推定（estimate）してしまうことになりますけれども、だけれども、バイアスの点では問題がないということです。

○司会　ちょっと専門的な話になってしまったかもしれませんけれども、IRT というのは、日本語で言うと項目同士の局所独立という条件があって、その前提の下での確率計算に基づいて理論が成り立っていますので、1 つの大問の中にある項目同士はどうしても独立と言えるのかどうなのかが問題になりますので、そのあたりをちょっと議論させていただきました。

それと、もともとは 2 値データの正誤に関する正答確率で理論が成り立っていたのですが、例えば 5 点満点であればそういう段階的な部分にも IRT が発展的に広がってきているということですね。ありがとうございました。拍手をお願いいたします。（拍手）

試験の日本的風土

東京工業大学
大学院社会理工学研究科
前川眞一

はじめに

- 試験の日本的風土

  日本で行われている試験・テストの特徴
  日本的試験文化
  日本的テスト文化

- 検査、試験、考査、テスト、（検定）
- quiz, test, examination

日本的テスト文化

- 年に一度、同一問題での試験の斉一実施
- 新作問題のみでの試験の実施
- 試験問題の公開
- 大問形式の利用
- 問題作成とテスト編集の融合
  テスト・心理測定の専門家の不在・不利用
- 素点・配点の利用（０点と満点）
  回をまたぐと比較不可能

日本的テスト文化（理由）

- 年に一度、同一問題での試験の斉一実施
  （同一問題でないと不公平）
- 新作問題のみでの試験の実施
  （プリテストは不公平）
- 試験問題の公開
  （教育的見地？情報公開？規制緩和？）
- 大問形式の利用
  （多肢選択式で思考力を測る努力？　細かいスペックの欠如？）
- 問題作成とテスト編集の融合
  テスト・心理測定の専門家の不在・不利用
  （権威主義）
- 素点・配点の利用（０点と満点）
  回をまたぐと比較不可能
  （心理測定学への無関心、尺度得点への不信感）

日本的試験文化の逆

- 年に一度、同一問題での試験の斉一実施
  （年に複数回、異なる問題で分散的に実施、IRT/CBT）
- 新作問題のみでの試験の実施
  （統計的性質（難易度等）の分かっている問題のみ）
- 試験問題の公開
  （試験問題の非公開・再利用、個人への成績通知）
- 大問形式の利用
  （細かいスペック・広い分野を測定、独立項目の多用）
- 問題作成とテスト編集の融合
  テスト・心理測定の専門家の不在・不利用
  （テスト理論を用いた難易度のコントロール）
- 素点・配点の利用（０点と満点）
  回をまたぐと比較不可能
  （テスト理論を用いた比較可能な尺度得点の利用）

世界標準

世界標準の特徴と疑問

- 年に複数回、異なる問題での分散実施
  - ◆受験生にとっては有り難いはず
    - ⋆事務の繁雑化
    - ⋆公平性・セキュリティの問題
- コンピュータ化（CBT）・自由記述の利用
  - ◆適応型テスト・HighTech、　ICT　の利用
    - ⋆同じ学力が測定可能か？
    - ⋆コンピュータで試験！？

世界標準の特徴と疑問

- 独立小項目の多用
  - 広い範囲をカバーできる
    - ⋆ 大問形式の方が思考力や活用力をより良く測る
    - ⋆ 作題の事情　（大問形式の方が容易？）
    - ⋆ 膨大な項目数に対応出来る作題体制
    - ⋆ 作題者等の感性　（独立項目は品がない！）
- 尺度得点を受験生へ
  - 得点が時期を越えて比較可能
    - ⋆ 尺度得点は信用できない
    - ⋆ 何問正答したかや素点を知りたい
    - ⋆ 満点と 0点に意味がある

7

歴史的背景

- 個別大学の入学試験に代わるものとしての共通第１次学力試験やセンター試験
  - 一年ごとに受験生の順位が付きさえすれば良い
- 将来の展望の欠如
  - 勤勉な日本人の学力低下なんてあり得ない？
- 心理測定への無理解、挫折
  - 進適・能研テストへの反省・萎縮
- 官製テスト
  - お上への信頼・説明責任の欠如

8

日本的テスト文化最大の問題点

- 年に一度、同一問題での試験の斉一実施
- 新作問題のみでの試験の実施
- 試験問題の公開
- 大問形式の利用
- 問題作成とテスト編集の融合
  テスト・心理測定の専門家の不在・不利用
- 素点・配点の利用（０点と満点）
  回をまたぐと比較不可能

テスト得点の性質

X = T + E

観測されたテスト得点　　真の得点　　誤差

日本的試験文化では
この部分の存在を認めない！

テストの信頼性 = X の中の T の割合

10

心理測定・テスト理論

- テストとは、能力・学力・性格・適性等の人間の精神的特性を**測定する道具**である。
  - 物差、温度計、体重計、電圧計等と同様に、道具の善し悪しが存在する。
- 古典的テスト理論　（大問）
  Classical Test Theory
- 項目反応理論（項目応答理論）　（項目）
  Item Response Theory (IRT)

11

我が国の大規模試験の現状

- 司法試験　（法務省）
  - 年間１回、複数年利用
- 公務員採用試験　（人事院・日本人事試験研究センター）
  - 年間１回、平均・標準偏差等による標準化
- 医師国家試験　（厚生労働省）
  - プール制（項目プール・項目バンク）へ移行中だったが・・・
- 情報処理技術者試験　ITパスポート試験　（経済産業省・情報処理推進機構）
  - CBT (IRT)で実施中
- 医学部共用試験　（東京医科歯科大学医歯学教育システム研究センター）
  - CBT (IRT)で実施中
- ビジネス日本語試験 BJT　（経済産業省・日本貿易振興会 JETRO → 漢検）
  - IRTで実施中
- 日本留学試験 EJU　（日本国際教育協会 AIEJ）IRT で実施中　尺度得点の利用
- 日本語能力試験 JLPT（国際交流基金・日本国際教育支援協会）IRT で実施中
- 法科大学院入学試験　（大学入試センターと日弁連）
  - 本邦初の競争原理の導入？
- 民間の多くの語学テストでは IRT が利用されている。

12

参考文献

- 村上隆(2001) 第2言語としての日本語能力テストの開発 －一般的な問題と固有の困難－ *計測と制御* 第 40 巻
- 荒井清佳・前川 眞一(2004)日本の公的な大規模試験に見られる特徴 －標準化の観点から－ *日本テスト学会誌*
- 木村拓哉(2006) 戦後日本に於いて「テストの専門家」とは誰であったのか？ *教育情報学研究*
- 柴山直(2008) 日本のテスト文化について *人事試験研究*
- 繁桝算男編(2014) *新しい時代の大学入試* 金子書房
- 若林昌子、杉光一成(2013) わが国の国家試験における試験問題公開の現状と傾向 *日本テスト学会誌*

ご清聴有難うございました。

13

【報告２】

入試選抜の測定問題

東京大学大学院教育学研究科長
南風原朝和

○司会　続いて東京大学の南風原朝和先生から、「入試選抜の測定問題」ということでご報告をお願いいたします。

○南風原氏　東京大学の南風原です。

こういうテーマを与えられましたので、入学試験というのは測定論だけで語ることのできる問題ではありませんけれども、その一方で測定論抜きに語ることもできないということで、私のほうでは測定ということに焦点を当ててお話ししたいと思います。

測定の目標は単純で、測りたい特性をできるだけ高い妥当性― validity といいますが―高い妥当性で測ることに尽きます。キーワードは「測りたい特性」と「妥当性」ですが、重要なことは、測りたい特性は何なのか、何を測るべきなのかということですね。この議論は非常に大事なところです。

もう一つは「妥当性」という言葉、何度も出てきますけれども、これをどう解釈するか。同じ土俵で話をしなくてはいけないので、妥当性という概念をきちんと理解しなくてはいけない。例えば「信頼性」という言葉とどう違うのか、そのあたりの区別もつけて議論をしていく必要があります。

先ほど荒井先生からもありました中教審の答申案では、この大学入学希望者学力評価テストと呼ばれるものにおいて「以下のものを測る」と提案されています。知識、技能ではなくて、「それを活用して自ら課題を発見し、その解決に向けて探求し、成果等を表現するために必要な思考力、判断力、表現力等の能力」、これがその提案における測りたいもの、測るべきものですね。まず、ここでの議論が 1 つ大事になります。これを測るべきなのか、知識、技能よりも例えば表現力というものを測るべきなのか、これは議論すべきところです。

その議論は非常に重要ですけれども、とりあえず、仮にこれが測りたいものだとすると、あとはこの測りたいこと、表現力や思考力というものを、いかに高い妥当性で測ることができるかが問題になります。

妥当性のとらえ方

少し大学の授業のようになって恐縮ですけれども、妥当性というのはスライド 4 のように、概念的には定義できるのではないか。妥当性に関する議論だけでも非常に錯綜している部分がありまして、私はこんなふうに考えるのが生産的で、議論する上では役に立つのではないかと考えています。

先ほど前川先生からもありましたけれども、まず測定値、実際にスコアとしてあらわれるものは、いくつかの成分から成り立っていると考えます。

まず 1 つ目が、先ほどの測りたい特性を反映する部分。construct という言葉が使われるので「c」としています。construct ―構成概念という言葉が使われますけれども、ともかく測りたいものが少なくとも一部反映されているだろう。ここで、測定値 x イコール c であれば、つまり測定値が、測りたいもののみを反映しているのであれば、何も測定論の問題はないわけですけれども、それプラス何々というのがあるので、測定の問題が生じるわけですね。

まず、ここで測定値と呼んでいるのは、テストのみならず小論文の評定であったり面接の評価点であったり、いろいろです。数値化するものすべてです。それイコール c であればいいわけですが、通常は何らかの誤差が加わるわけですね。誤差をここでは 2 種類に分けて、系統的な誤差と偶然的な誤差、前者はシステマティックということで「s」、後者はランダムなエラーということで「e」としています。系統誤差というのは安定していて何度も出てくる、もう一回測っても出てくるんだけれども測りたいものとは違うという、irrelevant なファクターだということです。毎回変動するわけではないけれども、stable だけれども irrelevant というのが s です。

ここに「分散」という言葉があります。これは統計用語ですけれども、ともかく測定値 x には個人差があるわけですね。高い人から低い人まで個人差がある。このばらつきの程度が分散です。全員が同点であれば何の評価もできないので、まず分散があることが前提ですけれども、重要なことは、その分散がどれだけ本物の分散かということです。誤差だけでも分散、個人差は生じるわけです。このシステマティックな、irrelevant な誤差だけでも分散は生じるわけです。だから全体の分散、目に見える高得点から低い点までの分散のうち何%が見たい部分の個人差を反映しているか、この割合が妥当性というふうに概念的にはとらえることができます。

したがって、これから測定評価の問題を考えるときにこの 3 分割を考えて、測定値がどれぐらい c を反映しているか、また、どうやってその割合を高めるかということが大事です。これは、スライド 4 のように分解すれば簡単で、赤いところ（c の分散）を大きくして、その他の黒いところを小さくするのが妥当性を高める努力ということになります。

それを言葉で書いたのがスライド 5 になりますが、妥当性を高めるには、まず、測りたい特性の個人差を反映するような項目が必要になります。それは「表現力とか判断力とか思考力が高い人はこう反応し、低い人はこう反応する」という、その見たいものを反映する項目をどう開発するかが、まず主要な努力の目標となります。例えばそれをするのに複数の教科を合わせたような合教科型、合科目型が有用なのか、あるいは PISA 型といわれる項目が有用なのか。ここは開発の問題になるわけですね。

一方では、それ以外の要因の分散を減らすことが必要になります。2 つ目ですね。測りたい特性以外の要因による系統誤差を減らす。例えば今の中教審の案のように知識、技能ではないと言うならば、知識、技能の個人差が反映されればこれは誤差になるわけですね。単純な知識、単純な技能だけで高い点がとれるのであれば、これは誤差になるわけです。また、入試対策で面接の対策というのがあるらしいですけれども、それで得点が上下する

ようであれば、その部分は誤差になるわけですね。見たいところではないわけですから。

それ以外に偶然の要因がありまして、誰が評価に当たるのか、面接や小論文で何を問うのか。面接や小論文の質問はランダムではないと考えられるかもしれませんけれども、初めから 1 つに決まるわけではなく、たくさんの可能な小論文の項目、面接の問いのうち 1 つがある種、たまたま選ばれるわけですよね。別の項目をやれば別の個人差が生じるわけなので、ここはランダムエラーと見ることができるわけです。

ともかく後者 2 つの要因をどれだけ減らすのか、あるいはどれだけそれが影響しているかを査定しながら減らしていくことが妥当性を高める努力ということになります。

なので、ある能力を測るときに、この 3 分割の観点から見ていくことが大事になるかと思います。

信頼性との関係

それから、先ほど出ました信頼性とどう違うかということですけれども、前川先生の式の中に X ＝ T ＋ E というのがありました。それと妥当性の式はどう違うかというと、T にはスライド 6 に示したように測りたくないけれども stable なものも入ってくるわけです。なので、信頼性というのは妥当性そのものではなく、プラス、いまの例では例えば単純な知識、技能みたいなことによる分散も入って信頼性と呼ばれます。だから、信頼できるけれども妥当でないという可能性があるわけですね。信頼できるけれども、ほとんど s の分散、irrelevant な分散によって信頼性が高まっている可能性もあるわけです。逆に、この赤いところ全体が小さい、信頼性が低いとしたら、妥当性は必ず小さいわけです。だから、信頼性が低いともう妥当ではあり得ないということで、そういう意味では信頼性は必要ということになります。

これを言葉で書いたのがスライド 7 になります。信頼性が低いと妥当性は高くならない。妥当性が大事なわけですけれども、高い妥当性のためには高い信頼性が必要です。言葉を換えると、信頼性の高さは情報の量です。ランダムでない、stable な情報の量なんですね。情報量が少なければ妥当ではあり得ないんだけれども、情報はたっぷりあってもそれが知りたい情報でない、ゆがんだもの、例えば入試対策の出来・不出来、家庭の経済的なものの高低、そういったものが反映されている可能性もあるということで、これが妥当性の問題となります。

中教審の答申案から

再度、中教審の答申案で、大学入学希望者学力評価テストではいくつかのねらいが書かれていますけれども、ここで 2 つ取り上げて、この 2 つに焦点を当てて測定論というか、情報量の観点から少し考察してみたいと思います。これが今日の話題提供のメインのテーマとなります。

スライド 8 のまず 1 つ目ですが、選抜性というのはいわゆる偏差値の高い大学、低い大学ということを意味しているようですけれども、選抜性の高い低いにかかわらず、多くの大学でそのテストが活用できるように難易度は広くとる。非常に難しいものから易しいものまで入れる。特に、選抜性の高い大学が入学者選抜の評価の一部として十分活用できる水準の、高難度の出題を含むものとするということが、このテストのデザインとして提案されています。難しい問題も含んで、いわゆる難関の大学でもこれを入学試験で利用でき

るようにするということですね。

そういう意味では、中教審の答申でも測定の精度や情報量というものを意識した案となっています。

2 つ目。これはマスコミでもよく取り上げられていますが、1 点刻みの客観性に囚われた評価から脱して、各大学の個別選抜―二次試験における多様な評価方法の導入を促進する観点から、大学及び大学入学希望者に対して、1 点刻みの点数ではなくて段階別表示、5 段階の 5 とか 5 段階の 4 といった段階別表示による成績提供を行う。

この 2 点について測定、また情報量という観点から少し考察してみたいと思います。

広範囲の測定と情報量

まず 1 点目は、難易度を広くとるということでありました。これが先ほどから出ている項目反応理論、Item Response Theory ― IRT の大事なところになるわけですけれども、項目反応理論の一つの大事な点は、能力や特性といったことと項目の出来・不出来とかテストの点数を分けて考える。だから縦軸には項目の正答・誤答、あるいはテストスコアが出てきますが、それとは別に、それらテストへの解答をすべて通した能力というものを別途考えることが 1 つ重要なポイントになります。

スライド 9 の中央の黒い曲線は 1 つの項目をあらわしていますけれども、能力が高くなるほどこの項目に正答する確率は上がっていくわけですね。それが項目の特徴をあらわしているわけです。先ほど山本先生の話にあった項目のパラメータというのは、その項目がどれだけ難しいか、どれだけ識別力があるかといったことを意味しているわけですが、この黒い項目は中程度の難しさをあらわしています。例えば左側の赤いこの項目は、能力が低くても結構正答できる。右端のものは能力が非常に高くないと正答できないということで、右側にある項目ほど難しい項目ということで、項目反応理論では、こういう形で項目の特徴を表現します。

先ほどの広範囲の難易度をとるということについて、ここでちょっとシミュレーション的に考えてみたのは、すべて中程度の難易度の項目、つまり真ん中の黒い項目だけでテストを構成した場合と、中教審の言う広範な難易度をとった場合、つまりスライド 9 の中央の黒の項目のみのテストと赤の項目を含めた全体のテストに分けて、情報の量を計算してみました。

この情報量というのは信頼性に非常に近い概念ですけれども、信頼性と違って情報量というのは、スライド 10 のように、あるレベルの生徒に対する測定の情報量というふうに、能力レベルに応じて情報を関数として表現できる、これが項目反応理論のもう一つの特徴なんですね。例えば中程度の難易度の項目のみでやると、情報量は黒の曲線のようになり、当然中程度の生徒の能力を識別する上では大変情報量が高いわけですけれども、高いレベルの生徒に対してはほとんど情報がないわけです。この子たちにとってみればほとんどできてしまうので、このレベルの生徒たちの能力を識別することはできないわけですね。逆に、低いレベルの子供たちにとってはとても難しい項目ばかりなので全員ほとんど 0 点になって、やはり情報がない。

それに対して難易度を広範にとったテストの情報量は、赤の曲線のようになります。当然中心レベルでは下がりますけれども、レベルが高いところでも割合情報がある。低いと

ころでも割合情報があるということで、そういう意味では、広範囲にとることによって高い水準の大学にも対応するというのは、ある程度実現できそうに見えます。

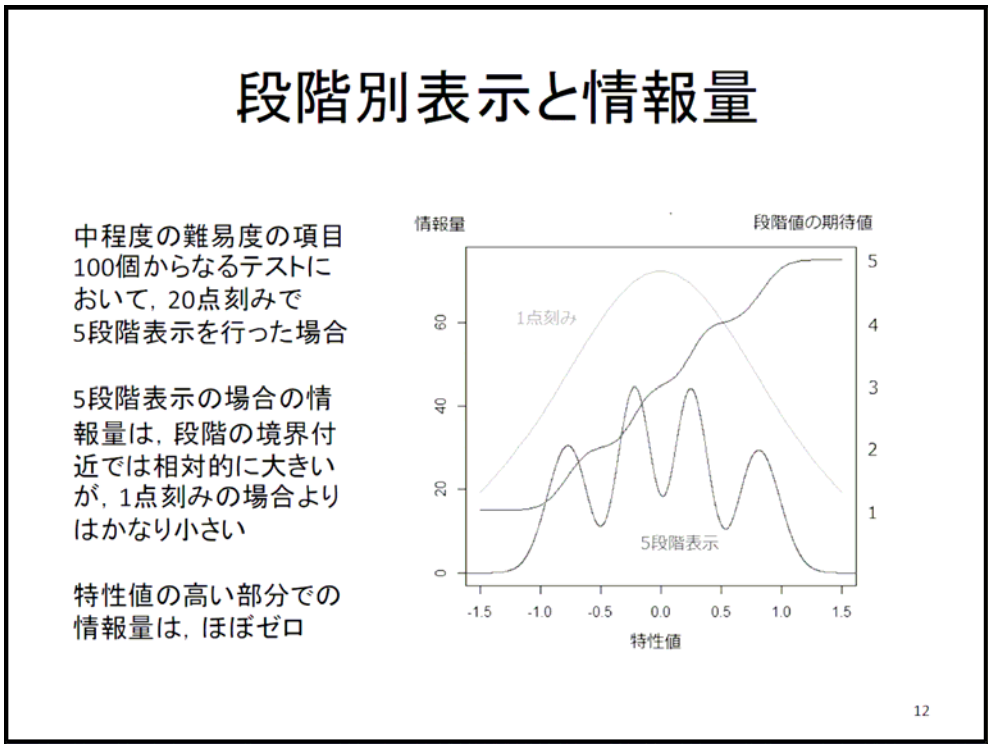
スライド 11 は同じことですけれども、先ほどの赤のテストと黒のテストの情報量の比をとりました。何倍情報量があるかという比をとったものです。1 倍であれば同じですね。これは広範囲にとった場合、中程度のみの場合に比べて何倍の情報量があるか。中心付近ではやはり 1 を下回りますけれども、高いレベルでは 2 倍、3 倍の情報量があるということなので、今、言いましたように、このようにすることによって選抜性の高い難関大学でも、このように範囲を広げれば対応できそうに見えます。

これが 1 点目についての暫定的な評価になります。

成績の段階別表示と情報量

2 点目は、段階別表示ですね。

スライド 12 が今日のスライドの中で一番大事なスライドになるかと思うんですけれども、とりあえずテストとしては、先ほどの中程度の難易度の項目を 100 個用意して、0 点から 100 点までのテストにしました。ただし、それを 20 点刻みにして、20 点以下は 1 点、20 から 40 点は 2 点というふうに 5 段階表示を行いました。そうすると、この黒の上昇する曲線は段階値の期待値なんですけれども、能力の低いところでは、右側の縦軸を見ると、ほぼ 1 点になります。能力が高いと、もう差がつかないので 5 点になります。段階の境界点があって、そこまでややフラットになり、境界点付近で上がってまたややフラットになり、これが段階別表示をしたときに、段階値が能力によってどう変化していくかという曲線になるわけですね。



もうここからはわかると思いますけれども、ある程度以上能力が上がっても 5 点、5 点、5 点なので差がつかないわけです。したがって、赤の曲線で示した 5 段階表示の情報量曲線で見ますと、特性値の高いところではほとんど情報量がありません。逆に低いところでは

この範囲ではほとんど 1 点しかとらないので、情報量がない、差がつかないわけですね。段階値の期待値を示す曲線のスロープが急になっているところで情報量がピークになります。5 段階だと境界点が 4 つありますので、4 つの山ができる情報量曲線になるわけですね。段階別表示の情報関数の特徴をあらわしています。

ただ、大事なことは、1 点刻みの場合に比べて段階別表示の情報量がどうなのかということですが、1 点刻みにすると、この上部にある緑の曲線が情報量曲線となります。段階別表示にすることによって明らかに情報のロスが生じるわけですね。1 点刻みではこれだけの情報量があったのに、段階別にすることによって、特性値の高い部分ではほぼゼロになってしまうということで、5 段階表示の場合の情報量は段階の境界付近では相対的に大きく―相対的にというのは、この中でということですね―なりますが、1 点刻みの場合よりは明らかに、かなり小さくなります。

スライド 13 は情報量の比をとったもので、段階別にしても、もし素点と同じだけの情報量があればこの 1 点のラインになるわけですけれども、ずっと下のほうになっていますよね。高いレベルではほとんど情報がないということで、情報量が大きいところでも相対的に 0.6 ということは、その逆数の倍の項目数が必要だということです。情報量が 0.5 しかないということは、同じ情報量を得るために 2 倍の数の項目のテストをしなくてはいけないということですね。この図からすると 5 段階表示をして 1 点刻みの場合と同等の情報量を得るためには、2 倍から 5 倍長いテストをしなくてはいけない、それだけのロスが生じるという定量的な計算になります。両端のほうでは、相対効率―情報量の比はほぼゼロとなるわけです。

スライド 14 は難しい項目でやった場合ですけれども、傾向としては同じで、ただ、高いレベルのところで同じ波を打ったことが生じることになります。

先ほど広範囲の難易度にすることによって高い水準のところにも情報量を残すことができそうだという暫定的な評価を言いましたけれども、5 段階評価にすることによってそれが台無しになって、高いレベルのところに情報があっても 5 段階、段階別表示にすることによって情報は完全にゼロになってしまう、そういう仕組みになるということです。

スライド 15 がその部分のまとめになりますけれども、選抜というのはいろいろな情報をもとに選抜するわけですが、その前の段階でのスコアが測定になります。情報を収集して提供するのが測定ですが、最終的には入学試験なので、2 段階表示をするわけですよね。合・否。しかし、その前で段階別にされてしまうと情報が失われてしまう、合否の判断が狂ってしまうわけですね。なので、測定の段階で段階別表示を行うことの問題をここに示したところです。

そういうことで、1 点刻みに対してということですけれども、東京大学ではどうなっているかというと、スライド 16 にあるように、1 点刻みどころか 0.0001 刻みで評価されているところであります。これがよいという意味ではありませんけれども。

段階別表示のその他の問題点

メインの話はほぼ以上ですけれども、段階別表示を行うことの問題点について少し追加で議論したいと思います。

スライド 17 の（ⅰ）の情報量の減少というのは、今、既に述べたことです。（ⅱ）（ⅲ）

（ⅳ）について簡単にお話しします。

（ⅱ）は段階に分ける、先ほどは 20 点、40 点というふうにそれこそ恣意的に分けましたけれども、その分け方の恣意性が問題になる。仮に分けるとしても、どこで分けるのかが問題になるということです。そこで参考になると思われるのが、アメリカでの、特にハイスクールの卒業の基準を満たしているかどうかというミニマムコンピテンシーテスト、これは満たしている・満たしていないという 2 段階表示になるわけですが、そこでの経験が、池田央先生が監訳された『教育測定学』の中の 1 つの章からの引用ですけれども、スライド 19 のような文章がありました。

「基準設定―どこで段階に分けるか―に関する文献が何か決定的な点を持っているとしたら、それは、コンピテシーテストで、擁護し得る基準を設定することの困難さについてである」、段階分けは難しいということだけでは合意されているということですね。

「Glass も Shepard も、コンピテンスは概念上完全な連続変数であると指摘する。それなら、コンピテンスをもつか否かで学生を 2 つの別個のカテゴリに仮に分ける分割点を設定することは、非現実的かつ非論理的である」

スライド 20 は省略しますけれども、基準設定に用いられている方法がいくつかあるんですけれども、どれも「なるほど、これですればいいな」という決定的なものがありません。1 つだけ紹介しますと、Angoff さんの方法は、ぎりぎり基準をクリアしている生徒を専門家が想定します。その生徒がこの問題に正答する確率はいくらか、次の問題に正答する確率はいくらか、これを足し合わせてそのぎりぎりの生徒の期待得点を算出し、それを分割点とするということで、かなり主観的な判断が必要になってきます。そういったいくつかの方法があるわけですが、方法を違えると不合格率が 30 倍違うといった結果も出てきて、そういう段階の分け方の問題が指摘されているところであります。

スライド 23 の（ⅲ）は、同一段階の中での個人差や変化が無視されて、段階間での差や変化が誇張されてしまう。つまり段階別表示では、同一段階の中では基本的に等質なもの、4 の人、5 の人というふうに等質なものとされますけれども、実態は同じ段階の中にかなり大きな個人差があるわけで、ある段階の中心付近に位置する者は、かなりの向上があっても同じ段階にとどまり、努力が結果に反映されない。

下にちょっと余計なことを書きましたけれども、ダイエットで 3 キロ痩せたのに「やや肥満」、5 キロ痩せたのに「やや肥満」、変わらなければやる気をなくしますよね。やはり努力が反映されるような入学試験が必要ではないかと実感したところがありましたので、ちょっと書いておきました。

スライド 24 ですが、少ない段階数で、例えば 5 段階で 4 とか 3 とか別段階に評定されると、実際にはわずかな差異でもそれが拡大されて合否判定に決定的な影響を与えてしまう。段階の境界付近にいる者は、学力の変化がなくても誤差変動だけでも上下してしまう。それがまた決定的な影響を与えてしまうということですよね。

次はちょっと強調したいところですけれども、1 点刻みの差は、中教審答申にもあるように、実質的な意味はありませんが、それは皆さんわかっているわけです。東大の 550 点満点の 0.0001 に実質的な意味があるとは思っていないわけですよね。1 点とか 0.0001 だからこそ問題はないと思うわけです。しかし、段階になると、実質的な意味のない境界付近

の差が実質的な意味をもつかのように、3 と 4 には「それは実質的な意味があるでしょう」と思ってしまう。しかし、実際には 550 点のうちの 0.0001 しか違わなくても 3 になったり 4 になったりするわけですよね。それが不当に大きな意味を持ってしまう。

また、5 段階の評価なので、これがかなり固定化されて、この人は 4 の人、3 の人というふうに固定化されて、不当に差別化されることも懸念されるのではないか。

スライド 25 の段階別表示の問題点の（ⅳ）です。選抜は大学ごとにいろいろ柔軟にやったらいいと思うんですけれども、柔軟にやろうとしてテストの得点に重みをつけて、他の、例えば面接等の情報を組み合わせようとした場合や、大学独自の基準で段階分けをしたいといった場合でも、既に段階分けされているとそういった多様な柔軟性を阻害してしまうというのが 4 つ目の問題として挙げられます。

例外的に段階別表示が必要な場合、推奨される可能性がある場合として、これは柴山さんの論文からとったものですけれども、例えば小論文の採点の場合は、100 点満点の細かい評定は、まず評定の能力からして無理があるだろうということで、より信頼性を高めるためには、むしろ 5 段階等にするのがよいのではないかということです。そういう段階別評価が有効な場面もあると思いますけれども、もともと情報量豊かにスコアリングされているものを段階別にすることには問題があるということでありました。

大学入試の日本的風土

最後に、本日のシンポジウムのメインテーマに挙げられています大学入試の日本的風土については、私の中ではここがかなり日本的だと思っているのが、それぞれの大学が独自に膨大な時間、人、コストをかけて個別学力試験を作成し、採点する。そして書店に行くと大学ごとの赤本が東京大学、京都大学というふうにざっと並んでベストセラーになっている。ここは、例えばハーバード大学の問題集とかないですよね。ここは非常に日本的なところで、かつ、やはりこの膨大な時間というのが、研究者、教育者の膨大な時間を取ってしまっていることは大きな問題として感じるところでありますので、このままでいいと思っているわけではありません。改善は必要だと思いますけれども、そういった風土が新テストによって変えられるかどうか、これは検討すべきところかなと。

ただ、「風土」という言葉を使った時点でもう変わらないという感じもしますが（笑）、あるいは「大学入試の日本的……」と言っても、今、言っていることはかなり東京大学で象徴的にやっていることなので、「東大的風土」が変わるかどうかは大きなところかなと思いました。

最後のスライド 28 は小まとめで、1 つ目は、測定論という観点からの議論も必要になるだろうということ。2 つ目には、やはり理論や技術もこれから開発、向上していかなくていけないけれども、そのためにかかる時間ということもあるので、実行可能性ということ、時間軸をよく考えなくてはいけないだろうというまとめであります。

以上です。（拍手）

○司会　南風原先生、どうもありがとうございました。

非常にわかりやすく解説していただきましたけれども、何か事実確認的なご質問がありましたらお受けいたします。よろしいでしょうか。

広範囲の難易度というのは、私はそのことが答申案に盛り込まれたときに、情報量の部分では選抜性の高いところは確かに上がると思いますけれども、いわゆる共通試験等の枠組みの中では高難度を持つ問題というのはせいぜい 1 題 2 題ぐらいしか出せないとなると、内容的妥当性といいますか、項目の領域代表性的なところでどうしても限界が出てくるのではないかと感じました。ということで、1 つの試験の中で広範囲の難易度をカバーするというのはなかなか厄介なことではないかと思ったんですけれども、その点いかがでしょうか。

○南風原氏　そうですね。今の内容的なこともありますけれども、例えば選抜性の高い大学にとっては、広範囲のテストになるとこの辺の易しい項目はほとんど使えない。全員ができてしまうので要らない項目がたくさん入っているわけですよね。だから 100 項目やっても使えるのは 10 項目、20 項目となるので、できればその大学にとっては 100 項目使えるようなテストにしたいわけですよね。広範囲のテストというのは、実は誰にとってもベストではないということは言える。

○司会　ありがとうございます。

その辺も今後、中教審答申が出ましたら専門家会議などの場で詰めていくことになると思います。一方で、センター試験が今、複雑過ぎるということで、科目数を減らすということで合教科・科目型などという提案も出てきていますので、あまりバージョンを増やすことはできないと言ったあたりの立場とのせめぎ合いという感じになるかなと思っております。

他に何かございませんでしょうか。

それでは、また後ほどパネルディスカッションのときに。

南風原先生、どうもありがとうございました。（拍手）

それでは、今日は長丁場ですので、ここでもちょっと休憩を入れさせていただきます。

〈午後 3 時 25 分　休憩〉

大学入試センターシンポジウム
「大学入試の日本的風土は変えられるか」
(2014.11.29，東京工業大学)

入試選抜の測定問題

南風原朝和
（東京大学大学院教育学研究科）

測定の目標

測りたい特性を，
できるだけ高い妥当性で測ること

重要なことは
・測りたい特性，測るべき特性は何なのか
・妥当性が高いということはどういうことなのか

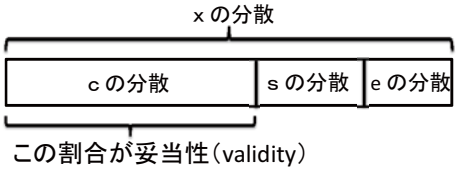
中教審答申案より

大学入学希望者学力評価テスト（仮称）では，
「知識・技能を活用して，自ら課題を発見し，
その解決に向けて探究し，成果等を表現するために必要な思考力・判断力・表現力等の能力」を中心に評価

この「測りたい特性」を，いかに高い妥当性で測ることができるかが問題

測定値の成り立ちと妥当性

測定値（x） ＝
測りたい特性を反映する成分（c）
＋ 系統誤差（s） ＋ 偶然誤差（e）



妥当性を高めるには

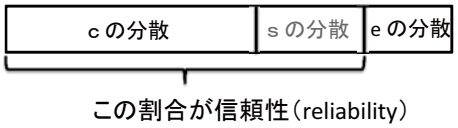
■ 測りたい特性の個人差を反映する項目を開発する
例：思考力・判断力・表現力等

■ 測りたい特性以外の要因による系統誤差を減らす
例：単純な知識・技能
入試対策

■ 偶然要因による誤差を減らす
例：誰が面接や小論文の評価にあたるか
面接や小論文で何を問うか

測定値の信頼性

測定値（x） ＝
真値（t） ｛測りたい特性を反映する成分（c）
＋ 系統誤差（s）
＋ 偶然誤差（e）



妥当性と信頼性の関係

信頼性が低いと
　妥当性は高くならない

⇒ 高い妥当性のためには，
　　高い信頼性が必要

信頼性の高さは，情報量をあらわす

⇒ 情報量が少なければ妥当でありえない
　ただし，情報量は多くても，それが
　知りたい情報でない可能性も
　（妥当性の問題）

7

再び，中教審答申案より

大学入学希望者学力評価テスト（仮称）では，

・選抜性の高低にかかわらず多くの大学で
　活用できるよう，広範囲の難易度とする。
　特に，選抜性の高い大学が入学者選抜の
　評価の一部として十分活用できる水準の，
　高難度の出題を含むものとする。

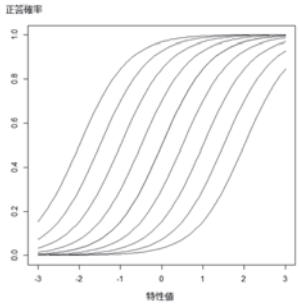
・「１点刻み」の客観性にとらわれた評価から脱し，
　各大学の個別選抜における多様な評価方法の
　導入を促進する観点から，大学及び大学入学希望者
　に対して，段階別表示による成績提供を行う。

8

難易度の範囲と情報量

すべて中程度の難易度の項目（黒の曲線）のみの場合と，
難易度が広範にわたる項目（赤の曲線）の場合，
情報量はどうなるか

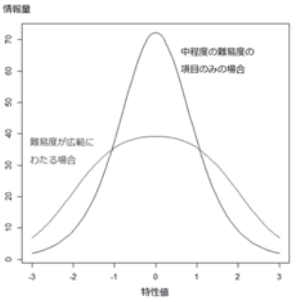
項目反応理論（Item Response Theory; IRT）を用いて計算



9

難易度の範囲と情報量

難易度の範囲を広げることによって，中央部分の情報量は減少するが，特性値の高い部分での情報量は，確かに増大する

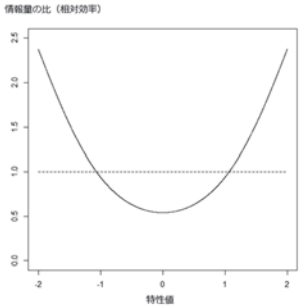


10

難易度の範囲と情報量

難易度が広範囲の場合の情報量は，中程度の難易度の項目のみの場合に比べ，特性値の高い部分では，2倍以上

したがって，選抜性の高い大学にも対応できそうにみえる



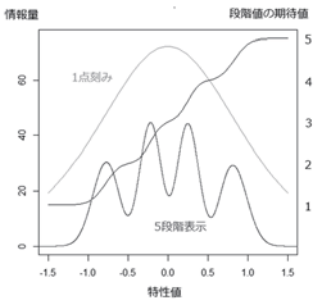
11

段階別表示と情報量

中程度の難易度の項目100個からなるテストにおいて，20点刻みで5段階表示を行った場合

5段階表示の場合の情報量は，段階の境界付近では相対的に大きいが，1点刻みの場合よりはかなり小さい

特性値の高い部分での情報量は，ほぼゼロ

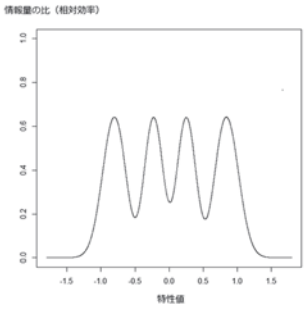


12

段階別表示と情報量

5段階表示の場合の情報量の，1点刻みの場合の情報量に対する比（相対効率）は，特性値の分布の中央付近で20%～60%
⇒ 5段階表示で同程度の情報量を得るには，約2～5倍の項目数が必要

両端では，相対効率はほぼ0%

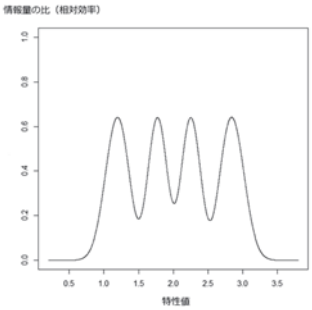


段階別表示と情報量

特性値の高い部分の測定に焦点をあてるために，難易度の高い項目100個からなるテストにおいて，20点刻みで5段階表示を行った場合

特性値の高い部分でも，情報量は1点刻みのときの20%～60%程度

難易度が広範囲だとこの部分の情報量はより少なくなり，結局，段階別表示では，選抜性の高い大学への対応は無理



選抜の前段階である測定（＝情報の収集と提供）において，得点そのものでなく，得点をいくつかの段階に分類した段階別表示を行うことの問題

ちなみに東京大学では，１点刻みどころか・・・

東京大学第2次学力試験 合格最低点（2013年度入試）

科　類	満　点	合格最低点
文科一類	550	348.5333
文科二類	550	342.5000
文科三類	550	347.2111
理科一類	550	315.7333
理科二類	550	302.7333
理科三類	550	370.3889

測定の段階において 段階別表示を行うことの問題点

(i) 情報量の減少（既述）

(ii) 段階への分け方の恣意性

(iii) 段階内での個人差や個人内変化の無視と段階間での差や変化の誇張

(iv) 選抜における測定結果の多様な活用を阻害

(ii) 段階への分け方の恣意性

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の2段階表示を行うテスト）の経験

以下，この項の引用は，

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　央他（監訳）教育測定学　下巻　みくに出版　pp.215-257.)
より

「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。」（p.226）

「Glass (1978) もShepard (1979, 1980) も，コンピテンスは概念上完全な連続変数であると指摘する。それならば，コンピテンスをもつか否かで学生を2つの別個のカテゴリに仮に分ける分割点（Cut-off score）を設定することは，非現実的かつ非論理的である。」（p.227）

19

基準設定に用いられている方法

Angoffの方法

基準を最低限クリアしている生徒を想定し，その生徒が，テストの各項目に正答する確率を推定し，全項目にわたる正答確率の総和を分割点とする。

Nedelskyの方法（多肢選択式項目用）

基準を最低限クリアしている生徒を想定し，その生徒が，テストの各項目の選択肢のうち，誤答選択肢であると指摘できなければならないものを選ぶ。そして，それらの生徒は，残りの選択肢の中からランダムに解答を選ぶことを仮定して，各項目に正答する確率を推定し，全項目にわたる正答確率の総和を分割点とする。

20

基準設定に用いられている方法（続）

境界線グループ法

当該のテストとは独立の情報によって，分割の境界線にいる生徒たちを選び，彼らに当該のテストを受けてもらう。その得点の分布の中央値を分割点とする。

グループ対照法

当該のテストとは独立の情報によって，基準をクリアしている生徒たちと，クリアしていない生徒たちを選び，彼らに当該のテストを受けてもらう。その2グループの得点分布を用いて，たとえば相対度数分布が同じになる得点を分割点とする。

21

使用する方法による結果の違い

極端な場合，方法の違いで，「不合格率」が約30倍違った。

22

(iii) 段階内での個人差や個人内変化の無視と段階間での差や変化の誇張

段階別表示では，同一段階内は基本的に等質なものとされる。しかし，実態は，同一段階内に大きな個人差がある。

ある段階の中心付近に位置する者は，かなりの向上があっても同じ段階にとどまり，努力が結果に反映されない。

cf. ダイエットと体重測定の例

23

少ない段階数で，別段階に評定されると，実際にはわずかな差異でも，それが拡大され，合否判定に決定的な影響。

段階の境界付近にいる者は，学力そのものに変化がなくても，誤差変動だけで上または下の段階に移行。

１点刻みの小さな差には実質的な意味はないが，そのことは了解されやすい。段階になると，実質的な意味のない境界付近の差が実質的な意味をもつかのように誤解される。

段階による不当な差別化とその固定化も懸念される。

24

(iv) 選抜における測定結果の多様な活用を阻害

テストの得点に重みをつけて他の多様な情報と組み合わせたり，大学独自の基準で段階分けしたりするなど，測定結果を多様に活用するうえで，測定の段階での，初めからの段階別表示は障害となる。

25

段階別表示が推奨される例

「もし，小論文の採点結果を選抜に使うとしたら，せいぜい5段階程度の分類にすべきであって，他の信頼性の高いテストの得点と同じように，たとえば100点満点で小論文を評価して，その結果をそのまま選抜に利用するなどは，かなり危険なことであるといわざるをえないのである。」

（柴山直，「日本のテスト文化について」，『人事試験研究』No.206，2008年9月）

26

大学入試の日本的風土

各大学が，独自に，膨大な時間・人・コストをかけて，個別学力試験を作成・採点

大学ごとの「赤本」が書店に並び，ベストセラーに

新テストによって，その風土が変えられるか？

27

小まとめ

中教審答申を含め，入試選抜に関する諸提案については，測定論の観点からの吟味が必要

理想を実現するための測定理論や技術の開発が求められると同時に，開発に必要な時間等を勘案して，実行可能性を判断することも必要

28

【報告３】

センター試験で何が測られるのか？

独立行政法人大学入試センター研究開発部長
大津起夫

○司会　再開いたします。それでは、大学入試センターから研究開発部長の大津起夫より「センター試験で何が測られるのか？」と題しまして、今、中教審の議論の中でもセンター試験は知識して測っていないという、何か少し偏った見方もあるのではないかという印象もありますので、その辺も含めてご報告をお願いします。

○大津氏　大学入試センターの大津です。よろしくお願いします。

タイトルは「センター試験で何が測られるのか？」とちょっと大きく出たんですが、今日は、センター試験の実情はどのようなものかということについて報告させていただきます。

理想的には、人間の知性というのはどのような構造を持っていて、そこのどういう機能を測ろうとしてどういうふうに試験が設計されているかといったレベルまでの話ができればいいとは思うんですが、そこまではよくわからないので、外から見たというか、統計的な特徴の話に限定させていただきます。

センター試験の受験率が過去どのように変わってきたか、私立のほうは出願したことはわかるんですが合否はわかりませんので、国公立大の合否とセンター試験の得点がどのようになっているかという状況について報告します。それから、先ほどから前川先生の話にもあったように、毎年結構点数が変わるわけです。共通尺度として使うことはなかなか難しい。それで、それぞれの科目の平均点がどのように推移しているかについて紹介します。あとはちょっと細かい話になるんですが、センター試験の科目の点数は、1 人が 1 つだけ受けるわけではありませんから、他の複数の科目の相関を計算することができます。それが全体としてどういう構造になっているか、どういう科目とどういう科目の関係が深くて、どういうところが別の動きをしているかという全体像を紹介します。

あと、簡単に科目選択数の地域差の話、それから英語科目得点と中教審や英語教育に関する有識者会議で話題になっています資格検定試験との関係で、わずかですけれどもデータをとってありますので、その紹介をしようと思います。

センター試験の受験率の変化

まず、センター試験の受験率の変化についてです。

共通 1 次試験は 1979 年―昭和 54 年にスタートしていますけれども、センター試験は平成 2 年― 1990 年からです。この第 1 回のとき、現役高校生の受験者は大体 25 万人です。それから浪人等の学生が 15 万 6,000 人ぐらい。このときの日本全体の高校 3 年生のうち、

センター試験を受けたのは 14 %にすぎません。ところが今現在、平成 26 年―今年 1 月です。現役生で受けた人は 41 万 7,000 人、42 万人近いわけです。一方、浪人はずっと減って大体 11 万人。現役の受験率は現在では 40 %を超えています。ですからこの二十数年、四半世紀の間に状況が大きく変わった。センター試験というものがかなりマイノリティ、特殊な一部の学生さんが受けるものから、半数にはいきませんけれども、高校生の 4 割が受けるポピュラーな試験にだんだん変化してきたということです。

特に受験者数が増加したのは、1990 年代です。だから平成 1 桁のときに受験者数がどんどん上がってきました。平成 10 年を過ぎると数としては安定しております。

これは内田さんが入研協セミナーのためにつくった資料ですけれども、18 歳人口はどんどん減ってきているわけです（スライド 4)。ところがセンター試験の受験者数は、このように平成 1 桁のときにずっと数は上昇して、それ以降は数としては一定しているんですね。ところが、高校生のほうはどんどん減っていますから、受験率は相対的に上がってきている、そういう状況です。だから年代にもよると思いますが、私以上ぐらいの人たちが想定していたセンター試験あるいは共通 1 次試験というものと現在のセンター試験だと、そのターゲットとする学生さんの層が全然違っていると言えるかと思います。

科目平均の年推移

次に、科目平均の年推移についてです。

センター試験というのは基本的に科目得点の平均が大体 6 割であるようにしようとしているわけです。なぜ 6 割なのかという話は出てくると思いますが、基本的に、ほとんどの場合 5 択あるいは 4 択の多肢選択型の出題をしています。それが非常に批判を浴びたりすることも多いわけです。多肢選択の問題の限界というか、欠点としては、要するに、極端に難しい設問は出せません。出せないというのは、出すことはできるんですが、試験として余りよくない。要するに、先ほど前川先生のお話にもありました、測定のツールとしてよくないということです。それはなぜかというと、非常に難しい問題、ターゲットとする学生さんにとって非常に難しい問題を出すと、その問題に正答した人のほとんどがまぐれ当たりによって正答した人になってしまう。そうすると、測定の尺度として見たときに、ほとんどノイズとしてしか機能しない。だから、ある程度の人ができる問題を集めないと、なかなかいい試験にならないわけです。

センター試験の場合は、ただつくりたい人がつくりたいようにつくっているわけではなくて、レビューの委員会もありますし、複数の委員会のチェックを受けているわけです。高校関係者のチェックも受けています。とはいえ初出問題、要するにプリテストなしで期待どおりの科目平均点を得るのは現実には非常に難しいということは、経験的にはわかります。特に理系の科目で平均点の変動が非常に大きいということがあります。

これは数学の例です（スライド 8)。これは今のカリキュラムですね。来年からカリキュラムが変わりますが。2006 年から 2014 年まで同一のカリキュラムで、これが数学ⅠA です。教科の中で一番ポピュラーな科目ですけれども、こんなに変わるんですね、平均点が。2013 年―去年はかなり難しかった。今年はそれよりは少しやさしくなっています。これぐらい変わってしまう。

例えば 2010 年度に受けた高校生の学力が低かったのかと言われると、そういうことは多分ないので、やはり問題の難しさがこれだけ変動していると考えざるを得ません。

数ⅡＢのほうはまだ少し安定しています。低いところで安定しているんですけれども、これぐらい。それでもやはり結構変わる。

もう一つ、これは理科です（スライド 9)。P は物理Ⅰで C が化学Ⅰ、B が生物Ⅰで E が地学Ⅰですね。これも、例えば化学を見てもらいますと、こう下がって上がってガンと落ちてまた上がって、一度落ちてまた上がる。地学もこのように、ここが下がってその次に上がって､これぐらいで推移して今年はかなり難しかったというようなことがわかります。どうしても変わってしまう。

数学も理科も、いずれもはっきりとした大問形式です。要するに、1 つのテーマがあってそれについて幾つかの小問がついているという形です。その小問の間の関連性が高い。それがために、1 つ間違えると他の問題も関連して間違えてしまう可能性がかなり高いので、余計、問題の難しさをうまくコントロールすることが結構難しくなっているのが実情です。これも複数の先生方に 6 割とれるように事前に慎重に考えていただいてはいるんですが、それでもやはり難しい、これが実情です。

科目ごとの合格者と不合格者の平均

次に、合否に関する情報を紹介します。

点がたくさんあって見づらいんですが、これらの点が 2013 年―去年の前期日程の学部とかコースです(スライド 10)。これ 1 つが 1 つの選抜の単位をあらわしています。横軸が合格者の平均点、この場合は英語筆記試験の合格者の平均点。縦軸が不合格者の平均点です。

当然ですけれども、合格者のほうが不合格者より成績がいいわけですから、合格者の平均のほうが高いです。190 点ぐらいとらないと合格しないようなところもありますが、ここら辺（合格者平均 150 点くらいのところ）で一貫して合格者のほうが 10 点か 20 点ぐらいでしょうか、こちらのほうが点数が高いことがわかります。センター試験の英語なんてやさしいからみんな満点だとかおっしゃる人もいるんですが、それでも結構上のほうでもやはり多少は差がついています。ただし、一番合格者と不合格者の差がつくのは真ん中、ボリュームゾーンのところですね。ここらあたりで一番差がついていると思います。

もう一つだけお見せします。

これは 2013 年の数学ⅠＡの点数です(スライド 11)。この年の数学ⅠＡは大変難しかったんですね。そのために、これを見るとずっとこういう形になっていて、成績上位層、成績のいいところで合格者と不合格者の差がたくさん出ている傾向があります。これがやさしいとまたさっきのように弓なりになって、上のほうでは余り差がつかないということが起きますけれども、この年は上のほうで差がつくという傾向が見られました。

これは省略します（スライド 12)。

科目得点相関

これは次の話題で、科目間の得点の関係です（スライド 13)。

いろいろ細工をしました。ここに示した科目は全部ではありません。受験者数の少ない科目は省いてあります｡地歴のＡ科目とか情報関係基礎とかフランス語とかドイツ語とか、そういうものは省いてあります。それで相関係数、要するにデータがどの程度リンクして動くか。その相関係数の大きさを〇であらわしてあります。当たり前ですけれども数ⅠＡと数ⅡＢの点数は相関が高い。片方が高ければもう一方も高い。

それから、これも当たり前と言えば当たり前ですが、これは英語筆記試験で、ここがリスニングです。別科目として見ていますから、英語筆記試験と英語のリスニングの試験はすごく関係が高い。全般を見てやりますと、このあたり（左上部分）はみんなそれなりに相関関係が高いんですね。理系の科目です。ここが数学で物理Ⅰ、理科総合Ａ、化学Ⅰ、生物Ⅰ、地学Ⅰ、それから英語の筆記という感じで並んでいます。

それから「倫理、政経」が結構いろいろな科目と相関が高い。一番多くの科目と相関が高いのは英語の筆記試験ですけれども、「倫理、政経」も結構相関は高い。意外と地歴の科目と理科が大きな相関を持っています。

ただし、理科とか地歴、公民の科目は第 1 解答科目の点数だけをとっています。2 コマあって、地歴、公民の中から最大２つとれるんですが、第 1 解答科目の点数だけをここでは使っています。そうすると、地歴、公民では 1 人の受験者に関して 1 つの科目のデータしかありません。本来、地歴間の相関は推定できないんですけれども、いろいろ統計的な仮定を置いてやって、それでも数学とか英語とか国語は共通に受験しているので、その情報を使ってある種の補完を行って、相関係数を推定しています。

表示の大きさについては､ゼロのときにゼロにしてしまうと違いが余りわからないので、ちょっと強調して、相関係数が 0.3 のときに大きさがゼロになるようにしています。それから相関係数が 0.8 のときに、ここに目盛りがついているんですが、半径が 1 になるように設定しています。ですから一番大きいところで大体相関係数 0.8 ぐらいです。一番小さいところでも相関係数は 0.45、それぐらいの関係はあります。

全体のまとまりがなるべくよくなるように科目順番を並べかえているんですが、ここら辺（左上部分）が理系の科目、地歴、それから政経、倫理、現社、一番反対側は、数学と一番関係性の薄い科目として国語が並んでいます。まあ当然かなという構造になっています。

これは別の分析をしたものです（スライド 15)。

表示している点はさっきと同じように一つ一つの学部ですけれども、横軸が英語の筆記試験、合格者の平均点です。縦軸が国語の得点ですね。完全に同じではないけれども、それなりの関係があります。ただし中間層、ボリュームゾーンですと点数はかなりばらける。ところが数学ⅠＡと数学ⅡＢだと（スライド 16)、個人は当然ある程度点数がばらけるんですが、集団で見るとはっきりとした曲線関係が出てきています。

科目選択の状況

このシンポジウムの文脈とはちょっと違うかもしれませんが、データを見ていて当事者としてちょっと気がかりなことは、科目数です（スライド 20)。要するに、東京、埼玉、千葉、神奈川、この 4 都県がかなり科目選択に関して特異な状況です。ずっと見てみますと 3 科目受験者が 37 %、7 科目受験者が 29 %です。ところが、例えば同じ大都市でも京阪圏、京都、奈良、大阪、兵庫では 3 科目受験者が 18 %、7 科目受験者が 56 %でした。それ以外の地方では 3 科目受験者 11 %、7 科目受験者 57 %でした。要するに、東京では特異的に受験科目数が少ない、そういう傾向はありますので、ちょっと偏っているのではないかなという心配はしています。

英語資格・検定試験との比較

あと資格試験の話ですが、これは本番の実際の受験者になかなか聞くわけにいかないの

でモニター調査の参加者に調査をしました。モニター調査というのは、センター試験の当日に大学 1 年生を集めて、アルバイトとしてその年のセンター試験の問題を解いてもらっています。本試と追試両方解いてもらっているんですが、参加申し込み時に英語の資格試験の点数を自己申告してもらいました。その自己申告した内容と本・追の英語試験の合計点の関係を見てみました（スライド 22)。

これは横軸が本試験、先ほどから誤差の話が出ていますけれども、縦軸が追試験です（スライド 23)。同じ作題チームが同じコンセプトのもとに、同じような問題構成で 2 セットの問題をつくっているわけです。それを同じ学生たちに受験させるとどうなるかということです。

見てわかるように、ここら辺（右上部分）に集中しています。実は非常に優秀な学生さんが多いので、筆記 200 点、リスニング 50 点、合わせた点数で 200 点超えが結構たくさんいるわけですけれども、それでも、相関係数を見ると 0.8 を超えています。こういうふうに偏ったデータでも 0.8 超えている。

これが（スライド 24、25）英検の点数でして、準 2 級だとこれぐらいの点数、2 級でこう、準 1 級がここで、1 級の方は 2 人です。当たり前と言えば当たり前ですけれども、級が上がるほど点数は高くなっている。ただし、準 1 級で見なし満点とか言われるんですが、少なくとも本試験と追試験両方あわせてやると、センター試験の満点は結構難しいです。現実には。

あと TOEIC と TOEFL についての情報がありますが、TOEFL はちょっと人数少なかったので TOEIC だけご紹介します（スライド 26、27)。TOEIC といっても、この場合、ほとんどは大学のプレイスメントテストでの TOEIC-IP の受験者です。横軸がセンター試験で 500 点満点、縦軸が TOEIC または TOEIC-IP のスコアです。六十数人いるんですが、こういう感じです。

これは回帰分析ではなくて、分布の嶺を推定したものです。真ん中をある種の統計的な方法で推定してみました。そうすると、TOEIC800 点で 9 割は超えると思うんですが、ちょっと満点は難しいのではないか。900 点ぐらいいかないと、センター試験の本・追合計の満点はなかなか難しいのではないかなと思いました。

あと、モニター参加者はかなり優秀な学生さんたちですけれども、注意すべきことは、本物の TOEFL-iBT は 368 人中 2 名しか受験経験者がいなかった。その程度の普及率だということです（スライド 28)。

まとめは、最後にこういうふうに書いてあります（スライド 29)。

以上です。

○司会　大津先生、どうもありがとうございました。(拍手)

何か、事実確認程度のご質問はありますでしょうか。

○義本氏　文部科学省の義本です。ありがとうございました。

特に首都圏で 3 科目受験が多くて関西と大きな違いがあるというのは、どういう背景があるんでしょうか。例えば私学との併願が多いから、そういうふうなことがあるんでしょうか。

○大津氏　恐らくはそうだと思います。関西も私学の受験者いるとは思うんですが、関西

ではセンター試験の受験率自体が関東に比べて低いので、そのために余り関西だと差が見えていない、3 科目受験者数がそれほど多く見えていないのかもしれません。そこまでは今、ちょっと詳しく数を持っていませんのですみません。
○山崎氏　埼玉医科大学の山崎と申します。非常に興味深いお話、ありがとうございました。途中で理科のお話をしていただいたところで、大問形式の問題だと小問の間で関連性が高いというようなことをおっしゃったと思うんですけれども、先ほどのディスカッションの中で、大問形式でも小問の間のディペンデンシーは高いという討議もあったように思います。実際にセンター試験で大問形式の中の小問の問題の相関が高いというデータをおとりになっているなら教えていただけないでしょうか。
○大津氏　ちょっと今、手元にはないんですが、例えば数学の場合でしたら一つ一つの問題を普通の形の IRT で一問一問バラバラに当てはめようとすると、ちょっと実情とは違うなということは明らかに見えます。それぐらいの答えになるんですが。
○山崎氏　理科でもそうでしょうか。
○大津氏　今、全部は見ていないので詳しくは言えませんが……
○山崎氏　実際は、そのデータはお採りになっているわけですね。
○大津氏　そうです。
○司会　他によろしいでしょうか。
　それでは、大津先生もパネルディスカッションに参加されますので、またそのときにお願いいたします。
　大津先生、どうもありがとうございました。（拍手）

大学入試センターンポジウム　平成26年11月29日
東京工業大学 社会理工学研究科 デジタル多目的ホール

センター試験で何が測られるのか？

大学入試センター 研究開発部
大津 起夫

概 要

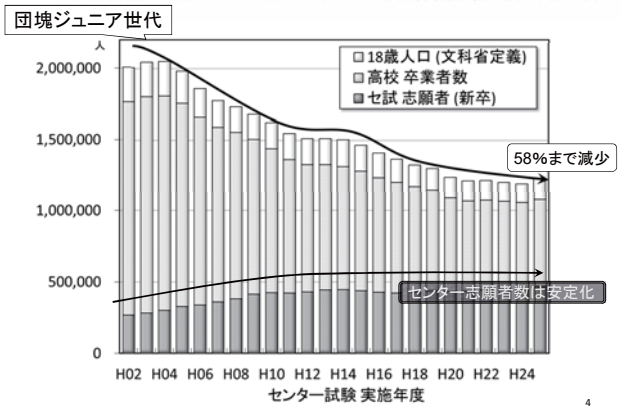
センター試験の実情はどのようなものか？

- センター試験の受験率はどう変わった？
- センター試験と国公立大の合否の関係
- 科目平均点の推移
- センター試験科目得点の相関関係
- 選択科目数の地域差
- 英語科目得点と資格・検定試験

センター試験 受験率の変化

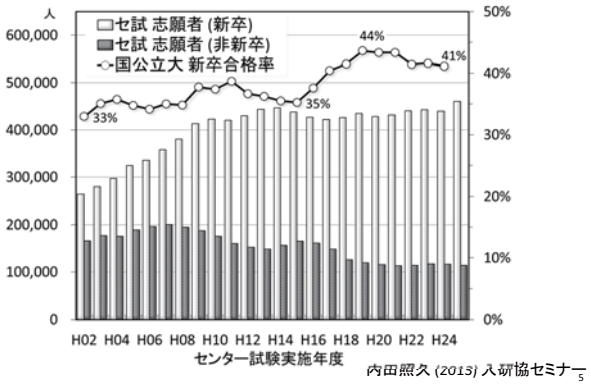
- センター試験第1回 (平成2年,1990)
  現役約25.1万人，既卒他15.6万人が受験
  現役受験率は約14%(現役志願率15%)
- 現在 (平成26年,2014)
  現役約41.7万人，既卒他10.8万人
  現役受験率は約40% (現役志願率 42%)
- 受験者数は1990年代に増加
  受験率は1990年以後最近まで増加傾向

18歳人口とセンター試験志願者数



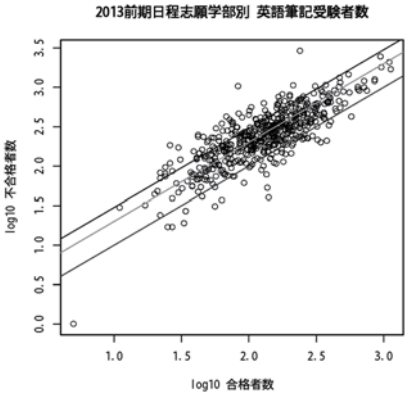
内田照久 (2013) 入研協セミナー

新卒合格率の推移



内田照久 (2013) 入研協セミナー

2013（平成25）前期日程 学部別志願倍率



対数尺度
学部・コースの数は540(複数の選抜単位を含む場合もある)

赤2倍、
緑3倍、
青4倍

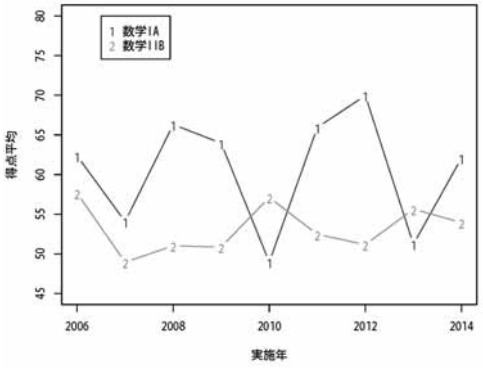
多くが志願倍率2～4の範囲

科目平均の年推移

- センター試験は科目得点の平均がおおよそ6割であるように，意図されている
- 多肢選択型の出題では，極端に難しい設問は，識別力を損なう（SN比が悪い）
- 複数の委員会によるチェックを受けてはいるが，初出問題で期待通りの科目平均を得ることは，現実には難しい
- 理系の科目で平均点の変動が大きい

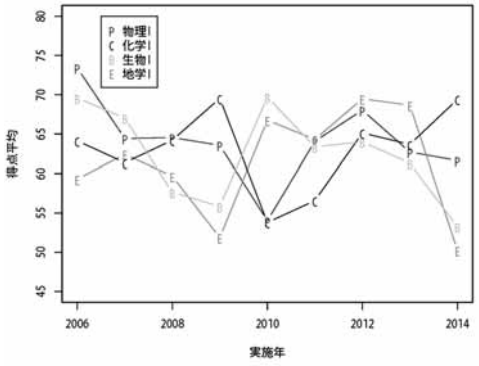
7

科目平均点の推移　数学
平成18-26(2006-2014)年

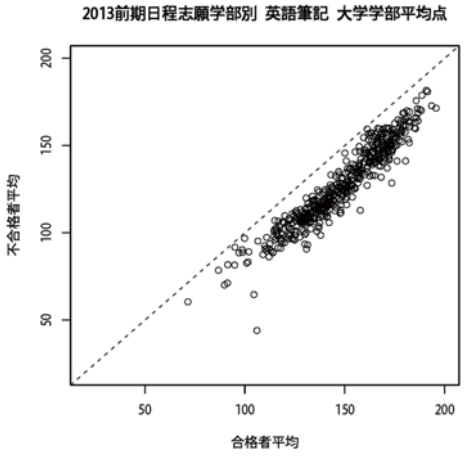


8

科目得点の平均点　理科
平成18-26(2006-2014)年



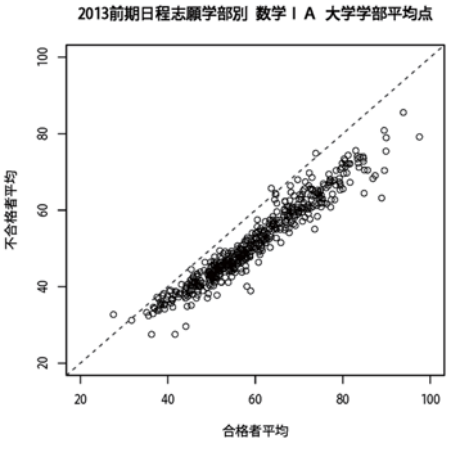
9



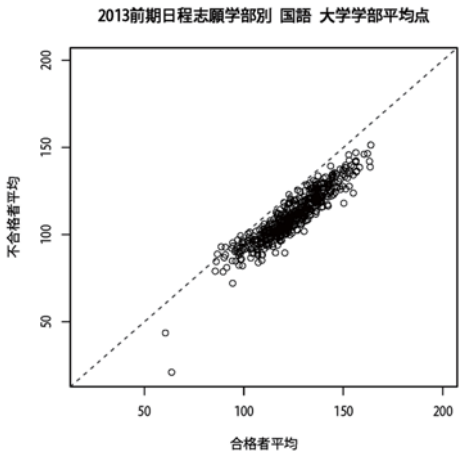
各点は大学の学部またはコースを示す

前期日程における学部・コース数 = 540

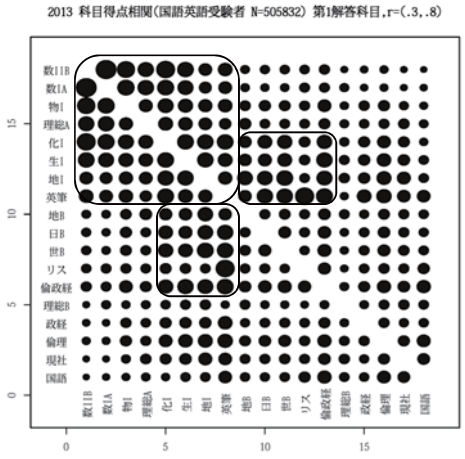
10



11



12

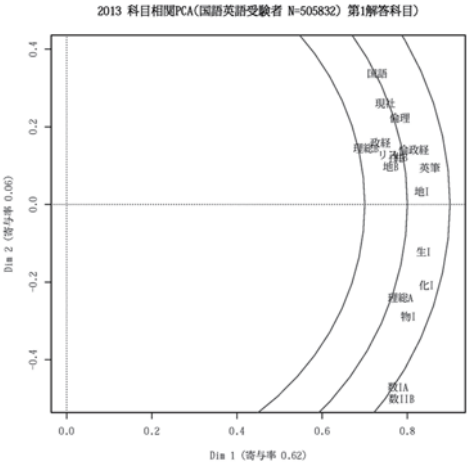


H25年科目得点の相関
地歴・公民・理科は第1解答科目

r=0.3が直径0，r=0.8が直径1となるように表示

欠測の場合は，他科目を考慮して推測（MARを仮定，Rのnormパッケージを利用）

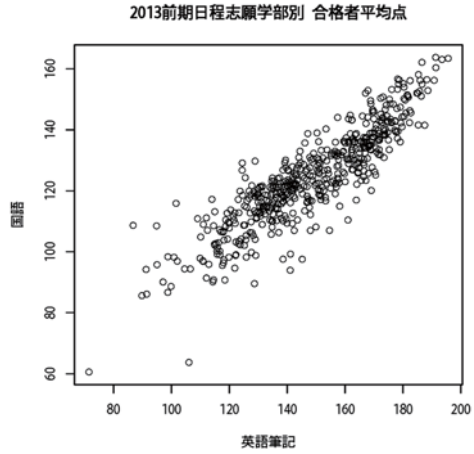
相関係数の範囲は，凡そ0.45－0.80



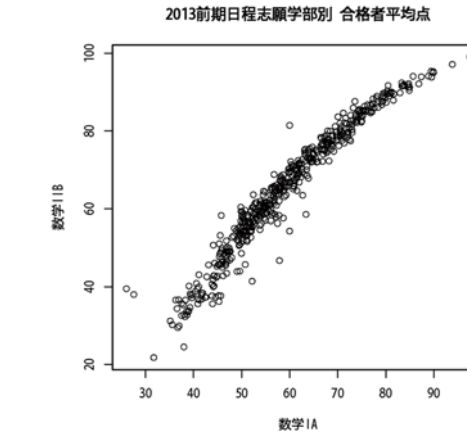
推定した相関行列に基づく主成分分析（PCA）の結果

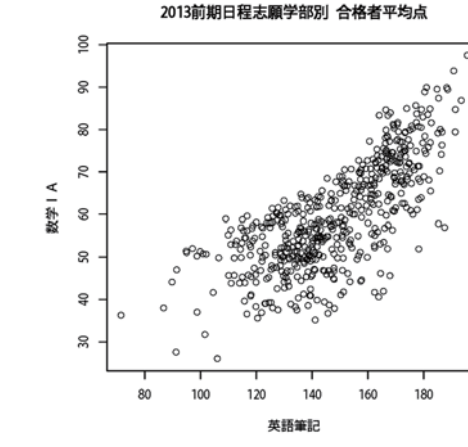
原点からの方向が近く，長さが1に近いと相関が大きい

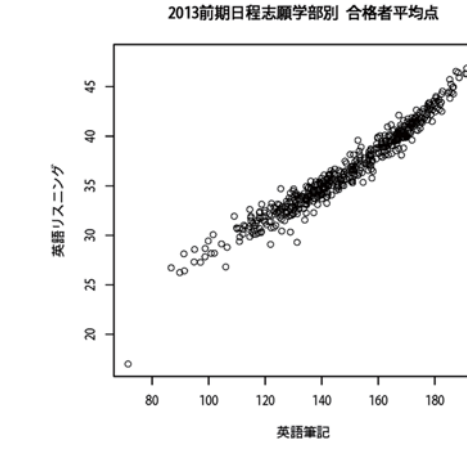
第１主成分(横軸）が過半の変動を説明



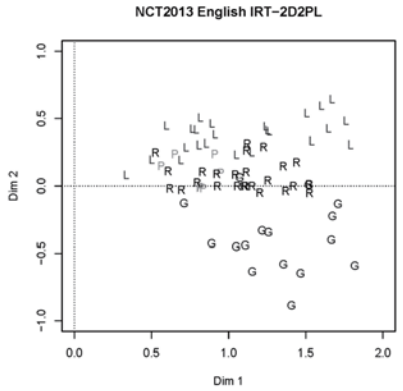
各点が各大学の学部またはコースを示す







英語設問項目の統計的特徴



平成25年　英語本試験（筆記，リスニング）解答の構造

2次元のIRT（2PL）によって得られた識別力パラメータ
ただし，リスニングの18、19は局所依存性が高いため，段階反応モデル(鮫島モデル)でまとめて一問とした。

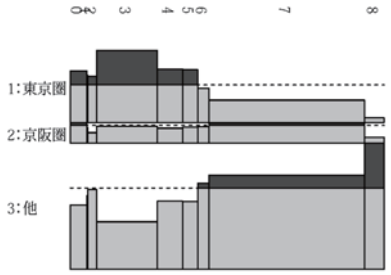
P:筆記第1問（発音）
G:筆記第2問（文法）
R:筆記第3～6問
L:リスニング

科目選択の状況

- 気がかりなこと：　選択科目の偏り
- 東京付近（埼玉，千葉，神奈川）で，3科目受験者が7科目受験者を上回っている
- 平成25年の地域別受験科目数
- 東京圏では3科目受験者が37%，京阪圏（京都，奈良，大阪，兵庫）では18%，他では11%

科目数	0	1	2	3	4	5	6	7	8	計
1:東京圏	11347	685	5489	58711	17924	10717	5112	46543	1187	157715
1:率%	7.19	0.43	3.48	37.23	11.36	6.80	3.24	29.51	0.75	100.00
2:京阪圏	4558	225	1234	13666	5000	3249	2464	42063	1495	73954
2:率%	6.16	0.30	1.67	18.48	6.76	4.39	3.33	56.88	2.02	100.00
3:その他	14177	1069	9429	38581	22873	13741	12875	195811	33119	341675
3:率%	4.15	0.31	2.76	11.29	6.69	4.02	3.77	57.31	9.69	100.00

2013　地域別科目数別受験者数



1:埼玉，千葉，東京，神奈川
2:京都，奈良，大阪，兵庫
3:それ以外

高校所在地に基づく区分

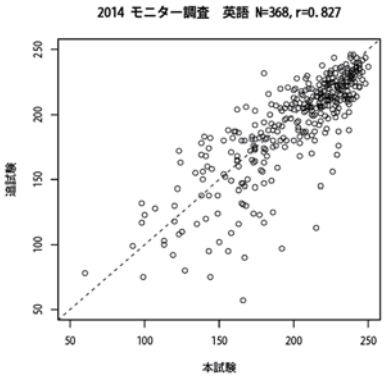
セルの面積が人数を示す

英語資格・検定試験との比較

- 難易度が標準化されている（はず）の試験と比較したい。（テストの目的や構成は，必ずしも同じではないが….）
- モニター調査の参加者を使う
  （センター試験，本追試験同日に，同年度の試験問題を時間遅れで解答させる）
  都内国立大学1年生（368名有効）
  参加申し込み時に，英語外部試験の受験歴，得点を記述させた（得点・合否は自己申告）。
- センター試験の英語得点は，本追の合計を使う。

大津(2014) テスト学会

2014モニター　英語本試験と追試験



モニター調査参加者の英語得点（筆記＋リス）

参加者は都内の大学1年生

横軸：本試験　縦軸：追試験

得点の相関は r=0.827

2014モニター:外部英語試験の受験経験

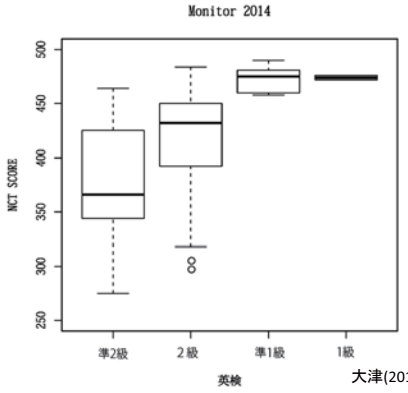
表 2: 平成 26 年モニター調査　英語外部試験受験経験とセンター英語試験成績

外部試験	人数	本試験平均点		追試験平均点		本追合計	
		筆記	リス	筆記	リス	平均	SD
英検 1 級	2(1)	192.0	48.0	190.0	44.0	474.0(492.0)	2.8
英検準 1 級	7(9)	194.1	48.9	184.0	45.1	472.1(466.8)	12.8
英検 2 級	48(28)	171.5	44.2	163.6	40.0	419.2(409.1)	45.5
英検準 2 級	17(10)	150.2	41.3	148.4	37.3	377.2(407.6)	50.6
TOEIC/TOEIC-IP	63(55)	158.8	43.0	154.7	38.4	395.0(396.7)	57.6
TOEFL-ITP	29(9)	174.8	44.8	170.8	41.7	432.0(491.3)	28.2
TOEFL-iBT	2(1)	196.0	50.0	190.5	46.0	482.5(438.0)	3.5
全体	368(348)	163.0	43.3	155.6	38.8	400.8(394.6)	67.8

英検は各級合格者の得点，英検以外は受験経験者のセンター試験英語得点を示す。英検 3 級以下は省略した。
括弧内の数値は，平成 25 年のモニター調査の結果を示す。

大津(2014) テスト学会

2014モニター：英検取得級とセンター試験



英検の受験はほとんどが大学入学以前。
英検の取得の再高級を示した。
3級以下は省略

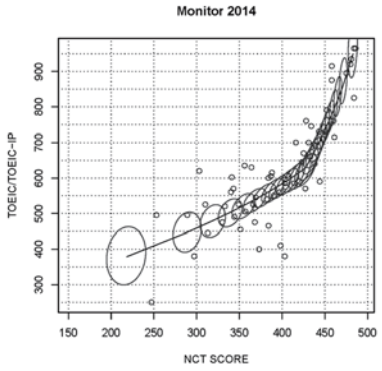
大津(2014) テスト学会

2014モニター：センター試験とTOEIC/TOEIC-IP

- TOEIC: 聴解と読みから構成される（センター試験と類似）
- 世界で年間延べ200万人超の受験者（集団受験の方がやや多い）
  source：国際ビジネスコミュニケーション協会
- モニター参加者のTOEIC受験は，多くが大学入学後。大学で実施されているプレースメントテストとしての受験が多いと推測される。
- 英検との関係について**は**，www.toeic.or.jp に資料あり。
- TOEIC-SW，TOEIC-Bridgeの報告はなし。

- TOEICスコア（TOEIC-IP含む）報告者63名
  TOEICスコア平均 619.1点(S.D. 153.5)
  センター本追計　395.0/500点(S.D.　57.6)
  相関係数　　r = 0.803
  （2013年は，N=55,r=0.768）

2014モニター：センター試験とTOEIC/TOEIC-IP



横軸：センター試験得点（2014） 本追試験英語得点合計　500点満点

縦軸：TOEIC/TOEIC-IPの自己申告点

分布の峰を折れ線で推定(非線形因子分析）

大津(2014) テスト学会

2014モニター：センター試験とTOEFL-ITP

- 留学資格に使われるTOEFL-iBT(オンライン4技能試験）の受験経験の報告は368名中2名のみ
- TOEFL-ITP（集団実施版；聴解，文法と表現，読解の3領域）
  上級レベル(LEVEL1)310点～677点，報告者は29名（うち1名は得点誤記と推定）
- TOEFL-ITP 28名，平均505.5点，S.D. 41.3
  センター試験との得点相関 r=0.699
  (2013年は N=9, r=0.730)
  2014年　TOEICとの共通受験者は1名のみ

まとめ

- 平成2年から平成26年の間に、
  現役志願率は15%から42%へ上昇
- 初出問題での科目得点の変動
- 英語は，国公立中位で合否に強く関与
  数学は上位で関与の程度が強い
- 全ての科目得点の間で正の相関がある
- 埼玉,千葉,東京,神奈川で3科目受験者が多い
- センター試験英語得点は資格・検定得点と強く関連

【報告４】

高大接続のこれから

中京大学現代社会学部長
村上　隆

○司会　報告の最後となりますが、「高大接続のこれから」と題しまして、中京大学の村上先生にご報告をお願いします。

○村上氏　村上でございます。よろしくお願いいたします。

大変充実した報告が続いておりますので、もう皆様お疲れではないかと思いますが、私の場合は理論的にそう難しい話ができるわけでもございませんで、専ら経験に基づいて、今、考えられている、いわゆる入試改革において今後どんな障害が起こりそうかといったことを中心にお話ししてみたいと思っています。

高大接続の改善には、当然入試改革が含まれるわけですけれども、今の時点でなぜこういうことが問題になるかというと、1 つは、教育に対して社会的要請が変化しつつある。これは大きな産業構造の変化とつながっているのではないかと思いますが、そういうことは今、論じるつもりはありません。その一方で、入試に対して利用可能な技術革新が起こってきていることへの期待もあるのではないかと思われます。

また、私自身は自分がかかわった公的試験の立ち上げに当たって、幾つか経験したエピソードがあるわけですけれども、それについてお話しして、その上で、これは今後の入試改革にかかわる議論の一つの焦点になると思いますけれども、CBT ―コンピューター・ベースド・テスティング―が今日問題になっていますが、それがいわゆる日本的テスト文化の間にどんな摩擦を興す可能性があるかといったことを中心に議論してみたいと思います。

くわえて、ある能力観といいますか、理念というものがある程度成熟してくるということがあるとして、それとともに、それに対応する技術がある程度整ったとしても、それが本当に、例えば今の中教審答申のようなもので期待されている、いわゆる真の学力の評価に本当につながっていくのかどうか。これは日本のテスト風土ということだけではなくて、やはり能力評価というのは、行財政の問題に大いに依存すると私は思っています。そのあたりについても少しお話ししたいと思います。

日本留学試験の経験

ところで、私が立ち上げを経験した試験というのは、実は先ほど前川先生からご紹介いただきました日本留学試験でございまして、この試験はたびたび話題になっております項目反応理論に依拠した得点等化を行った、恐らく我が国で最初の、いわゆるハイステークスな試験、いわゆる入学試験であるかと思います。CBT にはほど遠い紙ベースの試験で

ございますけれども、入試センター試験あるいは今、考えられているその後継試験が CBT 化されるとすれば、避けて通れないような問題が幾つか提起されていると思いますので、その点について多少ともご参考になればと思っている次第です。

日本留学試験、このスライドは Wikipedia からの引用でございますので省略させていただきますが、年間 2 回実施されている、それから 2 年間にわたって有効である、そのためには得点が相互に比較可能でなければならないという条件があったということ。かつ、この試験の受験者が日本人ではないという点において、先ほどから出ております国際標準というものに、やはり合わせなければならないということがあったとも思います。

そのために、特別研究推進費、いわゆる政策科研というものをいただくことができて、私自身がそのときは僭越にも研究代表者でしたけれども、今、この会場をざっと見回しても 6 名か 7 名の方に、研究協力者あるいはその後の実務に携わっていただく形でかかわっていただきました。そこで扱った問題は、やはり IRT による得点等化、それから問題作成体制の整備、それから、先ほどから出ておりますデータベースの構築。また、試験全体へのバックアップ体制、人的体制というものがあります。それからもう一つは法的問題への対応というのが、私自身はそのとき、海外でこの試験に対する訴訟が起こされるようなことを危惧しておりましたので､それについても研究しなければいけないと思った次第です。

得点等化ですけれども、どんなことが問題になったのか。特にテストのソフト側の面と言ってもいいかと思うのですが、公的試験というのはどうしても非常に偉い方の委員会が上についてまいります。そういう人たちのご意見であるとか、あるいは作題者のご意見、あるいはユーザー側のご意見がいろいろな形で抵抗として、一種の摩擦につながるということがございます。でも、これは今日のお話の中では省略させていただきます。

そのとき何を理解してもらうべきだったのかということですが、このスライドに書きましたのは、先ほど申し上げた政策科研を申請した際についてきたコメントであります。まず第 1 に、欧米で開発された手法があるのであれば、それをそのまま使えばいいのではないか。これは今日の問題点の 1 つでございます。

それから、我が国にもそういう得点等化を可能にしている試験は既に存在するのではないか。確かにそれはございました。例えば、南風原先生たちの東京大学のグループでおつくりになった語彙能力の試験といったものは既に存在していたわけですけれども、これはハイステークスな試験というわけにはいかなかっただろうと思います。

その他にも、学術研究として行われた例があるのではないか。これはそれこそ、そうですね。レベルの高いものが存在しておりましたが、現実場面で使われる試験というわけではなかった。また、データベースを科研でつくっておいて、それを事業に転用していいのか。まあ、ありそうな話ではあります。目指しているのは、あくまでも研究段階で、実用化までは無理だと答えました。

今もこういった反論は、とりあえず出てくるのではないかと思うんですが、その中でもまず、IRT というものに対して一定の誤解があったということ、これは申し上げておかなければならないと思います。

項目反応理論 IRT に関わる誤解と課題

IRT という理論が、あるいはプログラムがあれば得点等化はできる。それから試験の大

部分の問題は解決できるのではないかと思われていた偉い方々、これは計量心理学の専門家ではない方々ですけれども、そういう方々がいらっしゃった。失礼ながら、私はそうした方々を半可通と呼ばせていただきますが、例えば、これはちょっと極端な例ですけれども、IRT というもの、これは latent trait という、今まで今日のご報告の中では単に trait と呼ばれている場合が多かったと思いますけれども、この latent trait という言い方、これは要するに、直接観測することはできないけれども、項目反応から推測される一つの能力の次元といったものを指しているわけです。けれども、この latent という言葉は、潜在している、今はあらわれていないけれども将来開花するかもしれない、そういう可能性としての能力のようなものを指すという辞書的意味は、確かにあるんですね。つまり、同じデータでも IRT というプログラムを通すことによって潜在能力の指標になり得るのではないかという風に思っておられた方がある、これはとんでもない誤解だと思いますけれども、まずはそういうところを正すことから始めなければならなかったのです。

私はたまたま今の科研の代表者であるとともに、幾つかの点で外部とのインターフェイスの役をしておりました。実際にテクニカルな面でご活躍いただいたのは、今日もここのおいでになっている何人かの先生方ですけれども、私の方は、インターフェイスとしていろいろな経験をしたということを、今日はお話ししているわけです。

ともかく何か計算すれば複数回試験の間の得点を比較できると思っておられた方が多くて、これはちょっとこの分野の研究者の誇大宣伝にも原因があるなと、実は思っています。それから、この手のものは数学者にはほとんど認めてもらえない、ここもおもしろいところです。

ともかく何をわかってもらわなければいけないかというと、得点等化が可能になるためには IRT という数理モデルはあったとしても、得点項目がある程度共通している 2 つの試験の間か、あるいは受験者が共通している 2 つの試験の間でしか得点調整は可能でないということなのです。ただし、受験者は時間がたてば能力が上がったり下がったりしますから、非常に接近した時点で同じ受験者に 2 つの試験を受けてもらった場合に限られます。ただし、それでは、いきなり良問をそろえるのが難しいから、予備テストといって、実際の受験者集団とほぼ同等と思える少数の受験者にとりあえず受けてもらうことが望ましいということになります。すると、ある一定数の人たちに、その後で使われる問題を見せなければならなくなる。しかし、第 1 の条件、つまり 2 つの試験に同一の問題を含めるというのは、端的に言えば問題の再利用ということですから、予備テストもできるだけ秘密が保てる条件で行って、基本的に問題は非公開にするしかありませんが、IRT をすごく推進していらっしゃると思っていた先生が「試験をやったら全部の問題を直ちに公開するのが常識である」とおっしゃったりするので、これも大変困りました。

もう一つは、やはり IRT の理論構成上、今日のご報告の中にもあったと思いますけれども、基本的には小問主義、つまり多数の小問からなるテストにすることは不可避です。このことは、常に問題供給の体制を構築しておかなければいけない、そしてそのデータを蓄積しておかなければいけないことを意味します。それから、問題を一部分は公開しなければならないとすれば、その公開可能性の範囲をどう決めるかということがありますし、さらに予備試験ができるのかできないのか、さらに、これはある意味で最後の手段ですが、

モニター試験という、留学生ではない人たちを使って問題の難易度を推定するといったことが起こってきたときに、そのセキュリティをどうするのかといった多くの厄介な問題につながっていきます。

この辺に非常に神経を使わざるを得ないのは、これは東アジア的試験風土とおっしゃる方もありますけれども、過去問に非常にこだわる人たちが多いということにも起因するかも知れません。過去問を何としても手に入れたい、あるいは試験問題を見たらすぐにコピーする慣習があるような、語学学校などで聞いてみるとそういうことがあるということから、どうしてもセキュリティの問題に神経を使わなければならなかった。

試験の実施体制の課題

これ以外にも、やらなければいけないことが幾つかありました。たとえば、海外で訴訟を起こされるのではないかということは、私はかなり本気で心配しておりましたが、今までのところ、こうしたことは起こっていないようです。

ともかく、全体としてどのぐらいヒトとカネを投入しなければいけないか。その時点で私は口を開けば「ヒトとカネ」と言っておりましたけれども、では、具体的にどのぐらいのものが必要なのかはよくわかっておりませんでした。もちろん、技術的な面では、先ほどご発表いただいた前川先生を初めとする現場の方々、非常なご苦労があったと思っています。

一つの例として、やはりアイテムライター集団というものをどうしても持っていなければいけない。大学の先生に片手間で問題をつくっていただくような状況ではとても追いつきませんので、外部にアイテムライター集団を抱える必要がありました。

こういったことが比較的スムーズにいったのは、日本語教育という分野は比較的海外事情に通じた人が多い、あるいは海外で教えてきた、あるいは海外で教育を受けてきた人たちが多かったので、こういったスタイルをとり得たんですが、他方で、この後で申し上げるような、どうしても問題をよりオーセンティックなものにしたいという要求が強くて、ここのところの妥協が結構難しいということがありました。トレーニングをしたアイテムライターは、大体大学院生レベルの人たちでしたけれども、この人たちでもスキルの習得は比較的早かったように思っています。

それから、余り考えていなかったことが幾つかございました。まず、だれが試験監督をするのか。もはや、大学教員は動員できないので大学職員の方々にお願いしたわけですが、私も試験場に入れてもらいましたけれども、やはりなかなか慣れないということに問題があった。

それから、外国人が受けるということは、日本の試験の常識が余り通用しません。試験場にある入学試験の張り詰めたような雰囲気は、余り感じられません。結果的に非常に不正行為がたくさん発生して、それの摘発に追われるようなことにもなりました。さらに言うと、これは、別の試験での話ですが、試験問題を先に送付しておくと、なんとそれを開封して配ってしまう方が存在するということもあって、グローバリゼーションというのはひょっとするとこういうことでもあるのかなと思ったこともありました。

日本人対象にしても、これから問題を撮影されてしまうとか、あるいは音声を録音されてしまうといったことは当然起こってくるのではないかと思います。

小問形式問題と authentic 問題

作題に関してですが、CBT 化された試験を考えるとすれば、どうしても多肢選択の問題になるだろうと思います。そして、先にも申し上げた小問形式というものが、試験の得点の信頼性、妥当性を確保する上で大変重要な側面だということもあるわけです。けれども、特にばらばらの知識・技能、ここのところが中教審答申などで非常に批判されているところだと思いますけれども、語学の場合には、こういうものを discrete points（離散的点）と呼ぶようです。試験範囲を公示するとすれば、こういうものを網羅的に列挙するのは比較的容易です。それを一覧表の形にしておけば、ほとんど自動的に問題ができるのではないかという考えもありました。初期の日本語テストの問題は、こういった形のものだったのですけれども、これでは真の言語能力は測れないのではないか、要するに、いま、入学試験で問題になっているような表現力であるとか思考力であるとか判断力であるといったものは測れないのではないかということになります。では、いわば真の学力に相当するものを測るためには、もっと authentic（最近は「真正な」と訳されているようです）問題とはどのようなものかと言うと、このスライドのようなものになります。2 ページにわたるこの程度の長さの問題ですが、これでたった 1 つの答えしか求められない。つまり、先ほど前に映した問題であれば 30 分のテストに 60 問とかそういったことができるわけですけれども、こういう問題は 1 問やるのに 2 分程度がかかりますので、まず、同じ時間かけても、信頼性が低下してしまうという問題もあるわけです。さらに、こうした問題の作題の難しさがありますし、それ以外にも、いろいろと問題が出てきます。

たとえば、実際には複数の問題を組み合わせて 1 つのテストを作っていく。それも複数のバージョン作っていくときに、例えば今のリスニングの問題などは、一定時間にちょうどおさめることが非常に難しくなってきます。ある教室では 2 分ぐらい時間があいてしまうとか、別の教室では本当にぎりぎりになるといったことも、ある程度の長さの問題を考えようということになると起こってまいります。

ところが、どうしてもこういった方向に行かざるを得ないのは、テストの内容が実際の教育課程に対して逆に影響していくということ、これを遡及効果と呼ぶようですけれども、どうもこのテストで測れていない真の学力が存在するのではないかといったことの根拠は、このことにあるのではないかと思います。テストの形式、内容が教育内容を逆に規定してしまう。つまり、テストの準備教育というのが教育の目標になってしまいますと、テストで点数をとることそのことが、やはりテストの点数に対して影響を与える。つまり、真の能力というものがあるとすれば、このテストは真の能力を測っているのではなくて、いわばこのテストのための能力を測っていることになってしまうのではないかということが、今回のみならず常に改革のモチベーションになってきたのではないかと思います。

日本語のテストについて言えば、私は日本語教育の先生に、次のような議論を吹きかけたことがありました。先ほど、authentic でないテストの例としてお見せした離散的な知識や技能のテストは、日本語の native speaker であれば、ほぼ間違いなく正解できるようなものです。つまり、これらは日本語能力の必要条件ではあるわけですね。もしも、こういう形のテストで高い得点が取れるような教育というものは、離散的な点を網羅的に丸暗記させるような、いわば使えない知識技能を憶えこませるようなものではなくて、実地に

authentic な言語使用を通じて獲得するようなものであることを、日本語教育の先生方が証明してくだされば、テストの方は作題が難しい割には信頼性の低い authentic なものを作らなくても、離散的な点を並べたようなものでいいのではないか、と言ってみたわけです。しかし、やはり学習者は手っ取り早く点数がとれる暗記主義に走るわけですし、結局、試験の形がいわば薄っぺらな使えない能力の獲得という方向になることを反証することは、実際問題として難しいようでした。

記述式問題・PISA 型問題と信頼性

そうするとさらに、一足飛びに、客観形式ではなくて記述形式の問題にしたほうがいいのではないかという主張も出てくるわけですね。CBT からはだんだん遠ざかってくるかと思いますけれども。これについては採点に主観が入る、採点者によって得点が変わるのではないかという話は常に起こります。けれども、私の経験からすれば、これは十分な打ち合わせをする、あるいは一定以上の点差がついたときに改めて協議をするといったことによって相当の信頼性が得られると思います。

むしろ問題になるのは、この種の試験では、まさに大問間の相関が極めて低いということだと思います。これはもう 30 年ほど前に日本で推薦入試が開始されたころに、たまたま 2 回の小論文を受験する受験者の得点を比べてみたときに気づかれていました。私は、これはまたもう一つ別の経験になるわけですけれども、名古屋大学では複数の問題を含む小論文試験、いわば小論文の問題をオムニバスにしたような論述的学力検査という名称の試験をやっていたわけですけれども、そこでもはっきり確認することができたわけです。

他方、これについては、やはり逆の遡及効果というものもあるわけで、最後に書きましたように、こういうことによってある種の受験者を排除することができる。つまり、他の試験ではそこそこ学力が高いのに何か記述させてみると A4 用紙の上の方に 2 行ぐらいしか書けないような人間はいるわけですけれども、そのような受験者を最初から排除することができるといった意味では、そのことの善し悪しはともかく、効果はあると考えると、こういう方向性はあるのかもしれないとは思います。

いずれにしてもこのあたり、中教審が示している一種の「理想」というものを考えたときに、越えるべきハードルは非常に多いのではないかと思っているわけです。

今日ぜひ伺いたいなと思っていたのは、やはり PISA 型テストについてです。中教審答申にも非常に明確に「これを目指す」と書かれておりますが、先ほどの山本先生のお話から私は理解したつもりですけれども、今の PISA 型テストは個人の評価を目的にするものではないように思っております。これはむしろ政策の評価のためのものではないかと。私も、とある進学校で学校評議員をしておりまして、そこはスーパー・サイエンス・ハイスクールの指定を受けている学校です。ここで文科省さんから PISA 型の評価方法を考えるようにという宿題をもらってきたという話がありました。私もにわか勉強で PISA 型の問題を検討させてもらったわけですけれども、個人を評価するということは、やはりこのタイプのテストでは諦めたほうがよろしいのではないか。非常に短い時間、一定の高校の 1 コマの授業ぐらいの時間で行われるとすれば、たまたま先ほど申し上げた複数の小論文の場合のように、たまたまその時間内に正答を思いついた生徒の割合が増える、それは決して個人の能力の信頼性の高い測定にはならないけれども、そういう中で正解に達する確率

が高まっていけば、例えばスーパーサイエンス的なアクティブラーニングをやった効果の測定にはなるだろう。つまり、平均値の変化を追うことは可能かもしれないけれども、こういった大きな問題によって個人を評価するということについて言えば、非常に大きな問題、難問が控えているのではないかと考えた次第です。

いささか駆け足になりましたけれども、これで終わらせていただきます。

ご清聴どうもありがとうございました。（拍手）

○司会　村上先生、ありがとうございます。

何かご質問ありますでしょうか。よろしいですか。

留学生試験のときに、問題作成の際にいろいろな意見が出てくるというのは実際に経験されたことなわけですよね。それで、やはりオーセンティックな問題と言いますか、最近の PISA 型の問題が推奨されるというのも同じ流れになるかと思いますが、そのような意見が多かったということでしょうか。

○村上氏　あの程度のサイズの問題ですと、問題数が少ないという意味で余り信頼性を高められないという問題はあるにせよ、比較的相互相関は高いので、IRT に載せていくことに決定的な困難があったわけではない。だからこそ、今、できているわけですけれども。

ですから、それではまだ今、問題にされているような、恐らく表現力というのは全く難しいでしょうから、それをやるとなると先ほどお示ししたのよりもう一段階大きな大問に向かわなければいけなくなって、このときに本当に難しい問題が出てくるのではないかと思っています。

○司会　記述式の問題なども新しい学力テストに含めるといったことが答申案には書かれているんですけれども、そういった採点者の信頼性の問題などについては、先生のご経験では、ある程度訓練すれば大丈夫という感じなんでしょうか。

○村上氏　実際には、大学教員はローテーションで採点を担当したりしますので、訓練というところまではなかなか難しいですけれども、事前の協議は非常に大事ですね。

それから、やはり採点したら分析して、ともかく差のあるところを見つけてもう一度協議をするといったことをすれば、相当の信頼性は確保できると思いますが、問題は、複数の問題間で相関が低い、場合によってはほとんど相関がないことであって、では、これは一体何を測っているのかということですね。いつか得意な問題が出てくれば、今まで余り高得点でなかった学生もその問題では高得点をとる可能性があるということですから、一体幾つ問題を用意したらいいのか、何回やったらいいのかについては、ちょっと答えがないような気がします。

○司会　採点する側の負担感も大きいと思いますし、採点の日にちも限られますから、大学入試の中で記述式などの試験をやるのは大変かなという印象もあるんですけれども。

○村上氏　語学試験でしたし、私どもは比較的小規模な学部でしたので、その点はそれほど問題にならなかったと記憶しています。ただ、もちろん入学定員のすべてをこういう方式でとなれば、現在の陣容では不可能だとは思います。

○司会　それでは、また後ほどパネルディスカッションのときに。

村上先生、ありがとうございました。（拍手）

高大接続のこれから

村上　隆
中京大学・現代社会学部

私が経験から語れること

- 高大接続の改善の提案には，当然，入試改革が含まれる。現時点では，特に，**教育への社会的要請の変化**への対応と，**入試にも利用可能と考えられる技術革新**への期待が，改革の提案につながっているように思われる。
- 私は，自分のかかわった公的試験の立ち上げにあたってのいくつかのエピソードをお話しし，今後の入試改革にかかわる議論の1つの焦点となると思われるCBT（computer based testing）が，日本的テスト文化との間に起こす可能性のある「摩擦」を中心に論じたい。
- そこから，能力評価の新たな理念が（ある程度）成熟し，それに対応する技術が（一定の水準で）整ったとしてもなお，それが期待されている「真の学力」評価につながるかどうかは，**日本の風土だけでなく，わが国の行財政のあり方に依存**することを示唆するつもりである。

私が経験から語れること

- 私が立ち上げを経験した試験とは，2002年から実施されている「日本留学試験」である。
- 「日本留学試験」は，項目反応理論（item response theory; IRT）に依拠した得点等化（test score equating）が行われた，おそらくわが国で最初の，いわゆるhigh stakesな公的試験であると思われる。
- CBTからは程遠い紙ベースの試験であるが，もし，入試センター試験（あるいはその後継試験）がCBT化されるとすれば，避けて通れない問題がいくつか提起されたと思われる。多少ともご参考になれば幸いである。

日本留学試験とは？

- 日本留学試験（にほんりゅうがくしけん、Examination for Japanese University Admission for International Students、略称EJU、日留試）は、独立行政法人日本学生支援機構が主催する、日本の大学（学部）や専修学校に入学を希望する外国人留学生を対象とした共通の入学試験である。それらの教育機関で必要とされる日本語能力および基礎学力の評価を目的とする。2002年から実施されている。
- 国内あるいは主要アジア諸国とウラジオストクで受験することができる。ただし、中国本土（香港は除く）では実施していない。
- 現在多くの大学・短期大学や、一部の大学院や専修学校などが外国人留学生の入試選抜に利用している。これら利用校を志望する受験生は大学入試センター試験と同様、利用校の入学選抜試験出願時に日本留学試験の成績を提示する。試験は6月（第1回）と11月（第2回）の年2回実施され、何回受験してもかまわない。

（日本留学試験 Wikipedia）

日本留学試験とは？

- 試験科目は志望校の指定に従って選択する。文系学科は日本語、総合科目（社会）、数学コース1が、理系学科は日本語、理科2科目、数学コース2が一般的である。もちろん一部科目だけでよい学科もあるが、日本語はほとんどの学科で必須である。総合科目（社会）、理科、数学の問題の難易度はセンター試験よりも若干低く設定されている。大学によっては、日本留学試験の成績の他にTOEFLなどの成績を必要とする学科もある。
- 成績そのものの有効期間は2年である。日本留学試験は複数回分の試験成績を同時に比較して選抜に利用するため、問題の難易によって有利不利が生じる。このため、項目応答理論という統計学的方法によって得点を等化している。これによって得られた「尺度点」を成績として利用している。

（日本留学試験 Wikipedia）

2000～2002年度特別研究推進費（いわゆる政策科研）

- 「我が国の公的試験における得点等化の導入に向けた心理・教育測定的研究」
- 研究代表者は村上隆，研究協力者は，その時点で，計量心理学の最先端の研究者，入試センターにおける研究開発の担当者，さらに，法学研究者を含めた9名であった。
- 扱った問題は，IRTによる得点等化，問題作成体制の整備，データベースの構築，試験全体へのバックアップ体制のあり方，法的問題への対応，といったものである。

得点等化の実現について

- ここでは，IRT あるいは，その他の手法による得点等化の方法について，改めて解説する必要はないと考える。
- むしろ，得点等化を実施するにあたって，何が問題だったのか，これを実施するために，テスト関係者に何を理解してもらう必要があったのか，他方，IRT 以外にどんな準備が必要だったか，また，作題に当たる人たちから（意識的，無意識的に）どのような抵抗を受けたか，といったことについてお話したい。
- ただ，現時点では，記憶が薄れている部分も多い。もし誤りに気づかれたら，研究協力者をお願いした先生方でここにご列席の方もいらっしゃるので，ご叱正をいただきたい。

CBT 導入の効果，新たな問題
（あくまでも，村上の推測である）

- センター試験を今のまま実施する以上の負担には，大学側は耐えられない（であろう）。
- よって，新たな「到達度テスト」を，現行の体制で実施するのは，（センター試験が存続する限り）無理。
- 他方で，センター試験を「到達度テスト」に置き換えるなら，（今の構想だと）複数回実施は回避できない。
- **CBT化で，大学の負担は回避できる**（施設，セキュリティ等の問題は新たに発生すると思うが）。
- しかし，複数の試験間（あるいは受験機会間）の得点等化と，大量の問題項目作成の問題がおこる。
- であるとすれば，CBT化に際して，問題作成の（ある程度の）自動化も期待されている？

まず，何を理解してもらうべきか

- 先の政策科研の申請の際の，文部省（当時）の意見

（1）欧米で開発された手法をそのまま使えばいいのでは？

→ この点が今日の話題の1つ

（2）我が国にも得点等化を可能にしている試験はすでに存在するのではないか？ → あったが試作程度

（3）学術研究として行われた例はあるのでは？

→ レベルの高いものが存在したが，high stakes な試験ではない。

（4）科研で開発したデータベースを（日本留学試験という）事業に転用することの問題

- このあたりが常識的反応であった（今もある？）と思われる。

IRT さえあればいいのか？

- 項目反応理論（IRT）さえ導入すれば，得点等化のみならず，試験の大部分の問題は解決できると思っていた（計量心理学の外部の）人々が少なくなかった（半可通）。
- IRT があれば，潜在能力が測れるという大誤解 ← 理論が潜在特性（latent trait）という「直接観測不可能な能力尺度」の存在を前提に組み立てられているところからそう思われた？
- ともかく，コンピュータで何か計算をすれば，複数回の試験の間の得点を比較可能にできると思っている人は，相当多かったと思う。
- これは，研究者の「誇大宣伝」にも原因があると考える。

数学者には認められない（？）IRT

- 意外にも，数学の問題作成者からは，IRT はまったく理解してもらえなかった。
- 数学の現実への適用については，きわめて慎重な人が多いのは事実。
- 他方で，能力評価については，楽天的。あるいは，そこをあまり問題にしたくない？
- いわゆる小問主義への強い抵抗感。
- 基礎教育としての数学と応用分野との対応は，大学教育でも大問題であり続けているが・・・。

データが満たさなければならない条件があることについて

- 問題項目か受験者か，どちらかに共通のものが含まれていなければ，等化は不可能である（間接的なつながりでもよいが）。
- どちらもできないなら，予備テスト，それもできないなら（一応，試験と無関係な人を対象とした）モニター試験。
- IRT の理論構成上も，（ある程度の）小問主義は不可避。
- ともかく，絶えざる問題供給の体制の構築とデータの蓄積，さらに，問題の公開可能性の範囲，予備試験，モニター試験等における問題のセキュリティといった新たな問題が発生する。
- 過去問への（異常な？）執着という「東アジア」的試験風土も障害の1つ

研究しなければならなかったこと

- 日本の現実（たとえば，日本語学校の教員は，試験問題らしきものがあったら，ただちにコピーするのが習慣）の中で，どう IRT 使用のための前提条件を満たすデータを収集するか？
- データベースの整備。作題の体制の確立。
- 海外で訴訟が起こされたときの対応（どうやら，杞憂だったらしいが）。
- ともかく，どのくらいヒトとカネを投入すればいいのか？ （このあたりの知識と経験は，当時の村上には決定的に足りなかった。）
- もちろん，IRTのプログラムを実際に運用して得点等化を行うにあたっては，前川先生をはじめとする現場の方々の大変なご苦労があった。

作題体制の確立：
アイテムライターの募集と養成

- もはや，大学教員が片手間で問題を作っていては追いつかない。大量の問題案を供給するためには，外部にアイテムライターの集団を抱えることは必須であった。
- 日本語教育の場合，比較的海外の事情に通じた教員が多かったことから，小問主義も得点等化もすんなり受け入れられた（もし，これが国語教育だったらどうなるだろうか）。
- それでも，**authentic な問題**（後述）への要求は強く，測定サイドは大きな妥協を強いられたが，現状は，だいたいいいバランスになっているように思われる。
- アイテムライターのトレーニングコースには，少しだけお付き合いしたが，大学院生レベルの応募者の問題作成スキルの習得は速かったように思う。

対処していなかったこと：
事務サイド等のご苦労

- 以下の点は，事前にはあまり考えられていなかったことである。
- 誰が試験監督をするのか。 → もう教員は動員できず大学職員，やはり不慣れが目立った。
- 受験者が外国人であること。日本の試験の常識は通用しない。試験場に，あの入学試験の張詰めた雰囲気はない。結果，大量の不正行為の摘発。
- 海外実施体制。試験問題を実施前に開封して（気に入った）日本語学校に配ってしまうという事例が，先行する日本語能力試験ではあった。
- グローバリゼーションとはこういうことでもある！
- しかし，日本人対象でも，受験者による問題の撮影や録音は，行われることを覚悟すべきであろう。

作題について：前提となることがら

- CBT 化された試験は，（比較的多数の）多肢選択形式の問題項目からなると思われる。
- これ以上の，高等教育の基礎となる学力を反映した performance を，コンピュータベースで評価することは，10年後においても困難であると予想する。
- もちろん，入力はコンピュータ，採点は人間という形で実施される記述形式の試験はありうる。さらに，人工知能によるプログラムが採点にあたるといった技術革新も考えられなくはないが，複数の採点者のうちの少なくとも一人は人間でなければならないであろう。
- 多数の多肢選択の問題項目をもちいることにより，試験の信頼性と妥当性も（記述式の大問形式の問題より）高められる。

蛇足ながら

- 小問形式の問題項目は，個々には大きなランダム誤差を反映するとしても，複数個の反応を何らかの方法で合成することにより，全分散中の誤差分散の割合を減じることができる。すなわち，項目数を増やせば，試験の点数の信頼性が高められる。 ー まぐれ当たりは続かない。
- 小問形式の問題項目は，測定しようとする能力の範囲（レンジ）が広い場合にも，それを（比較的）偏りなくカバーするように作題することができる。これは領域代表性と呼ばれる測定の（内容的）妥当性（の一つの側面）を高めることにつながる。 ー ヤマがかけにくい。

小問主義と相性がいいのは discrete points

- Discrete points とは，いわばバラバラな知識（技能）。言語能力の試験では，文字・語彙・文法等について，「項目」が列挙される。
- かつての日本語能力試験には，級（1～4）ごとに，漢字，語彙，文法事項の一覧表（シラバスと称していた）が存在していた。
- こういうものが用意できるなら，そこから問題項目はほとんど自動的に作ることができるという希望が出てくる。
- しかし，そういう問題だけのテストには，authentic でない等，多くの批判が集中することになる。
- その間の，妥協点を求めることが必要になる。しかし，少なくとも省力化からは遠ざかる。→ 村上（2001）

問題IV　次の文の＿＿にはどんな言葉を入れたらよいか。1・2・3・4から最も適当なものを一つ選びなさい。

(1) 試験の＿＿静かにしてください。
1 うち　*2 あいだは　3 まに　4 うちに

(2) アパートは部屋の中を見た＿＿借りるかどうかを決めたい。
1 うえに　2 うえから　3 うえにも　*4 うえで

(3) お借りした本を楽しく＿＿いただきました。
1 読まれて　2 お読みに　*3 読ませて　4 読まされて

(4) 天候に＿＿あすの午後2時から試合を行います。
*1 かかわらず　2 よると　3 よれば　4 かかわり

(5) 電話帳で調べた＿＿そういう名前の会社はなかった。
*1 ところ　2 だけに　3 からに　4 ばかりでは

図1　小項目主義の客観テストの例

これでは「真の言語能力」は測れない？

- テスト理論との相性がいいのは，短い多肢選択項目であるから，discrete points を問うタイプの項目が好まれる。しかし，それには作題サイド，さらには試験のユーザーサイドに不満が生じることになる。
- 言語教育で言えば，作題サイドは，言語知識ではなく言語運用能力を問いたい。
- 同様に，入学試験では，（離散的な）知識だけではなく，思考力・判断力・表現力（！）等も問うことが求められる。
- 信頼性を犠牲にしても，（より本質的な）妥当性を高めるという目標を作題サイドはもちろん，ユーザーサイドも期待する。

言語の知識ではなく運用能力とは？

（1） 言語能力は，単なる知識の集積ではなく，実際場面においてコミュニケーション活動を行いうる能力として捉えられる。この結果，言語テストによる測定の対象は，言語の知識ではなくコミュニケーション能力となる。

（2） これに対応して問題項目は，いかにもテストのために設定された人工的な場面ではなく，自然で現実的なもの(これが authentic ということである)が求められる。

（3） コミュニケーション活動そのものに，認知心理学，社会言語学的な分析が加えられ，非常に多面的に区別されるようになる.

- その結果，個々の問題項目は実際のコミュニケーション場面を設定した上で，より現実的な言語活動の遂行を想定した反応が用意される。

Authentic な問題(例）

【聴　解】

テープ II　学生が，指導教授のところに，サークルの顧問になってほしいとお願いに来ています。
サークルの顧問は，どうなりますか。

（ノックの音・トントン）
教授：はい，どうぞ。
学生：失礼します。
教授：何か用？
学生：あのー，山田先生，ちょっとお願いが……。
教授：何？
学生：あのー，先生にサークルの顧問をお願いしたいんですが……。
教授：サークル？
学生：あ，はい，「国際交遊会」っていう，留学生の会なんです。
教授：ほう，面白そうな会のようだね。
うーん，でも，顧問はちょっとね……。
学生：だめですか。
教授：うん，今年は，外国出張が多くて，留守が多いからね。
学生：はあ。

教授：田中先生は？もう，聞いてみた？
学生：はい，だめだそうです。ですから，何とか先生にお願いできないかと……。
教授：うーん，ちょっと，余裕ないなー。他をあたってみてよ。
学生：そうですか（がっかりした様子）。
じゃ，また，だめだったら，ご相談にのっていただけますか。
教授：うん，相談ならいいんだけど，顧問はね，やっぱり，ちょっと……。
学生：わかりました。どうも，ありがとうございました。

サークルの顧問は，どうなりますか。

1 山田先生に顧問をしてもらう
2 田中先生に顧問をしてもらう
3 別の先生に顧問を頼む
4 山田先生に顧問を頼む

図2　言語運用能力を測定しようとするテスト項目の例

客観形式のテストへの脅威

- お気づきのように，authentic な問題項目への傾斜は作題そのものを著しく困難にする，多肢選択形式のテストは，単純な知識の有無を問う場合ですら,選択肢の設定次第で難易度が著しく変わりうる。
- 誤答の選択肢(迷わし，distracter と呼ばれる)の作り方によっては,受験者の判断を不安定なものにしたり(信頼性の低下)，できる受験者をかえって誤答に導いたり(妥当性の低下)することがある。
- 作題者が思いもかけなかったところに，正答を選択する手がかりができてしまうことがある等，信頼性，妥当性の低下を招く可能性が一層高まる。
- それでも，こうした問題に移行せざるを得ない理由は確かに存在する。 → 遡及効果

遡及効果： テストで測れていない**真の学力**が存在するという考えが生まれる根拠

- テストの形式や内容が教育内容を逆に，規定してしまうことを遡及効果（wash-back effect）と呼ぶ。
- 事実，テストの形式と内容が，そのテストに対する準備教育を担う教育機関の教育の方法と内容に大きな影響を及ぼすことは疑う余地がない。
- さらに，授業の目標がテスト対策に焦点化され，そのための指導技術，学習方法が「進歩」すると，習得されるものが「特定のテストで高得点をとる能力」に偏っていくのも事実であろう（実証されているか？）。
- これは，評価方法の劣化とみなされることになり，何らかの改革へのモチベーションになる。
- 言語能力の場合，これは知識から運用へという流れであった。 → 大学入学者選抜では，「思考力」，「判断力」，「表現力」。

それでもなお，客観形式のテストへの脅威

- この種の問題は，当然，1問あたりの解答所要時間が長くなる。このことは，必然的に1つのテストあたりの項目数を減少させ，小問主義の客観テストに比して，信頼性の低下をきたすことになる（前述のように，妥当性への影響は複雑で，一概には言えない）。
- また，こうした項目の所要時間は不揃いとなり，項目を入れ替えて複数の版を用意して行うような得点等化の作業に不便をきたす。
- すなわち，運用能力に的を絞ったテストは，形の上では小問主義の客観テストではあるが，テスト理論的観点ら見たその長所の多くが失われることになる。
- **さらに，記述式への移行が考えられると，妥当性の観点から，さらに大きな問題が発生する。**

記述式テストの問題点

- 従来行われてきた小論文に近いものを想定する。
- しばしば問題になるのが，採点に主観が入ること，具体的には，複数の採点者間で評価が一致しないことであろう。しかし，この問題は，採点者間の十分な打ち合わせと，不一致の際の審議，さらには採点者数を増やすことによって解決でき，相当の信頼性は得られる。
- 問題はむしろ，複数の大問の間の相関が著しく低いことである。この点は，小論文試験の採用の初期から（複数回受験者の結果等を通じて）指摘されてきたことであるが，解決を見ていない。このことは，常に3つの大問で構成されていた名古屋大学教育学部における「論述的学力検査」において繰り返し確認された。

記述式テストの問題点

- この問題は，もし仮に高度な人工知能による機械採点が（完全に）可能になったとしても，未解決の問題として残る可能性が高い。
- つまり，たまたま取り上げられたテーマへの相性によって，同一受験者の点数に大きな変動が起こりうる。すなわち，トータルにみれば信頼性の低い評価方法であるのは（少なくとも現時点では）確実である。
- もちろん，こうした問題間の相関の低さは，かなり選抜性の高い大学での入試だったからだとは考えられる。また，こうした問題のもつ（ポジティブな）遡及効果も否定はできない。
- 名古屋大学の同僚から，「最近，読むに耐えないレポートを書く学生がいなくなった」という感想を聞いたことは忘れられない。

これからの入学者選抜の中で

- 今，中教審等で提案されている方向性には，一定のメリットがあることも事実であるが，「理想」の実現には，越えるべき多くのハードルがある。
- 議論を進めるためには，求められている「真の学力像」が，（想定されている社会像とともに，）もっと具体化される必要がある。そこではじめて，現実の評価方法，特に，CBT で可能なことと，さらにその実施にあたっての問題点がより明確に見えてくるのではないか。
- その上で，この先に待ち受ける複数のハードルを乗り越えるためには，十分な財政的，人的態勢がなければならない。それがどこまで作れるかが，究極の問題であると考える。

ご清聴，ありがとうございました

- 参考文献

小嶋秀夫・村上 隆（1991）名古屋大学教育学部における論述的学力検査 1990年度科研費総合研究（A）「大学入試における実技・面接・小論文等の評価に関する研究」報告書, 31-60.（研究代表者：岩坪秀一）

村上 隆（2001）第2言語としての日本語能力テストの開発 ―一般的な問題と固有の困難― 計測と制御, **40**, 576-580.

村上 隆（研究代表者）（2003） 我が国の公的試験における得点等化の導入に向けた心理・教育測定的研究 平成12年度～平成14年度 科学研究費補助金（特別研究促進費(1)）研究成果報告書

CBT のサンプル・デモンストレーション

大学入試センター研究開発部助教
大久保智哉

○司会　今日はコンピュータ・ベースド・テスティング ― CBT が話題に上っておりますけれども、大学入試センターでも CBT の開発研究をしておりますので、大久保智哉先生にちょっとデモンストレーションしてもらいたいと思います。

○大久保氏　大学入試センターの大久保が発表させていただきます。

本研究、特にシステム開発につきましては大学入試センターの私と、宮埜に加えて東京工業大学の室田真男教授、前川眞一教授に共同研究していただいております。

これは実際に、約 1 年前に大学入試センターのモニター調査の中で実験として、CBT を使って英語の 4 技能試験の一部を実施したときの写真です。CBT による問題とはどのようなものか、実際にここでお見せしたいと思います。

英語：動画利用リスニング問題

もちろんコンピュータですので、こういった動画を見せることもできます。通常のリスニングと違って、多少語彙のレベルが上がっても映像がありますので、受験者は情報を補完しながら、より現実的な英語の状況に近いようなことを、テストとして行うことができます。

今回の設定は、2 回動画を流します。そして 2 回の間に 1 回 30 秒のインターバルを設けて、「次の動画の後、こういう質問をしますよ」ということを表示しておいて、2 回目の動画をこのように流すようになっています。当然こういった設定は幾らでも変えることができます。そして、「中身はどれですか、選んで進みなさい」と出題します。

英語：発音録音問題

これは動画を使った例ですけれども、実際に録音させるような問題もできます。現在の入試センターの問題ですと、発音記号を示して，どれが等しいですか？とか、アクセントはどこにありますか？という、本当に「できるか」ではなくて、知識をシンプルリコール(単純想起)できるかということに焦点を当てているのですが、(CBT による問題では) どう採点するかはまた別の問題になるものの、技術としては、このタブレットに受験者が読み上げた部分を記録する。そして，その音声を後で聞いて、きちっと伝わっているか、伝わっていないかといったことで点数化するようなことができると思います。

英語：長文資料読解問題

他にも、資料の読解問題をカラーで見せたり、かなり長い文章を出して、全部が全部回答に必要な情報ではなくて、全く関係ない文章も混ぜて、そこから斜めに読み出せるか？

必要な情報だけをピックアップできるか？まさに我々が普段情報を得る過程でやっているような作業を、紙に印刷すると、どうしても枚数が増えてしまいますけれども、コンピュータではそういったことも簡単にできるようになります。

英語：動画リスニングとライティングの複合問題

あと、ライティング問題もテキストとして記録できます。この問題では動画を見てもらって、その内容についてまとめなさいという問題です。こういった動画とライティングのコンビネーションみたいなことも可能です。これをどう採点するかはまた別の問題になりますが、技術的にはこういうことが可能になります。この動画自体は AFP 通信から許可をいただいてテストに二次利用することができるようになっていますが、動画というのは本当にいろいろな素材がありますし、動画を見せることによって、単語レベルが多少低くても実際英語をしゃべるときには、雰囲気でというか、「こういうことを言っているのではないかな」とうまく推論して、言語ですから、推論しながら理解していくといったことも、可能になるのではないかと考えております。

大規模試験でのCBT運用方法案の提示

では、実際こういった問題をどのようにやるのかが問題になると思いますが、今、示されている中教審の答申案ですと、高校 2 年生、高校 3 年生が場合によっては全員受ける。そうすると 200 万人近くになるわけです。大量に同時的に受ける可能性もあります。そうした場合に重要になってくるのは、やはりコストとセキュリティの問題だと思うのですが、今、一つの案として考えているのは、ここに示す図のようなものです。

実際にテストの場所では、このタブレット本体にデータを入れて、セキュアな形で、アプリケーション形式にしてしまって、それを実際にテストの実施場所、高校でもいいし大学の講義室でもいいし、いずれにしてもネットワークもシステムのインフラ設備もない所に持っていって、そこでテストを受けてもらいます。受けてもらってそのデータを回収し、それを日本全国何箇所か配置したデータの中継地点、ローカルオペレーションセンターと私は呼んでいますけれども、そこへタブレットを持って帰って、そこでデータを吸収して、そこから中核センターまではセキュアな回線で回答データを送受信することが考えられるのではないかと思います。こうすることで建物を借りたり準備したり、また、管理する人間の教育だとかそういった管理・運営コストも大分減らすことができますし、何よりも CBT という形で実際に運用できるのではないかと、一つの案として今は考えております。

タブレットの一括管理技術の提示

では、何十万台、何万台というタブレットを実際どのように管理するかという問題ですが、そのためにタブレットを一括管理する技術とか仕組みも研究開発しております。今ここでお見せしているのは 100 台のタブレットにサーバーから命令を出して、1 番から 10 番までには数学の問題、11 番から 50 番ぐらいまでは英語の難しい問題を入れるというようにそれぞれタブレットに指定して、サーバーから一括で命令を出して、同時的にそれぞれのタブレットに異なる問題を入れているところです。

IRT によって尺度化された試験であれば、問題が変わってもそれを解いた正誤データとその問題の難易度をもとに共通尺度上で表現することができるようになりますので、IRT を使うとそういったことも可能になります。

1 つここで申し上げたいのは、この IRT、先ほどから出ている統計理論のことですが―

IRT とコンピュータの相性が良いです。コンピュータを使う以上、必ず壊れます。私、100 台を 1 年間管理してみて、1 台が原因不明の不具合が出ました。これはデバイスの問題だと思います。それが試験中でなかったからよかったのですが、試験中に起こることだってありうるわけです。そうなれば必ずフェールセーフが必要で、そうなったときにも別のテストフォーム、別のテストでも比較可能なスコアが出るような方法が存在するのと存在しないのとでは全然違います。年に一度きりの一斉試験でコンピュータを使うのであれば、トラブル 0 での実施は、ほとんど無理だと私は思っています。

解答データ送受信画面の提示

さらに問題の回収もこういった画面で効率的にできますし、画面にお示ししたのは、回答データの管理もこのような形で一括管理できます。だれがどんな問題に対してどういうふうに答えたか、そのときのテストの ID はどんな ID で、何番目のブロックで問題は何番目だったか、そのような記録がすべてデータベース化されます。

試験問題の管理・編集についてもこのように、問題が何回曝露されて誰が受けて、どういうモデルを使ったときのパラメータがどうであるかといったことがデータベースとしてすべて管理されます。

先ほどは、動画を見せたり記録をするという英語 4 技能の一つの例として示したのですが、後半お見せするのは 2 つです。コンピュータを使うので、今まで紙と鉛筆ではできなかったことができるのではないかという可能性を示唆する意味で 2 つ例を挙げさせていただきます。

CBT による数学連問形式の例示

先ほどから「大問形式」という言葉が出ていますけれども、大問形式、一番イメージしやすいのは数学の問題だと思います。大問形式とは 1 つのトピックを中心に、さまざまな側面から複数の問いが出題されるような形式です。1 つ大きな問題は、前の項目の正誤によって次の項目に影響を及ぼしてしまうことです。たとえ能力のある人でも「1 問目の答えに幾つを足したら 6 になりますか」これだと 1 問目を間違えたら 2 問目も間違えてしまいます。そういった実験的独立性が大問形式では欠如します。次元的独立性というのもありますけれども、ここでは実験的独立性を扱います。

大問形式にはいろいろメリットがあるという研究者もいるのですが、IRT に馴染まないということが先ほどまで議論されてきました。ただし、一部に関してはコンピュータで大問形式を、このようにステップに区切ることによって前戻りできないような形にすれば、そして次のステップに進んだときに前の問題の答えをしっかり教えてあげて、修正可能な状況まで戻してあげれば、大問形式に似た状況で、局所独立性を回復できます。これは紙のテストではできません。コンピュータで問題提示をコントロールするからできることです。

このように、コンピュータを使うと問題をクリアできる可能性は増えていくと思います。ただし、これに関しても、先ほど議論があったようにプリテストをするなど実験をしてしっかりデータをとっていかなければいけないと思います。

CBT による自記式解答とコンピュータ採点技術の提示

最後にお示しするのは地理の関連問題です。これは何かというと、かなり長いレポートです。実際にお見せするとこんな感じで、A 国がどんな国からどんなエネルギーを輸入し

ているか延々に説明が書いてあります。国名は伏せられているのですけれども、A はポーランド、B はロシアで、全体の文章を読むとポーランドがすごくロシアからエネルギーを輸入しているよということが前半に書いてあって、一方で、ポーランドがエネルギーの開発として原子力を導入したり、ロシアから輸入していたエネルギーをウクライナから持ってきたり、カタールからパイプガスを引っ張ってきたりしていますよと書いてあります。「なぜこんなことをしているのか書きなさい」という問題ですけれども、これも自記式、すなわち記述問題になっています。

現在のところ、これは自動的に採点できるものではありませんが、国名、単語ぐらいは形態素解析をしてマッチングをかけて正解、不正解ぐらいのことは、現段階で、すでにできています。

例えばこれ、「ロシアへのエネルギー依存度を引き下げることがポーランドの政策上、重要であったため」というのが答えですけれども、この問題を多肢選択式にしてしまうと非常に簡単になります。読んでいれば「あ、そうか。ポーランドがロシアからのエネルギー需給を引き下げたいのだ」とすぐわかってしまうのですけれども、書かせることによってそういったことを減らす可能性はあります。

ただ、これをやっても長い文章を自動的に採点することは、先ほど村上先生のご報告にもありましたけれども、私は個人的に、10 年たっても無理だと思います。なぜかというと、それができるのであれば英語の自動翻訳が当然できるようになるわけです。今、ここで想定しているのは極めて短い文章もしくは単語を拾えているか、もしくは部分的にハイライトできるか、そういったものを想定して研究開発を進めています。

ただし、研究開発をした後、実際データをとってそれを確認して信頼性、妥当性、そこまでしっかり確認して、それが使えるものなのか、使っていいものなのか、コストはどうなのかという検討はしなければいけないはずです。とりあえず今回のこの説明では、こういった技術を一部紹介させていただきました。

これは、どういった能力を測っているかというスライドです。

ご清聴どうもありがとうございました。（拍手）

○司会　大久保さん、どうもありがとうございました。

簡単なご質問を受けたいと思いますけれども、いかがでしょうか。よろしいですか。

これはまだ研究開発の段階ですから、これが新しいテストの問題にそのままいくことはないとお考えいただければと思います。

それでは、パネルディスカッションの準備がありますので休憩します。

〈午後 4 時 44 分　休憩〉

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』

2014年11月29日(土) 於 東京工業大学 デジタル多目的ホール

CBTのサンプル
デモンストレーション

発表者:
大学入試センター 研究開発部 大久保智哉

システム研究開発:
大久保智哉[1]・室田真男[2]・前川眞一[2]・宮埜壽夫[1]
[1]大学入試センター・[2]東京工業大学

発表の概要

1. 情報端末を利用した英語4技能試験の問題例
2. 大規模共通試験でどのように情報端末を利用して試験を運用するのか
3. 大規模なCBTを支えるいくつかの技術の紹介
4. 数学などでの情報端末を用いた工夫例
5. 出題形式の発展可能性

タブレット端末を用いたCBT ・実験風景

英語4技能試験の問題例:動画聴解問題

動画終了まで
0:21

©AFP通信

英語4技能試験の問題例:発音録音問題

準備時間終了まで
0:13

録音が始まったら，赤色の下線が引いてある英文のみを適切に音読しなさい。

1st February 2012

Dear Mr. John Smith,

It is a great pleasure for me to recommend Mr. Taro Yamada to your renowned internship program. I teach Mr. Taro Yamada who is currently pursuing Ph.D. at

Skip

Pause

英語4技能試験の問題例:英文資料読解問題

東京工業大学

Home > Prospective Students > International Students > International Research Students > Application Procedures for MEXT Scholarship Research Student

Application Procedures for MEXT Scholarship Research Student

Application Procedures for Tokyo Tech Recommended Monbukagakusho (MEXT) Scholarship Student (Research Student Program) commencing October 2014

Recipient of the Monbukagakusho (Ministry of Education, Culture, Sports, Science, and Technology - MEXT) Scholarship will enroll in Tokyo Institute of Technology (Tokyo Tech) **as a Research Student in this program starting from October 2014.**

Scholarship recipient who has demonstrated excellent academic achievement and who has passed the entrance examination during his/her Research Student status period may be eligible to enroll in Tokyo Tech's Master's or Doctoral Programs and extend the duration of his/her MEXT Scholarship under certain conditions, as noted herein:

- Recipient may advance to Tokyo Tech's Master's program, provided that he/she successfully passes the entrance examination for April 2016 or earlier. In this case the student is eligible to apply for a MEXT Scholarship extension.
- Recipient may advance to Tokyo Tech's Doctoral program, provided that he/she successfully passes the entrance examination for October 2015 or earlier. In this case the student is eligible to apply for a MEXT Scholarship extension.

In principle, an applicant who has obtained either an official recommendation from a university with which Tokyo

解答終了時間まで
3:54

Submit

課題文

資料

Skip

Pause

©東京工業大学

情報端末を利用した英語4技能試験の問題例

英語4技能試験の問題例：英文資料読解問題

解答終了時間まで
3：51

Submit

課題文

資料

次の選択肢のうち，資料にある奨学金制度に応募資格がある学生は誰か？もっとも適切な選択肢を一つ選びなさい。ただし，選択肢に書かれていない条件については，すべて応募条件を満たしているものとしなさい。

1. 1975年4月生まれ．大学では日本文学を専攻し，大学を卒業した．大学院でも日本文学を専攻する学生．
2. 1981年9月生まれ．大学では物理学を専攻し，大学院でも同様に物理学を学ぼうとする学生．
3. 1980年10月生まれ．大学では日本文学を専攻し，大学を卒業した．大学院ではこれまでに学んでいない新しい分野にチャレンジしようとする学生．
4. 1985年4月生まれ．大学生では日本史を専攻し，卒業した．大学時代の学業が評価され，すでに他の奨学金の受給を受けている学生．

Play

情報端末を利用した英語4技能試験の問題例

英語4技能試験の問題例：記述問題

解答終了時間まで
3：01

Submit

動画の内容を英語で30単語以内にまとめなさい．なお，会話中に出てくる「Twitter」とは，短文(Tweet)を投稿できるミニブログのようなサービスである．投稿された短文は他者によって見る(followする)ことができる．

©AFP通信

00:00 / 00:50

現在の語数
3

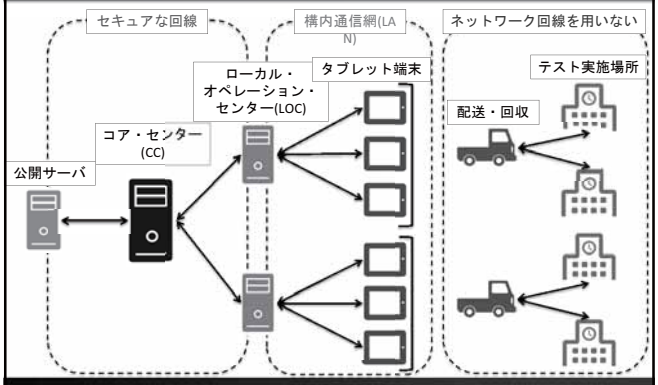
Kouichi Wakata is

Skip

Pause

大規模共通試験でどのように情報端末を利用して試験を運用するのか

CBTの運用デザイン(案)



大規模なCBTを支えるいくつかの技術の紹介

試験問題の情報端末へのセッティング

大規模なCBTを支えるいくつかの技術の紹介

回答データの回収

現在 1 件の結果があります

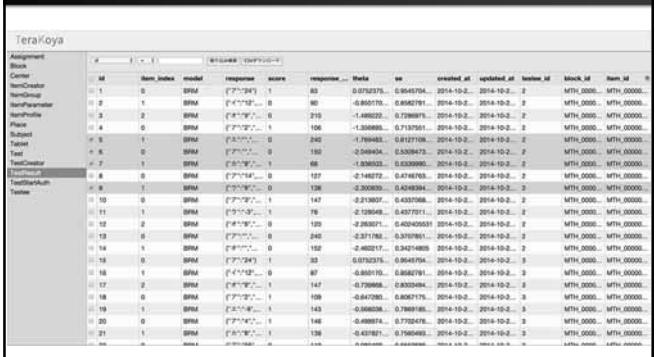
結果を送信

結果を消去

戻る

大規模なCBTを支えるいくつかの技術の紹介

解答データの管理



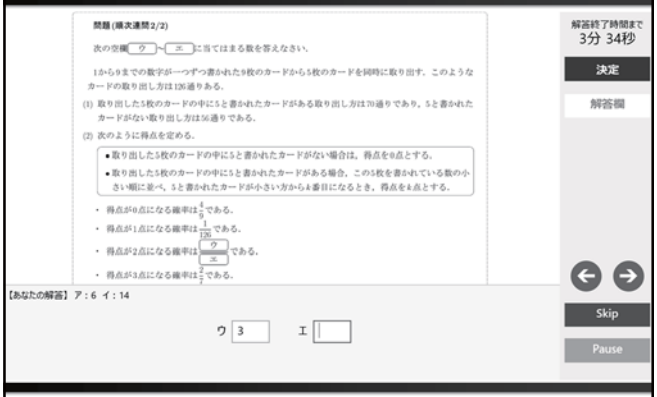
試験問題の管理・編集

大問形式と連問形式

- ✓ 大問形式(数学IA・IIBなど)
  1. 1つのトピックを中心に様々な側面から複数の問いが出される形式（石塚, 2003）
  2. 前の項目の正誤によって、現在の解答の正誤に影響を及ぼす場合がある
  3. 現在の項目を正答できる能力があるにも関わらず，前の項目の正誤に依存して誤答
- ✓ 大問内の項目において局所独立性がない場合
  1. 大問内の項目を独立な項目として扱うと，能力推定値の信頼性が過大評価される (Sireci, Thissen & Wainer, 1991)
  2. 項目における次元的独立性の欠如
  3. その前の正誤状況に依存する実験的独立性の欠如
- ✓ 連問形式
  1. 大問をいくつかのステップに区切り，順次的に解答させる形式
  2. 受験者が解答を入力し次のステップに画面遷移後，直前のステップの正答を提示した上で後続の空欄について解答を求める形式
  3. この形式は情報端末上でのみ利用可能

大問形式と連問形式

大問形式と連問形式



連問形式 ステップ1

問題(順次連問1/3)

次の空欄 ア ～ イ に当てはまる数を答えなさい．

aを定数とし，xの2次関数

$$y = x^2 + 2ax + 3a^2 - 6a - 36$$

のグラフをGとする．Gの頂点の座標は，$(-a, 2a^2 - 6a - 36)$である．Gとy軸との交点のy座標をpとする．

$p = -27$のとき，aの値は$a =$ ア イ である．ただし、 ア ＞ イ とする．

連問形式 ステップ2

問題(順次連問2/3)

次の空欄 ウ ～ エ に当てはまる数を答えなさい．

aを定数とし，xの2次関数

$$y = x^2 + 2ax + 3a^2 - 6a - 36$$

のグラフをGとする．Gの頂点の座標は，$(-a, 2a^2 - 6a - 36)$である．Gとy軸との交点のy座標をpとする．

$p = -27$のとき，aの値は$a = 3, -1$

ステップ1の[ア][イ]正答が提示される

Gがx軸と共有点を持つようなaの値の範囲を表す不等式は

ウ $\leqq a \leqq$ エ (1)

である．

連問形式 ステップ3

問題(順次連問3/3)

次の空欄［ オ ］～［ カ ］に当てはまる数を答えなさい．

aを定数とし，xの2次関数

$$y = x^2 + 2ax + 3a^2 - 6a - 36$$

のグラフをGとする．Gの頂点の座標は，$(-a, 2a^2 - 6a - 36)$である．Gとy軸との交点のy座標をpとする．

$p = -27$のとき，aの値は$a = 3, -1$である．

Gがx軸と共有点を持つようなaの値の範囲を表す不等式は

ステップ2の［ ウ ］［ エ ］
正答が提示される

$$-3 \leqq a \leqq 6 \quad (1)$$

である．

aが(1)の範囲にあるとき，pは，$a = 1$で最小値-39をとり，$a =$［ オ ］で最大値［ カ ］をとる．

自記式解答の問題例（地理及び関連領域）

解答終了時間まで
3分 45秒

決定
課題文
資料
解答欄

問題

資料1は，ある国(A国)のエネルギー政策について2012年前後に書かれたレポートの一部を改変したものである．資料中では，3つの国名がA国，B国，C国として名前が伏されている．

A国は，このレポートが作成された時代に，なぜレポートにあるようなエネルギー政策をとっていたのか？その理由を説明するために【ア】には国名を，【イ】には20字以内で文章を入れ，文章を完成させなさい．ただし，【イ】には国名以外はすべて文章にある語を用いて解答すること．なお，国名は略称でも構わない．

【ア(A国の名称)】は【イ(20文字以内)】ことが政策上重要であったため．

ア
イ

Skip
Pause

自記式解答の問題例（地理及び関連領域）

解答終了時間まで
3分 16秒

決定
課題文
資料
解答欄

1.2 エネルギーの国外依存の状況

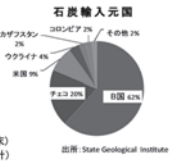
石炭は，2008年より輸入が輸出を上回る純輸入国となったものの，国内に豊富な埋蔵量を有しており，国内消費量に対する純輸入(輸入－輸出)の割合は4.3%に過ぎず自給率は高い．

他方，原油は殆ど(97%)を輸入に依存しており，天然ガスについても76%を輸入するなど，輸入依存度が高い．特に，原油，天然ガスともB国への依存度が高い(輸入に占めるB国の割合は原油94%，天然ガス85%)．

(単位：千トン)

生産	76,728
輸入	13,608
在庫増	4,422
総供給	94,758
国内消費	84,791
輸出	9,966
損失・統計誤差	1
総需要	94,758

石炭確認可採埋蔵量 57 億トン(2011年末)
(出所：BP統計)



Skip
Pause

自記式解答の問題例（地理及び関連領域）

2 解説

正答するのに必要な能力・知識

1. 資料を読んでA国がポーランドでB国がロシアであるということを判断するための**地理の知識**(C国はドイツである)
2. ポーランドの主要エネルギー政策に記載されている国名・地理をおおよそ理解できる程度の**地理の知識**
3. 資料を読んでポーランドがロシアへのエネルギー依存度が高いということを理解できる**資料読解力**
4. そのうえで，
   (a) ポーランドがロシアへのエネルギー依存度が高いという内容
   (b) ポーランドがエネルギー供給先の多様化・エネルギー種類の多様化(シェールガス開発，原子力開発)を図っているという内容

   から，同国のエネルギー政策がロシアへのエネルギー依存度を引き下げるためにおこなわれているものであると推論できる能力[1](**推論能力**)
5. なお，資料の量が多いため，**地理および関連教科に対する総合的な理解**がないと資料読解に際して不利になるであろう．

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』

CBTのサンプル
デモンストレーション

ご静聴ありがとうございました

パネルディスカッション

Kentaro Yamamoto 氏 ・ 前川眞一 氏 ・ 南風原朝和 氏 ・
大津起夫 氏 ・ 村上　隆 氏
総合司会　大塚雄作　大学入試センター試験・研究副統括官
入学者選抜研究に関する調査室室長補佐

○司会　再開します。それでは、パネルディスカッションを始めます。

それぞれのご報告の後で質問も受けましたけれども、35 分間という時間の関係もありますので、まず最初にフロアのほうから質問とか、こういう点はどうなんだというコメントめいたものでも構いませんけれども、幾つか受け付けて、それに関連して報告者の先生方に補足的にお話ししてもらおうかと思います。

いかがでしょうか。

○質問者 A　大変勉強になりました。

南風原先生のお話の中で、中教審が報告された内容についても非常にわかりやすいお話をいただいたんですけれども、先生のお話を伺うと、中教審の答申の新しいテストで、広範囲の難易度の問題を用意して選抜性の高い大学でも使えることと、段階別で成績評定を行うことは、私の理解では非常に矛盾していて同時に達せられないように思えたんですけれども、これは一体どういうことなのか、中教審の考えはどういうことなのかがよく理解できなかったので、その辺をお話しいただければと思います。

○司会　わかりました。

他にいらっしゃいますか。

○里見氏　文部科学省の里見でございます。本日はありがとうございました。

PISA 型の問題が、ある教育政策の下で学んだ人がどのぐらい達成できたかを集団として把握するもので、個人の能力を測るものではないということが 1 つ大きな論点だと思います。例えば、山本先生が示されておられた問題であるとか、大久保先生に見せていただいた問題を見ますと、社会的な文脈の中で知識等を実践的に活用しながら回答することが求められる問題に思えますが、それで個人の能力を測ってもよいのではないかと思えるんですが、それはなぜ望ましいことではないのか、PISA 型の問題では個人の能力を測定することはできないのか、その点について教えていただけないでしょうか。

○司会　もう一つぐらい出れば。よろしいですか。

では、また後で皆さんのほうから受け付けますので、何かありましたら考えておいていただければと思います。

まず、広範囲の難易度を含むことと段階別評定というのは矛盾することなのかというご質問だったと思います。これは南風原先生にご回答いただければと思います。

○南風原氏　私も、測定論的には矛盾していると思います。1 つ目の点では難関大学でも通用するように難しい項目を含めるとしておきながら、その高いレベルのところは全部 5 段階の 5 としてしまうのであれば、せっかく作ったものを帳消しにしてしまうので、測定論的には矛盾していると思います。

そういう意味では現在、センター試験も恐らくかなり広範囲の測定ができているのではないかと思いますが、それでも足りない部分を難関大学等では個別試験で足して、例えば東京大学の場合だったらセンター試験を 110 点に換算して、残り 440 点は個別試験でやる、そういう案分が自分たちの大学にとってはよいだろうということで来ているわけですね。そういう形の補充の仕方をしてきているわけです。

なので、段階別表示というのは恐らくそういう精度の議論、情報量の議論だけでは出てこない、別の観点から出てきていることだと思います。それは議論のプロセスをもっと近くでご覧になっている文部科学省の方や大学入試センターの方にご説明いただければと思いますけれども、ここまで出ている中から推測するに、各大学では多様な方法で選抜をしてほしいというときに、例えばもう 5 段階の 5 ということで全員イコールであれば、もうそれは使いようがないわけですよね。多様な方法に頼らざるを得なくなるわけで、そういうふうに仕向けている部分もあるのかなとは思います。

そうした、いわば測定論を超えた配慮というか、ねらいがあれば意味はあるかもしれませんけれども、測定論的には矛盾していると思います。

○司会　この点について山本先生、何かご意見ありますか。ETS が実施している SAT（Scholastic Assessment Test）などはどうなのかなということなんですが、段階別評定といっても、例えば、SAT などでは標準得点に直して、小数点以下は示さずに、20 点から 80 点の間で表現するから 61 段階評定という感じで提示していますよね。それだったら素点と大差なく、情報量もそんなに落ちないというとらえ方もあるのかと思いますけれども、さすがに 5 段階評定というのはちょっと情報量という点では厳しいかなと思います。

それから、PISA の日本での報告ですと、やはり何段階かにレベル分けされていて、それぞれ何％いるかという形で報告されているんですけれども、それも後ろのほうの問題とも関係してくると思うんですが、その数段階のレベルで例えばある集団において最上位に何％いるかというのは、それなりに情報量があると思いますし、安定すると思うんですけれども、一人の個人がどのレベルかという話になると、南風原先生がおっしゃったように情報量に問題が出てくると思いますし、段階の境界付近の不安定さといった問題も出てくるかなというのが私の印象なんですけれども。

その辺、アメリカでの考え方はどんなところにあるのかということなんですが。

○山本氏　特に私たちが留意しているのは、測定誤差の大きさをどう表現するかという点ですね。例えば、測定誤差が大きかったら 5 点ごとにレポートするとか、そういうときに 1 点ごとの判断を避けた方がよいということを伝えるために、5 点ごとにレポートするといったことをします。

それから、このまま後者の質問の回答につなげてしまいますけれども、PISA はラージスケールアセスメントのうちの 1 つですけれども、ラージスケールアセスメントというの

はもともと政策に関して結果を出そうということですが、PISA 型のアイテムが個人のレポートに使えないわけではないんですよ。例えば、一つの測定対象とすべき構成概念（construct）、ドメインの中で、アイテムの数はかなり多く、80 問とか用意されていますけれども、その 80 問を全部 1 人の人に与えるのは無理です。ということで、BIB スパイラルといいますけれども、バランスト・インコンプリート・スパイラル・デザイン（Balanced incomplete spiral design：螺旋式釣合型不完備ブロックデザイン）を使って一部のアイテムを各個人に与えています。それをもとにして相対的な尺度上に乗せています。

そういうことで、個人個人としては全く別のアイテムのセットをつくりますから、そこで ASA を使うんですけれども。でも、PISA 型のアイテムを使って個人の能力を出すことは確実にできます。

PISA 型のアイテムというよりも、応用力を主に見た、記憶力をアセスメントしていないテストもかなりあるんですね。例えば私たちのつくったアダルトリテラシーとか、PIAAC でも使いましたけれども、PIAAC は、今度そのアイテムも使って新しく、完全にコンピュータベースで、それと同時に適応型テスト（adaptive test）ですから、コンピュータベースで個人の能力を推定するテストです。違うのは、自分個人の覚えたことを持ってきて回答を出すという問題はかなり少なくなって、すべて必要な情報はページに載っていて、そのページの情報をどのように使って回答をつくるのか、そういう能力を測ろうとしたものです。そのようなアイテムをもとにして個人ごとの能力のレポートもできます。だから当然、PISA 型のアイテムでもできます。

○司会　ありがとうございました。

村上先生、お願いします。

○村上氏　これはぜひ伺いたいと思っていたんですが、1 人の個人に対してクラスター 4 つ、さらにもう一つですか、5 つぐらいのクラスターの試験を受けさせているわけですね。そのときに、クラスター間で相関がどのぐらいあるものなのか。つまり、1 つのクラスターの中でのいわゆる内的整合性というのはラッシュモデルへの当てはまりであると理解できるのですが、異なるクラスターの間でどのぐらい相関があるのか、もしわかりましたら教えていただければと思います。

○山本氏　例えば 6 つのクラスターの中で、ドメインの中では相関性はかなり高いですね。例えば数学と科学の相関性は大体 0.85 とか 0.8 とかそのぐらいです。かなり高い。それをもとにして、それがあるからこそ、例えば新しいデザインに関して先程お示ししたように、33 ％はサイエンスとマスをとって、33 ％はサイエンスとリーディングをとるんですけれども、結局サイエンスとリーディングをとる人はマスをとっていないわけですね。だけれども、サイエンスとマスの相関性がありますから、その別々な相関性をもとにして補完することはできますからね。その相関性をもとにして、推算値（Plausible Value）といいますけれども、実際に得られた得点からその得点に関する多次元的事後分布（posterior multivariate distribution）を推定して、それに基づいて補完（impute）する値が算出できます。

○村上氏　それだけあれば、ほぼ一つの次元で 1 人の個人を測定することは可能であろうということですね。原理的には。

○山本氏　ええ。1 つのクラスターに大体 10 問から 15 問ぐらい入っているんですよ。だから、例えば記憶だけのアイテムだったらもっと多く入ると思いますけれども、ただ、応

用力を見るとそのときにはやはり 10 問とか 15 問ぐらいになってしまう。30 分で。
○村上氏　今のお答えで、私は大変驚きました。そういうことであれば PISA 型の問題で個人の能力を測ることは可能だろうと思います。

私はあくまでも自分の経験でしか話せないのですが、日本でよくあるような小論文型の問題は、評価者の間での相関は非常に高いけれども、異なる問題間での相関が極めて低いということ。この事実は、やはりそれらが一つの次元のものを測っているとは考えにくいわけですから、要するにそういうタイプの大問で能力測定は難しいのではないかと考えていたということ。

それから、PISA 型の問題をより選抜的にするというか、選抜性の高いレベルの高校で数学の問題をつくろうとしたときに、これは一定時間の中でこの問題の回答に到達できる可能性はかなり偶然的なものがあって、その偶然を高めることはできるだろう、つまり集団としての能力を測ることはできるかもしれないけれども、個人の能力を測るという点に関してはかなりやばいのではないかという印象を受けていたということで、先ほどのように発言させていただきました。

（発言者注：　この時点では、このように反応したが、改めて考えてみると、山本氏の実施されている PISA 型試験の受験者の能力の幅の広さといったことが、この高い相関のもとになっているとも考えられる。他方、私は比較的選抜性の高い大学、すなわち、受験者が受験以前の段階で相当程度選抜されている大学入試における小論文試験の経験を話しているので、やや議論の土壌が食い違っていたように思われる。ただ，これを言い出すと、現行の多くの選抜性の高い大学における２次試験の「数学」にも同様の問題があるように思われる。いずれにしても、現行の大学入試センター試験の問題を PISA 型に切り替えれば問題解決というわけにはいかないと考える。）
○司会　ちなみに、このサンプル問題はどのぐらい時間がかかるんですか。すべてやるのは結構大変そうな気がするんですけれども。
○山本氏　シミュレーションのアイテムはかなり時間がかかりますよね。10 分はかかっていると思います。最初のほうのページ（スライド 54）に説明が書かれているのですが、設問に含まれる条件の設定方法、例えば、CO_2 濃度とか光の強さなどの変数についてスライダーを動かしてどういうふうに変えるのかとか、まずそういうことを知る必要もあります。

それと同時に、理由などの記述式（スライド 59）のスコアリングというのは信頼性がすごく低いんですよ。例えば前に 1 度やったのは、ニューヨークの高校生の卒業試験（exit test）ですごく重要なテストがあるんですけれども、そのライティングエッセイのリライビリティはもうすごく低いことが報告されていました。エッセイの問題の相関性というよりも、スコアラーの相関性が低いのかもしれません。そこで見えたのは、スコアラーの相関性というのがまず極端に低かったということです。
○村上氏　私の経験では、そこの部分は何とかなる。これはやはり何人かで 2 つを比べたときに、もう絶対こっちのほうがよくできているよねという印象はあるわけですが、では、次の問題を見たときにそれと同じ傾向になっているかというと、そうではないというのが私の経験です。
○司会　基本的に、今、山本兼太朗先生が出してくださったサンプル問題は最低でも 10

分ぐらいはかかるだろう、もうちょっとかかりますかね。ということは、決められた時間の問題数を減らざるを得ないということですよね。集団の平均正答率とかそういったものは一人一人の被験者が測定機会になるので安定した結果にはなるんだけれども、1 人の被験者の能力を推定するという意味で、問題数が少ないと安定した結果が出にくい、そういう矛盾がどうしても出てくるということはあるのではないですかね。

○山本氏　ありますね。

例えば、協調的問題解決（Collaborative Problem Solving）という新しいドメインをやりましたけれども、その時点では結局、どのようにインタラクトするかという問題なんですけれども、30 分で 1 つの設問しかないんですよ。その中でどのように経路をたどるか、それが問題なんです。その経路ごとに見ていくと大体 30 問とか 40 問になりますけれども、だけれどもその時点で多分つながるのは、その設問ごとの相関性が低いんですね、その時点では。だから深くいったときにどれだけ相関性があるか、そこは言えます。

○司会　この問題に関して何かコメントありますか。

○南風原氏　今のことに関連して、先ほど妥当性の 3 つの箱の話がありましたね。それに当てはめるとどういうことになるかということで、今、小論文で表現力を見たいとか構想力を見たいといったことが求められているわけですけれども、今ここで一致しているのは、問題が変わると個人差の順位が変わってくる。どの問題を出すかによって誰が合格するかが違ってくるということですよね。それを妥当性のあの図式に当てはめると、測りたいのは文章力であったり表現力であったりするわけですけれども、実際にはそれとは違う、例えばエネルギー問題であればエネルギーに関する興味や知識が測られている。それを測りたいわけではないわけですね。第 2 問になると今度は人権問題が問われる。人権への関心、ずっと考えてきたこと、そういったことが問われているわけで、どちらも文章力とか表現力とか含まれていますけれども、かなりの部分エネルギーへの興味、知識、人権への興味、知識が測られているということで、そこがエラーになるということなんですね。そういう意味で妥当性が低い。

先ほど休憩時間に、どの問題が出されるか、誰が評定に当たるかということはシステマティックエラーではないかという質問をいただきました。これは見方にもよりますけれども、例えばエネルギーに関する知識、人権に関する知識という意味では安定した成分ですので、そういう意味ではシステマティックで stable だけど irrelevant な成分と言ってよいと思いますけれども、たくさん出し得る問題の中でたまたまエネルギーの問題が出る、たまたま人権の問題が出るという意味ではランダムな部分がある。いずれにしても妥当ではない部分ということです。

いま議論している評定間相関の問題も妥当性の枠でとらえると、測るべきものを測っているかどうかという観点で統一的にとらえられるかなと思いますので、補足しました。

○司会　ありがとうございます。

他に何かございますでしょうか。

○義本氏　文科省の義本と申します。

このシンポジウムのテーマ「大学入試の日本的風土を変えられるか」という問題に多分関連する話なんですけれども、恐らく大学入試の選抜の問題と、例えば就職あるいは公務員を受けるというのはかなり違う風土があるような感じがするんです。

例えば就職であれば、むしろ面接とかやりとりを用意して、今のこのテスト理論でいけば妥当性というよりもむしろ相性とかいうことですけれども、それで回している。公務員も一応試験がありますけれども、面接での対応とかが採用においてかなりの比重を占めるわけです。

ですから、公平性、妥当性の問題になるのか、テスト理論専門の先生方にこういう話をぶつけるのはちょっと筋が違うのかもしれませんけれども、どのようにバランスをとるのかということかなという感じもするんですけれども。つまり、一定の妥当性、あるいはテスト理論的には測れない能力、例えばその人の能力とか、あるいは高校時代どんな意欲を持ってやったかとか、あるいはその気持ちとかいうことについては、やはりテストだけではなくて、いろいろなものを組み合わせてやるという方針が恐らく答申の中にはあると思うんですけれども、その辺のバランスをどう組んでいくのかという点について先生方にお考えを伺えればと思っております。

○司会　ありがとうございます。

○池田氏　教育測定研究所の池田です。

私は、先ほど南風原先生がおっしゃったことをちょっと別の角度から情報提供したいと思います。結局、テストの場合、被験者の数は非常に多くとれるんですけれども、問題の数は限られた時間ですので受験者数に比べれば全然少ないわけですね。

それで、私の先生であるクロンバック先生がクロンバックのαというのを考えていたわけですけれども、そのクロンバックのαは人を対象にして、それと問題との関係を挙げていたわけで、問題数が増えればだんだんαが高くなるわけですね。その残りの部分がエラーとして処理されるわけですけれども、そのエラーというのは表現が悪いので、実はその中にはいろいろな要素が含まれている、それを分析していこうと思ってβとかγとか考えていたのではないかと思うんですけれども。それは実際には受験者数のように増やすことは無理なので、どうしても有限の、数の少ない問題で処理しなければいけない、したがってそれは、個数を増やすことを諦めて、むしろ分散分析の考え方を取り入れて、少ない個数、例えば評定者であれば 5、6 人ぐらいまでが限度でしょうから、そういったものを分散・共分散のような考え方で分析するという立場をとり始めたわけですね。ですから、従来エラーと言う方法で処理されたものを、そのエラーの要素を評定者間の分散とか、それから何回受験したかの回数の分散とか、そういう形で要素に分けて、分散分析の考え方と拡張していこうとして、いわゆる一般化可能性理論が生まれてきた。そういう要素があるということを補足したかったわけです。

○司会　ありがとうございます。

池田先生の師匠がクロンバックであれば、私の師匠、最初に教育測定の授業を受けたのが池田央先生の授業でした。また貴重な情報をいただきまして、ありがとうございます。

それでは、もうあと 10 分切りましたので、義本審議官からも出てきましたこれからの日本の大学入試、今、新しい学力テストは入試センターが実施主体となるという方向性が出てきておりますけれども、それに加えて、個別大学での試験はそのバランスをとるような形で実施していくべきだというようなことが答申に盛り込まれております。それを実現するために、学力試験が公平であるといった従来の入試観から、答申には「公正性」という言葉も出てきておりまして、その観点から多面的な方法による入試が提言されており、

入試というのはそういうものだということを共有していく中で新しい入試を確立していくといったことがうたわれております。

その辺について、今度は逆に村上先生のほうから、これからの学力評価、新しいテストについてでも構いませんけれども、今後の入試の方向性について一言ずつコメントをいただければと思います。

○村上氏　今、民間企業への就職あるいは公務員試験ではまた別の観点があるのではないかというお話がありましたけれども、それは全くそのとおりだと思います。

ただ、一方で、公務員試験の一部分あるいは教員採用試験、あるいは非常にたくさんある資格試験、これらは全く今のところ個別の、私の言った言葉で言えばディスクリートポイントの知識、技能で行われているのは事実なわけですね。ある試験について申し上げると、二十何領域かあります。これはもうほとんどすべて暗記で決まる。さらに言うと、1領域でも0点があったら他が幾ら点がよくても不合格にするという驚くべき資格試験が私の身近なところにはあるわけで、そういうものについて考えると、実は個別知識型の入学試験と非常に相性がいいというか、結局そういうふうにしてきた人間が結構成功することになるみたいなところが一面ではありますね。

それとは別の面で、いわば面接によってわかるとされている人物評価に関しては、話し始めると恐らく30分ぐらいかかってしまいそうなので、また別の機会にお話しさせていただきたいと思います。

○大津氏

センター試験で何を測っているか、余り説明したことになっていなかったのかもしれませんけれども、逆に何を測っていないかを考えると、はっきり根拠があるわけではありませんけれども、恐らく対人的な能力、特に立場や価値観の違う人を言葉で説得して動かす能力、そういう側面を測っているかというと、多分測っていない。

企業の方とか、特に外国との交渉に携わっているような商社の方等が強く要求して今の勉強ではだめだとおっしゃるのは、恐らく安西先生も念頭に置いておられるのは、そういうことかなと思うんですが、ただ、それを試験で測れるかとなると、どう考えてもそれは難しいのではないか。むしろ別の方法を考えたほうがいいのではないかなという気がしています。

何が試験でカバーできるか、分をわきまえるというか、その限界は必要かなとは思います。答えになっているかわかりませんが。

○南風原氏　東京大学で長く学生たちを見ていると、難しい試験を通ってきているんだけれども、例えば卒業研究とかになるとパタッとアイディアが出ない。自分自身で問題を探すということは考えたこともないし、できないということで、そこから非常に苦労して、結局卒業できないといったこともあったりします。ですので、現在の東京大学の試験はそういった学生がパスしてしまう側面を持っているわけですね。

一方、アメリカのアドミッションズオフィスによる評価を見ていると、そういう標準テスト、精度の高いテストに加えて高等学校で何をしてきたかといったことを多面的に評価している側面があると思います。だから今の学力試験一辺倒で測られていない大事な能力があるということは、その通りだと思います。それが1つ。

他方、そういう問題発想力が十分でない学生や協調性のない学生でも、現在ならば東京

大学で学ぶことができるわけですね。そういった学生を切るのか、彼らが学ぶ機会を取り上げてしまうのか、では誰が東京大学の教育を受ける権利があるかと考えるとまた難しい問題も出てくる。

答申案の中にも「障害の有無にかかわらず」とありましたけれども、では、どういう障害を想定しているのか、今、入学試験のときに合理的な配慮をして、例えば発達障害等があっても何とか真の能力と言われているものが発揮できるような方向で努力していることと、今、話題になっているような意欲とか交渉力とか説得力といったことを評価していくこと、これ全体をどのように考えていくのか、真剣に議論すべき問題がたくさんあるなと考えています。

○前川氏

公平性ということですけれども、先ほどの私のスライドにも書きましたように、今は年に 1 回、同じ問題で同じ時期に同じ環境のもとで受けた試験、センター試験ですね、それと 2 次試験の問題を重みを付けて足した点数を使って、その上のほうから何人か採るという形でやっているんですけれども、それが今、日本では入試に関して公平だと思われているんですが、最近、大学は入学者のポリシーを明確にしろとか言われていて、例えば「うちの大学は女性と男性の割合を五分五分にします」みたいなポリシーを作ったとしたら、女性を 5 割、男性を 5 割採るわけですから、当然成績の順だけでは採れなくなると思いますし、あと「うちは大学院まで行く人を優遇します」みたいなポリシーを作ったとしたら、やはり単に順序だけでとるわけではなくなると思います。それから「同窓生の子弟を優遇します」みたいなポリシーだって、スクールカラーを強調するためにはいいポリシーかもしれませんけれども、そういったポリシーに基づく試験をやると、今、やっているような重み付き得点で上から並べてとるという 1 つの作業だけでは採れなくなるのかもしれないという気がします。

それから、段階評定に関してですけれども、例えば 5 段階にしたら、5 段階でないとき、東大の教育学部に入るためには 450 点以上必要。東工大の教育学部は、そんな学部は無いんですけれども、350 点以上とかいうのが、5 段階にしてしまうと両方とも 5 だったらいいみたいなことになると、東工大としては嬉しい。（笑）そういうところもあるのではないかなという気がしないでもないです。○司会　山本先生、日本の大学入試事情はある意味で浦島太郎状態かもしれませんけれども、今日の議論を聞いて何か感想とか、まとめして一言お願いします。

○山本氏　まとめになるかどうかはわかりませんけれども、大学入試を考えたときに、ただ能力だけではなく、大学入試の目的として、大学で伸びる学生を選ぶのか、それとも大学でどう伸びるようにするのか、そんなことを考えて、結局は妥当性（validity）の問題だと思うんですね。だから大学入試の目的をもっとはっきりさせて、例えば社会に出て伸びる人をつくりたいのか、それとも大学の中で伸びる人、伸びるというのは大学の中で成功するというのをどう定義するか、その定義を明確にして、今までよく聞かれていたのは、大学に入ったら自由に何でもやるとか、そういうふうな考え方ではなくて、大学のあり方を同時に考えて、それに即して大学入試を改善していく、大学の目的にかなったものをつくっていかなければならないと思います。そういう意味で、応用力というのは一番続くかもしれないとも思います。ただ、それと同時に、高校までに習った能力は絶対必要ですか

ら、それをないがしろにはできません。だから、習う能力がなかったら当然その時点で能力がないわけですから、能力があるということ自体、習う力があることがわかるわけですね。そういうことで、結局、最終目的からバックアップして、それで大学入試を考えたほうがいいのではないかと考えます。

○司会　ありがとうございます。

私は、大学評価・学位授与機構にもいたことがあって、そのころからディプロマポリシー、カリキュラムポリシー、アドミッションポリシーを各大学で立てろということで、そういう意味で大学の教育の中でどういう学生像を掲げて、それを育てるかという目標をしっかり立てろという動きが広がってきているんですけれども、その辺がまだなかなか、明確な形では共有できていない部分もあるのではないかと思います。

そういう意味で入試改革が、これは測定論だけでないというのは、村上先生の言葉をかりればやはり入試というのは遡及効果があって、入試が変わることによって大学教育も高校教育も変わっていくという側面も大きなことではないかと思います。その点で、今の学力試験中心の入試が「公平性」ということでやられていると答申では書かれていますが、本当にそれだけなのか、私は、前任は京都大学でしたけれども、やはり学力が下がってきているといったことが、京都大学の先生方が授業をやって実感していて、大学の先生方の問題意識として出てくる部分もあるわけで、そうなると、やはり学力を上げるための試験をやらざるを得ないというのが、今までの学力試験中心の入試の要因であるのかなと思ったりします。つまり、「公平性」で学力試験をやっているというよりは、そういうニーズに基づいて学力試験中心にならざるを得ない、今までいろいろな入試改革を今までやられてきたわけで、一芸入試もあり、AO 入試も入ってきた、そういった中で結局、学力が問題だということにいつも返ってくるのは、そういう大学の先生方のニーズや思いというのが大きいような気もしているわけです。

ただ、やはりそれを乗り越えていかなければいけない時代に入ってきた部分も確かにありますので、これは我々入試センターの課題でもあるわけですけれども、当面新しい学力テストをどういう形で進めていったらいいのか、これは非常に多くの問題が隠れた部分にも大きく潜んでいます。例えば、南風原先生が言ってくださいましたけれども、特別措置の問題、つまり障がいを持った受験生対象の共通テストというのは、大学入試センターに来てみると本当に涙ぐましい努力をしていることがひしひしとわかるわけですね。そういう側面からすると、図表が出てきたり長い文章が出てきたりというのは、点字に訳したりするのはすごく大変で、点字はボリュームが広がりますから、この前、初めて点字の会社へ行ってみてわかったんですけれども、小さいコンサイス英和辞典が 100 巻の点字の本になっているんですね、そのぐらい大きな形になったりするという問題も含めて、時間の制約と信頼性・妥当性確保のための量の関係など、いろいろと矛盾するような問題を同時に抱えていかなければいけないという辛さを少しでも皆さんにも共有していただければと思ったりもします。その中で、基本的に大事なことは学生が育っていくことだと思いますので、その原点に立ち返って入試センターの立場から入試改革も進めていきたいと思いますし、また各大学、そして今日は予備校や塾の関係者の方も来ていらっしゃると思いますけれども、いろいろな立場から試験をそのためにどうやっていったらいいのか、今後も共に考えていければと思います。

今日は長時間にわたりご参加いただきまして、本当にありがとうございました。

講演者、報告者の先生方にいま一度拍手をお贈りください。（拍手）

とりわけこの東工大の会場を、最近は、これも大学評価の悪影響の 1 つではないかと思うんですけれども、こういうシンポジウムが至るところで開催されていて、こういう会場を確保するのがものすごく大変なんですね。そういう意味で、前川先生にはこの会場を確保していただいて、昨日もここで泊まり込みで荷物の管理をしてくれたりということで、本当にありがとうございました。（拍手）

また、今日まで、裏方で入試センターの研究支援係の事務職員がバックアップしてくれておりまして、無事にここまで来られたのは研究支援係の応援があったからだと思っております。

それから、前川先生の指導大学院生にもお手伝いいただきました。本当にありがとうございます。いろいろなサポートがあってここまで来ることができました。

雨はどうでしょうか、足元が滑りやすく、今日は玄関で滑って救急車でお 1 人病院に担ぎ込まれたという事件も実はあったので、くれぐれもお気をつけてお帰りください。また何かありましたら、アンケート用紙が入っておりますのでぜひそれに簡単に書いていただきまして、出口で係が受け取りを待っていると思いますので、お渡しいただければと思います。

それでは、本当に今日はどうもありがとうございました。（拍手）

大学入試センターシンポジウム 2014 後記

大学入試センター試験・研究副統括官
入学者選抜研究に関する調査室室長補佐
大塚雄作

シンポジウムの趣旨と概要

本報告書は、2014 年 11 月 29 日（土）13:00 ～ 17:30、東京工業大学大学院社会理工学研究科デジタル多目的ホールで行われた、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』（主催：大学入試センター入学者選抜研究に関する調査室、後援：文部科学省・日本テスト学会）の全記録である。折から、教育再生実行会議、中央教育審議会などを中心に、大学入試改革が議論されていたこともあって、雨模様の中にもかかわらず、文部科学省関係者 9 名、大学関係者 97 名、高校関係者 10 名などを含む 255 名の参加者を得て、充実した報告と熱心な議論が行われた。【参考】として、シンポジウムに関するアンケート結果を本後記の後ろに付けているが、いろいろと至らぬ点にも気づかされる一方、全体的には概ね満足が得られたシンポジウムを開催することができたことに、まずもって、参加者の方々、また、下支えして下さったスタッフの皆さんに心よりの御礼を記して表しておきたい。

このシンポジウムは、荒井克弘試験・研究統括官の趣旨説明にあるように、「大学入学者選抜が、学力エリートの選抜から、これからの知識社会を支える対象層を選抜し、教育し、社会的に配分していくという役割に変わってきており、その意味で、「選抜」というよりも教育プロセス、教育課程の積み上げという部分が強調される必要があり、そのために、大学入試は可能な限り妥当なそして信頼性の高い教育測定に重点が置かれなければならない」という観点から、教育測定の専門家をパネリストとして登壇いただいた。

まず、ETS の山本兼太朗先生に、PISA 調査に関わる分析等の基本的考え方について、教育測定の専門的な立場から講演をいただいた。入試をテーマにするなかで、PISA 調査を取り上げたことに違和感を感じた参加者もおられたようであるが、新たな PISA 調査に向けての準備のための緻密なフィールド調査の計画やトレンドを把握するための測定理論的背景に基づいた分析手法が示されたことによって、調査と選抜という目的の違いこそあれ、教育プロセスの流れのなかで新たな入試を開発していく際の流れの一端と留意すべきポイントのいくつかを窺い知ることができた。

続いて、前川眞一先生からは、新しいテストを導入するにあたり、世界の試験の動向を踏まえつつ、複数回の受験機会の準備や CBT の導入といった課題に対して、日本の入試風土をどう変えていく必要があるかについて紹介していただいた。南風原朝和先生からは、「公正性」の概念の基本となる信頼性と妥当性についてわかりやすく解説していただくと共に、成績提供の際の段階表示について、得られる情報量がいかに低くなってしまうことになるかということを中心に、公正なテストを目指すことの重要性について強調していただいた。大津起夫研究開発部長からは、大学入試センター試験の実際のデータに基づいて、科目得点間の相関分析の結果、また、受験者層の特徴などについて紹介していただいた。

村上隆先生からは、留学生のための日本語試験に携わられた経験に基づいて、オーセンティックな文脈に沿った測定が求められる風潮のあるなかで、複数回受験を保障する試験システムの導入の難しさなどについて浮き彫りにしていただいた。最後に大久保智哉助教から、研究開発の一環として開発中の CBT のサンプルデモがあり、CBT の具体的イメージと可能性について共有する機会をもった。それぞれの報告は、教育測定論のしっかりした基盤に基づき、それだけに専門的に難解な部分も含まれつつも、充実した内容のものであった。

その後、質疑応答、ディスカッションを 30 分程度行って、シンポジウムは終了したが、このシンポジウムでの議論は、残念ながら、ほぼ 1 ヵ月後の 2014 年 12 月 22 日に中央教育審議会から出された答申『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～（巻末参考資料参照・以降「高大接続答申」）』にインパクトを及ぼすには、時期的に遅きに失した。しかし、実際に入試を変えていく段階において、まずは十分に検討しておくべき基本的課題の多くが取り上げられたと思われる。

比較的近年においても、入試改革に関しては、平成 11 年（1999）中央教育審議会答申『初等中等教育と高等教育との接続の改善について』、平成 12 年（2000）大学審議会答申『大学入試の改善について』などにおいて、上記答申と同様の趣旨で、多面的な評価や複数回受験などが提案されていたにもかかわらず、その一部は結局実現されぬまま今日に至り、そしてまた同様の提言が昨年末の答申になされているという点は留意しておくべきことであろう。理念的によいということで、何度となく取り上げながらも、実際の具体的手段に結び付かないということがあるのは、一つは入試に関わる教育測定理論の基本が共有されていないということであり、また、入試がどうあるべきという入試観がその理念に追いついていないということなどが背景にあるのだろうと思われる。現実に、今回の答申案においても、私自身の目から見ると、相矛盾する提言が並列されていて、それをうまくバランスさせる入試改革は至難の業ではないかと正直感じさせられる部分が多く含まれてしまっている。本シンポジウムで、教育測定に関わる専門家の方々から提起いしていただいた、ある種テクニカルな側面からの入試改革に関わる課題について、改めて噛みしめてみていただければと思う。

「公正性」と多面的な評価における「妥当性」の課題

当日のシンポジウムでは、十分に整理しきれなかったが、以下に、私自身の個人的な観点からではあるが、答申に含まれていると思われる矛盾点や問題点について、今後の参考のために簡単に取り上げてまとめておくことにしたい。

まず、高大接続答申では、新しい大学入試は「公正性」に基礎を置くべしとされている。本答申については、全文を巻末に参考資料として掲載しているが、その「多元的な評価に向けた意識改革と、新たな評価手法の蓄積･共有」という項において、「個別選抜における評価に当たっては、画一的な一斉試験で正答に関する知識の再生を問い、その結果の点数のみに依拠した選抜を行う従来型の「公平性」「客観性」と、多数の受験生に対して短時間で合否判定を行うための効率性を重視するあまり、面接、集団討論、小論文、調査書、その他による多元的な評価を重視しない傾向がある。この点に関しては、客観性とは何かについての意識改革と併せて、個別選抜を行う側が、自らの都合のみにより選抜する方法

ではなく、一人ひとりの入学希望者が行ってきた多様な努力を受け止めつつ、入学者に求められる能力を「公正」に評価し選抜する方法へと意識を転換し、アドミッション・ポリシーに示した基準･方法に基づく多元的な評価の妥当性･信頼性を高め、説明責任を果たしていく必要がある。」と述べられている。

「公正性」という言葉には、それこそ、従来型の「公平性」「客観性」という要素も含まれていると見ることもできるし、社会的格差の是正であるとか、社会的正義と言った社会的な観点からの公正という捉え方もあるし、多義的でいろいろな捉えられ方があるが、そのなかで答申では、「信頼性」、「妥当性」という、教育測定が具備すべき二つの特徴を兼ね備えているということが強調されていることがわかる。これらの点は、教育測定の専門家からすれば、むしろ当然のことであり、異論を差し挟む必要のない入試等の測定における基本的前提でもある。また、知識のみならず、知識活用力（思考力・判断力・表現力等）を身に付けていくことや、「主体的な学び」につながる学習意欲を醸成していくことは、高校から大学教育にかけて重要な教育目標に位置づけられるということも異論のないところであろう。

しかしだからと言って、知識活用力や学習意欲などに関わる個人の特性を多面的に測定して、その結果を選抜に利用するということは、ある意味で短絡的に過ぎるし、何よりも、答申で強調されている「公正性」という観点からすれば種々の矛盾が生じることになるのである。例えば、「学習意欲」などに関わる測定が各大学の個別入試で実施されるということになると、その観点での点数を上げるためにどんなことが受験を目指して行われることになるか、想像するだに恐ろしい部分もあるが、大学入試はそういうインパクトを及ぼすということが常に内包されていることを念頭に置いておく必要がある。たとえ心理学的には妥当性の高い「学習意欲」の測定手法が開発されていたとしても、その手法が入試に採用された途端に、受験準備という波及効果を経て、すぐにその手法の「妥当性」は崩れ去っていってしまうのである。「妥当性」はまさに、測定手法の属性として不変のものではなく、利用目的が変われば変化するものであるし、また、受験者の準備のあり方、測定の対象となる集団の違いによって、変わってしまうという特徴を持っているのである。

さらに、「妥当性」は、一定の「信頼性」が確保されないとそれが高い測定にはなり得ないという性質を持っている。その点で、多面的な評価として想定されている「面接」であるとか、グループワーク等の「観察」であるとか、「小論文」であるとか、そういった類のいわゆるパフォーマンスに関する評価については、採点者の信頼性が基本的に低いという問題点をクリアしていかねばならないのである。入試の場合、限られた時間、限られたマンパワーの下で行われることで、この点はフィージビリティの課題にもつながることでもある。その一方で、では、信頼性を高い手法を採用するというところに力点を置くと、特に、測定の対象となる特性が複雑になればなるほど多面線を有することになり、妥当性を確保しにくくなるという、信頼性と妥当性の矛盾という測定論的な基本は存外共有されていないことの一つではないかと思われる。

知識活用力についても同様のことが言える。学習意欲ほどではないにしても、この点に関して答申で取り上げられている合教科・科目型の試験であるとか、総合型の試験というのは、さまざまな要素を含み込むことになり、結局、それによって何を測定しようとしているのかが不明瞭になる、言い換えれば「妥当性」が低くなる可能性が高いという問題点

が常につきまとうことになる。多側面の問題項目を含むということは、「妥当性」ばかりでなく、内的な整合性が確保しにくくなるということにもつながり、その結果、「信頼性」を確保することにも困難を生じることになる。

また、合教科・科目型、総合型の問題は、基本的に、広範囲の内容領域をカバーする必要も生ずること、リード文や図表なども併せて提示されるなど、問題解答により多くの時間が必要とされる可能性が大きくなることなど、望むらくは十分な試験時間が必要とされることになるが、一方で複数回の受験機会を準備する要請なども含めて、試験にかける時間はできる限り効率化を図ることが求められているという矛盾も抱えることになる。一定の問題数に絞らざるを得ないとなると、勢い「信頼性」も低くなるという問題点もクリアすべき課題となる。

さらに、合教科・科目型、総合型の問題というのは、それを作成するのが容易とは言えず、毎年引き続き行う必要のある入試において問題作成体制を長い間維持していくのはそう簡単なことではないというフィージビリティの問題にも遭遇することになるであろう。

なお、「知識活用型」の問題に関しては、山本兼太朗先生の講演において、PISA の CBT の問題例が示された。入試においても、そこで測定されるべき能力は、今後さらにより高次のものが取り上げられていくことになるであろう。その際に、まず、PISA のように、国別、地域別など、多くの被調査者の平均的なレベルを知るための「調査」と、個々の受験生の力量を識別しなければならない選抜のための「入試」との違いを認識しておく必要があるだろう。PISA 調査は、知識活用型の「問題」と「受験生」の両者の次元におけるサンプリングの工夫によって、国ごとの平均値に関しては、内容的な偏りなども小さくなるように工夫され、また、標準誤差なども小さくなるように配慮されている。確かに、個人個人が取り組んでいる問題は異なっているという状況にあったとしても、項目反応理論（IRT）を利用した統計的な推定値として、一つの尺度上に個人の能力を位置づけることも可能である。ただ、その推定値を大学入試などの選抜に利用するということに社会的なコンセンサスが得られるまでには、IRT の測定対象能力の一次元性の前提や局所独立の仮定などからの現実の乖離の問題、問題項目の非公開の問題、項目パラメータなどの算出に関わる信頼性やフィージビリティの問題などの多くの課題が山積しており、ある程度は頑健な結果が得られるであろうということがあったとしても、日本の試験風土においてはかなりの時間とステップが必要とされると思われる。

「情報量」と段階別表示による成績提供

高大接続答申では、共通テストとして、「高等学校基礎学力テスト（仮称）」と、「大学入学希望者学力評価テスト（仮称）」の二つの大規模なテストの実施が提言されている。そして、それぞれのテストは、多様な受験者を想定して、その選抜のための識別力を確保するために広範囲の難易度の問題を含むこと、また、「一点刻み」の入試の悪影響から脱するために、「段階表示」による成績提供などが提言されている。

この点については、南風原朝和先生からの報告に詳しいが、やはり答申にはいくつかの矛盾点が含まれている。IRT のモデルに従えば、各問題項目は、被験者の能力に関わる特性値の関数として、情報量という概念が提起されている。ある能力レベルにおける測定の情報量が大きいということは、その能力レベルでの誤差が小さく、識別力が大きいということに対応していると考えておけばよい。そして、能力レベルごとに、テストに含まれる

各問題の情報量をテスト全体で合計したものをテスト情報量、測ろうとしている特性値に情報量を対応づけた関数を情報量関数などと呼んでいる。テストは、情報量関数の大きい能力レベル付近の測定精度が高いということになる。各項目レベルでは、その項目の難易度を表す能力レベル付近で高い値を示すことが知られている。例えば、易しい問題ばかりであれば、能力の高いレベルの受験生にとってはみんながができてしまって、識別力が低くなり情報量も低いということになり、言い換えれば、易しい問題をいかに集めても、高レベルの能力の識別には有効ではないということになる。能力の高いレベルには、それ相応の難しいレベルの問題が必要となり、その意味で、確かに、広範囲の多様な受験生が想定される場合には、項目の難易度も広範なテストを用意する必要があるということにもなる。

しかし、試験というのは、時間的な制約があり、広範な難易度の項目を網羅的に含むようなテストは現実的には実施することが難しいという矛盾点が出てくるのである。そこで、それぞれの難易度の項目数を減らせばよいということも考えられるが、項目数を減らせば、一つの能力レベルにおける情報量を大きくしていくことができにくくなるわけで、テスト情報関数は全体的に情報量の低いものにならざるを得ないということがある。いわゆる適応型テストと呼ばれるコンセプトは、まさに、能力レベルの識別にふさわしい情報量の大きな項目を次々に提示することによって、少数の項目で効率的に能力推定をしようというものである。ただ、適応型テストをいきなり大学入試に導入するには、やはり多くの壁が聳えていて、現時点ではまだ現実味があるとは言えない段階である。

そこで、それを解決するためには、難易度の異なる何セットかのテストを用意し、それぞれのセットに一定の項目数が含まれるようにし、それらをその難易度にふさわしい能力レベルに対して提供するシステムを作ることが考えられよう。ただ、このような形のシステムは、どのレベルの試験を選択するかで、大学や高校の序列化に直結してしまい、それこそ日本の試験文化の中では実施しにくいという矛盾点もつきまとうのである。ただ、ユニバーサル化の時代にあって、現行のセンター試験は科目数が多すぎて複雑という言説もしばしば聞かれることではあるが、いくつかの難易度のテストを用意して、それを選択することのできるようなシステムが受け入れられるのであれば、この矛盾はある程度解消していくことができるであろう。それが複雑であったとしても、多様性に対応できる柔軟なシステムを導入できるのであれば、我が国の先進性を示す機会にもなるという意味で、一考の価値のあるポイントではないかと思われる。

段階表示に関しては、南風原先生の報告で詳しく言及されたように、試験の素点を少数の段階で評定することによって、テストのもつ情報量がかなり減縮してしまうということで、選抜に利用するという目的のためには必ずしも望ましいことではない。無論、各大学の個別入試において、多様な選抜に利用する情報を収集して、それぞれを段階評定して、それらを合成するなどして選抜に活用するという手順が想定されているのかと思われるが、いずれにしても、その合成得点の「一点刻み」は何らかの形で残ることになるという点には留意しておく必要があるだろう。また、そのような合成得点に共通テスト結果の段階評定を利用する仕方は、大学の状況によって違ってくるのであって、各大学へは点数の形で提供し、それを段階化するなどの次の工夫は個別試験の段階で各大学の事情に即して決めていくということが自然であろう。

段階評定に関しては、どの程度の数の段階にしたらよいのかということも、それぞの事

情や価値観によっても違ってくることであり、どの大学にも一律のこれがよいという分類の仕方があるわけではなく、せいぜいあり得るのは、情報量を減退させない形でのいくつかの標準化手法による変換得点の提供といったことが無理のないところであろう。IRT などが利用されるようになれば、得点は項目への解答パターンの数に相当する段階を想定することができ、それは 100 点満点の素点の段階に比べようもなく細かく算出されることにもなるし、また、そこで推定される特性値は基本的に標準得点であり、いわゆる偏差値に相当する指標であるということも、とりあえず共有しておきたい測定の基本である。

「複数回受験」に関わる諸課題

答申では、年「複数回」の実施ということも、その提案の重要ポイントの一つとされている。複数回実施ということは、まず、その実行可能性という点で、現行の実施の形を根底から変えるやり方を考えていかねばならないであろう。この点に関しては、前川眞一先生、村上隆先生の報告で言及されていたように、日本の入試風土では馴染みではないいくつかの点をクリアしていく必要がある。一つは、複数回の試験の得点が同等のものであるということで、それぞれの試験を等化する（equating）必要が必ず生じるということであり、そのために、問題項目などは非公開のデータベースに大規模に蓄積されていく必要があるという点である。等化の際には、IRT などの測定理論に依拠することが考えられるが、IRT は項目のパラメータを推定するために、予備調査が必要とされたり、あるいは、本番の試験に、総合得点の算出には利用されないダミー項目を一部に含んでおく等の工夫が必要とされる。それらのこと自体、日本には馴染みのないことであるが、それ以外にも、非公開であるはずの問題について予備調査をしなければならないという点、また、IRT は基本的に一次元性の能力を前提としており、多面的かつより高次の能力を対象とする場合への理論的展開はまだ発展途上にある段階にあるといった矛盾点も内包されいているということをまずは確認しておくべきであろう。

また、現行のセンター試験では、遅くとも翌日の新聞などでは、試験問題が公表され、自己採点によって、大学出願を各受験生レベルで決定していくというプロセスをとることになっているが、複数回受験などの場合には、試験問題を公表することができなくなると考えておくべきである。試験問題の公表ができなくなるということは、現在は、公表された試験問題を素材にして、さまざまなところで議論が起こったり、また、高校や大学においても教材として利用されたりということも散見されるが、そうした教育への活用もできなくなる。答申には、高校教育、入試、大学教育の一体的改革という理念が標題にも掲げられているが、問題を非公開にすることによって、入試の重要な教育的機能の一部が十全に発揮されにくくなる矛盾を抱えることにもなるという点はあまり気づかれていないことではないかと思われる。

さらに、複数回の試験得点の等化のために、IRT などを利用するにしても、得点は標準得点の類で表されることも必定であり、自己採点などの機会も失われることになり、出願のためのステップとして別の方法を策定していく必要も生ずるであろう。いずれにしても、そういった現行の入試の在り方に関する観念を廃棄して、新たな入試のコンセプトを提案し広く社会に共有していくことが必須となる。

もう一点、高校と大学の接続という教育的観点が関わる大学入試の際に考慮しておくべきことは、項目データベースに保持すべき項目や IRT の適用のために必要となる項目パラ

メータなどは、教育の改訂などに応じてその都度収集し直していかなければならないということである。日本では 10 年ごとくらいに学習指導要領が改訂されるなど、試験の対象となる内容領域の範囲が変えられていくことになるが、その度に、試験問題の項目の入れ替えや項目パラメータに関わる大規模な事前調査の必要が生じ、必ずしも効率のよい形で維持してくことは容易でないということも一つの矛盾点として留意しておくべきであろう。

さらに、答申では、英語などに関しては、4 技能の試験をすべしということで、外部の英語資格検定試験の活用が提言されており、そのために、各種の資格検定試験の得点の対応表などを作成し公表することが求められている。異なる目的の試験の得点の対応表は、そもそも測定論的には意味のあるものではない。目的の違う、あるいは、測定したい対象が異なる、あるいは、受験者の想定される母集団が違うといったことがある状況では、得点の等化という言葉に対して、対応づけ（concordance）という言葉が使われることが多いようであり、その手法の研究開発も進められているところではあるが、基本的に、対応づけにも誤差が含まれることになり、高得点者などについて一芸入試的な活用方法は有力であるとは思われるが、「公正性」という点からすれば、言い換えれば、妥当性や信頼性という点からすれば､種々の問題が残されるということは十分に自覚しておくべきであろう。

この点に関してもう一つ検討しておくべきことは、複数回受験は、一発勝負のプレッシャーから受験生を解放するというねらいも含まれているのであるが、逆に、高校教育の担当者や受験生からの声の中にしばしば見られるのが､複数回になることによって､しかも､基礎学力テストと選抜のための学力評価テストの二本立てとなると、高校生が試験漬けになってしまわないかということである。複数回受験は、その実施時期のコンセンサスを得ることであるとか、テストセンターの設置などの試験そのものの日常化のための環境作りであるとか、また、それに関する高校を初めとする社会の理解であるとか、そういった試験の周辺的な部分で、社会全体の協力体制を築いていくための雰囲気作りがまずは求められるということであろう。これもまた、日本の入試風土の転換が求められている点でもあろう。

CBT の導入

複数回受験にも関連して、答申では、CBT（Computer-Based Testing）の導入についても提言されているところである。本シンポジウムでも、山本兼太朗先生から、PISA における CBT 問題の例も紹介され、また、大学入試センターで進めている研究開発の一端についても大久保智哉助教からの報告があり、今後、われわれが避けては通れないメディアの変革を見据えて、入試についても CBT 化を含めた利用メディアの変革への対応は常に念頭に置くべき重要なポイントである。

おそらく、現行のマークシート方式を、CBT に置き換えるという程度であれば、デバイスの維持管理及び装置やテストセンターなどを整える初期投資などの問題がクリアさえすれば、それ程大きな問題もなく移行していくこともできるのではないかと思われる。ただ、リスニングで利用されている IC プレーヤーでさえ 1 万台に 1 台程度の不具合の申告があり、これはこの種の危機では非常に小さな確率ではあるものの、50 万人の受験生がいれば確実にそういう不具合が発生することにもなり、CBT などでの利用が想定されるタブレット PC などのより高次のデバイスはその確率はさらに高まる可能性が大きいだけ

に、そういうリスクマネジメントはしっかり講じておく必要があろう。

それに加えて、CBT 導入のメリットとして答申で考えられていることは、共通テストにおいても記述式問題を含めるといった提案がなされている点に関わることであろう。記述式などの問題は、50 万人規模の試験においては人による採点ということではほとんど実現可能性のないことであり、少なくとも CBT が導入され、受験生によって入力されたデジタル的な回答を自動採点するシステムが導入されることが前提となるだろう。自動採点システムは既に英語のライティングやスピーキングに関して一定のものが開発されてもいるが、しかし、短答式の記述回答であったとしても、現行の 50 万人規模、今後の受験生数の減少傾向を考えても、数十万オーダーの受験生からの多様な回答（誤答とは判定困難な想定し得ない回答など）に対応する自動採点システムを構成することは、そう容易なことではないと思われる。また、答申でねらっている多面的な能力といったレベルの力は、自動採点システムのアルゴリズムには乗りにくい部分もあり、そういった課題克服も残されていくことになるだろう。

また、自動採点システムが開発されたとして、そのアルゴリズムは公開されることはないであろうが、しかし、入試で利用される採点システムの特徴は、受験準備の段階でさまざまな角度から分析され、受験テクニックとして受験生に伝えられることになるという、入試の波及効果、インパクトといった点も考慮に入れておく必要があるだろう。記述式の問題に関しては、フランスのバカロレアの哲学の問題がしばしば話題として取り上げられるが、正解のないような抽象的な問題に対しても、大規模な入学試験で取り上げられることによって、受験テクニックがむしろ先行してしまい、ねらいとする哲学的思考力を測定するという意味では、ごく上位の一部に限られてくるという報告もある（坂本尚志, 2012）。

また、シンポジウムでも議論されていた点として、山本兼太朗先生の言葉を借りるのであれば、「アダプテーション（adaptation）」の問題も留意しておく必要があるだろう。自分が普段利用しているコンピュータを試験場に持ち込んで、CBT が実施できるというようなことであれば、このアダプテーションの問題はある程度回避できると思われるが、その場で初めて接するデバイスに関しては、その操作への慣れの程度に個人差が発生するということである。つまり、アダプテーションという個人差のファクターが、新たな体系的誤差要因として、そこで測定される得点の妥当性を低める可能性のあることに留意しておく必要がある。

もっとも、この問題は、紙筆検査における記述式においても、文字のきれいさといったファクターが無視し得ないということもあって、CBT だけに問題とされることではないし、また、一人ひとりにタブレット型 PC を与えての授業が小学校から導入されたり、電子教科書なども普及しつつあるなかで、重大な誤差要因として取り上げる必要もなくなる時代はそう先のことではないということかもしれない。その点で、順当な手順を考えるのであれば、まずは練習問題などの形で普段の授業の教材に組み込まれて利用を広げるといったところから、段階的な戦略を講じる必要もあるだろう。

もう一点、CBT を利用するということは、複数回の受験という体制も組み込まれることになるだろうし、また、今のような一斉の試験実施ではなく、各地のテストセンターなどに適宜受験生が集まって実施する形を取るなど、現行とは違った体制での試験が実施されていくことになるであろう。そのような分散型の試験実施体制では、小さな不具合など

への対応など比較的柔軟に扱いやすくなるであろう課題と、また、そのことで問題漏洩の確率が大きくなるなどの新たに生じる課題などを整理しておくことも肝要であろう。この点も日本の大学入試ではいままで経験してきていないとであり、動きながら修正して行かざるを得ないという部分もあるとは思われるが、大学入試という事業の性格上、十分なシミュレーションを怠るわけにはいかないであろう。

いずれにしても、現時点でまだ十分に先が見通し得ないことは、それを導入する際の初期投資、維持管理などに要する費用や、先にも触れたことであるが、膨大な項目データベースを準備するのに必要な労力など、受験料などでカバーしきれないであろう部分のリソースがきちんと手当てされるのかどうかといった点での不安である。入試という測定ツールで何が測定されているのかという点に関しては、大津起夫研究開発部長からの報告にもあったように、それぞれの能力にはそれなりの相関が観測されることになり、言い換えれば、現行の試験でも、CBT などによる試験でも、測ろうとしたい特性に関する情報はそれなりに得られるということもある。それならば、コストパフォーマンス的に、厳しい国家財政の中から、莫大な予算をそれに投入しなければならないという説得的な理由を社会に伝えていくことができるのかどうかは、大いに疑問の残るところでもあり、メディアの浸透の経緯などにも留意しつつ、多少なりとも時間的余裕を持った取組が必要とされることになるであろう。

入試改革をどう動かしていくのか

以上、答申に対して、おそらくは、かなり否定的なトーンが伝わったのではないかと思うが、もちろん、現行のセンター試験が最良の入試システムであるとは決して考えているわけではないし、時代の流れに即して、入試を変革していかねばならないという点についてはまったく異存のないことでもある。しかし、今の日本的入試風土のなかにあって、矛盾点を多く含んだままの入試の変革だけが先走りしても、おそらくは途中で挫折してしまうことは目に見えていることでもある。変革したときの入試像に関して、高校・大学の関係者、入試実施主体、そして、受験生を中心とするステークホルダーなど、その全体がそれでやっていけるという効力感を持てることと、それが今の入試制度に比してよりよい教育につながっていくという雰囲気作りがまずは大切なことであり、その点も考慮しつつ、長期的な戦略を立てることが肝要であろうと思われる。

東（2001）は、メリトクラシーの進展による階層化を緩和する手段として、大学入試にくじを導入するという提案をしているが、「くじの導入」ということと、現答申に基本的理念としてとりあげられている妥当性・信頼性に依拠した「公正性」という考え方とは相容れない部分があり、それはおそらく今でも社会には受け入れられない大学入試手法であるだろう。しかし、多面的な評価であるとか、合教科的な試験であるとか、信頼性や妥当性の点である種の問題を抱えている入試は、ある意味で、「くじの導入」ということに限りなく近い手法と言えなくもないという点は自覚しておくべきであろう。逆に、妥当性・信頼性に依拠した「公正性」という理念を棄てて、入学試験は「運」の要因が大きく含まれているものであり、せいぜいその程度のものだという前提が共有されたとすれば、現行の答申の提案もかなり実現可能性の高いものとして位置づけられることになるであろう。あるいは、「公正性」という言葉を別の見方から捉えてみるならば、「くじ」という手法はその理想的手段に位置づけられるかもしれず、それはそれで一つの選択肢となり得ると

いうことかもしれない。ただし、「くじ」などが導入されれば、統一的な学習指導要領などによる教育課程の体系性なども薄れることになるし、かえって学ぶ目標のとらえどころも曖昧になることから、学習意欲なども必ずしも確保しにくいという事態も招きかねないといったことも考えると、答申で前提とされている妥当性・信頼性に依拠した「公正性」という理念を覆すことはそう容易くできることでもないだろう。

そうであれば、例えば学力評価テストに関して、合教科型の試験であるとか、また、知識活用型の問題であるとか、CBT などといった新たな実施形態にいきなり本番の試験を変える以前に、高校での日常的な授業であるとか自習活動で利用できるような学習材のなかに、まずは本答申で提案されているようなことを一つひとつ組み込むところから始めていくといった手順が求められるのではないかと思われる。少なくとも、PISA 調査などを参考に、高校レベルの学力調査などの導入あたりから、CBT 等も含めて、実験的に試行していく手順なしに、いきなり入試そのものに新たな試みを仕込んでいくというのは、むしろ日本的入試風土の下では風当たりが強いであろうことは想像に難くない。それだけにより具体的かつ段階的な入試改革のプロセスを共有していくことが、まずは求められることではないかと思われる。

それと平行して、入試改革の基本は、教育測定を含む、大学入試に関わるリテラシーの底上げも必要となるであろう。新たな方法を導入した際に、追跡研究をはじめ、いろいろな角度からその手法に関わる検証をしていく必要もあり、それは原則的に各大学の状況に即して行っていくことになることから、各大学のアドミッションズ・オフィスに携わる人材の充実など、これは答申にも言及されていることではあるが、重要なステップとなっていくであろう。

いずれにしても、入試改革は、一朝一夕には実現することではない。少なくとも、大学入試センターのような試験実施機関とでもいうべき機関が試験を作って実施するからと言って、高校や大学、社会が、その利用価値を共有してくれなければ長続きするものではない。その点で、現行のセンター試験は、いろいろな批判を受けながらも 30 年以上にわたっての蓄積がなされてきており、ある意味での日本の入試文化の一端を担う位置づけも可能である。その足跡をきちんと踏まえ、その過程で積み重ねられた知見を新テスト開発に十分に活用するという基本的姿勢に常に立ち返りつつ、今後の入試改革の荒波に乗りだしていきたいものだと思う。

文　献

東洋　(2001). 子どもの能力と教育評価　第 2 版（UP 選書 198）　東京大学出版会

坂本尚志（2012). バカロレア哲学試験は何を評価しているか？―受験対策参考書からの考察―　京都大学高等教育研究, No.18, 53-63. 京都大学高等教育研究開発推進センター

【参考】アンケート集計結果

シンポジウム終了時のアンケート結果の概要を以下に示す。

◇回収数：146 票　（回収率：60 %）

①性別：男　106（73.6 %）　・　女　38（26.4 %）

②年齢：～ 29 歳　11（7.6 %）・ 30 ～39 歳　30（20.8 %）・ 40 ～49 歳　35（24.3 %）・ 50 ～59 歳　54（37.5 %）・ 60 歳～　14（9.7 %）

③所属：国公立高校　5（3.4%）・ 私立高校　8（5.5%）・ 塾・予備校関係　27（18.6%）・ 国公立大学　35（24.1 %）・ 私立大学　30（20.7%）・ 民間企業　26（17.9%）・ マスコミ関係　3（2.1%）・ 省庁・役所等　3（2.1%）・ その他　8（5.5%）

④職種：教員　49（35.0%）・ 事務系職員　43（30.7%）・ 専門職　21（15.0%）・ 公務員　4（2.9%）　・ その他　23（16.4%）

	平均	標準偏差	有効回答数	1(%)	2(%)	3(%)	4(%)
				あてはまらない←→あてはまる			
（1）わかりやすかった	3.09	0.63	146	1.4	11.6	63.7	23.3
（2）新たな発見があった	3.56	0.57	146	0.0	4.1	35.6	60.3
（3）興味深かった	3.66	0.60	146	1.4	2.7	24.0	71.9
（4）構成は適切であった	3.29	0.69	146	1.4	9.6	47.3	41.8
（5）有益であると思った	3.49	0.60	145	0.7	3.4	42.1	53.8
（6）集中して参加できた	3.28	0.69	146	0.7	11.6	46.6	41.1
（7）時間の長さは適当であった	3.05	0.75	146	1.4	21.2	47.9	29.5
（8）開催日時は適切であった	3.32	0.68	146	0.7	10.3	45.2	43.8
（9）総合的に満足できた	3.45	0.62	144	1.4	2.8	45.1	50.7

◇主な自由記述（抜粋）

- 評価の難しさ、多様さがよく認識できた。入試選抜の難しさも分るが、本題の部分の議論があまりなされていなかった。（男・60 歳～・その他・満足度 4）
- 「学力評価のための新たなテスト」の位置づけが明確になりました。単なる測定論だけではないこと（男・40 ～ 49 歳・私立大学・教員・満足度 4）
- 改めて、IRT 導入の重要性は再認識できたが、いざ、センター試験で CBT ＋ IRT 理論をいきなり導入するとなると、非常にムリがあると思う。紙ベースの長所、コンピューターの長所を活かせるような柔軟性のあるものへの対応が必要ではないか（男・30 ～ 39 歳・塾・予備・専門職・満足度 3）
- いろいろな考え方、立場があると改めて感じました。それぞれの“立ち位置”をもっと考える必要があると思います。新しい学力テストがどうなるか大変な問題です。（男・50 ～ 59 歳・国公立大・教員・満足度 3）
- 「入試」＝「テスト」という発想こそが日本の「入試の風土」である。今求められているのは、「テスト」＋調査書、面接、論文を含む総合的評価への方向性。その中で「テスト（CBT)」はどこまで測定論的に厳格でなければならないのか。（「センター試験」実施組織としてはいたしかたない）　（男・60 歳～・国公立大・教員・満足度 3）
- 理論的な知見と実務的な経験の両方を聞くことができたのがよかった。特に南風原先生の報告

は中教審の答申とテスト理論の話を組み合わせており、専門外の方であっても分かりやすい内容であったのではないかと思われる。(男・～29歳・民間企業・専門職・満足度4)

- CBT サンプルのデモの話が将来の導入向け　サンプル・デモンストレーション実施　印象に残りました。ありがとうございました。(男・50～59歳・国公立大・事務系職・満足度3)
- ・合教科・科目型について：テストは、高校までのカリキュラムとの関係を考えて、科目別（多少再編があってもよいが）に行うべきではないでしょうか。その上で、例えば生物の問題で英語の資料が使われたり、統計のスキルが必要といったことはあっていいと思います。

  ・小論文や論述試験について：IB のように、複数の領域を用意し、各領域で多くの問題を用意して受験者がテーマを選択できるようにした方がいいと思います。(主に個別試験だと思いますが、センターが問題を準備できないのでしょうか)

  ・複数回について：大学入学希望者学力評価テストの複数回化は、イギリスで（の）議論なども参考に慎重に判断して欲しいです。今回の改革で学力低下につながらないように工夫が必要だと思いました。

  本日は多くのことを学ばせていただき、ありがとうございました。(男・50～59歳・民間企業・専門職・満足度4)
- 大学入試の何を変えたいのか、何故変えるのか、は仮定せずに、何ができそうなのかを議論していたように思います。大学といってもどのような入試問題をどのような仕組みでという人材像と制度と、枠組みを仮定して議論しないと一つの解も見えないかもしれないと思います。啓発的な意味では大きな意義がありましたので、ありがとうございました。(男・50～59歳・国公立大・教員・満足度4)
- 具体的な問題の審査を基にした具体的なデータを示して頂きたかった。つまり現行のテストのあり様が、具体的に　どのように変わろうとしているのかを知りたい。ありがとうございました。(男・60歳～・私立高校・教員・満足度4)
- とてもバランスのよい構成だと思いました。ありがとうございました。(女・40～49歳・私立大学・教員・満足度4)
- センター主催のシンポジウムなのに、新テストに対する否定的な意見が多く聞けたのは意外でした。大変勉強になりました！（男・40～49歳・国公立大・事務系職・満足度4)
- "中教審の答申に、真向から否定する内容で、とても興味深かった。中教審には新しいテストの構築に向けて、実現性を考えてほしい。(女・40～49歳・塾・予備・事務系職・満足度4)
- （基調講演、報告 2、4 は特に）たいへん勉強になりました。CBT サンプルデモも興味深いものでした。会社での業務（実務）で出てくる用語の再確認と、今、行っている仕事へ確信を持つことができました。(男・50～59歳・塾・予備・その他・満足度4)
- 入試を一律に論じることの限界を感じました。日本の若者が学び、元気になる制度、システムとして高大接続を議論・検討する必要があります。それは、「入試」ではないと思います。(男・50～59歳・塾・予備・事務系職・満足度4)
- テストの分をわきまえるべきだという大津先生の発言が印象に残りました。答申案はないものねだりをしていると思う。(男・30～39歳・塾・予備・その他・満足度3)
- 「大学～学力評価テスト」が何を測定しようとしているか？「能力」なのか「学力」なのか？その測定する「問題」事例はイメージできるが、その「問題」をどのように測定するかによって「能力」か「学力」か、あるいはエラー（偶発性）かが決まるような気がします。(男・50～59歳・塾・予備・その他・満足度1)
- ・発表資料を配布してもらえたのは良かったです。　・今回「大学入試」についてなので、省庁関係・公的機関の方の発表もあれば興味深いと思います。　・発表者の方の研究（今回の発表に関係するもの）の紹介（スライド）がある

と尚よかったです。　・パネルディスカッションは有意義でした。（女・40 ～ 49 歳・その他・専門職・満足度 4）

- 大変有益な内容でしたが、もう少し時間があるとよかったかと思います（特にパネルディスカッション）。パネルディスカッションはある程度テーマを限定してもよかったのではないでしょうか。（男・30 ～ 39 歳・民間企業・専門職・満足度 4）
- 新しいテストについて、最新の情報を知ることができ、良い機会であった。（男・40 ～ 49 歳・私立高校・教員・満足度 4）
- 非常にレベルの高い講演者をそろえ、タイムリーなテーマでした。大変勉強になりました。今後とも宜しくお願いします。（男・50 ～ 59 歳・私立大学・満足度 4）
- 最後のパネルディスカッションの時間が短いのが残念であった。CBT の専門性の高い内容のシンポやセミナーを開催して頂けるとありがたいです。（男・50 ～ 59 歳・国公立大・教員・満足度 4）
- 学生の能力を測る観点と入学者選抜の観点では、一度に両方行おうとすると足りない部分や必要ない部分がそれぞれ存在し、一体どこまでの範囲を測り、どこまでを利用するのか議論が必要と思うし、また試行錯誤することになるのではないか。いずれにせよ、私もしっかりと勉強したいと思います。今回はありがとうございました。（男・50 ～ 59 歳・塾・予備・事務系職・満足度 4）
- 高校が生徒を有名大学へ何人入れたか国公立大学へ何人入れたかで評価されるような風土が変わらない限り、なかなか高大接続の現状は変わっていかないのでは？（男・50 ～ 59 歳・国公立高・教員・満足度 3）
- 高校で国語の教員をしております。センター試験の国語、特に現代文の試験内容は他にない質の良いものだと思っております。文章の本質を教えるための適切な教材として位置づけることができ、とても助かっております。教育現場に新鮮なメッセージを送るという意志にも賛同いたします。大津先生の「人間の知性はどうなっているのか」という話題にはとても興味があります。今後入試改革がどう進むか、現場はとても気にしています。改革の意図が正しく伝わるような情報提供、公開をお願い致します。（男・～ 29 歳・私立高校・教員・満足度 4）
- ・要するに、PISA の比較は尺度化の方法次第。こう言っては何ですが、技術的に万全ではない方法論で出た結果で世論が動かされ、それを基に政策が決められてきたということが明らかにされたという意味で、Dr.Yamamoto の講演には意義があったと思います。結局のところ、PISA が「実生活における活用能力」を測定しているという証拠は何もないと思いました。こういった問題を受験テクニック的に攻略しようと思ったら、今の入試問題よりも相当にたやすそうですね。　・素点の利用が問題という前川先生の御指摘は同感です。　・南風原先生の講演されたような考え方を基本的教養として物事を考えられる人材が多数輩出されることが、入試の議論には必要だと思います。

・大津先生の講演から、東京・首都圏を見て日本の教育を論じてはいけないことがよく分かりました。　・村上先生の講演で頭の中のモヤモヤが少し晴れた気がしました。もう一度、内容と論点を整理して考えてみたいと思いました。（男・50 ～ 59 歳・国公立大・教員・満足度 4）

- このような内容のシンポジウムは今後もとても大切ですし、議論されるべき内容だと思います。しかし、もっとはやい時期に開催されるとなおよかったかなと感じました。状況的に。（女・30 ～ 39 歳・国公立大・教員・満足度 4）
- 何故センター試験（及びそれに代わるもの）が必要なのでしょうか。大学をいくつかのレベルに分け、各レベルの試験をしよう、といった議論にならないのでしょうか。PISA のように、各レベルの教育水準といったものを測るものと、受験生の入学選抜を行うものを同じ土俵で論じてはいけないと思います。「どう応用できるか」

を論ずるべきと思います。とても勉強になりました。ありがとうございました。（男・50 ～ 59 歳・国公立大・事務系職・満足度 3）

- ・中教審答申の問題点は理解できるが、新テストは実施することが既に決っている。入試センターが実施母体だと思うが、どのように実施するのか道筋が不明である。結局各大学（国立）は大学独自の学力入試を捨てないのではないか？（男・60 歳～・国公立大・教員・満足度 3）
- 楽しい研究会であったが専門性が高くわからない部分もあった。（男・50 ～ 59 歳・国公立大・教員・満足度 4）
- 大学入試センターで開発した CBT テストのデモ問題を拝見できて興味深かった。（女・～ 29 歳・国公立大・事務系職・満足度 3）
- ―専門的な内容も、非常にわかりやすく具体的にお話ししていただき、学ぶことが多いでした。ありがとうございました。―大学入試改革については、私立大学も含めて、方向性が見えてくればいいのに・・・と感じております。（女・50 ～ 59 歳・私立大学・教員・満足度 4）
- ・CBT による運用面。特にコストと地域による格差問題　・大学入試の日本的風土として予備校（塾）や対策問題集でまかなえないテストが可能でしょうか（男・50 ～ 59 歳・その他・専門職・満足度 3）
- ・現実を踏まえた改善が必要だと思う。　・大学入試だけでなく、社会の文化・しくみを考えないと、実現しないかもしれない。（男・50 ～ 59 歳・国公立大・事務系職・満足度 3）
- ・各報告の発表予定時間は事前に周知してほしい。専門的用語が多用されて、理解に苦しむ場面があった。概ね有益な内容だったと感じましたが、チラシ（お知らせ）などに、報告のタイトルだけでなく、内容の概略などもあらかじめ記載しておいて頂けると、参加者も事前に下調べをしたり、予備知識を付けていったりする対応がとれ、より理解することが出来たと思いました。（女・国公立大・事務系職・満足度 2）
- センター側が出題についての悩みをできる限り正直に伝えていただけた印象を持った。教育の現場（私自身は国語なんですが）にいる者としては、生徒達の記述表現力を、ひいては、そもそも表現力を涵養できるテストを実現していただきたい、と切に願います。（男・40 ～ 49 歳・国公立高・教員・満足度 3）
- 「学力の経年変化」について、今後のとらえ方を考えてみたいと感じました。複数回、複数年度、問題非公開、尺度得点の実現により、これが可能になるのかどうか。ただ、「学力」概念そのものも、着目の仕方が変わるものだと思います。IRT、学力の経年変化を追う、学力観の変遷について、組み合わせた議論を深めて頂きたいです。（男・30 ～ 39 歳・民間企業・専門職・満足度 4）
- これだけのメンバーが集まったところを拝見できるのはすごいと思いました。テストの妥当性、これが一番重要だと思います。（男・40 ～ 49 歳・民間企業・その他・満足度 4）
- 現状の入試が抱えている問題を測定論という観点と実際の種々の試験と比較して考えることができました。公平性を確保することの難しさと、入試センターと各大学の役割分担を明確にすることが新しい学力テストに求められる要件ではないかと感じました。（男・30 ～ 39 歳・塾・予備・専門職・満足度 4）
- タイトルは「大学入試の日本的風土は変えられるか」というものであったが、結局そのタイトルとは焦点がずれてしまい、個別の事象の解説に終始したのは残念であった。一方、入試センターの業務の大変さなどがよく伝わり、改めて敬意を表する気持ちになった。（男・50 ～ 59 歳・私立大学・教員・満足度 3）
- IRT、CBT などを含めて、それと日本的風土を重ね合せたとき、必ずしもうまくいかない例が多く出され、それをどう克服し、うまく統合していくか、考えるとむつかしい問題で、これからテスト研究者がどう解決していったらよいか、考えさせられる問題が多かった。（男・60 歳～・民間企業・専門職・満足度 4）

参 考 資 料

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）

平成２６年１２月２２日

中央教育審議会

目　次

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　～

はじめに　－　高大接続改革が目指す未来の姿

本答申は、教育改革における最大の課題でありながら実現が困難であった「高大接続」改革を、初めて現実のものにするための方策として、高等学校教育、大学教育及びそれらを接続する大学入学者選抜の抜本的な改革を提言するものである。

将来に向かって夢を描き、その実現に向けて努力している少年少女一人ひとりが、自信に溢(あふ)れた、実り多い、幸福な人生を送れるようにすること。

これからの時代に社会に出て、国の内外で仕事をし、人生を築いていく、今の子供たちやこれから生まれてくる子供たちが、十分な知識と技能を身に付け、十分な思考力・判断力・表現力を磨き、主体性を持って多様な人々と協働することを通して、喜びと糧を得ていくことができるようにすること。

彼らが、国家と社会の形成者として十分な素養と行動規範を持てるようにすること。

我が国は今後、未来を見据えたこうした目標が達成されるよう、教育改革に最大限の力を尽くさなければならない。

生産年齢人口の急減、労働生産性の低迷、グローバル化・多極化の荒波に挟まれた厳しい時代を迎えている我が国においても、世の中の流れは大人が予想するよりもはるかに早く、将来は職業の在り方も様変わりしている可能性が高い[1]。そうした変化の中で、これまでと同じ教育を続けているだけでは、これからの時代に通用する力を子供たちに育むことはできない。

この厳しい時代を乗り越え、子供や孫の世代に至る国民と我が国が、希望に満ちた未来を歩めるようにするため、国は、新たな時代を見据えた教育改革を「待ったなし」で進めなければならない。

---

[1] キャシー・デビッドソン氏（ニューヨーク市立大学大学院センター教授）の予測によれば、「2011 年にアメリカの小学校に入学した子供たちの 65％は、大学卒業後、今は存在していない職業に就く」とされている。

１．我が国の未来を見据えた高大接続改革

（１）今後の教育改革が目指すべき方向性と現状の課題

（初等中等教育から高等教育まで一貫した、これからの時代に求められる力の育成）

新たな時代を見据えた教育改革を進めるに当たり重要なことは、子供たち一人ひとりに、それぞれの夢や目標の実現に向けて、自らの人生を切り拓き、他者と助け合いながら、幸せな暮らしを営んでいける力を育むための、初等中等教育から高等教育までを通じた教育の在り方を示すことである。

子供たちに育むべきこのような力を言い換えるならば、それは「豊かな人間性」「健康・体力」「確かな学力」を総合した力である「生きる力」にほかならない。

このうち「学力」については、戦後からの長い間、「自分で考え自分で実行する」型の教育と、体系的な知識を注入する型の教育との間で議論が繰り広げられてきた。過去の学習指導要領の改訂に際しても、「ゆとり」か「詰め込み」かのような二項対立的な議論がなされてきた。

こうした二項対立を乗り越え、平成 19 年の学校教育法改正により、「基礎的な知識及び技能」「これらを活用して課題を解決するために必要な思考力・判断力・表現力等の能力」「主体的に学習に取り組む態度」という、三つの重要な要素（いわゆる「学力の三要素」）から構成される「確かな学力」を育むことが重要であることが明確に示されたところである。

こうした「確かな学力」の育成を目指し、特に小･中学校においては、学力の三要素を踏まえた指導の充実が図られるよう、多くの関係者による実践が重ねられてきた。全国学力･学習状況調査において、主として「知識」に関する問題[2]だけではなく、主として「活用」に関する問題[3]も出題されていることなどが、関係者の意識改革や各学校における授業改善に大きな影響を与えている。また、現行の学習指導要領に基づく、学級やグループで話し合う活動や、調べたことや考えたことを発表し合う活動等を重視する「言語活動」、各教科や総合的な学習の時間等における探究的な学習といった、学力の三要素に対応した学習方法についても、評価の在り方と併せて実践が重ねられ充実が図られており、国内外の学力調査の結果[4]にも、そうした実践の成果が表れてきていると見ることができる。

高等学校教育及び大学教育においては、そうした義務教育までの成果を確実につなぎ、それぞれの学校段階において「生きる力」「確かな学力」を確実に育み、初等中等教育から高等教育まで一貫した形で、一人ひとりに育まれた力を更に発展・向上させることが

---

[2] 身に付けておかなければ後の学年等の学習内容に影響を及ぼす内容や、実生活において不可欠であり常に活用できるようになっていることが望ましい知識・技能などを中心とした出題。

[3] 知識・技能等を実生活の様々な場面に活用する力や、様々な課題解決のための構想を立て実践し評価・改善する力などに関わる内容を中心とした出題。

[4] OECD 生徒の学習到達度調査（PISA）、全国学力・学習状況調査等

肝要である。

（高等学校教育、大学教育、大学入学者選抜における課題）

高等学校については、現行学習指導要領において、知識・技能の習得に加えて、思考力・判断力・表現力等の能力や、主体的に学習に取り組む態度の育成を目指しており、その実現を目指した関係者による努力が重ねられている。大学教育についても、中央教育審議会答申等において、初等中等教育段階における「生きる力」の育成を踏まえ、「学士力」をはじめとする育成すべき力の在り方や、その育成のための大学教育の質的転換について提言されてきており、学生が主体性を持って多様な人々と協力して問題を発見し解を見いだしていく能動的学修（以下「アクティブ・ラーニング」という。）の充実などに向けた教育改善が図られつつある。

しかしながら、我が国が成熟社会を迎え、知識量のみを問う「従来型の学力」や、主体的な思考力を伴わない協調性はますます通用性に乏しくなる中、現状の高等学校教育、大学教育、大学入学者選抜は、知識の暗記・再生に偏りがちで、思考力・判断力・表現力や、主体性を持って多様な人々と協働する態度など、真（しん）の「学力」が十分に育成・評価されていない。

また、特定の分野に強い関心をもち、その向上に夢を賭けて卓越した力を磨いている高校生や、「世界にトビタテ！」[5]の精神でグローバルな課題に積極的に向き合う活力のある高校生、身近な地域の課題に徹底的に向き合い考え抜いて行動する高校生などが評価されずに切り捨てられがちである。

こうした状況では、それぞれの夢を育み、その中で自らを鍛えるとともに、秘められた才能などを伸ばすことはできず、未来のエジソンやアインシュタインとなる道や、世界を舞台に活躍する潜在力、地方創生の鍵となる問題の発見や解決を生み出す可能性の芽なども摘（つ）まれてしまう。

高大接続を実現するための方策は、「はじめに」に述べた未来の姿を実現するための一環とみなされるべきものである。高等学校、大学ともに進学率が高まり、多様な進路が開かれる中で、一人ひとりの生徒・学生に必要な力を身に付けるためには、上記のような教育改善の更に先にある、新たな時代に対応するための教育の在り方や高大接続の在り方を見いだすことが不可欠である。

そうした観点から高等学校教育と大学教育の現状を振り返ると、現行の大学入学者選抜の大きな影響下で、それぞれ下記のような課題を抱えている。

選抜性の高い大学へ生徒が進学する高等学校においては、国内外で活躍する次世代リーダーの育成に向けて、スーパーグローバルハイスクール、スーパーサイエンスハイスクールなどの取組や、国際通用性を高める観点からの国際バカロレアのプログラム導入、

[5] 海外での異文化体験や実践を焦点にした留学を推奨し、学生時代により多様な経験と自ら考え行動できるような体験の機会を提供することを目指し、官民共同による留学支援制度「トビタテ！留学 JAPAN 日本代表プログラム」などの取組が展開されている。

「総合的な学習の時間」を活用した課題探究の鍛錬、ユネスコスクール等における持続可能な開発のための教育の実践など、これからの時代に必要な力の育成を見据えた積極的な取組も多く見られる。その一方で、学校の教育方針が選抜性の高い大学への入学者数を競うことに偏っている場合には、高等学校教育が、受験のための教育や学校内に閉じられた同質性の高い教育に終始することになり、多様な個性の伸長や幅広い視野の獲得といった、多様性の観点からは不十分なものとなりがちである。こうした教育では、大学入試に必要な知識・技能やそれらを与えられた課題に当てはめて活用する力は向上させられたとしても、自ら課題を発見し解決するために必要な思考力･判断力･表現力等の能力や、主体性を持って、多様な人々と協働しながら学んだ経験を生徒に持たせることはほとんどできない。

そうした生徒がそのまま選抜性の高い大学に入学した場合、一定の知的な能力を持っていたとしても、主体性を持って他者を説得し、多様な人々と協働して新しいことをゼロから立ち上げることのできる、社会の現場を先導するイノベーションの力を、大学において身に付けることは難しい。

「従来型の学力」について中間層の生徒が多い高等学校では、知識量の多寡で進学先の難易度が決定される環境において、受験勉強が学習への動機付けになってきた。しかしながら、少子化の進展等により大学への入学が一般的に容易になっているため、それに対応して、従来のような受験勉強がそれほど必要でなくなっている。そうした中では、今まで以上に、社会で自立して生きていくために必要な力の獲得を目標として設定し、学習意欲を喚起する必要があるが、そうした動機付けを十分に行わず、自主的にはほとんど学習せず目標を持てない生徒を多数、選抜性が中程度の大学に送り出してしまっている例も多い。そうした場合、一人ひとりの知識・技能や思考力･判断力･表現力等の能力を伸ばす余地はあるにもかかわらず、学生に主体性や学修のための明確な目標が不足しているため、大学においてもそれができないままになっている。

「従来型の学力」の習得に困難を抱えている生徒が多い高等学校では、家庭環境や所得格差等の問題も背景として、必要な力を育む以前に、まずは通学させ卒業させることで手一杯であるという状況も多い。そうした中で、生活指導や教育相談、将来を見通した進路指導等の支援を熱心に行っている高等学校もあるが、入学者選抜が機能しなくなっている大学に漫然と送り出される場合も少なくなく、そうした大学においては、思考力･判断力･表現力等の能力どころか、その基礎となる知識・技能自体の質と量が、大学教育に求められる水準に比して不十分な段階にある学生が多いことが深刻な問題となっている。

こうした現状から課題として浮かび上がってくることは、高等学校においては、小・中学校に比べ知識伝達型の授業に留(とど)まる傾向があり、学力の三要素を踏まえた指導が浸透していないことである。ここには、一般入試においては、一斉かつ画一的な条件で実施される試験で、あらかじめ設定された正答に関する知識の再生を一点刻みに問い、その結果の点数で選抜する評価から転換し切れていないこと、またＡＯ入試、推薦入試の多くが本来の趣旨･目的に沿ったものとなっておらず、単なる入学者数確保の手段となっ

てしまっていることなど、現行の多くの大学入学者選抜における学力評価が、学力の三要素に対応したものとなっていないことが大きく影響していると考えられる。

また、高等学校の進学率が９８％に達する中で、高校生の進路が多様化し、教育課程や授業内容の在り方も多岐にわたり、高等学校教育として生徒に共通に身に付ける学力が確保されていないことも大きな課題となっている。

大学教育については、我が国の大学生の学修時間は米国と比べて依然として短く[6]、特に社会科学系において学修時間が短い傾向が顕著である[7]。授業の形態についても、一方的な知識の伝達・注入のみに留まるものが多く見受けられる。こうした現状について、大学教育において学生にどれだけの付加価値を付けて社会に送り出せているかという観点からは、依然として社会からの厳しい評価があり、国民、とりわけ学生や経済界は、大学教育の現状に満足しているとは言い難い[8]。さらに、大学教育の場が、多様な学生が切磋琢磨する環境となっておらず、また、自分が将来社会で活動することと大学で受ける教育がどのように関係しているのか、明確でないことが多い。その結果、主体性を磨くことなく、自ら目標を持ってそれを実現していく力を身に付けないまま、社会に出る学生も多い。

大学において育成すべき力とは何かを明らかにした上で、大学入学者選抜や高等学校教育との連携の在り方を変えていかなければ、大学入学のその先を見据えた、自らの人生を切り拓くための目標を高校生に持たせることも難しい。

また、大学入学者選抜については、前述のように、知識の記憶力などの測定しやすい一部の能力や、選抜の一時点で有している能力の評価に留まっていたり、丁寧な評価よりも学生確保が優先されるなど、高等学校教育で培ってきた力や、これからの大学教育で学ぶために必要な力を評価するものとなっていない。そうした背景には、年齢、性別、国籍、文化、障害の有無、地域の違い、家庭環境等の多様な背景を持つ高校生一人ひとりが、高等学校までに積み上げてきた多様な経験や能力を度外視し、１８歳頃における一度限りの一斉受験という画一化された条件において、知識の再生を一点刻みで問う問題を用いた試験の点数による客観性の確保を過度に重視し、そうした点数のみに依拠した選抜を行うことが「公平」であるという、従来型の「公平性」の観念が社会に根付いていることがあると考えられる。

---

[6] 1 週間当たりの学修時間が 11 時間以上の学生が我が国は約 15%、米国の学生は約 59%（東京大学 大学経営・政策研究センター「全国大学生調査」（平成 19 年）、NSSE（National Survey of Student Engagement））。

[7] 社会科学系においては、1 週間の授業に関する学修時間は、0 時間の者が約 2 割（東京大学 大学経営・政策研究センター「全国大学生調査」（平成 19 年））。

[8] ある新聞社の世論調査では、日本の大学が世界に通用する人材や社会、企業が求める人材を育てているかとの質問に、6 割を超える国民が否定的な回答をしている。また、経済団体の調査によれば、企業の大学教育へのニーズと大学が教育面で特に注力している点に認識の差異や隔たりがある。さらに、大学生の 5～6 割が「論理的に文章を書く力」や「人に分かりやすく話す力」について大学の授業の有効性を否定的に捉えているという調査結果もある。

（２）高等学校教育、大学教育を通じて育むべき「生きる力」「確かな学力」の明確化

「生きる力」や「確かな学力」の定義そのものについては、累次の答申等や関係法令において明示されている[9]ところであるが、大学におけるその在り方[10]を含め、学校段階に応じた具体的な在り方については、初等教育から高等教育を貫く視点に立って、今一度捉え直してみる必要がある。

とりわけ、高等学校や大学の段階に進むに従い、身に付けるべき力の在り方は小･中学校段階とは質的に変化していくものであり、特に、卒業後どのような進路を選ぶにしても、国家及び社会の形成者として自立して生きるための力を育成するため、社会とのより密接な関係を意識した学習が求められるようになる。このような観点も踏まえつつ、高等教育までを通じて育成すべき「生きる力」「確かな学力」の意義を明確にした上で、幼児教育、小・中学校で積み上げられてきた教育の成果を、高等学校、大学における教育で確実に発展させていくことが必要である。

こうしたことを踏まえ、高等学校教育、大学教育を通じて育むべき「生きる力」を、それを構成する「豊かな人間性」「健康･体力」「確かな学力」それぞれについて捉え直すと、以下のように考えることができる。

①　豊かな人間性

高等学校教育を通じて、国家及び社会の責任ある形成者として必要な教養と行動規範を身に付けること。大学においては、それを更に発展・向上させるとともに、国、地域社会、国際社会等においてそれぞれの立場で主体的に活動する力を鍛錬すること。

②　健康･体力

高等学校教育を通じて、社会で自立して活動するために必要な健康･体力を養うとともに、自己管理等の方法を身に付けること。大学においては、それを更に発展・向上させるとともに、社会的役割を果たすために必要な肉体的、精神的能力を鍛錬すること。

③　確かな学力

学力の三要素を、社会で自立して活動していくために必要な力という観点から捉え直し、高等学校教育を通じて(ⅰ)これからの時代に社会で生きていくために必要な、「主体性を持って多様な人々と協働して学ぶ態度（主体性･多様性･協働性）」を養うこと、(ⅱ)その基盤となる「知識･技能を活用して、自ら課題を発見しその解決に向けて探究し、成果等を表現するために必要な思考力･判断力･表現力等の能力」を育むこと、(ⅲ)さらにその基礎となる「知識･技能」を習得させること。大学においては、それを更に発展･向

---

9 平成８年中央教育審議会答申「21 世紀を展望した我が国の教育の在り方について（第一次答申）」など。

10 平成 20 年 12 月 24 日中央教育審議会答申（「学士課程教育の構築に向けて」）では、各専攻分野を通じて培う「学士力」として学士課程共通の学習成果に関する参考指針を提示している。また、平成 24 年８月 28 日中央教育審議会答申（「新たな未来を築くための大学教育の質的転換に向けて」）では、「生涯学び続け、主体的に考える力」の育成を提言している。

上させるとともに、これらを総合した学力を鍛錬すること。

本答申における「学力」とは、上記の三要素から構成される「確かな学力」のことを指す。なお、特に「多様性」については、生徒、学生に、多様性を受容し尊重する力を育んでいく必要があるが、そのためには、高等学校や大学の側(がわ)において、多様な生徒、学生が多様な環境の中でともに学ぶことのできる場を用意する必要がある。

高等学校、大学それぞれの段階において育むべき「生きる力」「確かな学力」が確実に育成されるようにするとともに、両者をつなぐものとして双方に極(きわ)めて大きな影響を与える大学入学者選抜の段階において、これらの力を念頭に置いた評価が行われることが必要である。また、こうした教育目標を生徒・学生自身に自覚させ、学習への動機付けを行い、意欲を喚起することも必要である。

また、グローバル化の進展の中で、言語や文化が異なる人々と主体的に協働していくためには、国際共通語である英語の能力を、真(しん)に使える形で身に付けることが必要であり、単に受け身で「読むこと」「聞くこと」ができるというだけではなく、積極的に英語の技能を活用し、主体的に考えを表現することができるよう、「書くこと」「話すこと」も含めた四技能を総合的に育成・評価することが重要である。
また、英語のみならず、我が国の伝統文化に関する深い理解、異文化への理解や躊躇(ちゅうちょ)せず交流する態度などが求められることにも留意が必要である。

なお、小・中学校において学力の三要素を踏まえた教育が定着してきている背景には、全国学力・学習状況調査など、知識・技能等を実生活の様々な場面に活用することや、様々な課題解決のための構想を立て実践し評価・改善することなどを含めた学力を評価する手法と、「言語活動」といった思考力・判断力・表現力等の能力や学習意欲を育むための学習・指導方法の具体的な在り方が明確化され、各学校に導入されたことがある[11]。高大接続における改革の方向性も、改革のための具体策との組み合わせによって示していくことが重要である。

（３）高大接続改革の意義

こうした育むべき力についての考え方を踏まえつつ、上記（１）に示した現状を、高等学校教育、大学教育、大学入学者選抜の改革による新しい仕組みによって克服し、少年少女一人ひとりが、高等学校教育を通じて様々な夢や目標を芽吹かせ、その実現に向けて努力した積み重ねを、大学入学者選抜においてしっかりと受け止めて評価し、大学教育や社会生活を通じて花開かせるようにする必要がある。
特に、１８歳頃における一度限りの一斉受験という特殊な行事が、長い人生航路における最大の分岐点であり目標であるとする、我が国の社会全体に深く根を張った従来型

11 学習活動そのものを直接評価する「パフォーマンス評価」など、複雑な学びを筆記以外の方法で評価する方法の開発も、こうした学力の三要素を踏まえた教育の定着に大きく貢献している。

の「大学入試」や、その背景にある、画一的な一斉試験で正答に関する知識の再生を一点刻みに問い、その結果の点数のみに依拠した選抜を行うことが公平であるとする、「公平性」の観念という桎梏（しっこく）は断ち切らなければならない。大学入学者選抜は、一時点の学力検査によってその後の人生を決定させるためのものではない。先を見通すことの難しい時代において、生涯を通じて不断に学び、考え、予想外の事態を乗り越えながら、自らの人生を切り拓（ひら）き、より良い社会づくりに貢献していくことのできる人間を育てることが高等学校教育及び大学教育の使命であり、これからの大学入学者選抜は、若者の学びを支援する観点に立って、それぞれが夢や目標を持ち、その実現に必要な能力を身に付けることができるよう、高等学校教育と大学教育とを円滑に結び付けていく観点から実施される必要がある。

そのためには、既存の「大学入試」と「公平性」に関するする意識を改革し、年齢、性別、国籍、文化、障害の有無、地域の違い、家庭環境等の多様な背景を持つ一人ひとりが、高等学校までに積み上げてきた多様な力を、多様な方法で「公正」に評価し選抜するという意識に立たなければならない。

現在ほぼ横ばいで推移している我が国の１８歳人口が、平成３３年頃からは減少に転じると予想される中、我が国社会の持続的な発展を実現していくためには、高大接続の改善が不可欠であり、もはや一刻の猶予もない。本答申においては、上記のような考え方に基づく改革の方向性を、改革実現のための具体的な方策とともに示している。国や高等学校、大学等の関係者、関係機関のみならず、社会全体で高等学校教育、大学教育、そしてそれを接続する大学入学者選抜の一体的な改革に向けた気運が醸成され、具体的な取組が強力に推進されることを期待する。

なお、本年７月には文部科学大臣から、小中一貫教育の制度化など今後の学制の在り方について、及び教員の資質能力と学校組織全体の総合力の向上について、中央教育審議会に諮問が行われており[12]、また、本年１１月には、初等中等教育における教育課程の基準等の在り方について、諮問が行われたところである[13]。高大接続特別部会における審議の内容は、これらの検討事項にも深く関連するものであることから、それぞれの検討の過程において、本答申の提言を十分に踏まえた議論が行われるよう期待するとともに、国においてはこれらの議論の成果を一体的に推進し、教育改革全体の将来像の中で、新しい時代にふさわしい教育への転換が図られるよう求めるものである。

---

[12] 平成 26 年 7 月 29 日に文部科学大臣から「子供の発達や学習者の意欲・能力等に応じた柔軟かつ効果的な教育システムの構築について」及び「これからの学校教育を担う教職員やチームとしての学校の在り方について」諮問が行われ、前者については、平成 26 年 12 月 22 日に答申が行われた。

[13] 平成 26 年 11 月 20 日に文部科学大臣から「初等中等教育における教育課程の基準等の在り方について」諮問が行われた。

（４）高大接続改革を推進するに当たって留意すべき点

高大接続改革をめぐっては、関係者の間にいくつかの誤解があり、それが改革を妨げる一つの要因ともなってきた。そうした誤解を生んでいる点について、改めて留意点として記しておくこととする。

「高大接続」とは、高校生の全てを大学教育に接続するということではない。高校を卒業して就職する生徒、専修学校等に入学する生徒、その他の進路を歩む生徒たちの人生も、大学進学と同様にそれぞれ花開くべきものである。高大接続を議論する際には、高等学校卒業生の多様な進路を踏まえ、国家及び社会の責任ある形成者として、自立して生きる力を高等学校教育において確実に育むという視点が重要である。

あわせて、高等学校卒業後、生徒がどのような進路を選択するにせよ、経済的な理由のみによりそれが左右されることのないような配慮も必要である。

また、「高大接続」の改革は、「大学入試」のみの改革ではない。その目標は、「大学入試」の改革を一部に含むものではあるが、高等学校教育と大学教育において、十分な知識・技能、十分な思考力・判断力・表現力、及び主体性を持って多様な人々と協働する力の育成を最大限に行う場と方法の実現をもたらすことにある。

「高大接続」改革は、知識・技能の習得を無視する改革ではないという点も重要である。「知識・技能」、「思考力・判断力・表現力」、「主体性･多様性･協働性」のすべてを十分に向上させることを目指すものであり、改革によって高校生、大学生が身に付けられるようになる力は、十分な水準の知識・技能はもちろんのこと、自分で目標を持って他者と協力しながら新しいことを成し遂げていく力までも含むものである。

「高大接続」は、新しい時代にふさわしい高等学校教育と大学教育を、それぞれの目標の下に改革し、子供たちがそれぞれの段階で必要な力を確実に身に付け、次の段階へ進むことができるようにするためのものである。

２．新しい時代にふさわしい高大接続の実現に向けた改革の方向性

高大接続改革を実現するためには、高等学校教育及び大学教育を、上記１．（２）に示したような力を育成するにふさわしい教育内容、学習･指導方法、評価方法、教育環境へと大きく転換させなければならない。

また、こうした改革のための現実的問題として大きく立ちふさがるのが、大学入学者選抜の在り方である。現在直面する最大の課題は、高等学校教育と大学教育とを接続する重要な役割を果たすべき大学入学者選抜において、上記のような育成すべき力の在り方を踏まえた評価がなされていないことである。

接続段階での評価の在り方が変われば、それを梃子（てこ）の一つとして、高等学校教育及び大学教育の在り方も大きく転換すると考えられる。高等学校教育改革、大学教育改革の実効性を高めるためにも、大学入学者選抜の改革に社会全体で取り組む必要がある。

このような観点から、以下の改革に一体的に取り組む。

◆　高等学校教育については、生徒が、国家と社会の形成者となるための教養と行動規範を身に付けるとともに、自分の夢や目標を持って主体的に学ぶことのできる環境を整備する。そのために、高大接続改革と歩調を合わせて学習指導要領を抜本的に見直し、育成すべき資質･能力の観点からその構造、目標や内容を見直すとともに、課題の発見と解決に向けた主体的･協働的な学習･指導方法であるアクティブ・ラーニングへの飛躍的充実を図る。

また、教育の質の確保・向上を図り、生徒の学習改善に役立てるため、新テスト「高等学校基礎学力テスト（仮称）」を導入する。

◆　大学教育については、学生が、高等学校教育までに培った力を更に発展･向上させるため、個々の授業科目等を越えた大学教育全体としてのカリキュラム･マネジメントを確立する（ナンバリングの導入等）とともに、主体性を持って多様な人々と協力して学ぶことのできるアクティブ・ラーニングへと質的に転換する。

◆　大学入学者選抜においては、現行の大学入試センター試験を廃止し、大学で学ぶための力のうち、特に「思考力・判断力・表現力」を中心に評価する新テスト「大学入学希望者学力評価テスト（仮称）」を導入し、各大学の活用を推進する。

◆　各大学が個別に行う入学者選抜（以下「個別選抜」という。）については、学力の三要素を踏まえた多面的な選抜方法をとる[14]ものとし、特定分野において卓越した能力を有する者の選抜や、年齢、性別、国籍、文化、障害の有無、地域の違い、家庭環境等にかかわらず多様な背景を持った学生の受け入れが促進されるよう、具体的な選抜方法等に関する事項を、各大学がその特色等に応じたアドミッション・ポリシーにおいて明確化する。このために、アドミッション・ポリシー等の策定を法令上位置付

[14] 選抜性の高低に則し改革すべき点については、別添資料２のイメージ図の通り。

けるとともに、大学入学者選抜実施要項を見直す。

◆　さらに、各大学が、新たな大学入学者選抜実施要項に基づく新たなルールに則(のっと)って改革を進めることができるよう、大学にとって改革のインセンティブとなるような財政措置等の支援を行う。

（１）各大学のアドミッション・ポリシーに基づく、大学入学希望者の多様性を踏まえた「公正」な選抜の観点に立った大学入学者選抜の確立

大学入学者選抜の改革を進めるに当たっては、「大学入試センター試験」の抜本的改革が必要であるが、それは全体の改革の一部にすぎない。

何よりも重要なことは、個別選抜を、画一的な一斉試験で正答に関する知識の再生を問う評価に偏ったものとしたり、入学者の数の確保のための手段に陥らせたりすることなく、「人が人を選ぶ」個別選抜を確立していくことである。「人が人を選ぶ」個別選抜の確立とは、高等学校教育で身に付けた「生きる力」「確かな学力」をいかに大学教育で発展･向上させ、社会へと送り出していくかという観点から、大学の入り口段階で求められる力を多面的･総合的に評価するという、個別選抜本来の役割が果たせるものにすることである。

また、そうした評価に転換するためには、大学入学者選抜を含むあらゆる評価において、画一的な一斉試験で正答に関する知識の再生を問い、その結果の点数だけを評価対象とすることが公平であると捉える、既存の「公平性」についての社会的意識を変革し、それぞれの学びを支援する観点から、多様な背景を持つ一人ひとりが積み上げてきた多様な力を、多様な方法で「公正」に評価するという理念に基づく新たな評価を確立していくことが不可欠である。

その際、画一的な一斉試験による大学入学者選抜だけを取り上げて「公平性」を論ずるのではなく、一人ひとりの人間の生涯を通して見た時に、多様な背景を持った学習者一人ひとりの能力が最大限に磨かれるように教育の機会が均等に与えられるという意味での「公正性」を確立していくべきであり、その一部として大学入学者選抜における「公正性」を理解すべきと考えられる。

①　各大学の個別選抜改革

（アドミッション･ポリシーに基づく個別選抜の確立）

各大学は、求める学生像のみならず、各大学の入学者選抜の設計図として必要な事項をアドミッション・ポリシーにおいて明確化することが必要であり、高等学校及び大学において育成すべき「生きる力」「確かな学力」の本質を踏まえつつ、入学者に求める能力は何か、また、それをどのような基準･方法によって評価するのかを、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。

現行法令上、アドミッション・ポリシーの策定が明確に規定されていない点も課題で

あり、法令上の位置付けを検討する必要がある。

アドミッション・ポリシーの策定に当たっては、各大学の強み、特色や社会的役割を踏まえつつ、大学教育を通じてどのような力を発展・向上させるのかを明らかにした上で、個別選抜において、様々な能力や得意分野、異なる背景を持った多様な生徒が、高等学校までに培ってきたどのような力を、どのように評価するのかを明示する必要がある。

また、「確かな学力」として求められる三要素を総合的に評価する視点を担保するため、どのような評価方法を活用するのか、学力の三要素全てを評価の対象としつつ、特にどういった要素に比重を置くのかを、大学入学希望者に対して明確に示していくことが求められる。

具体的な評価方法としては、下記②に示す「大学入学希望者学力評価テスト（仮称）」の成績に加え、小論文、面接、集団討論、プレゼンテーション、調査書[15]、活動報告書、大学入学希望理由書や学修計画書、資格・検定試験などの成績、各種大会等での活動や顕彰の記録、その他受検者のこれまでの努力を証明する資料などを活用することが考えられる。「確かな学力」として求められる力を的確に把握するためには、こうした多元的な評価尺度が必要である。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。

その際、英語については、高等学校教育において育成された「聞くこと」「話すこと」「読むこと」「書くこと」四技能を、大学における英語教育に引き継いで確実に伸ばしていくことができるよう、アドミッション・ポリシーにおいても四技能を総合的に評価することを示すこととし、「大学入学希望者学力評価テスト（仮称）」における英語の扱いも踏まえつつ、四技能を測定する資格・検定試験の更なる活用を促進すべきである[16]。

具体的な評価の在り方について、特に、スーパーグローバル大学等をはじめとする、国内外で活躍できる次世代リーダー等の育成を目指す大学においては、リーダーとして活動するために必要な力とは何かを明確に示し、大学の使命としてその育成を目指すとともに、多様な学生が切磋琢磨（せっさたくま）する環境作りが不可欠である。特にこうした大学を含め、選抜性の高い大学の学生については、これまでのように知識・技能やそれらを与えられた課題に当てはめて活用する力に優れていることは必要ではあるが、それらだけではまったく不十分であり、「主体性・多様性・協働性」や「思考力・判断力・表現力」を含む「確かな学力」を、高い水準で評価する個別選抜を推進することによって、年齢、性別、国籍、文化、障害の有無、地域の違い、家庭環境等にかかわらず、多様な背景を持った学

15 調査書には、(2) に示す通り「高等学校基礎学力テスト（仮称）」の結果が記入されるが、同テストについては、あくまで高等学校段階における学習成果を把握するための参考資料の一部として用いることに留意。

16 「英語教育の在り方に関する有識者会議」報告書も参照のこと。

生の確保に努める必要がある。

また、選抜性が中程度の大学における大学入学者選抜の現状を見てみると、個別選抜で二科目前後の特定科目を課す形態が多いが、大学独自の作問が負担となっていることの影響などから、知識量のみを問う問題となっていることが多い。今後は、「大学入学希望者学力評価テスト（仮称）」を積極的に活用しつつ、思考力・判断力・表現力等を含む「確かな学力」を総合的に評価する個別選抜へと転換する。

ＡＯ・推薦入試が本来の趣旨･目的に沿ったものとなっていないなど、入学者選抜が機能しなくなっている大学においては、下記（２）に示す「高等学校基礎学力テスト（仮称）」[17]の結果を含めた高等学校の学習成果を、調査書の活用等により確実に把握することや、活動報告書の提出や面接の実施等により、大学教育に求められる水準の学力を担保する。

なお、個別選抜全体の中では、アドミッション・ポリシーを踏まえて、多面的･総合的な能力を有する者のみならず、科学や芸術などの特定の分野において卓越した能力を持つ者が、適切に評価される仕組みも重要である。各大学の教育方針に応じて、そうした才能が適切に評価されるよう、アドミッション・ポリシーにおいて、科学オリンピックや各種大会等での活動や顕彰の記録をはじめとした高等学校段階までの様々な活動履歴等も含めて評価することを明確にした上で、大学教育での更なる成長につなげられるような個別選抜の在り方が確保されるべきである。そうした観点から、特に優れた資質を持つ高校生に、大学において高度な指導を受けてさらなる挑戦をする機会が与えられるよう、大学への飛び入学制度について、高等学校の卒業程度認定制度の創設を含め、さらなる活用が図られるべきである。

専門高校についても、主体的に自分の目標を持って専門性を育み、専門科目について高い知識･技能を獲得している生徒が、広範囲の教科･科目の知識が求められる選抜性の高い大学に進学できない場合もある。教育の場に多様性をもたらすためにも、こうした生徒に対応した個別選抜が、高等学校の進路指導や大学入学後の教育課程の多様性の尊重に向けた質的な転換とともに実施されるべきである。

また、上記のような改革の方向性と、「生きる力」「確かな学力」の本質を踏まえた上で、各大学のアドミッション･ポリシーに基づき、下記②に示す新テストに加え、思考力・判断力・表現力を評価するため、自分の考えに基づき論を立てて記述する形式の学力評価を個別に課すこともあってよい。

（多元的な評価に向けた意識改革と、新たな評価手法の蓄積･共有）

個別選抜における評価に当たっては、画一的な一斉試験で正答に関する知識の再生を問い、その結果の点数のみに依拠した選抜を行う従来型の「公平性」「客観性」と、多数

17 「高等学校基礎学力テスト（仮称）」は、入学者選抜への活用を本来の目的とするものではなく、進学時への活用は、調査書にその結果を記入するなど、あくまで高等学校の学習成果を把握するための参考資料の一部として用いることに留意。

の受験生に対して短時間で合否判定を行うための効率性を重視するあまり、面接、集団討論、小論文、調査書、その他による多元的な評価を重視しない傾向がある。この点に関しては、客観性とは何かについての意識改革[18]と併せて、個別選抜を行う側(がわ)が、自らの都合のみにより選抜する方法ではなく、一人ひとりの入学希望者が行ってきた多様な努力を受け止めつつ、入学者に求められる能力を「公正」に評価し選抜する方法へと意識を転換し、アドミッション・ポリシーに示した基準・方法に基づく多元的な評価の妥当性・信頼性を高め、説明責任を果たしていく必要がある。

こうした多元的な評価に対応した具体的な手法としては、主として複雑な課題に知識・技能を活用して探究し表現することを求める「パフォーマンス評価」、そうした複雑な課題の達成度を数段階に分け、達成度を判断する基準を示す「ルーブリック」、様々な学習過程や成果の記録等を蓄積して学習状況を把握する「ポートフォリオ評価」等が着実に開発されているところである。今後、高等学校教育及び大学教育におけるそうした評価の導入を積極的に推進するとともに、初等中等教育関係者と大学関係者とが協力して具体例を蓄積し共有し、新たな手法も研究・開発していく必要がある。さらに、入学後の学生の成績や活動実績、留年・中退率、卒業後の進路等について追跡調査を行い、評価基準・方法の妥当性を検証していくことも必要である。

こうした評価には事務的な負担が伴い、高い評価能力が要求されることから、国は、評価のノウハウを集約したセンターにおいて、多元的な評価に対応した資料の蓄積・共有、新たな手法の研究・開発を行うとともに、各大学におけるアドミッション・オフィスの強化や、評価の専門的人材の育成、教職員の評価力向上に対する支援を行うことが急務である。

②　入学希望者に求められる学力を評価する新テストの導入

毎年５０万人以上が受験する大規模な試験である現行の大学入試センター試験は、大学入学希望者の基礎的な学習の達成度を判定するという本来的な役割のみならず、高等学校教育における質の保証が課題視される中で、高校生の一定の基礎学力の確保に大きな役割も果たしてきたと評価することができる。

一方で、大学入試センター試験は「知識・技能」を問う問題が中心となっており、これからの大学入学者選抜において評価すべき「確かな学力」の在り方や、下記（２）に示す、高等学校段階の基礎学力を評価する新テストの導入なども踏まえると、「知識・技能」を単独で評価するのではなく、「知識・技能」と「思考力・判断力・表現力」を総合的に評価するものにしていくことが必要である。

このため、現行の大学入試センター試験を廃止し、下記のような新テスト「大学入学

18 「小学校、中学校、高等学校及び特別支援学校等における児童生徒の学習評価及び指導要録の改善等について（通知）」（平成 22 年文科初第 1 号）においても、学習評価について、客観性にとらわれ過ぎ、試験の点数のみに依拠した評価から脱却するため、「客観性」ではなく「妥当性、信頼性」という文言を用いることにより、指導改善や、きめ細かい学習指導の展開、児童・生徒一人ひとりの学習の確実な定着を目指していることにも留意。

希望者学力評価テスト（仮称)」を新たに実施する。

「大学入学希望者学力評価テスト（仮称)」の在り方

◆　大学入学希望者が、これからの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主たる目的とし、「確かな学力」のうち「知識･技能」を単独で評価するのではなく、「知識･技能を活用して、自ら課題を発見し、その解決に向けて探究し、成果等を表現するために必要な思考力･判断力･表現力等の能力」（「思考力・判断力・表現力」）を中心に評価する。

◆　「教科型」に加えて、現行の教科･科目の枠を越えた「思考力・判断力・表現力」を評価するため、「合教科･科目型」「総合型」の問題を組み合わせて出題する。具体的な作問に向けた検討の状況を見据えつつ、将来は「合教科･科目型」「総合型」のみ[19]とし、教科･科目に必要な「知識･技能」と「思考力・判断力・表現力」を総合的に評価することを目指す。

◆　解答方式については、多肢選択方式だけではなく、記述式を導入する。

◆　大学入学希望者に挑戦の機会を与えるとともに、資格試験的利用を促進する観点から、年複数回実施する。実施回数や実施時期については、進路を決めるに当たり、入学希望者が他者からの指導に受動的に従うのではなく、自ら考え自ら挑戦できるようにすることを第一義として、高等学校教育への影響を考慮しつつ、高等学校･大学関係者を含めて協議する。

◆　「１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、大学及び大学入学希望者に対して、段階別表示による成績提供を行う[20]。

◆　ＣＢＴ方式での実施を前提に、出題･解答方式の開発や、実施回数の検討等を行う。

◆　特に英語については、四技能を総合的に評価できる問題の出題（例えば記述式問題など）や民間の資格・検定試験の活用により、「読む」「聞く」だけではなく「書く」「話す」も含めた英語の能力をバランスよく評価する[21]。また、他の教科･科目や「合

19 今後、高等学校の教科･科目の構造が見直され、既存の教科･科目の枠を越えた「思考力・判断力・表現力」を育成･評価する教科･科目が設置されることになれば、既存の教科･科目の枠を越えた「思考力・判断力・表現力」を評価する問題が「教科型」として設定されることも考え得る。

20 段階別表示の具体的な在り方や、あわせてどのようなデータ（標準化得点や、パーセンタイル値に基づき算出されたデータ等）を大学に提供することが適当かについては、別途、専門家等による検討を行うこととする。

21 「英語教育の在り方に関する有識者会議」報告書（平成 26 年 9 月 26 日）も参照のこと。「大学入学希望者学力評価テスト（仮称)」独自の問題作成を行うべきか、民間の資格・検定試験に全面的にゆだねるべきかについては、四技能を踏まえた作問の質に加えて、日本人の英語力の現状を踏まえたテスト開発の在り方、各試験間の得点換算の在り方、受検料など経済格差の解消、受検機会など地域格差の解消等に関する具体的な検討が必要であり、今後、学校関係団体、試験団体、経済団体、大学入試センター等が参加して設置された「連絡協議会」において速やかに検証が行われるよう求める。

教科･科目型」「総合型」についても、英語についての検討状況も踏まえつつ、民間の資格･検定試験の開発･活用も見据えた検討を行う。

◆　選抜性の高低にかかわらず多くの大学で活用できるよう、広範囲の難易度とする。特に、選抜性の高い大学が入学者選抜の評価の一部として十分活用できる水準の、高難度の出題を含むものとする。

◆　生涯学習の観点から、大学で学ぶ力を確認したいものは、社会人等を含め誰でも受検可能とする。また、海外からの受検も可能とするよう、実施時期や方法について検討するものとする。

◆　入学希望者の経済的負担や受検場所、障害者の受検方法を考慮するなど、受検しやすい環境を整備する。

こうした新テストの実施に向け、特に「合教科･科目型」及び「総合型」における問いの設計については、①その問いにおいて、どのような「思考力・判断力・表現力」を評価するのかを明確化し、②明確化された力が、高等学校におけるどの教科・科目等においてどのような力として主に育成されているのかを特定し、③特定された教科・科目等において育成される力を、他教科・科目等のどのような文脈に当てはめていくことが効果的かを検討しつつ、教科・科目等の組み合わせを決定・作問する、というプロセスのイメージが考えられる。

具体的には、例えば①言語に関する「思考力・判断力・表現力」について、②こうした力を主に育成する国語・英語を、③他教科・科目（例えば理科）と組み合わせ、理科の文脈の中で言語に関する「思考力・判断力・表現力」を評価する問いを作問する、といったことが考えられる。「合教科･科目型」「総合型」の問で評価される力としては、言語に関する「思考力・判断力・表現力」のほか、数に関する「思考力・判断力・表現力」、科学に関する「思考力・判断力・表現力」、社会に関する「思考力・判断力・表現力」、「問題発見・解決力」、「情報活用能力[22]」なども想定される[23]。

こうした「合教科･科目型」「総合型」の作問については、思考力・判断力・表現力等を評価する各種の問題（ＰＩＳＡ調査、全国学力･学習状況調査の主として「活用」に関する問題、文部科学省が実施している情報活用能力調査、各大学の個別選抜における総合問題・小論文、高等学校の総合的な学習の時間における課題、大学入試センターにおける「新しい試験の開発に関する研究」[24]等）に関する知見を有する専門家を、民間も含めて結集し、早急に検討を進める。

---

22 情報及び情報手段を主体的に選択し活用していくために必要な①情報活用の実践力、②情報の科学的な理解、③情報社会に参画する態度。こうした能力を、情報技術を用いて評価することが考えられる。

23 別添資料４参照。

24 教科ごとの知識・技能とは異なる、問題解決や課題遂行に必要となる基本的な能力や適性、実践的な言語運用能力や数理分析力等を評価する新しい試験の在り方に関する研究で、具体的な問題を試作し、モニター調査による識別力等の分析・評価等に取り組んでいる。

なお、「合教科・科目型」「総合型」が評価する「思考力・判断力・表現力」の育成は、現行学習指導要領に基づく各教科等の指導内容としても謳われており、「思考力・判断力・表現力」を育成する指導の充実と「合教科・科目型」「総合型」の導入を、現行学習指導要領下で並行的に進めていくことは、まずは可能である。

ただし、こうした指導を飛躍的に充実させ定着させるためには、学力の三要素を踏まえた高等学校教育課程の抜本的な見直しが必要であり、次期学習指導要領に向けては、高度な思考力・判断力・表現力を育成・評価するための教科・科目構成の在り方や、「思考力・判断力・表現力」を育成するための学習・指導方法の飛躍的充実についても検討を進める必要がある。

（２）高等学校教育の質の確保･向上

高等学校教育については、「国家及び社会の責任ある形成者として、自立して生きる力」の確実な育成、またそのための教養と行動規範の涵養に向けて、教育内容、学習･指導方法、評価方法、教育環境を抜本的に充実させなければならない。

その際、初等中等教育分科会高等学校教育部会が平成２６年６月に取りまとめた「審議まとめ」において提言しているように、全ての生徒が共通に身に付けるべき資質･能力の育成という「共通性の確保」と、多様な学習ニーズへのきめ細かな対応という「多様化への対応」を両者のバランスに配慮しながら進める必要がある。

このうち、「共通性の確保」という観点からは、下記の新テストを導入する。また、「多様化への対応」という観点については、高等学校が、高校生の能力、適性、興味･関心、進路希望等の多様化を受け止めて必要な対応を行うのみならず、年齢、性別、国籍、文化、障害の有無、地域の違い、家庭環境等にかかわらず多様な生徒を積極的に受け入れ、多様な学習環境を創り出すべきである。

①　高等学校段階の基礎学力を評価する新テストの導入

全ての高校生について、身に付けるべき資質･能力を確実に育み、生徒の学習意欲の喚起、学習の改善を図ることができるよう、高等学校段階の基礎学力を評価する新テスト「高等学校基礎学力テスト（仮称）」を導入する[25]。

「高等学校基礎学力テスト（仮称）」の在り方

◆　高校生が、自らの高等学校教育における基礎的な学習の達成度の把握及び自らの学力を客観的に提示することができるようにし、それらを通じて生徒の学習意欲の喚起、

[25] このテストで評価する学力を「基礎学力」としているが、これは「大学入学希望者学力評価テスト（仮称）」で評価する学力よりも低い学力という意味ではなく、高等学校教育で高校生が共通に身に付けるべき学力という意味である。

改善を図る。

◆　上記以外にも、結果を高等学校での指導改善にも生かすことや、進学時や就職時に基礎学力の証明や把握の方法の一つとして、その結果を大学等が用いることも可能とする。

ただし、進学時への活用は、調査書にその結果を記入するなど、あくまで高等学校段階における学習成果を把握するための参考資料の一部として用いることとする[26]。

◆　高校生の個人単位又は学校単位での希望参加型とする[27]が、できるだけ多くの生徒が参加することを可能とするための方策を検討する。

◆　対象教科･科目については、実施当初は「国語総合」「数学Ⅰ」「世界史」「現代社会」「物理基礎」「コミュニケーション英語Ⅰ」などの高等学校の必履修科目[28]を想定して検討する[29]（選択受検も可能）。

英語等については、民間の資格･検定試験も積極的に活用する。

◆　出題内容については、高等学校で育成すべき「確かな学力」を踏まえ、「思考力・判断力・表現力」を評価する問題も含めるが、学力の基礎となる知識・技能の質と量を確保する観点から、特に「知識･技能」の確実な習得を重視する。また、高校進学率約９８％に達する高校生の知識・技能が広範にわたっていることに鑑み、高難度の問題から低難度の問題まで広範囲の難易度とする。

◆　解答方式については、多肢選択方式を原則としつつ、記述式の導入を目指す。

◆　高校生の主体的な学習を促進する観点から、在学中に複数回（例えば年間２回程度）受検機会を提供し、高等学校２年及び３年での希望に応じた受検を可能とする[30]。

実施時期については、夏～秋を基本として、学校現場の意見を聴取しながら検討する。

◆　各学校･生徒に対し、段階別表示による成績提供を行う[31]とともに、各自の正答率等も併せて表示[32]する。

---

26 初等中等教育分科会高等学校教育部会の審議まとめ（平成 26 年 6 月）においては、本テストの進学時への活用は、現在学力不問となっている推薦・ＡＯ入試を念頭に置いたものとされている。今後、大学入学者選抜について、一般入試、推薦入試、ＡＯ入試の区分を見直すことを踏まえ、今後の詳細な制度設計については、学校生活への影響も勘案しながら進められることが必要である。

27 実施場所については、高等学校単位の受検の場合は高等学校で、個人の受検の場合は都道府県毎に会場を設ける方向で検討。

28 高等学校学習指導要領を踏まえた問題とする。また、学習の達成度を測る性質の問題とし、選抜的なものとはしない。

29 保健体育、芸術、家庭、情報及び職業に関する各教科は、実習等による幅広い学習活動によって評価される比重が高く、一般に多肢選択式や記述式のテストになじみにくいこと等にも配慮して検討する。

30 高等学校１年生からの受検を可能とするかは、学校現場の意見を聴取しながら検討する。

31 テスト結果については、学校や生徒の序列化にならないよう、その取り扱いについて十分注意する。

32 学習指導上の困難を抱える学校では、希望に応じてテストの一部問題の活用等の工夫を行う。

◆　ＣＢＴ方式での実施を前提に、出題･解答方式の開発等を行う。

◆　家庭の経済的負担等を考慮するなど、生徒が受検しやすい環境を整備する。

◆　「高等学校卒業程度認定試験」との関係についても検討する。

②　高等学校の教育内容や学習･指導方法、評価方法等の見直し

高等学校における教育内容については、「国家及び社会の責任ある形成者として、自立して生きる力」を育む観点を一層重視することが必要であり、そのための教養と行動規範を涵養することを含めた取組の充実を、「高等学校基礎学力テスト（仮称）」の導入と並行して進める。あわせて、学習・指導方法についても、言語活動の積極的な導入をはじめ、生徒が受け身でなく主体的・協働的に学ぶことを促す方法へと進化を図る。

高等学校の学習指導要領については、さらに、多様な若者の夢や目標を支援できる高等学校教育の実現を目指し、①「何を教えるか」ではなく「どのような力を身に付けるか」の観点に立って、②そうした力を確実に育むため、指導内容に加えて、学習方法や学習環境についても明確にしていく観点から抜本的に見直す。

具体的には、高等学校の学習指導要領を通じて、全体としてどのような資質･能力を育成しようとしているのかをより明確化するとともに、例えば、以下のような見直しを行う。なお、育成すべき資質･能力の明確化に当たっては、教育基本法や学校教育法の目的･目標のほか、ＯＥＣＤのキー・コンピテンシーや、国際バカロレアが目指す論理的思考力や表現力、探究心等の育成などの考え方も参考にしつつ検討する。

◆「思考力・判断力・表現力」を育成するための、課題の発見と解決に向けた主体的・協働的な学習・指導方法の飛躍的充実
◆英語において四技能を系統的に育成するため、小学校から高等学校までを通じて達成を目指すべき教育目標を、「英語を使って何ができるようになるか」という観点から、四技能に係る一貫した具体的な指標の形で設定すること
◆国家や社会の形成者となるための教養と行動規範、また自立して社会生活を営むために必要な力を、実践的に身に付けるためのカリキュラムを充実させること
◆高度な思考力・判断力・表現力を育成・評価するための新たな教科・科目を検討すること
◆大学の卒業論文のような課題探究を行う「総合的な学習の時間」の一層の充実に向けた見直し
◆特別支援教育の充実のための見直し

具体的な教育課程の在り方については、本年１１月の「初等中等教育における教育課程の基準等の在り方について」の諮問を受けて更に検討する。

また、これからの高等学校教員には、課題の発見と解決に向けた主体的・協働的な学びを重視した教育を展開するとともに、生徒の多様な学習成果や活動を適切に評価することなどにより、これからの時代に必要な資質・能力を身に付けさせ、生徒一人ひとり

の可能性を伸ばしていく観点から指導を行う力量が求められる。そのために、きめ細かな指導体制の充実を図るとともに、開放制の原則[33]の中でもこうした力が身に付くよう、教員の資質・能力の向上に向け、教職課程を改善し、研修・採用等の方法を整備する。特に、大学の教職課程において、教員に必要な資質・能力を育成するとともに、現職教員について、各主体の研修においてこうした指導力を身に付けるプログラムが整備されるよう、必要な環境整備を図る。

具体的な在り方については、現在行われている教員の養成・採用・研修の改善についての議論の中で更に検討する。

加えて、新たな評価方法の研究･開発を行い、生徒の多様な学習成果や活動を評価する方法に転換する。

進路指導についても、そうした評価を踏まえつつ、単なる知識･技能の習得度に基づく指導を行うのではなく、多面的･総合的な評価に基づき、生徒一人ひとりの将来目標の実現を支援する観点に転換する。

あわせて、調査書及び指導要録の様式等についても、新たな高等学校教育の在り方を踏まえ、生徒の多様な学習成果や活動が反映されたものになるよう改訂する。

（３）大学教育の質的転換の断行

大学教育においては、高等学校教育において培われた「生きる力」「確かな学力」を更に発展･向上させるよう、教育内容、学習･指導方法、評価方法、教育環境を抜本的に転換する。

「主体性･多様性･協働性」を育成する観点からは、大学教育を、従来のような知識の伝達･注入を中心とした授業から、学生が主体性を持って多様な人々と協力して問題を発見し解を見いだしていくアクティブ・ラーニングに転換し、特に、少人数のチームワーク、集団討論、反転授業、実のある留学や単なる職場体験に終わらないインターンシップ等の学外の学修プログラムなどの教育方法を実践する。

大学において育成すべき力を学生が確実に身に付けるためには、大学教育において「教員が何を教えるか」よりも「学生が何を身に付けたか」を重視し、学生の学修成果の把握・評価を推進することが必要である。

このため、各大学においては、大学教育で身に付ける力等を明確にした上で、ナンバリングの導入等も含め、個々の授業科目等を越えた大学教育全体としてのカリキュラム･マネジメントを確立し、教育課程の体系化･構造化を行うことが求められる。このような各大学の取組を推進するためには、下記３．①に示すとおり、アドミッション・ポリシーと合わせて、学位授与の方針、教育課程編成・実施の方針の一体的な策定を法令上位置付けることが必要である。

---

[33] 国・公・私立のいずれの大学でも、教員免許状取得に必要な所要の単位に係る科目を開設し、学生に履修させることにより、制度上等しく教員養成に携わることができること。

また、大学全体としての共通の評価方針（アセスメント・ポリシー）を確立した上で、学生の学修履歴の記録や自己評価のためのシステムの開発、アセスメント・テストや学修行動調査等の具体的な学修成果の把握・評価方法の開発・実践、これらに基づく厳格な成績評価や卒業認定等を進めることが重要である。さらに、評価に係る専門的人材を育成することも必要であり、国は、こうした取組に対して支援を行うことが必要である。

認証評価制度についても、教育環境等の外形を中心にした現在の評価方法から、学生の学修成果や各大学における成果把握と転換の取組（内部質保証）といった、成果を重視した評価に改善することが必要である。

さらに、大学教育の質的転換を進める上では、学生同士が切磋琢磨（せっさたくま）し、相互に刺激を与えながら成長する場を創ることが重要である。このため、年齢、性別、国籍、文化、障害の有無、地域の違い、家庭環境等にかかわらず、多様な背景を持った教職員や学生を受け入れることによって、大学の構成員の多様化を進め、主体性を持って多様な人々と協働するとともに創造性を磨くことのできる学習環境を実現するとともに、多様な学生に対応できる教育カリキュラムを用意しなければならない。

なお、大学への入学についても、高等学校卒業後に入学する道だけではなく、編入学や転入学、社会に出た後の学び直しも含めた社会人入学など多様な道を開くことにより、容易に進路を変更でき、生涯を通じて学修に取り組める環境を実現する。

また、大学入学後の初年度における教育については、初年次教育、導入教育、リメディアル教育等の様々な概念が混在している。高大接続の観点から、高等学校教育の質の確保・向上とアドミッション・ポリシーに基づく大学入学者選抜の確立の上に、その意義をもう一度見直すならば、初年次教育は、高等学校で身に付けるべき基礎学力の単なる補習とは一線を画すべきであり、高等学校教育から大学における学修に移行するに当たって、大学における本格的な学修への導入、より能動的な学修に必要な方法の習得等を目的とするものとして捉えるべきである。

こうした大学初年次教育の展開・実践は、高等学校教育の成果を大学入学者選抜後の大学教育へとつなぐ、高大接続の観点から極（きわ）めて重要な役割を果たすものであり、その質的転換を断行するには、高等学校教育、大学教育の新しい姿を確立するとととともに、これらの教育で育成すべき力を円滑に接続するための研究開発が必要である。

なお、大学教育において育成が求められる「確かな学力」としては、各分野における専門教育の在り方も重要であり、特に、技術革新の激しい時代の中で、高度な専門知識等が必要な職業分野に求められる人材を養成するためには、学部段階の教養教育及び専門教育で身に付けた能力を、大学院段階の高度な専門教育で更に伸長させることが求められる。

そのため、大学と産業界との間で人材像の共有を図りつつ、学部教育から大学院教育まで一貫した視点で教育課程の編成等を行うなど、大学院教育までを見通した改革の観点も必要である。

上記の改革を実現するためには、学長のリーダーシップの下での戦略的な大学経営が

必要であり、来年４月から施行される改正学校教育法の趣旨も踏まえ、大学のガバナンス改革を推進する必要がある。

（４）新テストの一体的な実施

「高等学校基礎学力テスト（仮称）」と「大学入学希望者学力評価テスト（仮称）」とは、目的や性格の違いがある一方で、ＣＢＴの導入や両テストの難易度・範囲の在り方など、共通に検討すべき事項が多く、一体的な検討が必要である。

出題範囲についても、「高等学校基礎学力テスト（仮称）」は、６教科の必履修科目について、主として学力の基礎となる「知識・技能」を評価するものであり、「大学入学希望者学力評価テスト（仮称）」は、主として「思考力・判断力・表現力」を評価するものである。両者はテストの目的だけでなく出題範囲についても異なっているが、高等学校から大学への学力の円滑な接続を図るために、両テストの難易度をできるだけ連続的にすることが必要である[34]。

国においては、一体的な検討を行う専門家会議とその事務局体制を早急に立ち上げるとともに、両テストの円滑な実現に向けて、一体的な実施体制を構築することが必要である。

新テストの実施主体については、共通一次試験や大学入試センター試験等、高等学校教育の達成度を把握する試験や全国的な大規模の試験の実績・ノウハウを有する大学入試センターを、高等学校及び大学の学力評価や生徒・学生の学びを支援する観点から抜本的に改組した新たなセンターとする。新センターは、新テストの実施と方法開発、個別選抜やアドミッション・オフィス強化等の方法開発などの支援、面接や集団討論等を含むテスト方法開発などの支援、調査書の評価等を含む評価に関する方法開発などの支援、専門的人材の育成、入学者選抜や学力評価についての新しい方法の開発、これらの事項に関わる国内外の調査等を目的とし、名称についても、その機能を体現するものに変更する。

なお、「大学入学希望者学力評価テスト（仮称）」については、大学が共同実施する性格のテストであるということを踏まえながら、大学を含めた具体的な実施体制等を検討するとともに、「高等学校基礎学力テスト（仮称）」については、高等学校と密接に連携・協力して実施するための具体的な実施体制等を検討する必要がある。

---

[34] 別添資料５参照。

３．改革を実現するための具体策（「高大接続改革実行プラン（仮称）」の策定）

高大接続改革を実現するためには、国における制度改正のみならず、各高等学校や大学における教育や評価の在り方を抜本的に転換していく必要があり、そのための具体的施策や改革スケジュールの明確化が必要である。特に、生徒や学生の多様性を踏まえた「公正」の理念に基づく多元的な評価の在り方については、各学校における蓄積は十分ではないことから、国主導で先導的な評価方法の蓄積･共有や、新たな手法の研究開発を積極的に行っていく必要がある。

本答申では、そうした具体策やスケジュールについて、国や新テストの実施主体等に検討を求める事項の骨子を、以下の通り示すこととする。国においては、この骨子をもとに具体策やスケジュールの詳細を「高大接続改革実行プラン（仮称）」といった具体的な形で答申後速やかに策定・公表し、強力に推進することを求める。また、新しい時代に求められる教育の在り方を踏まえ、更なる検討が必要な点については、同プランに示されたスケジュールに基づき検討を進め、成果を得たものから順次公表するよう求める。

また、中央教育審議会においては、第８期以降の体制においても、高大接続改革の実現に向けた継続性を確保するため、「高大接続改革実行プラン（仮称）」の内容や進捗状況について国から適時に報告を受け、適切なフォローアップを行うことが重要である。

＜高大接続改革の実現に向けた、具体策とスケジュールの骨子[35]＞

①　各大学における個別選抜改革と教育の質的転換を実現するための、実効的な政策手段

国は、下記のような各大学が取り組むことが求められる事項について、どのような手段（法令改正、大学入学者選抜実施要項の見直し、評価、支援策）によってこれらの取組を促進するかを明らかにした上で、具体的な取組を推進することが必要である。

（各大学が取り組むことが求められる事項）

- アドミッション・ポリシーの明確化
- 個別選抜の改革（学力の三要素を踏まえた学力評価の実施、多元的な評価の推進等）
- 「大学入学希望者学力評価テスト（仮称）」の活用
- 高等学校の学習成果の適切な評価
- 特定分野において卓越した能力を有する者や多様な背景を持った学生に対する適切な評価
- 入学者の追跡調査等による、選抜方法の妥当性・信頼性の検証
- 評価方法の工夫改善、評価に関する専門的人材の育成・活用
- アドミッション・オフィスの強化をはじめとする入学者選抜実施体制の整備

[35] 全体のスケジュールについては、別添資料６に改革工程表のイメージを示している。

（法令改正）

各大学における個別選抜改革を推進するためには、各大学の入学者選抜の設計図であるアドミッション・ポリシーの充実が不可欠であり、各大学においては、それぞれの強み、特色や社会的役割に応じたアドミッション・ポリシーが策定されることが必要である。このため、国は、各大学におけるアドミッション・ポリシーの策定について法令上位置付けるよう検討すること。

その際、各大学においては、大学教育を通じて学生にどのような力を身に付けさせて卒業させるか、そのためにどのような教育を実施するか、教育を実施するに当たってどのような学生を受け入れるのかという一貫した観点から、アドミッション・ポリシーと合わせて、学位授与の方針、教育課程編成・実施の方針を策定することが必要であることから、これらの一体的な策定を法令上位置付けるよう検討すること。

また、各大学の個別選抜改革の取組に対する評価が適切に行われることも必要であることから、国は、法令で定められている認証評価の評価項目に入学者選抜を明記するよう検討すること。

（大学入学者選抜実施要項の見直し）

本答申の理念に基づく高大接続改革は、一部の大学や一部の選抜方法のみで推進されるのでは足りず、各大学が実施する個別選抜全体において実現されなければならない。その際、大学入学希望者が培ってきた力を確認し提示するマイルストーンとしては、多様な挑戦の機会が与えられることが望ましい一方で、いたずらに選抜が早期化・複雑化することにより、高等学校教育の本来の目的が大きくゆがめられる危惧もある。大学入学者選抜が双方の観点を踏まえた秩序ある形で行われるよう、新たなルールを構築していかなければならない。

このため、国は、適切なルールの下での入学者選抜全体の多面的・総合的な評価への転換を図るため、一般入試、推薦入試、ＡＯ入試の区分を廃止し、大学入学者選抜全体の共通的な新たなルールを構築する[36]ために、大学入学者選抜実施要項を抜本的に見直すこと。

具体的には、以下のような事項をルールとして盛り込むことを検討し、平成２６年度中に可能なものから見直しの方向性を取りまとめ、大学入学者選抜実施要項に段階的に反映させること。

・各大学のアドミッション・ポリシーに求められる観点
・アドミッション・ポリシーに基づいた個別選抜の具体的な方法や、選抜時の評価に活用する資料の種類等の受検者への明示
・個別選抜の実施時期
・「大学入学希望者学力評価テスト（仮称）」の積極的な活用と、応募条件として求める

36 各大学の特色等を踏まえたアドミッション・ポリシーに基づき、大学入学希望者の多様性を踏まえた大学入学者選抜が実施されることを前提とした上で、共通する事項についてルールを構築するものであり、全ての大学入学者選抜を画一化する意図のものではないことは言うまでもない。

成績の具体的な提示
・高等学校生活への影響にも十分配慮した、「高等学校基礎学力テスト（仮称）」の活用
・個別選抜における、学力の三要素を十分踏まえた学力評価
・特定分野において卓越した能力を有する者の選抜や、年齢、性別、国籍、文化、障害の有無、地域の違い、家庭環境等にかかわらず多様な背景を持った学生の受け入れ
・入学者の追跡調査等による、選抜方法の妥当性・信頼性の検証

なお、大学入学者選抜実施要項の見直しに当たっては、高校生をはじめとした関係者が見通しを持って対応できるよう配慮すること。

（評価）

前述のとおり認証評価における入学者選抜の評価を法令上位置付けた上で、認証評価における具体の取組を充実することが必要である。

このため、国は、認証評価機関と連携して、認証評価機関における見直し後の大学入学者選抜実施要項を踏まえた入学者選抜に関する評価の基準の見直しなど、新たなルールの遵守状況の適切な評価に取り組むこと。

さらに、ルールの遵守状況の評価のみならず、アドミッション・ポリシーと選抜方法との整合性や個別選抜の工夫改善の取組状況に対する評価等、各大学の独自の改革を促す評価の在り方についても検討すること。

また、国は、各大学の取組状況が広く社会においても共有・評価されるよう、本答申の趣旨を踏まえた具体的内容を反映させた大学ポートレートなどを通じ、情報公開の促進に取り組むこと。

（支援策）

新たなルールに基づく個別選抜への転換と、大学教育の質的転換を強力に推進するため、国においては、次のような支援に取り組むこと。

各大学における個別選抜改革が実現するかどうかは、アドミッション・オフィスの強化にかかっているため、国は、各大学のアドミッション・オフィスの整備・強化の在り方について検討を行い、具体的な支援策を取りまとめること。

前述のアドミッション・ポリシーの法令上の位置付けと合わせて、各大学のアドミッション・ポリシーの内容を充実するための取組を支援することが求められる。

このため、各大学の特色等に応じ、どのような力をどのように評価するのかを明確にした具体的なアドミッション･ポリシーの在り方について、平成２６年度中に先進的な事例を取りまとめた策定事例集を作成すること。

さらに、専門家による検討を踏まえて、求める学生像のみならず、各大学の入学者選抜の設計図として必要な事項を示す観点から、アドミッション・ポリシーに盛り込むことが求められる事項に関するガイドラインを策定すること。

国は、各大学におけるアドミッション・オフィスの整備・強化や、アドミッション・

ポリシーの明確化が実現されるよう、主体的に改革に取り組む大学にとってインセンティブとなるような財政措置の在り方を検討し、具体策を取りまとめること。

あわせて、国は、新たな個別選抜の在り方の開発支援を行うとともに、基盤的経費の配分における新たなルールの要件化や加算化、各種の大学改革のための補助金の応募条件における要件化の工夫など、主体的に改革に取り組む大学にとってインセンティブとなるような財政措置の在り方を検討し、具体策を取りまとめること。

また、各大学の選抜方法の改善やそのための検証の取組を支援するため、国は、新たな評価手法の研究開発を推進するとともに、評価に関する専門的人材の育成を推進すること。

② 新テストの制度設計、実施体制

「高等学校基礎学力テスト（仮称）」と「大学入学希望者学力評価テスト（仮称）」について一体的な検討を行い、「高等学校基礎学力テスト（仮称）」については平成３１年度から、「大学入学希望者学力評価テスト（仮称）」については平成３２年度から段階的に実施すること。

国は、高校生をはじめとした関係者が見通しを持って対応できるよう、実施までの具体的な制度設計、プレテストの実施等に係る詳細なスケジュールを策定し、公表すること。

国は、新テストについて早急に専門家会議を立ち上げ、対象となる教科･科目、「大学入学希望者学力評価テスト（仮称）」における「教科・科目型」、「合教科・科目型」、「総合型」等の具体的な枠組み、問題の蓄積方法、作問の方法、記述式問題の導入方法、ＣＢＴ方式の導入方法、成績表示の具体的な在り方などについて検討を行い、答申後一年を目途に具体的な内容について結論を得ること。

「大学入学希望者学力評価テスト（仮称）」における思考力・判断力・表現力を問う問題については、求められる力を、「教科型」において他教科の内容を掛け合わせつつ評価する問題と、「合教科･科目型」「総合型」として教科･科目の枠を越えて評価する問題の両方について、国が主導して検討を行い、平成２８年度中を目途に作問イメージを公表し、平成３２年度から実施すること。

新テストのプレテストが実施される時期を待たず、実施母体となる新たなセンターをできるだけ早く立ち上げること。

国は、新テストの普及のため、「大学入学希望者学力評価テスト（仮称）」については上記①に示した財政措置等の支援による改革のインセンティブを通じて、各大学における活用を推進すること。また、「高等学校基礎学力テスト（仮称）」については、できるだけ多くの生徒が参加することを可能とするため、関係機関への周知や、指導改善に生かせる分析結果等の各高等学校等への提供[37]、受検料の負担軽減策の検討、調査書様式例

[37] その際、学校や生徒の序列化にならないよう、その取扱いについて十分留意すること。

の見直し、企業への広報・周知等を通じて、積極的な活用を推進すること。

③　高等学校教育の改革

国は、今後の中央教育審議会における高等学校学習指導要領の見直しに関する議論の状況を踏まえつつ、育成すべき資質・能力の明確化と教科・科目の在り方等の見直し、課題の発見と解決に向けた主体的・協働的な学習・指導方法の飛躍的充実や、学習環境の在り方等、今後の高等学校教育の在り方について検討し、可能なものから速やかに具体策を示すこと。

国は、高等学校教員が、課題の発見と解決に向けた主体的・協働的な学びを重視した教育を展開するとともに、生徒の多様な学習成果や活動を適切に評価することができるよう、きめ細かな指導体制の充実を図るとともに、教員の資質・能力の向上に向け、教員の養成・採用・研修の改善を図るための具体策を示すこと。

国は、調査書や高等学校の指導要録の改訂に関する専門家会議を立ち上げ、生徒の多様な学習成果や活動が反映されるよう、調査書の様式の見直しや出願時提出資料の共通様式の策定、指導要録における観点別学習状況の示し方や、「高等学校基礎学力テスト（仮称）」の結果の示し方、大学での活用方策、関係書類の電子化などについて検討し、答申後一年を目途に具体的な改訂内容について結論を得ること。

④　評価方法の改革

国及び新テストを担う実施主体は、高等学校教育・大学教育・大学入学者選抜を通じた新たな入学者選抜方法・学力評価方法の開発、評価方法に関する専門人材の育成、教員の評価力の向上等に関する具体的な検討を行うこと。

あわせて、大学入学後の学生の追跡調査等、アドミッション・ポリシーに示した評価基準・方法の妥当性を検証する方法についても、具体的な検討を行うものとすること。

４．社会全体で改革を共有するための方策

新しい時代にふさわしい高大接続の実現という大きな改革を、我が国の社会全体で実現していくためには、教育関係者はもちろんのこと、子供たちやその保護者、企業、地域社会、その他、社会のあらゆる人々が改革を共有する必要がある。

特に、従来、企業内訓練等が担っていた人材育成機能が、雇用環境の変化により失われつつある中、特に高等学校及び大学において、これからの時代に求められる力を確実に育成し、子供たちを社会に送り出すことが、以前にも増して必要となっている。我が国の将来の社会構造の在り方や、そのために必要な人材像、これからの教育において育成すべき資質･能力の在り方も社会的に共有しつつ、特に、高等学校･大学の卒業生、大学院の修了者の就職先における人材開発・人事採用等の長期的展望と、本答申に基づく改革をしっかりと接続していく必要がある。

こうした改革を実効性あるものにするためには、学生の主体的な学びの確立に向けた環境整備が必要なこと[38]、OECD 諸国など諸外国における教育投資の状況[39]なども踏まえ、我が国の高等教育に真(しん)に必要な教育投資を確保していく必要がある。その際、社会全体で教育を支える環境を醸成するため、特に高等教育に対して、寄附の促進など民間資金の活用を含めた教育投資の確保を図ることも必要である。

同時に、選抜や評価に関する既存の「公平性」の意識を改革し、多様な背景を持った若者それぞれが、自分の夢や目標を持ち、その実現に必要な能力を身に付けることができるよう、それぞれの学びを支援する観点から、一人ひとりが積み上げてきた多様な力を多様な方法で「公正」に評価し選抜することが必要であるという意識を醸成するため、社会的な議論を深めることが必要である。　本答申を受けて、国が、「高大接続改革フォーラム」の全国実施など、理解啓発のための広報活動をあまねく展開し、答申の提言内容及び「高大接続改革実行プラン（仮称）」に対する社会への周知･理解を十分に広げるとともに、各団体等に要請を行うことを求める。

---

[38] 「平成 20 年科学技術人材養成等委託業務成果報告書」（日本物理学会キャリア支援センター）によると、The Times Higher Education の大学ランキング上位 5 校の教員一人当たり学生比の平均が 4.37 であるのに対し、東京大学・京都大学・大阪大学の当該比の平均は 6.19 となっている。

日本の国立大学は公的資金への依存度が高く、諸外国の大学では多様な事業収入の確保に向けた取組がなされている一方、日本の大学はそのような取組が進んでいないとの指摘がある。

[39] 「図表でみる教育（2014）」（OECD）によると、平成 23 年の高等教育段階の在学者一人当たり公財政教育支出（機関補助）は、OECD 平均が 9,221 ドルであるのに対し、我が国は 6,384 ドルとなっている。また、平成 23 年の全教育段階の在学者一人当たり公財政教育支出（機関補助）は、OECD 平均が 7,786 ドルであるのに対し、我が国は 8,106 ドルとなっている。

「National Account 2003-2010」（OECD）等によると、国民負担率は、OECD 平均が 49.8%であるのに対し、我が国は 38.3%と、租税負担率は OECD 平均は 34.8%であるのに対し、我が国は 22.0%となっている。（OECD 平均：平成 21 年又は 20 年、日本：平成 21 年）

第７期中央教育審議会委員

平成２５年２月１５日発令
（５０音順）

会　長	安西祐一郎	独立行政法人日本学術振興会理事長
副会長	小川　正人	放送大学教養学部教授、東京大学名誉教授
副会長	北山　禎介	三井住友銀行取締役会長
	相原　康伸	日本労働組合総連合会副会長、全日本自動車産業労働組合総連合会会長
	明石　要一	千葉敬愛短期大学学長、千葉市教育委員会委員、千葉大学名誉教授
	五十嵐俊子	日野市立平山小学校長
	生重　幸恵	特定非営利活動法人ｽｸｰﾙ・ｱﾄﾞﾊﾞｲｽ・ﾈｯﾄﾜｰｸ理事長、一般社団法人ｷｬﾘｱ教育ｺｰﾃﾞｨﾈｰﾀｰﾈｯﾄﾜｰｸ協議会代表理事
	浦野　光人	株式会社ﾆﾁﾚｲ相談役、公益社団法人経済同友会幹事、公益財団法人産業教育振興中央会顧問、一般社団法人ｱｸﾞﾘﾌｭｰﾁｬｰｼﾞｬﾊﾟﾝ理事長、一般社団法人日本経営協会会長
	衞藤　　隆	社会福祉法人恩賜財団母子愛育会日本子ども家庭総合研究所所長、東京大学名誉教授
	大島　まり	東京大学大学院情報学環教授、東京大学生産技術研究所教授
	尾上　浩一	公益社団法人日本ＰＴＡ全国協議会会長
	小原　芳明	玉川大学長
	帯野久美子	株式会社ｲﾝﾀｰｱｸﾄ・ｼﾞｬﾊﾟﾝ代表取締役、一般社団法人関西経済同友会常任幹事、大阪市教育委員会委員
	河田　悌一	日本私立学校振興・共済事業団理事長
	菊川　律子	放送大学特任教授（福岡学習センター所長）
	北城恪太郎	日本ｱｲ・ﾋﾞｰ・ｴﾑ株式会社相談役、公益社団法人経済同友会終身幹事、学校法人国際基督教大学理事長
	櫻井よしこ	ジャーナリスト、公益財団法人国家基本問題研究所理事長
	篠原　文也	政治解説者、ジャーナリスト
	白石　勝也	愛媛県松前町長
	高橋　香代	くらしき作陽大学子ども教育学部長、岡山県教育委員会委員
	田邉　陽子	日本大学法学部准教授
	長尾ひろみ	公益財団法人広島県男女共同参画財団理事長
	橋本　　昌	茨城県知事
	橋本　　都	八戸工業大学副学長、前青森県教育委員会教育長
	濱田　純一	東京大学総長
	早川三根夫	岐阜市教育委員会教育長
	平尾　誠二	神戸製鋼ﾗｸﾞﾋﾞｰ部ｾﾞﾈﾗﾙﾏﾈｰｼﾞｬｰ、特定非営利活動法人ｽﾎﾟｰﾂ・ｺﾐｭﾆﾃｨ・ｱﾝﾄﾞ・ｲﾝﾃﾘｼﾞｪﾝｽ機構理事長
	無藤　　隆	白梅学園大学子ども学部教授兼子ども学研究科長
	森　　民夫	長岡市長
	吉田　　晋	学校法人富士見丘学園理事長、富士見丘中学高等学校校長、日本私立中学高等学校連合会長

（３０名）

尾上浩一委員の発令は平成２５年８月２０日
北山禎介委員の発令は平成２６年２月　１日

第６期中央教育審議会高大接続特別部会委員

◎部会長，○副部会長

（委　　員）

◎安　西　祐一郎	独立行政法人日本学術振興会理事長
生　重　幸　恵	特定非営利活動法人ｽｸｰﾙ･ｱﾄﾞﾊﾞｲｽ･ﾈｯﾄﾜｰｸ理事長
浦　野　光　人	株式会社ﾆﾁﾚｲ代表取締役会長、公益社団法人経済同友会幹事、財団法人産業教育振興中央会理事長
金　子　元　久	筑波大学大学研究センター教授
○無　藤　　隆	白梅学園大学子ども学部教授、子ども学研究科長

（臨時委員）

相　川　順　子	一般社団法人全国高等学校ＰＴＡ連合会会長
荒　瀬　克　己	京都市教育委員会教育企画監
及　川　良　一	東京都立三田高等学校長、全国高等学校長協会会長
勝　　悦　子	明治大学副学長
小　林　　浩	リクルート進学総研所長、カレッジマネジメント編集長
近　藤　倫　明	北九州市立大学長
田　邉　恒　美	山口県教育委員会教育長
垂　水　共　之	岡山大学大学院環境生命科学研究科教授
土　井　真　一	京都大学大学院公共政策連携研究部・法学研究科教授
濵　口　道　成	名古屋大学総長
濱　名　　篤	関西国際大学長、学校法人濱名学院理事長
宮　田　裕　子	ﾕﾆﾘｰﾊﾞ･ｼﾞｬﾊﾟﾝ･ﾎｰﾙﾃﾞｨﾝｸﾞｽ株式会社取締役人事総務本部長
山　本　　繁	特定非営利活動法人ＮＥＷＶＥＲＹ理事長
吉　田　　晋	学校法人富士見丘学園理事長、富士見丘中学校高等学校校長、日本私立中学高等学校連合会会長

計　１９名

第７期中央教育審議会高大接続特別部会委員

◎部会長，○副部会長

（委　　員）

◎安　西　祐一郎	独立行政法人日本学術振興会理事長
生　重　幸　恵	特定非営利活動法人ｽｸｰﾙ･ｱﾄﾞﾊﾞｲｽ･ﾈｯﾄﾜｰｸ理事長
浦　野　光　人	株式会社ﾆﾁﾚｲ相談役、公益社団法人経済同友会幹事、公益財団法人産業教育振興中央会顧問、一般社団法人アグリフューチャージャパン理事長、一般社団法人日本経営協会会長
櫻　井　よしこ	ジャーナリスト、公益財団法人国家基本問題研究所理事長
○無　藤　　隆	白梅学園大学子ども学部教授、子ども学研究科長
吉　田　　晋	学校法人富士見丘学園理事長、富士見丘中学校高等学校校長、日本私立中学高等学校連合会会長

（臨時委員）

相　川　順　子	一般社団法人全国高等学校ＰＴＡ連合会顧問
荒　瀬　克　己	大谷大学文学部教授、国立高等専門学校機構監事、京都市教育委員会指導部顧問
及　川　良　一	国立音楽大学教授
勝　　悦　子	明治大学副学長
金　子　元　久	筑波大学大学研究センター教授
小　林　　浩	リクルート進学総研所長、リクルート「カレッジマネジメント」編集長
近　藤　倫　明	北九州市立大学長
垂　水　共　之	中国学園大学子ども学部教授
土　井　真　一	京都大学大学院法学研究科教授
濵　口　道　成	名古屋大学総長
濱　名　　篤	関西国際大学長、学校法人濱名学院理事長
山　本　　繁	特定非営利活動法人ＮＥＷＶＥＲＹ理事長

計　１８名

別添資料１

高大接続特別部会におけるこれまでの審議の経過

中央教育審議会では、平成２４年８月に文部科学大臣から「大学入学者選抜の改善をはじめとする高等学校教育と大学教育の円滑な接続と連携の強化のための方策について」の諮問を受け、総会直属の高大接続特別部会（以下「特別部会」という。）を設置し、検討を進めてきた。

また、高等学校教育の質の確保・向上については、平成２３年９月に初等中等教育分科会高等学校教育部会が設置されて審議が行われ、平成２６年６月には「審議まとめ」が取りまとめられた。特別部会においては、高等学校教育部会との合同会議の開催も含め精力的に審議を行ってきた。

平成２５年６月に、政府の教育再生実行会議が高大接続の在り方に関する審議を開始した際には、特別部会長が出席し、審議が円滑に行われるよう、特別部会の審議状況について報告を行った。教育再生実行会議においては、平成２５年１０月に第４次提言「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」が取りまとめられたところであり、特別部会はその後も議論を重ね、平成２６年３月に「審議経過報告」を取りまとめ公表した。

その後、パブリック・コメントに寄せられた意見や、関係団体・各界の意見等を踏まえつつ、更に審議を重ね、ここに本答申を取りまとめた。

第１回　平成２４年　９月２８日
- 部会長の選任等について
- 大学入学者選抜の改善をはじめとする高等学校教育と大学教育の円滑な接続と連携の強化のための方策について（自由討議）

第２回　平成２４年１０月３１日
- 大学入学者選抜の現状と課題について

第３回　平成２４年１１月３０日
- 大学入試における能力の判定の現状と課題について

第４回　平成２４年１２月１７日
- 入試方法の多様化や評価尺度の多元化等について

第５回　平成２５年　１月１５日
- 大学入学志願者の多様な能力・適性等の評価について

第６回　平成２５年　４月２４日
・部会長の選任等について
・高等学校教育の質保証をはじめとした高大接続の在り方について

第７回　平成２５年　５月２４日
・大学入学志願者の多様な能力・適性等の評価

第８回　平成２５年１１月　８日
・教育再生実行会議第四次提言を踏まえた検討課題について

第９回　平成２５年１１月２９日
・多面的・総合的に評価・判定する大学入学者選抜への転換
・大学の人材育成機能強化・高等学校教育と大学教育の連携強化

第１０回　平成２５年１２月１２日
・高大接続特別部会及び高等学校教育部会に共通する検討課題
※高等学校教育部会との合同会議

第１１回　平成２６年　１月２４日
・教育再生実行会議第四次提言を踏まえた検討課題について

第１２回　平成２６年　２月１９日
・教育再生実行会議第四次提言を踏まえた検討課題について

第１３回　平成２６年　３月　６日
・高大接続特別部会の審議経過報告について（素案）
・達成度テスト（発展レベル）（仮称）の考え方について

第１４回　平成２６年　３月２５日
・高大接続特別部会の審議経過報告（案）について

第１５回　平成２６年　５月２３日
・達成度テスト（発展レベル）（仮称）の在り方

第１６回　平成２６年　６月２０日
・答申（案）について

第１７回　平成２６年　７月２５日
・高等学校教育、大学教育、大学入学者選抜の一体的な改革の在り方について
・各大学の入学者選抜の在り方について

第１８回　平成２６年　８月２２日
・今後の国立大学の入学者選抜の改革の方向について
・高大接続の改善の方向性について

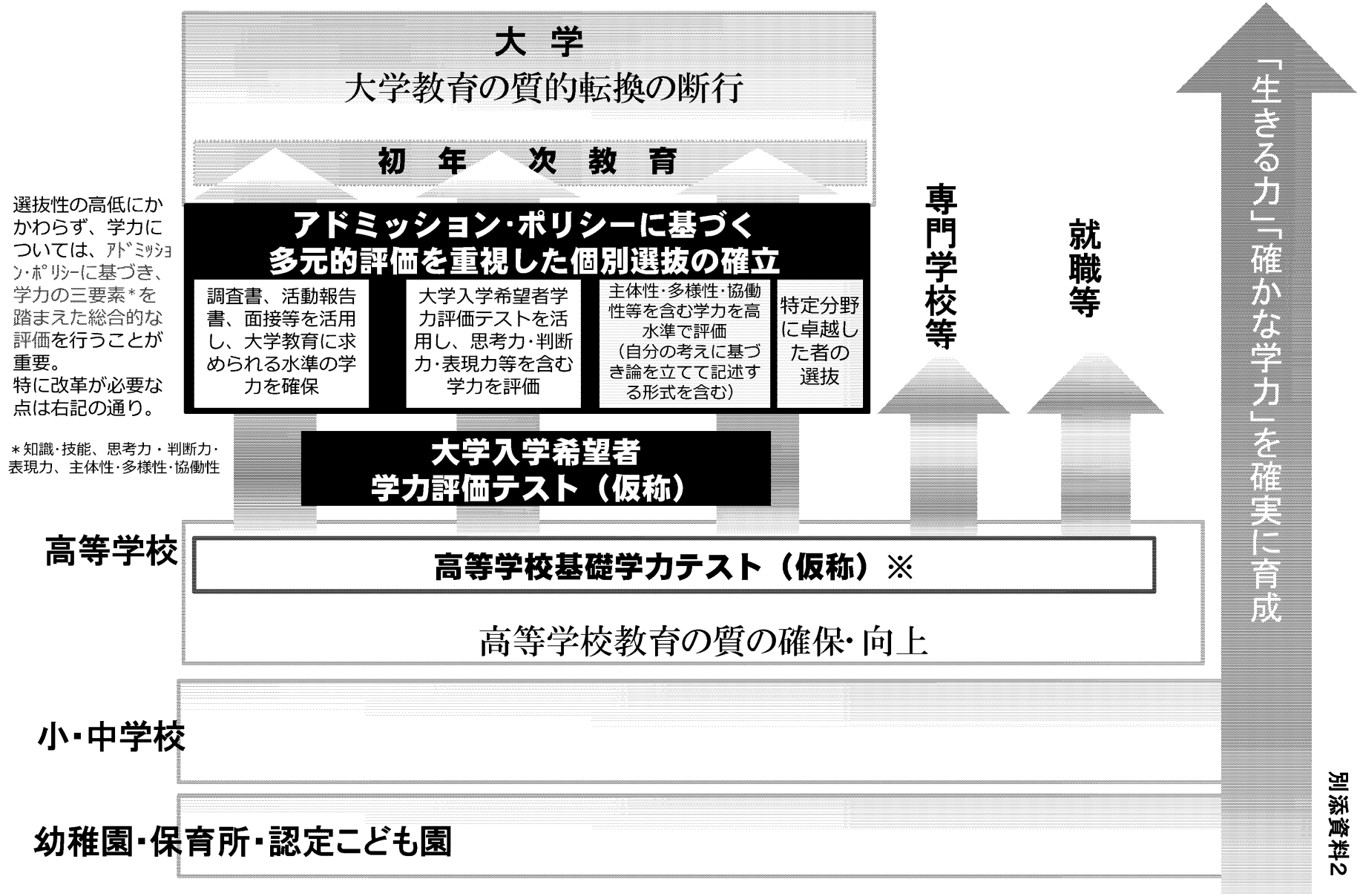
第１９回　平成２６年　９月１７日
・各大学の大学入学者選抜の在り方について
・高等学校教育、大学教育、大学入学者選抜の一体的な改革の必要性・背景、課題について

第２０回　平成２６年１０月１０日
・取りまとめに向けた要点の整理

第２１回　平成２６年１０月２４日
・答申（案）について

大学入学者選抜改革の全体像（イメージ）（案）

※「高等学校基礎学力テスト（仮称）」は、入学者選抜への活用を本来の目的とするものではなく、進学時への活用は、調査書にその結果を記入するなど、あくまで高校の学習成果を把握するための参考資料の一部として用いることに留意。



別添資料2

総称	**学力評価のための新たなテスト（仮称）** 別添資料３	
実施主体	**大学入試センターを、「学力評価のための新たなテスト（仮称）」の実施･方法開発や評価に関する方法開発などの支援を一体的に行う組織に抜本的に改組。**	
個別名称	**高等学校基礎学力テスト（仮称）**	**大学入学希望者学力評価テスト（仮称）**
目的・活用方策	○生徒が、自らの高等学校教育における学習の達成度の把握及び自らの学力を客観的に提示することができるようにし、それらを通じて生徒の学習意欲の喚起、学習の改善を図る。 ＜上記以外の活用方策＞ ○結果を高等学校での指導改善にも生かす。 ○進学時や就職時に基礎学力の証明や把握の方法の一つとして、その結果を大学等が用いることも可能とする。 ※進学時の活用は、調査書にその結果を記入するなど、高等学校段階の学習成果把握のための参考資料の一部として使用。	○大学入学希望者が、これからの大学教育を受けるために必要な能力について把握する。 「確かな学力」のうち「知識・技能」を単独で評価するのではなく、「知識･技能を活用して、自ら課題を発見し、その解決に向けて探究し成果等を表現するために必要な思考力･判断力･表現力等の能力（「思考力･判断力･表現力）」を中心に評価。
対象者	○希望参加型 ※　できるだけ多くの生徒が参加することを可能とするための方策を検討。	○大学入学希望者 ※　大学で学ぶ力を確認したい者は、社会人等を含め、誰でも受験可能。
内容	○実施当初は「国語総合」「数学Ⅰ」「世界史」「現代社会」「物理基礎」「ｺﾐｭﾆｹｰｼｮﾝ英語Ⅰ」等の高校の必履修科目を想定（選択受検も可能）。 ○高等学校で育成すべき「確かな学力」を踏まえ、「思考力･判断力･表現力」を評価する問題を含めるが、学力の基礎となる知識・技能の質と量を確保する観点から、特に「知識・技能」の確実な習得を重視。 ※高難度から低難度まで広範囲の難易度。 ○各学校・生徒に対し、成績を段階で表示 ※　各自の正答率等も併せて表示	○「教科型」に加えて、教科・科目の枠を超えた思考力･判断力･表現力を評価するため、「合教科・科目型」「総合型」の問題を組み合わせて出題。 ※　将来は「合教科・科目型」「総合型」のみによる「知識･技能」と「思考力･判断力･表現力」の総合的な評価を目指す。 ※　広範囲の難易度。特に、選抜性の高い大学が入学者選抜の評価の一部として十分活用できる水準の高難易度の出題を含む。 ○大学及び大学入学希望者に対し、段階別表示による成績提供
解答方式	○多肢選択方式が原則、記述式導入を目指す。	○多肢選択方式だけでなく、記述式を導入。
検討体制	○ＣＢＴの導入や両テストの難易度･範囲の在り方、問題の蓄積方法、作問の方法、記述式問題の導入方法、成績表示の具体的な在り方等について一体的に検討。	
実施方法	○在学中に複数回（例えば年間２回程度）、高校２・３年での受験を可能とする。 ○実施時期は、夏～秋を基本として、学校現場の意見を聴取しながら検討。 ○ＣＢＴ方式での実施を前提に開発を行う。 ○英語等については、民間の資格･検定試験も積極的に活用。	○年複数回実施。 ○実施回数や実施時期は、入学希望者が自ら考え自ら挑戦することを第一義とした上で、高校教育への影響を考慮しつつ、高校・大学関係者を含めて協議。 ○ＣＢＴ方式での実施を前提に開発を行う。 ○特に英語は、四技能を総合的に評価できる問題の出題や民間の資格・検定試験を活用。 ※　他の教科・科目や「合教科・科目型」「総合型」についても、民間の資格・検定試験の開発・活用も見据えて検討。
作問のイメージ	全国学力･学習状況調査のＡ問題（主として知識に関する問題）及びＢ問題（主として活用に関する問題）の高校教育レベルの問題を想定	知識・技能を活用して、自ら課題を発見し、その解決に向けて探究し成果等を表現するための力を評価する、PISA 型の問題を想定

「合教科・科目型」「総合型」について

思考力・判断力・表現力

知識・技能を活用して、自ら課題を発見し、その解決に向けて探究し、成果等を表現するために必要な思考力・判断力・表現力等の能力

（参考）

学校教育法第30条第2項においても、いわゆる学力の三要素の一つとして「知識・技能を活用して課題を解決にするための必要な思考力、判断力、表現力その他の能力」を示しているところである。 こうした力は、例えば、①概念・法則・意図などを解釈し、説明したり活用したりする活動、②情報を分析・評価し、論述する活動、③課題について構想を立て実践し、評価・改善する活動等を通じて育成されるものとされ、小中高等学校等における言語活動等の学習活動において重視されている。

教科・科目の枠を越えた「思考力・判断力・表現力」を評価するためには、個々の教科・科目の範囲にとどまらず、複数の教科・科目を教科横断的・総合的に組み合わせて出題することが必要。

※ 「教科を超える思考力・判断力・表現力」としては、たとえば以下のような力が挙げられる

- 言語に関する思考力・判断力・表現力（読解力、要約力、表現力、コミュニケーション力等を含む。）
- 数に関する思考力・判断力・表現力（統計的思考力、論理的思考力、図やグラフを描いたり読んだりする力等を含む。）
- 科学に関する思考力・判断力・表現力（モデルをつくって説明する力、計画を立てる力、抽象化する力、大ざっぱに推定する力等を含む。）
- 社会に関する思考力・判断力・表現力（合理的思考力、歴史や社会の問題を特定し、議論の焦点を定める力、矛盾点をあらわにする力等を含む。）
- 問題発見・解決力（答えのない問題に答えを見出す力、問題の構造を定義する力、問題解決の道筋を文脈に応じて定める力等を含む。）
- 情報活用能力（情報を収集する力、情報を整理する力、情報を表現する力、情報を的確に伝達する力等を含む。）

合教科・科目型の問題の設計のイメージ（案）

1）評価する思考力・判断力・表現力（上記※）を明確化。

2）明確化された思考力・判断力・表現力が、どの教科・科目等においてどのような力として主に育成されるか特定。

例えば・・・ 言語 ⇒ 国語・英語、 数 ⇒ 数学、 科学 ⇒ 理科、 社会 ⇒ 地歴又は公民

問題発見・解決力 ⇒ 総合及び各教科・科目、 情報活用能力 ⇒ 情報

3）特定された教科・科目等において育成される力を、他教科・科目等のどのような文脈に当てはめていくことが効果的かを検討しつつ、教科・科目等の組合せを決定し作問。

別添資料4

思考力・判断力・表現力を評価する調査問題の例　（参考）

◆全国学力・学習状況調査

教科に関する調査（国語、算数・数学、理科）のうち、主として「活用」に関する問題（いわゆるB問題）

◆OECD生徒の学習到達度調査（PISA）

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、以下の3側面が扱われる。

①生徒が各分野で習得する必要がある「知識領域」
②生徒が用いなければいけない「関係する能力」
③知識・技能の応用やそれが必要とされる「状況・文脈」

◆情報活用能力調査

情報活用能力を構成する次の3つの観点から出題。

①情報活用の実践力
②情報の科学的な理解
③情報社会に参画する態度

※調査問題の範囲は、各教科、道徳、総合的な学習の時間、特別活動等で実施することが想定される学習活動とする。

◆PISA問題解決能力調査

解決の方法が直ぐには分からない問題状況を理解し、問題解決のために、認知的プロセスに関わろうとする個人の能力（進んで問題解決に関わろうとする意志も含まれる）を測ることとしている。

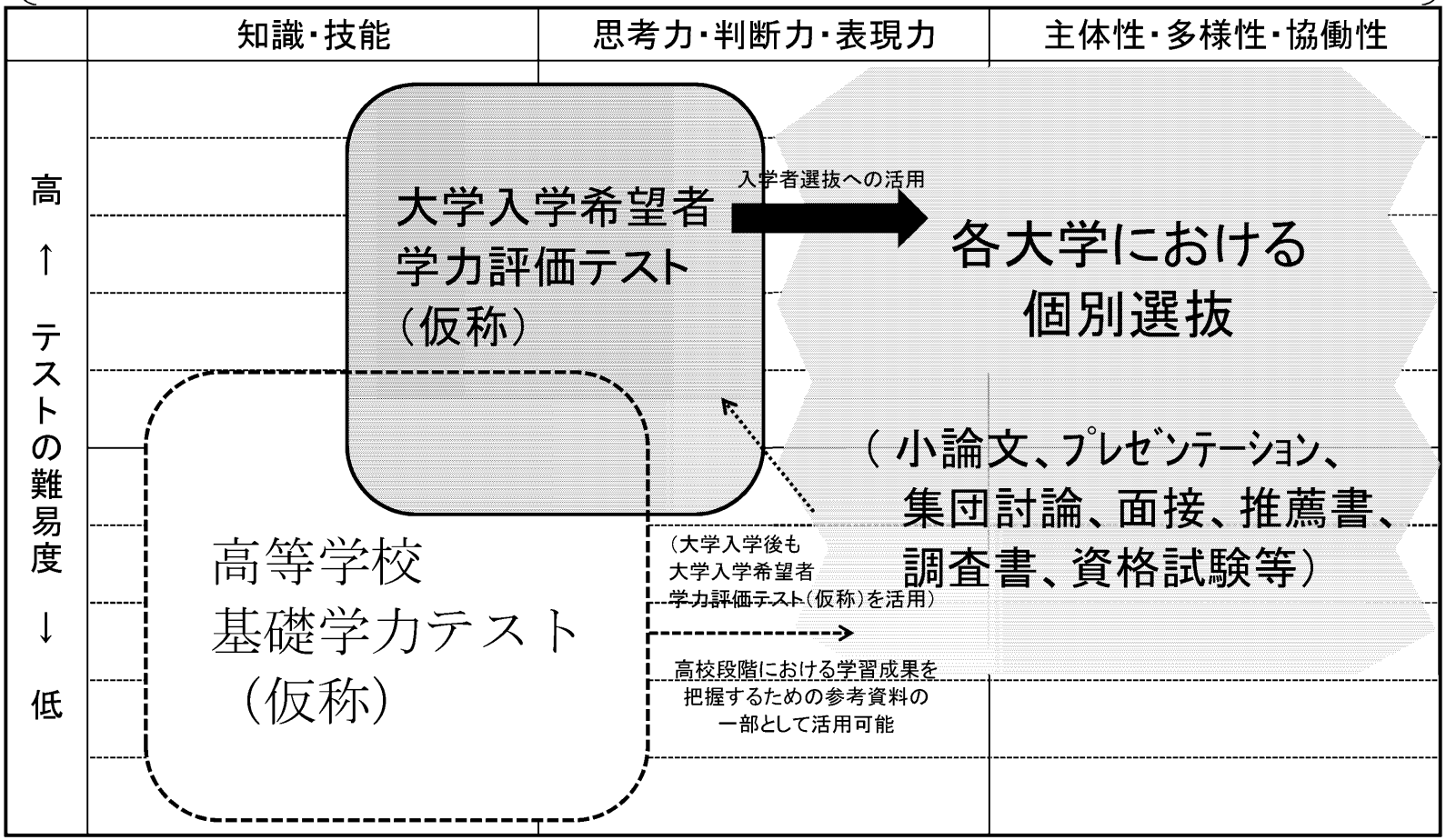
測定の対象となる認知的プロセスは、①探究・理解、②表現・定式化、③計画・実行　④観察・熟考。

◆大学入試センター「新しい試験の開発に関する研究」

教科ごとの知識・技能とは異なる、問題解決や課題遂行に必要となる基本的な能力や適性、実践的な言語運用能力や数理分析力等を評価。具体的な問題を試作し、モニター調査による識別力等の分析・評価等に取り組むなど、新しい試験の在り方を研究。

「高等学校基礎学力テスト（仮称）」と「大学入学希望者学力評価テスト（仮称）」の難易度と大学入学者選抜への活用方策のイメージ

一般入試・推薦・AO入試の区分を廃止し、入学者選抜全体において、
アドミッション・ポリシーに基づき大学入学希望者の多様な能力を多元的に評価する選抜へ抜本的に改革

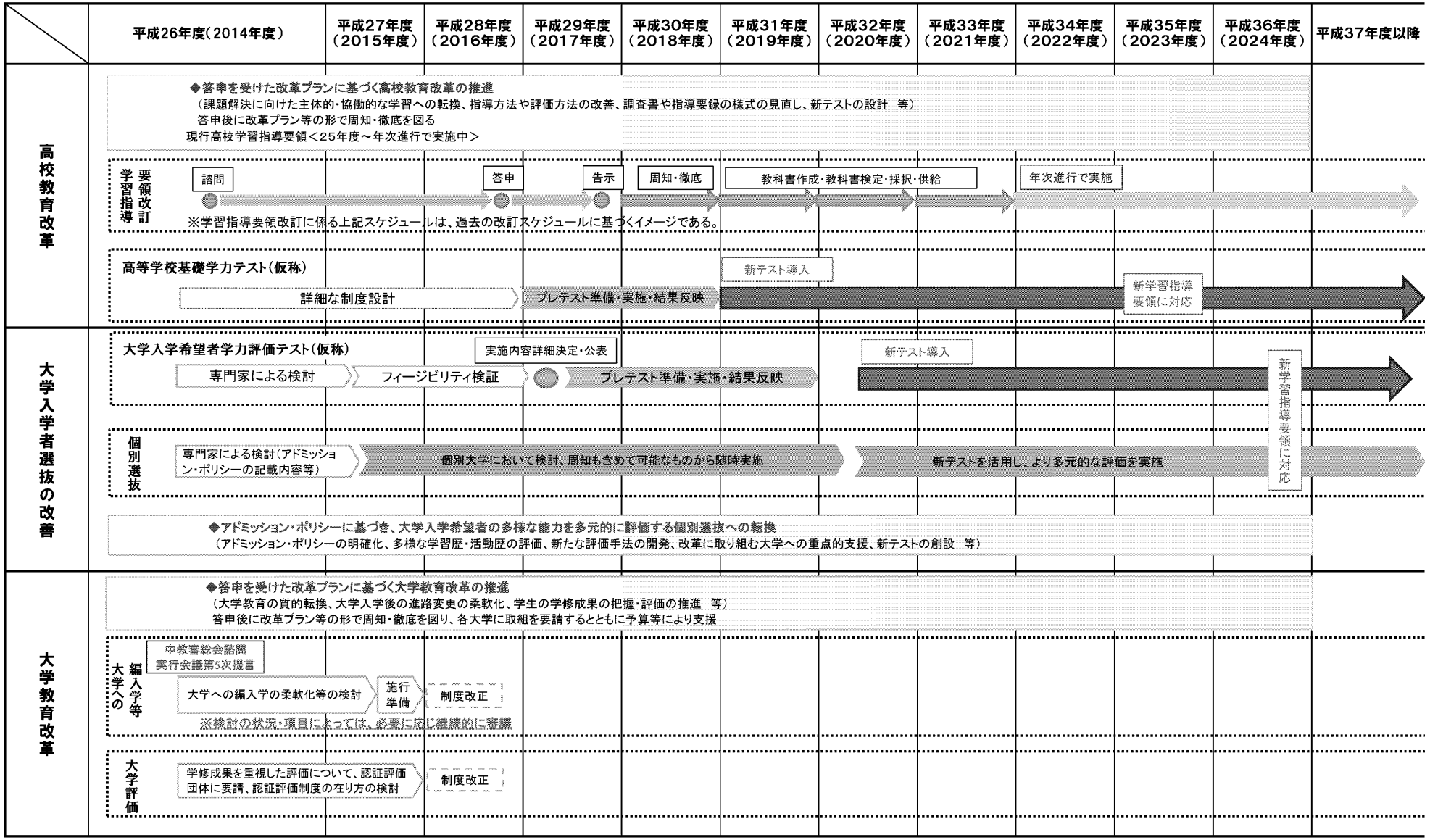


大学入学者選抜のための仕組み。
高校教育の質の確保・向上のための仕組み。

別添資料5

高校教育・大学教育・大学入学者選抜の改革スケジュール（案）

別添資料6



独立行政法人大学入試センター　入学者選抜研究に関する調査室報告書２
「大学入試の日本的風土は変えられるか」

発行　　平成 27 年 3 月 31 日

編集・発行　独立行政法人　大学入試センター　入学者選抜研究に関する調査室
〒 153-8501　東京都目黒区駒場 2-19-23
電話：03-3468-3311（代）

印刷　　株式会社　コームラ


NATIONAL CENTER FOR UNIVERSITY ENTRANCE EXAMINATIONS