講演３（翻訳）

テストと教育政策

Henry Braun（ETS，アメリカ）
翻訳　大澤　公一（東京大学大学院教育学研究科）

２００５年１２月１４日
於：東京大学教育学部棟　第一会議室

世界中の国々の政府は，人的資本の開発がその国の進歩や経済的成功にとって必要不可欠のものであると認識し，教育政策に力を入れるようになってきた。同時に，教育政策の実施においてテストが果たす役割は，ますます中心的なものとなってきている。しかしながら，測定の専門家としての観点からみると，テスト結果は適切に利用されていないことが多い。個々の特殊なケースに干渉することは別にして，テストの利用方法を改善するために測定のコミュニティーとして何ができるのだろうか。本稿では，既に提案されているテストの利用法に対して，より組織化され，先を見越した反応をもつべきであると主張する。その第一歩として，目標と手段の両方を含んだ教育政策のために，組織化の枠組を提案する。テストが果たすいくつかの役割を記述し，教育政策に対して期待される成果と意図せざる結果とを評価するための基礎として，システム妥当性の概念を導入する。システム妥当性に関する具体例を挙げ，最後に，異なる形式のテストが，いかにして従来に見られない方法で教育システムの生産性に貢献することが可能であるのかについて，提案を行う。

（本稿は講演内容および事後の質疑応答を事務局が要約的にまとめたものである。翻訳前の英文は後掲。）

1.　はじめに

本日は，テストと教育政策についてお話をしたいと思います。この話は，技術的なものでもなければ哲学的なものでもありません。しかし，私はテストと教育政策についての考え方を発展させていきたいと考えていますし，テストの重要性が日々増しているにもかかわらず，なぜテストが教育サイクルの中でまともに考慮されていないのか理解しようと思っています。

教育政策が一段と活発なものになってきていることを，私たちは教育政策が多くの国で政府の重要な関心事となってきていることから理解しています。アメリカでは「人的資本」という術語がよく用いられます。将来的には，人的資本の重要性は天然資源のそれを凌ぐことになるでしょう。日本は残念ながら豊富な天然資源に恵まれていません。そのため，国の健康や経済状態などの豊かさにとっては人的資本が最重要であると認識されてきました。

教育政策が政府活動の中で中心的な位置を占めるようになるにつれ，テストが教育政策やその実践の中で果たす役割はより重要なものとなってきています。これは，中心的な役割を果たすように志向されたテストは，（教育）システムに空いた穴（欠陥）に貼る「絆創膏」として見られることが多いからでしょう。教育改善のために私たちが行う必要のあることの中でも，テストは相対的に費用がかかりません。またテストは比較的簡単に実施することができて，かつ教育システムの様々なレベルに直接影響を与えることができます。従って，アメリカをはじめ世界中の国の政府は，政策の非常に自然な道具としてテストを考えているのです。

より優れたデザインの教育政策を実施することや，テスト結果がより教育に対して支援的で，有効な役割を果たすことは私たちがもつ普遍的な願望であると思われます。そうした願望を現実のものとするために，私たちはどのような形で貢献することができるのでしょうか？

私たちのような測定分野に身をおく人間は，テストのテ

クノロジーについてあれこれと考えてしまう傾向があります。どのようにテストに取り組もうか，どうやってテストを改善しようか，といったことをいつも考えています。私たちはこうした考察のことを，テストをより効果的なものとするためのある種の回答であると考えています。しかし，これらは回答のほんの一部に過ぎないのです。本日私が議論したいことは，測定科学者は（心理測定の専門家からテスト開発者やカリキュラム開発者などの測定の専門家に至るまで，非常に幅広い意味でこの言葉を使っています）政策に対してもっと注意を払わなければならず，また教育政策の中でテストが果たす様々な役割について理解しなければならない，ということです。

専門家たちがその理解を，自分自身の専門的な仕事の方向付けとして役立てるのみならず，科学者が見落としがちな政策的な事項により多く関わっていく契機とすることができれば幸いです。私たちは自分の専門的な仕事に集中してしまい，政策的な部分は他人に任せてしまいがちです。しかし，いまや教育政策の重要性はあまりに大きくなり，政治家たちには任せておけない重要な技術的課題が多く含まれるようになってきています（政治家たちが，自分の職掌領分をきちんとわきまえていると考えているとしても，です）。

現在のテストや教育に関する議論は，大学入試や進級試験など特定の問題に集中しすぎていると思います。さらに，人々は議論のテーマを同じ土俵で共有していないことも多く，議論が錯綜して解決策を見出すことができていません。ここで，政策に対する思考の一般枠組があれば，議論がよりよいものになると思います。その思考枠組を利用して，私たちはテストがどのように教育政策と調和するか理解することができます。またその思考枠組は，テストの改善や教育への貢献に対して，私たちがより合理的に取り組んでいくための基礎となります。

この場で皆さんにご提案するものは，さらなる議論の第一歩（たたき台）に過ぎません。この試みに対する皆さんのご意見を伺い，これをより良いものへと改善することができると考えております。

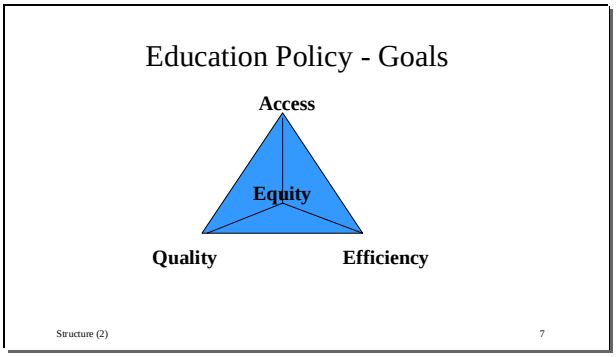
概要

まず，教育政策の構造あるいは政策の枠組から話を始めたいと思います。次に，テストが果たす役割とテストにまつわる問題点についてお話します。そしてテストと教育政策の両者について，システム妥当性（Systemic Validity）の概念を導入し，教育システムにおける生産能力（生産性）を構築するためのテストの利用方法について提案したいと思います。

2.　構造：目標，戦略，方法

教育政策の構造や枠組には，目標(Goals)，戦略(Strategy)，方法(Means)の３つの主な次元があります。目標とは政策の目的や標的のことです。戦略とは異なる方法・手段を用いて目標を達成するための全体計画のことです。方法とは意思決定や活動，そして資源の組み合わせであり，私たちはそれらを用いて政策を実行するわけです。

2.1.　目標：アクセス，質，公平性，効率性



教育政策には多くの目標がありうるのですが，ほとんどの目標に共通する大まかな基本的目標が４つあります。私たちは，その４つの基本目標：アクセス(Access)，質(Quality)，公平性(Equity)，効率性(Efficacy)によって，教育政策の目標についての思考を組織化することができます。ほとんどの国で，ほぼ全ての教育政策上の目標がこの４つの基本目標のどれかに適合すると考えられます。

2.1.1.　アクセス（Access；参加の権利，門戸）

英語の「Access」とは，どのように入るかという意味です。従ってアクセスは，例えば異なる教育水準や教育機関に入る生徒のための政府の目標を意味します。アクセスに関する政策は，時としてアクセス（門戸）を拡大する方向に動いたり，あるいはそれを制限する方向に動いたりします。東京大学のような入学競争の激しい大学ではアクセスは通常制限されており，入学許可に値する非常にレベルの高い学生を識別できる目標（入学試験の基準）が設定されるわけです。あるいは，ある人口集団の大多数に対して教育機会を開放しようとすることもあります。いずれにし

ても，私たちはアクセス（参加の権利，門戸）を問題にしているのです。

一方で，代替手段としてのアクセスが問題になることもあります。例えば，一時解雇されたり失業してしまった労働者を対象に教育やトレーニング，あるいはカウンセリングや適正な社会生活に向けた個人の取り組みを支援するための基金といった，アクセスに関連する資源を提供したい場合があります。これらは全てアクセスの側面であり，教育政策の観点からみたアクセスの目標であるのです。

2.1.2.　質（クオリティ）

質（クオリティ）についての考え方は人それぞれに違うと思います。ここでは3つの例を挙げたいと思います。一つは，学習環境における基準（Standard）です。教育政策の第一義は教育が行われるための資源や設定（環境・状況）が何なのかというところにありますので，教師の証明書，資格，能力や，学校・建物や教室の状態などを取り上げることができます。学級サイズや教室の数，生徒や教師が利用できる資源などは職場や学びの場の環境に直接関連していますから，それらはみな質的な側面であるといえるのです。

また，学習活動それ自身が学習成果の質に対する基準でもあるのです。それらは，例えば各学年に対して設定される達成基準であったり，大学卒業のための基準であったり，さらにはそれらの基準を満たすべき各コホートのパーセンテージ目標であったりするのです。例を続けると，中学校から高等学校に進学するとき生徒にある基準を達成してほしいとします。例えば，ある試験を受ける中学生の80%がその基準を満たしてほしいといった具合です。これらは全て，教育政策の質的な目標や質的な基準を構築するための一部分なのです。

質的な目標は，卒業生の能力の外部評価とも関係があるかもしれません。雇用者に調査を行なって，大卒者がどの程度仕事をこなすことができるのか，あるいは企業の中で効果的に機能するためにどの程度のトレーニングや研修が必要となるのかを聞いてみるのです。これらは質に対する外部評価であり，質的な目標と関連しているのです。

2.1.3.　効率性

効率性には様々な意味が含まれますが，教育における効率性とは人的・経済的資源が異なる教育レベルやセクターの間，あるいはその内部で適切に配当され利用されているのかといったことと関連してきます。つまり，政府の機能に関する限り，教育のために割り当てることのできる資源は限定されているので，私たちはそれらの資源を最大限に利用していると自信をもって言いたいのです。また，それらの資源を無駄遣いしていないと確信をもって言えなければなりません。例えば，コンピューターのハードウェアとソフトウェアに重点的に資源を投資したとしても，それらの効果的な使い方や，テクノロジーを創造的に学級における教育活動に統合する方法などを教員に教示・訓練しなかったとしたら，私たちは教育のための資源を効率的に利用できているとは言えません。

効率性の欠如は実に様々な場面で起こりうることなのです。私たちがよく用いる術語の一つに投資収益率（return on investment, ROI）というものがあります。ビジネスの世界では，投資収益率とは投資した資本に対して得られる利益の割合を意味しています。しかし，教育システムの中でこのような収益率を計算することは簡単なことではありません。実際，それはものすごく難しいことなのです。しかし，私たちは収益率の計算を間接的に行うことはできます。例えば，アメリカでは第９学年（高校１年）に入学した生徒のうちたった60%～70%しか，その後６年から７年の間に高校を卒業することができません。この統計は，私たちの投資収益率が極めて低いということを示唆しているのです。もし生徒が高校の卒業証書を取得できないようならば，時間的あるいは社会的な観点から，私たちの投資収益率は低いと分かるわけです。この低い投資収益率は，非常に乏しい所得の見通しとなって社会で表面化します。アメリカでは，高校を卒業していない者の就業機会やまともな賃金獲得の可能性はきわめて低いのです。よって，社会的，個人的な観点の両面から，極めて低い投資収益となっているのです。これらは全て，教育政策の枠組の中において効率性の目標が意味するものの例となっています。

2.1.4　公平性

公平性の概念は人によって違いますが，基本的には不平等でないことを意味します。例えば，どんな国であっても，質の良い教育とその門戸が，全ての関係者に平等に分配されることが重要な目標の一つであると考えられます。

アメリカでは，貧民街や貧しい地区に居住する子弟には教育機会が均等に与えられていません。その理由は，彼らが通う学校の資源が，中産階級や上流階級の子弟が通う学校の資源よりもずっと少ないからです。なぜそのようなことが起こるのかというと，アメリカではほとんどの場合，学校の教育予算は地元地域の財産税で賄われているから

なのです。貧しい地域の出身であれば税収入の基盤は弱いものとなりますから，教育に回すことのできる予算も限られてくるのです。逆に，裕福な郊外では税収入が多いため，それだけ教育に割り当てることのできる財源も大きなものとなるのです。そのため，同じ州であっても地区ごとに資源の量（財源）が異なるため，学習・教育機会に大きな格差が出てきてしまうのです。

さらに，公平性が意味するところには，機会の均等配分だけではなく，私の言葉でいうところの構造的不平等を減らすための積極的な努力も含まれます。今お話した教育予算の格差は構造的不平等の一例です。少なくともアメリカにおける構造的不平等の原因の一つとして，労働協約が教員側にどのような学校に赴任したいのかを一任していることが挙げられます。従って，非常に有能なベテランの教師は貧乏な学校に行きたがりません。そのため，最も優れた教育能力のある教師を真に必要としている生徒（財政資源の乏しい学校の生徒）には，そうした優れた教師が全くあてがわれないという構図ができあがります。この問題は，直接取り組まなければならない構造的不平等の問題の一つなのです。

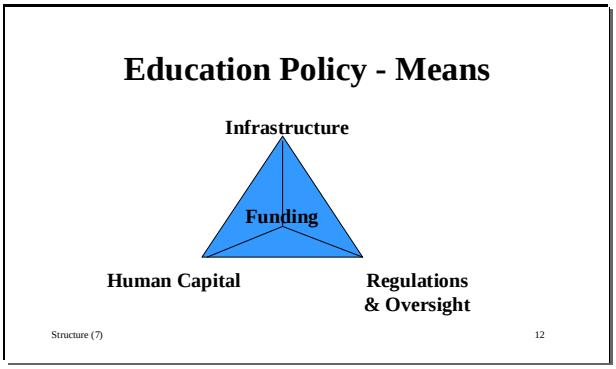
純粋に論理的な観点からみると，公平性は質（Quality）の一側面であると考える方がおられると思います。しかし，公平性の問題はその文化的，政治的な意味において重要な社会問題となっています。そのため，公平性は単独で私たちのフレームワークの一角を担うに値するのです。もう一つの理由は，公平性の概念は，それを強調しておかないと見落とされやすいということです。従って，私の議論の中では公平性は質やアクセスと同様に，教育政策の目標の一つとして重要な位置を占めているのです。

2.2.　戦略

今回の講演の準備をしているとき，戦略は各国に固有の特性があると感じました。そのため，戦略についてはあまり多くを語らないようにしたいと思います。というのは，戦略とは政治的，文化的条件によって規定されるものであり，同じ国の中でも時が経れば変わり行くものであるし，それらの様相は国によって全く異なるものだからです。日本は中央集権的な教育制度をもっており，多くの戦略は中央政府から降りてきて，各都道府県で施行されることと思います。これに対してアメリカでは状況は全く異なります。というのは，米国憲法の下では教育の法的権限は各州にあるからです。連邦政府は教育に対してごく限られた役割しか果たしません。ですから，州によって教育戦略がまったく異なったものとなっているのです。また，連邦政府はある特定の方向に向けて全ての州に対し影響力を行使しようとする点も異なります。このように，全体戦略や戦略の開発方法などは環境に依存するものであり，ここでは多くのことを語ることは避けたいと思います。

2.3.　方法：財源，基盤，人的資本，規制と監督

方法とは政府が目標を達成するために，直接的に，あるいは他の保管人への活動を通して行使する道具や意思決定，活動などを意味します。ここでは，4種類の異なる方法について述べたいと思います。それらは，財源(Funding)，基盤(Infrastructure)，人的資本(Human Capital)，そして規制と監督(Regulations&Oversight)です。



2.3.1.　基盤と人的資本

基盤や人的資本とは，教育において政府が目標を達成するために行う投資の基本的側面であると私たちが考えているものです。それらには，教育に用いられる物理的施設，つまりは建物や不動産などのハードウェアがまず該当します。また，教職員のほかにも必要となる資源，教科書，コンピュータ，その他の消耗品なども含まれます。人的資本とは教育システムの中で何らかの役割を担う人間のことであり，例えば教員，校長，学校レベル，郡レベル，あるいは中央省庁レベルでの指導者，また民営部門なども含まれます。日本でもアメリカでも同様に，民間企業は教育において重要な役割を担っています。例えば塾や教科書出版会社，テスト出版会社，カリキュラム出版会社などです。これらは全て基盤や人的資本に含まれ，教育目標を達成するための方法なのです。

2.3.2.　財源

財源とは当然お金のことで，それ以外の何者でもありません。財源にはいろいろと異なる側面があります。つまり，その出所ということです。日本の事情には詳しくありませ

んが，アメリカでは教育の財源は連邦政府，州政府，地方自治体からの複数の出所があります。さらに，私設基金や慈善基金があり，教育に対して年間何十億ドルという投資が行われています。そうした組織は，財源という観点から教育に対して重要な影響力をもつことになります。

財源の分配が異なる地域や学校，あるいは地域内の生徒のレベルでどのように行われるのか，という問題があります。アメリカでは，連邦政府は年間数十億ドルを教育資源として支出しますが，それらは主に貧しい生徒や，多数のそうした子どもたちのいる学校に通う生徒のために使われます。この財源はTitle Oneと呼ばれます。こうした数十億ドルを特殊な生徒集団に投入する特別な財源の流れがあるのです。この点が，私たちが財源やその出所，分量，そして分配方法について知っておかなければならない理由なのです。

財源が短期のものなのか，それとも長期にわたるものなのかといった財源の時間的範囲についても知っておかなければなりません。お分かりのように，日本やアメリカの教育システムを劇的に改善させることは，1年や2年では成しえません。教育システムが極端に複雑なのです。そこでは多くの慣性が働いていて，容易に改変を許しません。従って，もし変革を望むのであれば，そのための資源を長期間にわたって投入しなければなりません。少なくともアメリカでの経験から，こうした試みは政治的な色合いを帯びることが非常に多いのです。つまり，政府が変わるとそうした施策も変化し，改革に向けた動きはかなり弱体化します。ですから，財源投資の時間的な見通しという側面は，財源の問題を考えるときに非常に重要なのです。

2.3.3. 規制

規制とは法律のようなものですが，通常はより官僚的，行政的な意味合いをもちます。州政府や連邦政府が新しい法案を通すときは，「どのように物事が起こるべきか」という点について提案がなされます。しかし，そうした内容は決して十分に具体的なものとなってはいません。その結果，こうした法案は教育部や教育省などの適切な部署によって解釈されることになります。そして，具体的で詳細な手続きを記した何千ページにもおよぶ規制ができあがるのです。ここで，教育システムの各構成要素がどのように機能するか，という点を例に挙げます。例えば，早期幼児教育，初等教育，中等教育といった教育段階から私たちは何を期待しているのでしょうか？私立学校についてはどうでしょう？アメリカでは，多くの就学前教育が政府や自治体の学校区域ではなく個人によって提供運営されています。こうした私立学校を行政的に監督する規制はどのようなものなのでしょうか？どのような資格があればそうした教育を提供することができるのでしょうか？どのような規則が新しく必要となるのでしょうか？どのようなスタッフが必要なのでしょうか？アメリカにはたくさんの私立学校があります。宗教的な私立学校やチャータースクールもあります。中には無宗派のものもあります。つまり，多種多様の学校と，それらの運営を司る一連の規制が存在するわけです。

教育システムの構成要素やレベルの間にある関係に適用される規制もあります。例えば，教育システムのあるレベルから別のレベルまでどのように生徒が移動するのか，といったことです。異なるレベルの間でどのような共同関係が必要となるのでしょうか？規制は学校スタッフの資格証明や必要条件をも規定するのでしょうか？教員にはどのような資格が必要とされるのでしょうか？ベテラン教師は，教員資格の再申請を行う必要があるのでしょうか？アメリカの多くの州では教員免許の有効期限は終身ではなく，再認証する必要があります。彼らは定期的に更なる専門的訓練を積む必要があるわけです。教員は，5年や10年区切りで教員免許を更新するためのコースを取る必要があります。校長にも資格認証のための条件が設定されています。

2.3.4. 監督

それでは，学生，保護者，教員，経営者などの各グループにある権利と義務はどのようなものでしょうか。カリキュラムやアセスメントのための必要条件は何でしょうか。規制は，各学年に設定されるカリキュラムや詳細事項の水準，学生に課されるべきテストの種類，テストがもつべきカリキュラムとの関連性などについて，詳細に述べるのでしょうか？これらは全て規制によって規定され，監督とは規制と非常に密な関係にあります。しかし，監督と規制はやや意味合いが異なるのです。監督は，モニタリングとしての側面がより強いのです。言い方を変えれば，システムや組織の異なる部分が現実的にそれらの規制に従っているのか，そうでないとしたらなぜなのでしょうか？そしてもしそうなら，それらの規制は質，アクセス，効率性，公平性という目標の観点からみて，望ましい効果をもっていると言えるのでしょうか？よって，監督は私たちがどのようにシステムの機能をモニターするのかといったことと関連しています。監督は時として（説明）責任に関する意

思決定（結果に対する判断）と関連してきます。物事がうまく運んでいるときにシステムの構成員である個人に報奨を与えるべきかどうか，反対に物事がうまく運んでいないときには何をすべきか，教員を解雇するのか，校長を更迭するのか，強制的に学校改革に着手させるのか，などの意思決定を行います。これらは全て監督の一部なのです。

監督にはもう一つ重要な側面があります。これは公式の規制というわけではないのですが，監督あるいは（説明）責任の結果が，当局や一般大衆の様々なレベルにどのように伝わるのか，というものです。なぜならば，監督が最も強い影響力をもつのは，モニタリングの結果（説明）が大衆に理解できる形で，また彼らが何らかのリアクションを起こせる形で伝わったときだからなのです。

3.　テストの役割

今までの議論からお分かりの通り，この枠組は必ずしもテストを必要としていません。むしろ，テストの問題を考えることなしに教育政策の枠組を構築できてしまうのです。原理的には，私たちはテストを全く含まない教育システムを想定することができるのです。しかし，現実にはそうしたシステムはほとんどの国で存在しないでしょうし，これからも存在するとは考えにくいと思います。

政策の目標が社会によって規定される一方で，テストは教育政策の道具（手段・方法）の一つとなっています。テストは政策を実行するための手段の一つにすぎませんし，規制や監視・監督下で運用されるのが通常です。テストの役割は，規制や監督のための道具を提供することであり，ひいてはそれが政策目標を達成することを意味するのです。通常，テストは政策目標と直接のつながりをもちませんが，政策目標を達成するためのよりよい道具として開発されてきた，私たちが利用できる体系的なツール（道具や手段）の一環であるのです。

3.1.　アクセスの目標を達成するために

ある教育プログラムに入学するため，就職のための選抜，専門的な資格証書の付与などのために，テスト結果は明確な形で利用されています。これらはアクセスの目標を達成するための方法としてテストを用いている例です。選抜目的に対しては，どの学生が入学するにふさわしいかを決定するために，候補者のスクリーニングを行う手段の一つとして私たちはテストを利用しています。その他の例では，例えばe -learningにおけるテストを取り上げると，テストの役割は学生を締め出すことではなく，正しい教程に生徒を導き本人にとって最適なコースを選択させることにあります。そうすることで，生徒は学習の要求水準や，教育やトレーニングの目標と彼らのもつ個別の目標，そして達成水準の組み合わせとしての目標を達成できるのです。

テスト結果がアクセスにおいて意味をもつこともあります。適当な公的証明が存在しないとき，テストの結果をもって個人の習熟の証拠とする場合があります。例えば，アメリカではあるテストで十分な成績を収めることで，いくつかのコースの履修が免除されることがあります。彼らは次のレベルのコースに進むことができるのです。これは，テスト結果がアクセスに影響を及ぼす例といえます。しかし，繰り返しますがテストは目的ではありません。テストはあくまでも目的を達成するための手段の一つに過ぎないのです。

3.2.　質（クオリティ）の目標を達成するために

質的な目標はどうでしょうか。ここで，事情はややトリッキーなものとなります。といいますのは，私の考えではこれがテストに関係する問題の一つだからというのもありますが，私たちは質を定義する目的でテストを用いることがあるからです。これを行ってしまうと，私たちは危険な領域に足を踏み入れることになります。というのは，そこでテストそれ自身と教育目標とを混同し，同義語と考えてしまう危険性があるからです。それでも，理論的な観点からは，教育目標は決してテストを主として定義されることがあってはなりません。しかし学校側は，例えば80%の生徒が第10学年か第11学年の期末に州試験に合格してほしい，という表現で学校教育の質を定義しようとします。そのため，テスト結果が質的な目標の定義として直接用いられることになります。問題は，テストの出来（性質）があまりよくないとき，それをもって質的な目標を決定してしまうと，教育的な観点から非生産的状態に陥ってしまうことが多いということなのです。そのような場合には，テスト結果と教育目標との間に距離をとっておくのがよいでしょう。実際にアメリカでは，テストに批判的な人の多くは，特に多くのテストはそれほど出来（性質）がよろしくないという理由で，私たちがテスト結果という狭い視野においてのみ教育の質を定義していると主張しています。彼らの主張によれば，教育の（質的な）定義はもっと視野の広いものであるべきだというのです。この点については後ほど触れたいと思います。

仮にテストがよいものであれば，テスト結果を用いることの有効な側面としては，教育システムに質的に欠けてい

る部分に光を当て，問題点を指摘してくれるということなります。例えば，あるテストを行うことによって，ある地域のある学校の生徒たちは代数学の能力に問題があることが分かり，その事実が地域や学校経営者にとって，代数学の教育の質を向上させる必要がある，あるいは他の問題が存在するかもしれないといったことを教えてくれるかもしれません。その意味で，テストは非常に有効な警鐘としての機能をもっていると言えます。

3.3.　効率性の目標を達成するために

モニタリングを通して，効率性の観点においてもテストは何らかの役割を果たします。なぜならば，テストは私たちが教育システムの中でどの程度の効率性を達成できたのかを理解する助けになってくれるからです。例えば，国家規模のテストや大規模な調査を行なうことで，異なる地区や地域の間で教育システムの機能にどのような違いが見られるのかを比較することができます。アメリカではNational Assessment of Educational Progress（NAEP）という大規模な調査が，第4学年と第8学年の読解力と数学能力を対象に2年おきに実施されています（不定期に他の科目の調査も実施しています）。NAEPは全ての州の非常に大きな学生サンプルに対して実施されます。各州は，他州とNAEP成績の比較を行い，教育に投じた資源や資本が効率的に利用されているのかどうか，何らかの判断を下すでしょう。他州との比較は，自州の生徒たちが他州の生徒たちに比べて相対的にどの程度のパフォーマンス（成績）を残しているのかを比べることで行われます。各州は独自の試験制度をもっていますから，州試験のレベルで州間の比較を行うことはできません。しかし，NAEPを通せばそうした比較ができるのです。

集団レベルのテスト結果を用いた特殊な分析が，地域や学校，教員といった個別のユニット（単位）の教育成果への貢献度を評価する目的で行われることがあります。こうした分析は，ある学級や学校，あるいは地域の中の生徒集団についてのテスト結果を利用して，そのユニット（学級や学校など）の教育効果について何らかの知見を得ようとしているわけです。これは単純なことではありませんが，テスト結果の将来的な利用方法の一つであり，教育システムをモニタリングする中で何らかの役割を担うものです。

さらに，TIMSSやPISAといった国際調査を行うことで，国家間の比較を行うことができます。過去のPISAの結果を見ると，アメリカも日本も明らかに成績が良いとはいえません。他の国との比較において，それぞれの国がいかにしてパフォーマンス（学生の成績）を向上させるかといった議論が多くなされています。

従って，私の考えではテストは警鐘としての役割を担うことができ，教育システム内部で起こっている状況に対して，システム内部（例えば国内）の比較を行うことによってあなたはある程度満足することができるかもしれません。しかし，不意にフィンランドやシンガポール，大韓民国などとの比較に晒されると（システム間の比較を行うと），たいして満足できる成果を挙げていないことが分かるのです。

3.4.　公平性の目標を達成するために

最後に，テスト結果は公平性をモニターするために利用することもできます。例えば，地方と都市部の学生の間に，学習の達成レベルに関して構造的な格差が継続して観察されるのなら，そこには公平性の問題（構造的不平等）があるといえます。従って，テスト結果は公平性の欠落をモニターする手助けとなってくれます。また，テスト結果を基にして補習的な意味合いのテストを実施することで，公平性の格差を改善しようとする必要性のある学生に対して，何らかの教育的支援を与えることができます。ですから，テストは公平性の目標達成に向かって努力していく中で，非常に建設的な役割を果たすことができるのです。

4.　テストの問題点

ここで私たちがもつ疑問は，テストが教育政策の中で重要な役割を多く果たしうるのであれば，なぜテストは現在こんなにも広く批判されているのだろうかということです。この疑問に対する回答は，いくつかあります。

4.1.　回答（1）：テストの性質と適切性

第一の回答は，テストがその役割を果たすことに成功するか，というものです。テストの役割とは，目標を達成するための手段の一つです。テストが成功するかどうかは，目標に照らしたときの，テストの性質や適切性に依存します。実際，テストの性質や適切性といったものは，私たちが意図する目的や目標に対して不十分であることが多いのです。私たちが現在利用している多くのテストは，それほど良くできたものではありません。これは，例えばそうしたテストが専門的に開発されていなかったり，仮にそうであったとしても，経済的な制約などにより複雑な素材ではなく非常に単純な素材をテストするものとなってしまったりしているからです。ですから，生徒や教師はこれら

のテスト結果で簡単にゲームをすることができます。生徒はテストを受けてよい成績を残すことができますが，実際には教育的に進歩していないのです（テストの性質がよくないから）。これが第一の回答です。問題の一部分は，テスト自身の性質にあるのです。

4.2. 回答（2）：キャンベルの法則

第二の回答は，キャンベルの法則と私たちが呼んでいるものです。Donald Cambel博士はアメリカの著名な社会科学者で，彼の法則とは次のようなものです。「いかなる定量的な社会指標でも，それが社会的な意思決定の場でより多く用いられるほど，その指標は退廃への圧力を受けやすくなり，その指標が監視するはずの社会的プロセスを歪曲し，堕落させがちになる。」ここでキャンベルが言っているのは，いかなる定量的な指標（教育システムであれビジネスシステムであれ，あるシステムがどのように機能しているのかを私たちに教えてくれる量的な指標）であっても，それが力をもてばもつほど（例えば，その指標に付随する結果の社会的重要度が増せば増すほど），その指標はシステムをおとしめようとするものたちによって，より大きなプレッシャー（退廃への圧力と呼んでいます）にさらされることになる，ということです。つまり，彼らはその指標に関してうまくやっていきたいと願ってはいるものの，実際にはその指標が監視するプロセスに関してうまくやっていないのです。

ビジネスの世界を例に挙げたいと思います。ある企業の株価の観点から，ビジネスの経営陣の評価をしようとします。そのとき経営陣は，株価をあの手この手で操作して，あたかもビジネスが順調に行われているように見せかけることができます（実際にはビジネスが必ずしもうまく行っていないにも関わらず）。アメリカや日本は似たようなビジネスシステムであるかもしれません。アメリカではEnron，WorldCom，Tycoといった大企業の不正会計疑惑がありました。従って社会的指標，この場合は株価という経済指標は，あまりにも多くのものがそれに依存していたがために退廃への圧力に晒されてしまったのです。

この事例を教育システムに応用することはそれほど難しいことではありません。もし，校長や教員の（教育的）成功が，彼らの生徒がどの程度テストでよい成績を上げたのかに依存するとすれば，校長や教師には学生をしっかりと指導することが望まれるわけです。しかし，彼らは生徒たちを指導することなしに，生徒のテスト得点を直接上げてしまうという手っ取り早い近道を見つけてしまうこともあります。これは教育システムの退廃の一形態です。キャンベルが言うには，これは非日常的な出来事では決してないのです。彼の議論では，これは社会的，政治的システムの機能の中の自然な部分なのです。

社会的に影響力の大きいテストのリスク

ー　カリキュラムの狭窄

具体的な例として生徒，教員，あるいは校長に重大な影響をもたらすテスト（high-stakes testing）を取り上げると，システムの退廃としてカリキュラムの狭窄という問題に直面します。これは，テストの対象（出題範囲）にのみ教育の焦点が当てられてしまい，カリキュラム全体に教育が行き渡らない状態です。そして，もしテストが良いテストではなかったとしたら，例えばテストがカリキュラムの特定の部分に絞った多肢選択式の項目ばかりを含んでいたとしたら，テストに出ない部分のカリキュラム内容が多肢選択項目の出題範囲と同等か，あるいはもっと重要だったとしても（しかしそれらをテストすることは難しいとして），学校システムはテストに出題される範囲に教育の焦点を絞ってしまうことでしょう。分かりやすい例を挙げますと，事実に関する知識・情報や単純な論理的スキルをテストするのは簡単ですが，創造性や問題解決能力を測定することはずっと難しいものです。さらに，その種の情報を引き出すために設計されたテストは非常に高額なものとなります。もし経済的な制約がある場合，私たちは複雑な形式の課題をテスト構成から排除してしまいがちです。その結果，カリキュラムの狭窄につながるテストが出来上がってしまうわけです。

教員資源の不公平な配当

アメリカでは現在新しい責任法の下で，ある教育段階を超えようとしている（例えば進級や進学を控えている，など）特定の学級や学年の生徒たちに焦点が当てられ，多くの教員はそのもてる力の全てを要求水準のちょうどすぐ下にいる生徒たちに注ぎ込んでいます。なぜならば，彼らこそが集中的な教育の恩恵を最も多く受けることが見込まれ，その結果，次の教育段階に進むことができるようになるからです。非常に優秀な生徒たちのことで，教員が思い悩む必要があるでしょうか？彼らはクリアするべき要求水準を既に大きく超えているではありませんか。それでは，基準よりはるかに下にいる生徒たちは？彼らは要求水準に追いつくことは恐らくできないでしょうから，悩む必

要はありません。その結果が，非生産的で不公平な教員の人的資源の割り当てなのです。多くの研究によってこれが現実に起こっていると証明しています。

生徒たちの不公平な取扱い

これは，生徒の不公平な取扱いにも繋がることです。もし，学校や校長が一定の教育水準を満たす生徒の割合に対して（説明）責任を負う立場にあるとき，一部の生徒は退学を勧められたり，あるいはできの悪い生徒の多くは規律面で問題を抱えているため，テスト実施の直前に停学処分を受けたりすることがあります。そうすると，できの悪い生徒たちは（少なくともそのテストが実施されるときには）学校にいないことになりますから，学校全体の見かけの成績は上昇するわけです。このような行いを，私たちは教育的な観点からみて生産的であるとは言いません。

蔓延する不正

生徒や教師による不正も発生します。これらは，学校の教育に対してもつ（説明）責任という，現実世界の中で起こっているキャンベルの法則の具体例に過ぎないのです。

4.3.　回答（3）：質の悪いテストがより深刻な問題の呼び水となる

第三の回答は，第一と第二の回答の組み合わせとなります。その要点は，キャンベルの法則が述べるように，テストにまつわる問題の多く（おそらくはそのほとんど）が規制や監督に関連するテストの役割ゆえに不可避的であり，問題点がテスト固有のものであるのならば，ある特定の文脈における質の悪いテストがキャンベルの法則をより現実のものとしやすい，ということです。言い方を変えれば，キャンベルの法則はほぼ全ての社会的な文脈で有効であることが示されたものの，テストの品質がよければそれを受けるものたちがシステムの廃退に携わる可能性は低くなる，ということを私は言っているのです。

例えば，飛行機のパイロットになるためには二種類のテストを受けなければなりません。一つは筆記テスト，もう一つは飛行（実技）テストです。筆記試験については，非常に質の良いテストを提供することができるかもしれません。しかし，教科の内容を知らずしてもうまく勉強することで試験を切り抜けることができるかもしれません。しかし飛行テストに関しては，試験に合格するための方法は唯一つ，実際に飛行機を操縦する以外にはありえません。それができなければ，あなたは一緒に搭乗した人々と共にお墓に入ることになってしまいます。あなたが本当に行おうとしている教育活動やトレーニングに近いテスト，すなわち真正なテストというものには，ごまかしや小細工は通用しないのです。そのようなテストから得られる数量的な指標は，あなたが実際に行っていることと密接に繋がっているため，システムを廃退させることは困難なのです。

今，アメリカでは非常に有名な言い回しがあります。「教えるに値するテストを作ろう」これは言い方を変えれば次のようになります。「非常に出来がよく教育目標と明確に繋がっているため，生徒たちにそのテストのための教育を行うことが，そのままカリキュラム内容の教育を行うことに繋がる，そんなテストをつくろう」テストとカリキュラムの距離が離れれば離れるほど，キャンベルの法則による予測が現実のものとなりやすくなります。私の考えでは，この社会の中で行われるテスト批判の多くは，キャンベルの法則に絡んだものです。テストが理想ほどよくできていない場合や，特に教育目標をテストの観点から設定し，教育システムに（適切な能力が備わっていない場合でも）ある種の結果を求めて過剰な圧力がかかっているような状況でキャンベルの法則が働くことに対して，人々は反応しているのです。

5.　システム妥当性

それでは，こうした問題に対する答えとは何なのでしょうか。もちろん単純明快な答えなどはありませんが，ここで私たちがシステム妥当性（Systemic Validity）と呼んでいる概念を紹介しましょう。Systemicという形容詞はシステムに関係があります。Validityとは，ここではシステムにとって価値があり適当であるという意味です。この概念についてもう少し詳しく定義していきたいと思います。

教育目標を達成するためにテストの貢献度を高め，またテストを改善していくつもりがあれば，私たちは大まかに２つの方向性を選択することができます。一つはトップダウン的なもので，私たちが教育政策を施策するとき，それを達成するための手段を実行する中でシステム妥当性の発想を用います。私たちは目標を達成するためにテストを利用しますが，この点について少し言わせていただきます。二つ目は，ボトムアップ的なもので，システムがより良く機能し目標を達成できるよう，システムの能力構築を含んだインフラストラクチャーを，テストを利用して強化していきます。両者の方向性は共に実行可能なアプローチ方法ですが，教育システムの中でテストを最大限に利用しようと考えるならば，トップダウンとボトムアップの両方を利

用する必要があると思います。

南アフリカの研究者であるAnil Kanjeeと私が書いた論文の中で，システム妥当性の定義について以下のように触れています。「アセスメントの実践が組織的に（システムとして）妥当であるとは，アセスメントを行うことでアクセス，質，公平性，効率性のうち一つ以上の側面を，残りの側面や水準について過度の悪化を引き起こすことなく促進させることを支援する有用な情報が生み出されるときである」ここで私たちが得たい要点はこういうことです。政策を開発するときに4つの目標の中の一つだけに焦点をあて，その政策が焦点を当てた目標について現状を改善させる，という言い方をすることが非常に多いのですが，その際に，その目標を改善させることにより，ほかの目標に関連するそれまで以上に深刻でさえある問題を新しく作り出していることがある，という事実を無視してしまったり見過ごしてしまったりすることがあるのです。そのような場合，システム全体として考えると何も改善されていないということになるのです。

幾つか例を挙げたいと思います。アメリカではここ15年ほどの間，一学級の人数を減らす動きがあります。いまの学級の規模は大きすぎるというのです。一クラスに25～30人の生徒がいるのは多すぎで，教員一人当たり生徒20人が適当であろうというのです。これが合理的な人数であるということを示す研究もあります。それで，1990年代中後半のカリフォルニアで，次のようなことが言われました。「よし，我々の教育システムを改善しよう。規則を発行して，初等教育ではどの学年でも学級の人数が20人を超えないようにしよう」これは一見して良い案だと思われましたが，実際にこの法案をどのように施行すればよいのかを考えなければなりません。まず，教室の数をかなり増やさなければなりませんでしたが，そのための資金がありません。そこで彼らは，トレーラーをもってきて校庭に駐車しました。これで，確かに追加の教室（としてのスペース）を確保することはできましたが，当然理想的な状態ではありません。そして，別の問題が発生しました。教員を何処から調達して来ればよいのでしょうか？資格のある教師は十分にいませんでした。教室の数を35%増やしたとすると，教員も35%余計に必要になりますが，増員分を埋めるだけの教員はいません。そこで，無免許教員を採用したり，臨時の資格証書を発行したりしなければなりませんでした。その結果，教員の質が低下しました。学級のサイズは小さくなりましたが，多くの生徒が質の悪い無資格の教員から教育を受けることになってしまいました。これで状況が好転したと言えるのでしょうか。

この法律は一見すると聞こえが良いものですが，果たして質の高い教育という目標を達成することができたのでしょうか？上記の状況下で，生徒たちは次の教育段階に進むための準備をするために，より広く深く学ぶことができるのでしょうか。私は，答えはNoだと思います。

次に，1980年代のニューヨーク市の例を挙げ，この問題の別の側面を説明しましょう。ニューヨーク市にはニューヨーク市立大学という４年生カレッジと２年生カレッジの両方を有する独特の大学制度があります。何年も前のことですが，このシステムは非常に質の高い教育システムでした。大学は非常に厳格な入学要件，非常に優秀な教授陣を擁し，多くのノーベル賞受賞者を輩出しました。しかし，80年代か90年代を回るころ，ニューヨーク市に以前には見られなかった移民集団が入り込んできました。彼らの多くはメキシコやラテンアメリカのスペイン語圏の国々からやってきました。彼らは英語を話すことができず，大学入学要件を満たすことができませんでした。

誰もがシステムに対して等しく税金を納めますから，大学にも誰でもアクセスできるように政治的な圧力がかかりました。そこで制度を変更したのです。「これからは門戸を広く開放します。高校の卒業証書をもっている人なら誰でも受け入れます」残念なことに，アメリカでは高校の卒業証書は第三次教育（大学レベルの教育）を受ける準備ができている保証にはなりません。殺到する学生に大学の門戸を開放した結果，名目上カレッジに在籍しているだけの学生のための，初等レベルの読解，表現，算数などの補習教育に多くの大学資源が費やされることになりました。その結果，大学システムの質と効率性が低下したのです。アクセスという一つだけの目標に固執し，それが他の目標に対してどのような影響を与えるのかを無視してしまうことで，ニューヨーク市立大学の一件は私に言わせればシステム妥当性の欠如という結果となりました。結局のところ，大学システム全体にとってのみならず，学生たち自身にとっても教育的な観点からみて非生産的な結果となってしまいました。本当に大学で学ぶための準備が整っていた学生たちは，大学のシステムから教育的な恩恵を受けられなかったからです。現在，ニューヨーク市の大学制度は以前のような選抜を行なう政策に立ち戻っていますが，この長い期間で大学はすっかり破壊されてしまいました。

システム妥当性（続）

より一般的な教育政策をもつことについて私たちは話

をしているのであり，この点に関して私たちはシステム妥当性の発想をそのまま利用することができると思います。教育政策は，私たちが意図する4つの目標，アクセス，質，公平性，効率性の一つ以上の促進を，残りの目標に関して後退させることなく導くような意思決定や行動と結びつくとき，その教育政策はシステムとして妥当であると言えます。ここで私たちに要求されるのは，「意図した政策が特定の目標に焦点を当てていたとしても，常に広い視野をもっていなさい」ということなのです。その政策が意図していない他の目標に対してどのような影響を与えるのかを考察し，システムがその他の目標に悪影響を与えることなくその目標を達成できるようにしておかなければなりません。

システム妥当性という術語は，いろいろな意味を志向して用いられてきましたが，その多くは純粋なテスト場面で典型的に参照されてきました。Heyneman（1987），Hyneman & Hansom（1990），Frederiksen & Collins (1989)らは，テストとの関わりの中で，特にテストの波及効果（backwash effects）との関連でこの術語を用いてきました。つまり，テストが非常に良いものであれば，テストが要求する水準を満たすためにより高いレベルで教育を行う方向に教員を促していく，という意味においてそのテストはシステム的に妥当であるのです。これは正の波及効果であり，従って，ある意味でシステム妥当性に貢献し，あるいはその具体例となるのです。

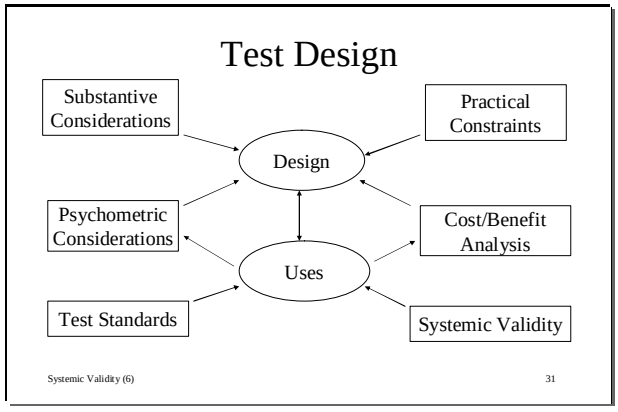
他に関連する概念は，ETSでの私の元同僚であるSamuel Messick(1989)による，結果妥当性の基準です。彼によると，テストの妥当性について考えるとき，テストの理論的・経験的正当化である構成概念妥当性に思考の範囲を制限する必要はなく，テストがどのように社会で用いられるのかといった観点について考えなければならないとしています。結果妥当性はテスト結果がどのように解釈されるのかということで，それはテストの妥当性の全体枠組の一部分であるべきだと議論していました。結果妥当性の概念はシステム妥当性と密接に関連しています。

システム妥当性をどのように強化するか

どうすればシステム妥当性を強化していくことができるのでしょうか？テストという観点から，どのようにシステム妥当性の概念にアプローチしていけばよいのでしょうか。最も明らかな方法が二つあります。一つはテストデザインを改善すること，つまりテストの質的な側面にもっと注意を集中することです。これは，私が前にお話したキャンベルの法則についての議論，つまり私たちが使うテストは現実的には本来あるべき姿に比べてそれほどよくできていないし適切でもなく，その事実からシステムを退廃させる負の影響の多くが発生する，という予測に基づいています。いま一つは，私たちは適切なテストの利用に関するガイドラインに従う必要があるということです。例えば，生徒や教師に対して影響力の大きな意思決定を行うとき，ただ一つのテストにのみ基づいてそれを行ってはならないということです。常に幅広い情報を利用するように努めなければならないのですが，私たちはとかくたった一つのテスト（結果）から重要な意思決定を行ってしまいがちです。当然，そのテストに過大な圧力をかけてしまうことになり，結果としてその指標（テスト結果）を廃退させる圧力となってしまうのです。

さらに三点目の指摘をするならば，計画立案にはより現実的なシナリオが用意されるべきであり，法律や規制を制定する前に，それらがシステムの能力の期待値が与えられたときに，時間の経過に伴ってどのように機能するのかを考察しなければなりません。ですから，カリフォルニアの例に話を戻しますと，学級の数が１/３増えればスペースや教師も１/３余計に必要になると気付くのに，それほど思考は必要とされなかったはずなのです。一体それらを何処から調達できるのでしょう（そんなことできるわけがないのです）。しかし，どういうわけかこの点に思考を至らせた人が皆無だったのです。

テストデザイン



テストデザインについてはどうでしょうか。ここでの話は，テストデザインのスキーマの一種です。システム妥当性についての理解がどのようにテストデザインに影響を与えるのかを示したいと思います。主要科目に関わるあれこれや，テストのコストといった現実的な制約，どのようにテストが実施されるのかなど，私たちには重要な検討事

項がいくつもあります。これら全てがテストデザインに影響を与えるのです。妥当性や信頼性などの心理測定学的な検討事項やコスト-収益分析も，当然のことながらテストデザインに影響を与えます。

それと同時に，テストの利用によってテストデザインが影響を受けることも明らかです。テストデザインはバックグラウンドで機能するプロセスであるべきだと主張する人もいます。しかし，私たちはテストの利用について考えるときに限ってテストの基準やシステム妥当性について話ができるのです。私の考えでは，テストの異なる利用法についてシステム妥当性の観点からもっと真剣に考えることができれば，結果として心理測定学的な検討事項と同様にテストデザインに対して影響を与えることになるでしょう。

具体例：（説明）責任*のためのテスト

具体例を挙げたいと思います。私たちがテストを（説明）責任*を果たすために用いるとき，つまり地域や学校，教員をテスト結果によって評価しようとするとき，そこに顔を出す疑問があります。評価対象は，報酬から処罰，放校（解雇）までの幅広い結論に至る評価活動を行い，教育において十分に有能な職責を果たしているのでしょうか？

この点に関する単純な理論的解釈は，社会的な影響力の大きいテストは，人をより勤勉な方向に向けて刺激するであろうということです。そのようなテストは，現存する教育資源の生産的な再配分をも結果としてもたらすことがあります。言い方を変えれば，政府はこのように言っているのです。「みてご覧なさい，私たちの教育システムは非常に効率がよくありません。お金をたくさんつぎ込んでいるのに，それに対する見返りがありません。みんなが重要な部分に注意を払っていないからです」。ここで社会的な影響力の大きいテストを導入すれば，公平性の目標を達成する手段の一つとなるでしょう。そこで人々は，私たちが本当に真剣であり，彼らが行っていることをしっかりと観察しているのだと気付くことになるからです。そして，彼らは自らの努力をより生産的な方向に再配分していくことでしょう。これが政治における行為の理論と呼ばれるものです。

これらの議論を踏まえて，テストの質と構成概念妥当性について話を進めて行きたいと思います。構成概念妥当性には，そのテストを利用することで，目標を達成できると言えるための理論的根拠をチェックする機能が含まれています。例えば，ある事柄（教育内容など）をマスターしたか識別するといったことです。構成概念妥当性への脅威として，二つの問題が挙げられます。一つは構成概念の過少表現（テストの出題範囲がカリキュラムに比べて狭い場合），いま一つは構成概念に無関係な分散成分（テストの目的と無関係なノイズをシステムに引き起こす可能性がある場合）です。構成概念に無関係な分散成分の例として，「標準テスト」を取り上げます。標準テストの原義は，そのテストが標準化された条件の下で実施される，つまり全ての受験者が全くの同一条件でそのテストを受ける，という意味です。これはテストの公平性の問題であり，標準テストにおいては，実施方法の違いがテスト結果の分散に寄与することはありません。テスト結果の分散には，テスト方法の違いが原因の分散ではなく，構成概念における分散（構成概念の程度の個人差）のみが反映されることが期待されます。

テスト実施に関連する問題も考慮しなければなりません。テストの版（フォーム，冊子）は複数用意されているのでしょうか？同一版を繰り返し使用するならば，それが社会的な影響力の大きいテストであればなおのこと，受験者は過去のテスト項目を暗記して試験に合格しようとします。機密保持の点で安全なテスト項目はない，これが世界中の社会的な影響力の大きなテストについて私たちが分かっていることなのです。従って，一つのテスト結果にのみ基づいて意思決定を行おうとするならば，この種の問題が常について回るのです。しかし，その一方でテストの版を複数用意することができる場合でも，それらは適切に等化されているのでしょうか？そうでないならば，受験する版の違い（項目の内容や困難度の違いなど）によって有利不利が出てくる可能性があり，それでは公平性の面で問題が出てきてしまいます。そして，テストの実施要綱が実施現場の状況に応じた様々な微調整を認めるものとなっていると，テスト結果の比較可能性に関して心理測定学的な問題が出てきます。

テストの利用は，評価のための唯一の根拠となるのでしょうか。テストが（説明）責任という重責を担う限り，キャンベルの法則から逃れることはできないのです。

*訳者注：accountabilityは日本語で「説明責任」と訳されることが非常に多いが，responsibility（責任）と意味はほとんど同じであり，前者では「説明する」という意味合いが強調されているに過ぎない。accountability（あるいはresponsibility)とは，（原因があってプロセスがあって，最終的に）引き起こされた結果に対して，当事者の影響（結果だけではなく，原因やプロセスに対する寄与も含む）がどの程度あるのかを明らかにし，それを説明し検討する（義務を負うこと），というほどの意味である。

そして，評価のための基準とは何でしょうか？テストに基づいて，私たちはどのように個人について判断を下すのでしょうか？どんなテスト結果も多かれ少なかれ不確実性にさらされている，という事実に対して少しでも理解が得られているのでしょうか？その事実を私たちはどのように考慮していけばよいのでしょうか？最後に，あるテスト結果とより規模の大きい（説明）責任システムとの関係はどのようなものなのでしょうか？システム妥当性の観点から，これら全てが考慮されなければならないのです。

システム妥当性の展望

ここでの議論は，政策そしてトップダウン型の観点から，テストを含んだ（説明）責任システムを提唱する場合，私たちは行為の理論について明確にしておかなければならないということです。例えば，そのシステムはどのようにして目標を達成するのか，行為の理論を支持する経験的な証拠は何なのか，そもそも経験的証拠があるのかどうか，といった観点についてです。

具体的な例を挙げたいと思います。アメリカの公的教育における根強い議論の一つに，ある学年で十分な学習成果を示さなかった（テストの成績が必要な基準を満たさなかった）生徒を留年させるべきかどうかというものがあります。

多くの州ではSocial Promotionといって，生徒のテスト成績が良くなくて進級基準を満たしていなくても，その生徒は留年することはなく次の学年に進級できるというシステムがあります。しかし現在の（説明）責任システムでは，多くの学区が達成度評価の基準としてより厳格なものを採用しており，基準を満たさない生徒は同じ学年をやり直さなくてはならないことになっています。

しかし，生徒を留年させることは彼らにとって助けにならず，状況が好転するどころかさらに下に落ちていくだけだと指摘する研究がいくつかあります。そして，それが高学年になればドロップアウトを奨励されるだけの結果になるのです。もしあなたの子どもが16歳で，第９学年の学習内容を２～３年繰り返しても留年してしまうようであれば，要するにその子はドロップアウトしてしまう見込みが非常に高いということになるでしょう。

原理的あるいは理論的に，次のようなやり方が良いと議論する人がいるかもしれません。基準を満たせない生徒には別のチャンスを与えるべきだが，それがうまく機能しないという経験的な証拠があるのなら，「それでは他の選択肢は何だ？」ということになります。おそらくその選択肢とは，その生徒は進級させるけれども，同年齢の仲間と同じ学習をするのに加え，特別な指導時間を取って学習内容についていけるようにしよう，というものです。これは，特別なケアも与えずに無条件で進級させる，あるいは留年させて更なる問題を引き起こすよりは，良い戦略であると思われます。

このような方法に関してはまだやるべき仕事を多く含んでいますが，私には賢明なものであると思われます。というのは，システムに関する考察を巡らせておかないと，あれやこれやで結局はシステム効率が非常に低いものになってしまうからです。システム妥当性に関して言えば，私たちは正負両面の意図せざる結果を考慮に入れ，またそのような意図せざる結果の生起確率やコストについてよく議論しなければなりません。それらを踏まえ，現実味のある代替案やシナリオを準備しておく必要があると私は考えています。

ビジネスの世界ではこうしたシステムがしっかりと確立されています。シェル石油はこの手のシナリオ立案に関するパイオニアの一つです。これを一般的な社会政策や教育政策に応用しない手はありません。実際には，少なくともアメリカでは，政治の一環としての教育談義で出てくる教育問題を，ごく単純な解決提案と共にみていくのには，それほど洗練された政策分析やシナリオ分析は必要とされていません。

今度は，テストを含む政策に焦点をあて，より相互作用的な政策デザインについて考えたいと思います。いくつかのシナリオを実行するうちに，私たちははじめに抱いていたアイデアがそれほど良いものでもなかったということに気が付きます。そして初期の計画を改良するのですが，これはテスト単体のことではなく，テストをその一部として組み込んだシステムです。その目的は，意図せざる負の結果（やその影響）を軽減させ，私たちが期待するものを達成できる確率を高めることです。こうしたデザインの変更には，構成概念妥当性の改善が含まれることがあります。そのテストに計り知れないほどの重要性があるのならば，私たちはより多くの資本を投入し，回答構成型の項目を増やし，より視野の広いテストを得る必要があるのです。さもないと，私たちは教育的な観点から非生産的な活動に終始してしまうことでしょう。このようなことが何度も繰り返し起こるというエビデンス（証拠，根拠）があります。また，評価活動をサポートするために複数のテストを用意する必要があるかもしれません。さらに，私たちは現実的な評価基準を設定する必要があります。ここで私の意図す

るところは，非常に非現実的な基準を設定してしまい，システムの中の人間がその基準を満たす能力がないにも関わらず，結果に対する（説明）責任だけは負わされているということに気が付くと，彼らは別の抜け道を探ってシステムをうまく乗り越えようとするだろう，ということなのです。

私たちがはっきりさせておきたいのは，本質的にも外見上も，自分たちのシステムは公正なものなのだということです。ある基準を設けてテストを不公平なやり方で使用したら，私たちはシステム妥当性を損なうことになります。次に，資源を適切に配置することにも十分配慮している，ということもはっきりさせておきたいところです。新しいカリキュラム，新しいテスト，新しい評価基準を導入したとしても，その新しいカリキュラムを教える教員を揃えるための投資を渋っているようでは，その後様々な問題を抱えることになるでしょう。適切な資源配分の欠落については，その（負の）効果が現われる前に事前に検討されておかなければなりません。

政策の急激な変更について検討するとき，一年でそれを全て行ってしまおうとするのではなく，変化を段階的に導入することを考える必要があります。何年もかけて段階的に変更を行うことで，システムが無理なく適応できるのです。非常に複雑なシステムにおいてあまりにも急激に事を起こそうとすると，意図せざる負の結果を招くことにもなりかねません。

これらは，物事をシステム全体の機能の中で考えることが，いかにより生産的で知的な政策を生み出すことの助けになるか，ということの例示です。少なくとも，アメリカで私たちが近年取り組んできた政策の多くは，システム妥当性の考え方に則っていませんでした。その結果，私たちはそうした政策から多くの意図せざる負の結果を招くことになったのです。

前に申し上げたように，こうした考え方はシステム妥当性の考え方をトップダウン的な取り組みとみなしています。言葉を変えれば，政策レベルからシステムに降りてくる類のものなのです。為政者が彼らの目標を達成するためにテストを利用する際，私の考えでは，物事をシステム全体の中で考える作業が必要となります。そうすることで，テストを最大限に活用することができるのです。これを怠ると，テストがシステムにとって非常に特異な要素となり，システム全体を攻撃するための根拠となってしまっても驚くことではないのです。

6.　生産能力（生産性）の構築

テストをより建設的に利用する方法があると私は考えています。それはテストを利用して生産能力，つまり正の教育目標を達成するためのシステムの能力（生産性）を構築することなのです。モニタリングのためのテストから得られたデータは，システムの生産能力を構築するために利用することができます。すると，トレーニングや支援を必要としている教師や学校などの個人やユニットが，テストデータを上手に利用することができるようになります。また，そうしたデータの質はその利用価値に影響を与えます。こうした生産能力を構築するためには，その能力に対して投資しなければならないというのが私の信念です。テストは確かに生産的な能力をシステムが手に入れるための安価な手段の一つではあるのです。

いくつか例を挙げたいと思います。アメリカの多くの学校では，データ駆動型（data-driven）意思決定と呼ばれる活動に従事しています。そこでは，州試験のデータを使って，学級や科目レベルでの弱点を発見するのです。そして，明らかとなった問題領域に焦点を当て，教育資源を配分するわけです。いま，ある学校の教員や校長が，初等教育段階の生徒の読解力が十分ではないという問題が分かると，彼らは読解力教育の質を改善するためにより多くの資源を割り当てなければならないと知るのです。初等教育の段階では，生徒のモニタリングを頻繁に行って問題を抱えている生徒を把握しておく必要があります。彼らが抱える問題は健康上のものであったり眼鏡が必要だということだったり，様々でしょう。アメリカの貧困地域では，子どもたちが抱える学習障害の多くは適切な健康管理が行われていないことに原因があります。栄養状態が悪かったり，他にもまだ私たちが把握できていない様々な障害を抱えているかもしれません。こうした問題を引き起こす潜在的な要因はたくさんあります。そして，テストの結果なしでは学校の校長は事態が手遅れになるまでこうした問題が存在するということにすら気が付きません。この問題は，テストをどのように利用してシステムの生産的な能力を構築していくかということの単純な例なのです。

また，多くの学区は州試験のデータからあまり教育効果を上げていない教員を発見し，彼らの質を向上させるための専門開発を行っています。これは，システムの生産能力のもう一つの例であります。さらに別の例として，TIMSSを取り上げます。これは数学および科学の国際学力調査であり，教科テストだけではなく教授法に関する調査項目を

も含んでいます。そうした調査項目では，各科目の教授方法が，国によってどのように異なっているのかを調べています。

多くの国がTIMSSの豊かなデータベースを利用しています。テストの結果だけではなく，教授法，カリキュラムなどに関するデータベースの情報を利用して，自国の教育への取り組み方を見直したり，教員に対して新たなトレーニングが必要であるといった提言を行ったりしています。時々アメリカで耳にするのですが，日本の数学教育は概念的な教授方法を重視しており，アメリカよりもカバーされるトピックの数はずっと小さいものの，各トピックは深い内容までカバーされ，生徒の方も深い理解が得られるということですね。これが本当なのか私は知りませんが，アメリカの研究者はそのように言っていますし，彼らはそうした研究調査の結果によってアメリカの数学教育に影響を与えようとしています。

アフリカやラテンアメリカなどの開発途上国では，TIMSSの資源を自国のカリキュラム改善に役立てようとしています。つまり，TIMSS試験や採点ガイド，回答構成型項目（自由記述型のテスト項目）などを，カリキュラムの改善や教員の研修方法の改善などに役立てる助けにしています。また，彼らはテスト結果を教員や研修機関を通じて各地に周知させ，そうすることで教員の教授方法がその国の伝統的なローカルスタンダードから国際標準へと徐々に移行し始めることを狙っています。

形成的アセスメント・学習のためのアセスメント

これらの例の良いところは，こうした資源が所与のものであるということです。つまり，アメリカのような先進国でも南アフリカのような開発途上国でも，ある国に要求される追加投資額は，国際調査に既につぎ込まれた投資額に比べれば比較的小さなもので済むのです。そのため，こうしたテスト資源を利用してシステムの生産能力を構築することは，非常にコスト効率が良い方法であるといえます。ですから，私の考えでは，トップダウン型ではなくボトムアップ型の取り組み方を通じて，テストはシステムの生産能力を構築するために利用できると思います。実際，その最も強力な方法が，形成的アセスメントとか「学習のためのアセスメント」と私たちが呼んでいるテストを通してのものなのです。

「学習のためのアセスメント」とは，そのデザインと実践の第一優先事項が，生徒たちの学習を促進させることであるアセスメントです。生徒に（説明）責任を負わせたり評定を下したりするのではなく，彼らが自分の強みや弱みを理解することを助けます。教員や生徒が自分自身やお互いを評価する中で，フィードバックとして利用できる情報をもたらしてくれるのであれば，アセスメント活動は学習活動の助けとなることができます。学習ニーズを満たすように教育活動を適合させる目的で様々なエビデンスが用いられるとき，そのようなアセスメントは形成的アセスメントとなります。これは2002年のBlack & Williamの著書である「Inside the Black Box」からきています。

この発想は，アセスメントを単なる規制や監督のための道具と考えるのではなく，教育方法にとって不可欠の要素であると考えよう，というものです。教員は当然の如く生徒にテストを与え続けますが，そうした行為は時として非公式的に（インフォーマルに）なされていたり，教員がアセスメントの結果を自らの教育実践を改造するためにどのように利用しているのかといったことは，常に明らかであるとは限りません。形成的アセスメントの背後にある発想では，こうしたアセスメントは実際にはより公式（あるいは形式）的なものとなるのではなく，学級全体あるいは生徒個人に対する教員のアプローチ方法を適合させる手段にとって，アセスメントデータの利用が必要不可欠な部分となるのです。形成的アセスメントは学級の中での日々のモニタリング活動となります。非常に興味深いのは，アメリカやイギリスでは多くの研究がなされており，それらによると，強力な形成的アセスメントがあれば生徒の学習活動に大きな影響を与え，0.5ポイントあるいはそれ以上の効果量を得ることができるということが示されました。これは，形成的アセスメントを行わない教員の生徒や学校と比較すると，教員が形成的アセスメントを効果的に使えるようにトレーニングされた学級の生徒は，標準偏差を1単位として0.5ポイント，あるいはそれ以上の進歩を期待することができるということなのです。

しかし，これ程の効果量を達成するためには（教育において0.5ポイントの効果量は非常に大きな値なのです），２年間から３年間にわたって教員開発に対して相当の投資を行うことが要求されます。教員が形成的アセスメントを快適に使えるようになるまでには時間を要しますし，生徒の素材を見るためのトレーニングも必要となります。信じがたいことかもしれませんが，少なくともアメリカでは，教員はテストについての教育をあまり受けていないのです。仮にテスティングについて何らかの講習を受けるにしても，それらはテスト結果をどのように解釈するのかという内容のものです。テストをどのようにデザインするか，

また彼らの教育方法に意味のある影響を与えるようなテスト結果の解釈方法についてのコースはありません。形成的アセスメントのトレーニングとはまさにこのような内容を取り扱うものであり，教職の研修初期において本来であれば習得しておくべき専門開発を教員に提供するのです。アメリカでは，残念ながら教員はそのような形成的アセスメントのトレーニングを受ける機会がありません。

形成的アセスメントは，教員の知識とスキルの両方を改善させることで教育システムの質や効率性に関する目標を達成することに貢献することができます。また，形成的アセスメントは教員が学習環境を調節するのを支援します。つまり，教員がニーズという観点から生徒を区別することができるようになり，異なるニーズを共有する生徒集団に合わせて適切に教育方法を調整することができるようになるのです。形成的アセスメントによって，教員は援助を最も必要としている生徒を把握して支援することができるため，結果的に生徒の自律的な教育活動への参加を促します。そのため，彼らの理解をも増幅させることになるのです。だからこそ，適切に実施された形成的アセスメントから0.5もの効果量を引き出すことに成功したのです。

7.　結論

日本やアメリカのような国の教育制度は極めて複雑かつ多相的であり，強力な伝統をもっています。そのため，そうした国の教育制度を変えていくことは困難です。教職員や校長の組合，政治家，ビジネス関係者などの，多くの異なる政治的・専門的な興味関心によっても教育制度は影響を受けます。真に消えずに残る変革を成し遂げることは非常に困難なのです。テストが規制と監督という重要な役割を日常的に果たしてきたことを指摘する一方で，テストにできることはまだまだあると私は考えています。テストを創造的に学習プロセスに取り入れることで，その質と効率性を大きく高めることができると考えていますし，またシステム妥当性の考え方を厳密に応用することで，テストの政策的利用を改善させることができると思います。私たちには，形成的な性質の強い「学習のためのアセスメント」と，（説明）責任を果たすための総括的評価としての意味合いが強い「学習のアセスメント」との間で，より良いバランスを保つことが必要とされているのです。

今現在，多くの国々では「学習のためのアセスメント」と「学習のアセスメント」との間に不均衡が発生いています。私たちが行っているほとんどのテストは規制と監督のための「学習のアセスメント」です。そこでの「学習のためのアセスメント」の相対的な割合は非常に低く，現在行われているアセスメントの内容はあまり良いものであるとはいえません。より多くの資源を「学習のためのアセスメント」につぎ込む必要があり，それは今日からはじめなければなりません。この領域はテクノロジーが，教員の専門開発と形成的アセスメントの両者を支援するという，重要な役割を果たすことができる部分だと私は考えています。例えば，Web配信は生徒たちの形成的アセスメントや診断的アセスメントに対する即時的，継続的なアクセスを可能にし，生徒個人に適合した方法での学習活動を促進させることができます。

こうした可能性を現実のものとするためには，私たちは教育政策の定式化に対してより専門的に取り組んでいく必要があります。これは，システム妥当性と投資収益率を注意深く検討することによって達成することができます。創造力を働かせて現存する資源を利用することで，私たちにはまだまだできることが残されています。大規模な国際調査の周辺に存在している資源が良い例で，それらは残念なことに十分に活用されていません。さらに，測定の専門家がもっと政治的なプロセスに関わっていく必要があります。科学者として，私たちは政治に関わることから顔を背けてしまいがちです。そうしたことは，他の誰かの（つまりは政治家の）仕事であると思っているのです。多くの国では，科学者たち（この場合は測定科学者たち）は政治的なプロセスに対して助言を求められることがあります。しかし，その多くのケースでは，彼らは政治家や為政者が既に決定した事柄について，それを肯定したり承認したりするためだけに呼ばれるのです。私たちは，今まで以上に政治的なプロセスに関わっていく必要があると思います。それは，そうしたプロセスや政策の立案活動は，教室や研究室の中で行われる活動よりも，テストの役割を形成する力が強いからです。私たちがより積極的な役割を担わないと，システム妥当性や生産能力に関する諸問題は，教育システムが必要とするしかるべき注意や関心を得られないことになります。その結果，資源を引き続き非生産的に投資し続けることになり，遅れをとることに対して不平不満を言い続けることになるでしょう。

ご清聴有難うございました。

8.　質疑応答

Q1：「学習のためのアセスメント」の具体例を教えていただけませんか。

A1：特定の領域では多くの例があります。例えば数学では，BlackやWilliamといった人々が，学習者に自分の位置をピンポイントに指し示すことを目的にデザインされた一連のテスト項目の開発に取り組んでいます。ALEKS（http://www.k12.aleks.com/）と呼ばれるコンピュータシステムによる数学のアセスメントがあります。カリフォルニア大学アーバイン校の認知科学者Jean-Claude Falmagneによって開発された，非常に洗練された適応型テストシステムです。内容は，代数学から微積分以前までをカバーしています。受験者である学生の強い分野と弱い分野という観点からプロフィールを生成します。その結果を基礎資料として，教師はその生徒に対してどの分野の支援が必要なのかを明確に理解することができるのです。段階的な評定は与えられず，次年度に進級するために利用されるテストではありません。このシステムは，学生や教師に彼らの努力（資源）をどのように配分すればよいのかを伝えるための，伝達手段の一つなのです。これは，現実世界における「学習のためのアセスメント」の一例といえるでしょう。

Q2：それらのアセスメントには診断的な機能が含まれているのですか？

A2：そうです。この分野ではDylan Wiliamが第一人者の1人だといってよいでしょう。3年前，彼はETSにやってきていて，診断テストと形成的テストの区別をつけるのが好きでした（それは微妙な違いです）。彼ならばこう言うでしょう。診断テストは，教員がその結果を生産的に利用して問題のある領域を示すことができない限り，教育的であるとはいえない，と。つまり，素晴らしい診断的な情報が得られても，それを利用して何かをする人が誰もいなければ，それは形成的アセスメントではないのです。大掛かりな医療検査を行う医者と似ていますね。その医者は非常に良い診断を行うことができるかもしれませんが，診断を行うだけで患者を治療しなかったならば，その医者は患者を助けたことにはなりません。もちろん，どんな診断テストの結果でも，それが利用されることが大前提だと考えたいものです。

Q3：全国の生徒が受けることのできるような，大規模な診断テストというものは現実的でしょうか。

A3：それは良い質問ですね。その点については，いまも議論がされているところです。多くの人は，総括的アセスメントあるいは「学習のアセスメント」のためにデザインされている大規模試験は，「学習のためのアセスメント」と同等に機能することはできないと信じています。このことは，もちろん問題の一部でありますし，皆さんには大規模試験ではない適応型試験であるALEKSを例に挙げました。生徒に要求されるものによって項目提示の流れが随時調整されるのです。単独の大規模試験に，全ての生徒にとって利用価値のある形成的な情報を提供することを求めるのには無理があります。また，そうした目的はテストデザインにそもそも含まれてはいません。もちろん，できる限りの努力はしようとしていると思います。例えば，複数の下位尺度から構成される数学の大規模なアセスメントがいくつかあります。ミドルスクールの数学では，数の性質や演算，測定，データ解析，幾何学などがカリキュラムに含まれますが，そうした個々の単元に対応する下位尺度を定義することができます。そのような下尺度のレベルでテスト結果を提供すれば，形成的アセスメントでこそありませんが，診断的な目的に利用できる総括的アセスメントを行う助けになる，と考える人がいます。そこで問題となることは，異なる下位尺度から得られる結果は，ひとたび測度の非信頼性の補正を行ってもそれほど信頼性が高くなるわけでもなく，相互に高い相関関係をもっているということです。異なる下位尺度から，性格の異なる有用な情報がどの程度得られるのかは定かでありません。

SAT(Scholastic Achievement Test)を例に取り上げたいと思います。SATはETSが提供するアメリカの大学入学試験の一つです。２，３年前までは，SATでは受験生の言語推論能力，数学的推論能力のスコアと各領域のパーセンタイルが報告されていました。しかし現在では，ルールスペース法に基づいた診断的プロフィールも提供しています。このことについては多くの批判があります。SATは受験生の言語推論能力と数学的推論能力を測定するためにデザインされた試験ですから，それぞれの尺度は一次元的な尺度を構成しています。そのため，一次元モデルからの乖離や下位尺度についての拡散的な情報量に関しては限界があります。ルールスペースモデルを利用すれば，ある種の有用な情報を引き出すことが可能になると考えられます。そのため，私たちがより良い診断的アセスメントを利用できるようになるまでの当座の測度として，ルールスペース法の応用的実践をCollege Boardに推奨しています。私は，いわゆる典型的な大規模アセスメントが形成的アセスメントとして機能することはできないということは認めています。なぜならば，両者のデザインにおける原則があまりにも異なっているからです。

形成的アセスメントを行いたいのであれば，異なる観点から取り組んでいかなければなりません。テストデザインについて，最終地点（目標）から出発地点まで遡って考えていくと，次のような疑問点に順次答えていくことになります。まず，何がこのアセスメントの目的なのでしょうか？私たちは，どのような種類の情報がほしいのでしょうか，あるいはどのような種類の意思決定を行いたいのでしょうか？そうした意思決定を行うために，私たちはどのような種類のエビデンス（証拠）が必要なのでしょうか？そのエビデンス（証拠）を得るために，どのような課題が必要なのでしょうか？それらの課題（テスト）を組み立てるために，私たちはどのような種類のアセスメントを用意できるのでしょうか？仮に，生徒たちを言語推論能力に基づいてランク付けすることだけを単純に考えているのならば，あなたが最終的に手にするテスト手段は（SATとは）随分と異なったものとなるでしょう。つまり，テストの目標が大きく異なると，アセスメントの内容や構成もまったく異なるものが出来上がるのです。水陸両用車両のアナロジーを考えてみてください。水陸両用車両は陸上でも水上でも稼動しますが，いずれの状況下でも中途半端なパフォーマンスしか発揮できませんね。

Q4：効率性についてコメントをしたいと思います。Colemanレポート以来，教育資源をより効率的に使うためのある種のガイドラインとして総括的テストを用いるという長い歴史があり，この問題に対して継続的に研究が行われてきました。例えば，教師の数に対する投資は，彼らの質に対する投資に比べて効率性が低いということを示すエビデンス（証拠）もあります。これは，総括的テストを政策の基礎としてどのように用いるのか，という問題だと思うのですが。

A4：総括的アセスメントをシステムの効率性に対する予備的判断の基礎として用いる際に生じる問題点として，総括的アセスメントは評価を行う時点までの生徒たちの教育履歴の全過程を反映するため，例えば最近2，3年間の効率レベルを評価判断するための基礎としては十分でない，といったことが挙げられます。このことは，総括的アセスメントを応用している学校が，生徒がどの程度成長してきたのか，最近３年間の彼らの成長は何だったのか，といったことを理解しようとする理由の一つになっています。つまり，成長の度合いに基づいた比較を行うことで，より焦点を絞った，より公平なやり方でシステムの効率性や有効性についての判断を行おうというのです。

Colemanの研究は興味深いものです。彼の報告は1966年に初めて出版されました。その議論の中に，生徒の到達度をより予測できる変数は，学校ではなく家庭の背景であるとあります。しかし，彼の注意はテスト得点の良くない生徒たちに集中しており，学校ではなく生徒の背景により大きな変動があったというところに問題の一部があったと私は考えています。Colemanの分析には根本的な部分にある種の欠陥が存在すると指摘している研究があまたあります。教育プログラムに関しての効率性や投資収益率について理解しようとするとき，それを十分に行えるだけの記録が乏しいのです。LevinとMcEwanが著した新しい本（第２版）では，非常に良い分析が行われています。Henry Levinはティーチャーズカレッジの経済学と教育学の教授で，この件に関して彼らは非常に良い分析を行っており，学校レベルとプログラムレベルの両方で教育部門における投資収益率の推定を試みています。基本的に，ひとつだけこれ，という回答はないと彼らは言っています。従って，為政者や一般大衆がこの質問をいつも繰り返しているように，この問題は非常に重要なオープンクエスチョン（答えが不明で，様々な回答可能性の余地を残す類の疑問）であるのです。このような類の非常に大まかな比較をすることは私たちにできる最善のことなのですが，もっと深い部分で人本主義的な調査によるフォローアップを行わないと，こうした計量的な指標によって私たちが誤った方向に導かれてしまうか，あるいはそうした指標がもつ情報を適切な形できちんと与えられるか，その確率は五分五分ということになってしまいます

Q5：教育の生産関数（educational production function）の発想についてどのようにお考えでしょうか。この考えにはテクノロジーへの注目が欠落し，物理的な側面に焦点を当てているといえると思います。一つの例として，日本の公立学校の学級サイズはアメリカのそれよりもずっと大きいものですが，そのような状況でも，達成レベルは逆転しています。そこにはテクノロジーや経済的な条件の違いがあるのかもしれません。

A5：アメリカのテネシー州で行われた有名な社会実験があります。1980年代の中後半期に行われた，テネシー学級サイズ実験です。生徒や教員を異なる処遇条件に無作為に配置しました。ある処遇条件では，学級のサイズが18人か19人以下に制限されました。別の条件では，23人から25人に制限されました（これは通常の学級サイズに近い人数です）。また別の条件では，通常の学級サイズに加えて教

員補佐も割り当てられました。このように，異なる学級条件下で生徒のパフォーマンスを比較することがこの実験の目的でした。その結果，15人～17人の最小クラスの生徒がより学習効果が大きかったという知見を見出しましたが，同時に，学習効果が最も大きかったのは貧しいマイノリティーの子弟たちであることも分かりました。学校に行く準備も満足にできない貧困層から来た子どもたちは，人数の少ない学級の中ではより個別の注意を引くことができ，そのせいで学習効果が大きくなったのです。しかし，その他大勢の生徒たちにとっては，教師の注目は教育効果にとってそれほど重要な要因ではありませんでした。ここでの議論は，今私たちは（この実験で得られた）経験的な証拠に基づいて話をしているのだということです。つまり，もし学級サイズに対して投資を行うのであれば（それは巨額なものとなるでしょうが），カリフォルニア州のように「みんなを小さな学級に入れよう」とただ声を上げるのではなく（これは意図せざる結果を招きやすくなります），投資を行う対象をよくよく絞らなければならないということです。

経済学者であるRichard Rothstein著の「Class and Schools」という書籍があります。その議論の一部は，アメリカにおいてはテスト単独で教育を改善させることはなく，社会的に影響力の大きいテスト単独ではアメリカの教育水準を世界一に引き上げることはできないというものです。ここにいる皆さんにとってはもう自明のことだと思います。彼は，アメリカとその他の国々との間の成績ギャップのみならず，合衆国内に存在するギャップもまた，栄養状態，住宅事情，健康などに関連した様々な不平等の関数であり，こうした不平等は後になってから教育格差に反映されると議論しています。

彼は国全体の初等教育における学級サイズを減らすためのコストを検討し，その結果，何10億ドルという莫大な規模の試算に至りました。次に，彼はアメリカのすべての子供たちにまともな朝食や夕飯，十分な健康管理，必要ならば眼鏡や補聴器などを提供し，子どもたちが養子にならないように肉親が十分な資金と持ち家を有するためのコストを試算したところ，学級サイズの減少コストと同じ程度の莫大な金額となりました。彼は，エネルギーや資源は，学級サイズの減少のような形で間接的に教育システムに投資するよりも，子どもたちやその親が生活する環境に対して集中的に投資した方が，コストがより効率的になるだろうと議論をしています。私は，全ての人が彼の議論に賛同するとは思えません。物事を本気で改善させたいと願うのならば，これはどのようにその改善がなされるべきなのかという長い議論のはじまりなのだと私は考えています。これは方法や手段に関する問題に留まらず，戦略の問題でもあります。だからこそ，戦略に関する議論は文脈や環境に本当に依存すると私は言っているのです。日本とアメリカでは，健康や食糧に対するアクセスという観点での公平性の様相が異なるからです。日本がアメリカと同様のギャップを抱えているとは考えがたいのです。

ETSでの私の同僚の１人であるPaul Hollandは，国家委員会の一員として食糧不安について推計しようとしていました。様々な水準の食糧不安にある家族や子どもたちのカテゴリーを把握しようとしていたのです。私には食糧不安の問題はありません。去年は，いつでも十分な食糧を確保することができました。それどころか，過剰なほどの食糧があったのです！彼らの研究は過去６ヶ月あるいは12ヶ月の間に，少なくとも１ヶ月間継続的に食糧の確保が困難であった家族の割合を考慮に入れました。食糧不安が深刻な家庭の学童期の子弟の数は膨大であり，腹を空かせたまま学校に来る子どもたちが，果たして十分な学習活動ができるのかどうかを考えてみなければなりません。学校には様々な種類の昼食制度こそ普及していますが，朝食制度はあるにしてもローカルな水準でしか実施されていません（制度化はされていません）。

シカゴである研究が行われました。初等教育の学習水準かなり落ちこぼれている5000人の生徒のうち，40%～50%は健康状態に問題を抱えていたことが分かりました。彼らには，眼鏡や補聴器，あるいはその他の健康補助器具が必要でした。また，学習を著しく阻害する喘息などの病気にも罹っていました。日本ではそれほど問題にはなっていないと思いますが，私たちにとっては非常に重大な問題なのです。なぜならば，教育から甚だ離れたところで生徒のための資源が大きな不平等を抱えている，という事実に連邦政府は気が付いていない，あるいは意図的に気が付かない振りをしているからです。ですから，これは現在進行形の議論なのです。こうした理由から，私は教育の生産関数に基づく議論の多くは，極めて不毛かつ特別何かの役に立つこともないと思っています。

Q6：診断テストについて触れられましたが，私もその重要性についてはまったく賛成いたします。しかし，診断テストの実施コストは非常に高く，現在のテストの効果の水準については定かでありません。そのようなテストは，アメリカではどの程度の量が出回っているのでしょうか。ま

た，それらのテストを作成しているのはどのような方なのでしょうか。どのような人がそうしたテストを実際に使っており，どのような基準で利用するテストを選択しているのでしょうか。

A6：私はそのような市場調査をしたことはありません。私がALEKSを例に取り上げたのは，それがかなり長い間利用されてきたという実績と，世間によく知れ渡っているということによります。いわゆるローカルイニシアティブというべきテストもいくつかあります。例えば，オースティンのテキサス大学ダナセンターのUri Triesmanは形成的アセスメントの領域で研究活動を続けてきています。また，カリフォルニア大学バークレー校やロサンゼルス校でも，CRESSTという研究センターがあります。ローカルイニシアティブにもいくつか種類があって，その他にもSuccess for All and Reading Firstといったものもあります。これは非常に厳格な構造のテストですが，ある意味では形成的アセスメントの傘下にあるテストだと言えます。

私たちは，テストのことを良質の形成的アセスメントであると考えてしまいがちですが，テストから得られる情報を活用できる教師がいなければ，真に大きな効果を得ることはできません。残念なことに，教員のトレーニングにはテストを実施するよりもずっと多額の予算が必要になります。例えば，ALEKSのようなシステムが一度正しく運用されだせば，生徒一人当たりのコストは非常に小さなものなのです。しかし，ALEKSを利用するためのトレーニングを教員に行うことに対しては，低コストは当てはまりません。私たちのシステムが抱える欠陥の一つは，教師にこの種の形成的アセスメントの評定をさせることや，時間をかけて他の教師と協働したり生徒の素材を採点することが，彼らの専門的開発に対して大きな価値を与える，ということを認識することができないことなのです。

アメリカにはレベルの高いAdvanced Placement program（AP）というものがあり，高校生が大学レベルのコースを履修することができます。学年末に実施されるAPの試験はいわゆる（説明）責任のテスト（学習のアセスメント）です。ほぼ全てのアセスメントは多肢選択式の項目と回答構成型（自由記述型）の項目からなっており，生徒たちは小論文や長文問題に取り組まなければなりません。しかし，全国から多くの教員たちが採点のために一同に会し，学生のパフォーマンスについて議論する機会を得ます。採点者の待遇は満足のいくものではなく，寄宿舎に寝泊りし，十分な報酬もありません。しかし，彼らはAPの採点経験を非常に有益で刺激的だと感じており，愛情をもって臨んでいるのです。これが，そこで彼らにこの（形成的アセスメント）モデルを採用させ，それをより一般的な形で利用した理由なのです。Dylan Williamは，アメリカで今まさにこれを実践しています。彼は，教育方法の専門的開発と結びついた形成的アセスメントプログラムは，学校における生産能力（生産性）を構築する手段なのであると，学校側が認識するのを支援しようとしているのです。

以上

講演３

Testing and Education Policy

Henry Braun
Educational Testing Service
Princeton NJ USA

Abstract

Recognizing that human capital development is essential to progress and economic success, governments around the world are becoming more active in education policy. At the same time, testing is playing an increasingly central role in policy implementation. From the point of view of measurement specialists, however, test results are often improperly used. Aside from attempting to intervene in specific cases, what can the measurement community do improve test use?

In this talk, I argue that we must have a more organized and proactive way of responding to proposed applications of testing. As a first step, I propose an organizing framework for education policy that encompasses both goals and means. I then describe the different roles that tests play and introduce the concept of systemic validity as a basis for evaluating education policies with respect to expected outcomes and unintended consequences. These ideas are illustrated with some examples and the talk ends with some suggestions on how different forms of testing can contribute in new ways to the productivity of education systems.

Introduction

Today, I want to talk about testing and education policy. This is not a technical or philosophical talk; however, I really hope to develop a way of thinking about testing and policy and to try to understand why testing is so poorly regarded in many education circles even as it is becoming more and more important. So we have a little bit of a paradox that I want to discuss and perhaps suggest some solutions.

I think that we all understand that education policy is becoming much more active; it's a critical governmental focus in many countries. The term "human capital" is often used in the United States. In the future, human resources will surpass natural resources in importance. Japan, unfortunately, doesn't have so many natural resources so that it has always regarded human capital as essential to its well-being, economic and otherwise.

At the same time, as education policy is becoming more central to governmental work, testing is playing a more important role in education policy and in practice. This is partly because centrally directed testing is often seen as a "quick fix" (American slang: It connotes a rapid solution to a problem). This is because in the end, testing is actually relatively inexpensive as compared with other things that we have to do to improve education; it is relatively easily implemented and could directly influence multiple levels of the education system. Therefore, many governments, in the United States and other countries, see testing as a very natural instrument of governmental policy.

I think there is a universal desire for the implementation of better-designed education policies and for test results to play a more supportive and appreciative role. How can we contribute to making this hope a reality?

Those of us in the measurement community tend to think about the technology of testing: How we go about doing the testing, and how we can improve it. We consider these as some sort of answer to making testing more effective; however, this is only a part of the answer. What

I want to argue today is that measurement scientists—I use this term to include a very broad range of people, from psychometricians to measurement specialists to test developers and even curriculum developers—must pay more attention to policy and to understanding the different roles that tests play in education policy.

I hope that they will use this understanding not only to inform the direction of their technical work but also to guide their greater involvement in policy issues, which scientists typically overlook. We tend to focus on our work and leave policy to others. But I think education policy is now so important and involves so many serious technical issues that we can't afford to leave it to the politicians, even though they think they know what they are doing.

I want to argue here that too often, the discussions about testing and education are very narrowly focused on a particular issue—for example, testing for entrance into university, for promotion from one grade to another, etc. Further, people are really not talking about the same things, and so the conversation becomes very confused and doesn't produce many solutions. I believe we can do better if we have a general framework for thinking about policy. We can use the framework to understand how testing fits into education policy and use that as a basis for a more rational approach to improving testing and its contributions to education.

What I am going to propose here is just a first iteration. I really would like to hear the opinions of the audience, the students and the faculty, regarding this attempt, which of course could certainly be improved.

Overview

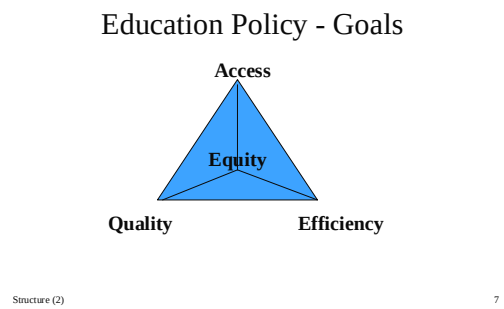
I'll begin my talk with the framework or the structure for education policy. Then, I'll talk about the roles of testing and some of the problems with testing. I want to propose an idea called systemic validity for both testing and education policy and how we can use testing to build productive capacity in the education system. I'll then present a few conclusions.

Structure- Goals, Strategy, and Means

The framework or the structure for education policy has three main dimensions—Goals, Strategy, and Means. Goals are the purposes or targets of the policy; I will discuss each of these in detail in a few minutes. Strategy is the general plan for achieving the goals by using different means, and Means are the combinations of decisions, activities, and resources by which we carry out the policy.

Goals –Access, Quality, Equity, and Efficiency

While education policy can have many goals, I would like to argue that for most purposes, there are roughly four basic goals that we can use to organize our thinking about policy goals; Access, Quality, Equity, and Efficiency. I think in most countries, almost all goals can fit one of these four vertices pretty well.



Access

The English word "access" means how to enter; so Access might mean, for example, governmental targets for students entering different education levels or types of institutions. Sometimes, the policies of Access are related to expanding access, and sometimes, they are related to restricting access. So for selective institutions like the University of Tokyo, access is usually constrained, and we try to set goals in order to identify very high-level students who will be admitted to the university. In other cases, we aim to expand the educational opportunities for large groups in the population; however, in either case, we are talking about Access.

We might also be talking about targets for alternative pathways; for example, adults who have been in the labor force and have now either been laid off or their jobs have disappeared. We want them to have access through

alternative pathways to education, training, or other resources that are related to access—such as counseling or funding to support individuals to move in the right direction. So all of these are aspects of Access; they are goals of access from the educational policy perspective.

Quality

There are lots of different ways in which people think about Quality. Here are three kinds of examples. One would be standards for the context of learning; that is, the first issue for policy is what are the resources and the setting in which education takes place. So we can talk about teacher credentials, the qualifications of the teachers, the school conditions, the nature of the buildings and classrooms, etc. The class size, the classes, and the availability of resources for students and teachers are all aspects of Quality as they are related to the context of the work.

We also have standards for the quality of the learning itself, the learning outcomes. They might have to do with, for example, the standards that we set for each grade or the standards for university graduation, and they might have also to do with the targets for the proportions of each cohort achieving those standards. So continuing with the example, we might say that to go from junior high school to high school, we would like students to achieve a certain standard, e.g. 80% of the junior high school students taking these exams should achieve the standard. All this would be part of establishing a quality standard or a quality goal for education policy.

Another quality goal might be related to the external evaluations of the capabilities of the graduates; this implies carrying out surveys of employees, asking how well university graduates are generally performing in their jobs, and asking how much training they need before they are able to function effectively within your company. These would be external evaluations of quality and I'll again relate this to a quality goal.

Efficiency

Efficiency has a general meaning, but in education, we might say that it has to do with the appropriate allocation and use of human and financial resources within and across educational levels and sectors. This means that we want to be sure that we're making the best use of resources since resources for education, as for every governmental function, are limited. And we want to make sure that we are not wasting them. For example, if we are investing very heavily in hardware and software for computers but are not instructing the teachers on how to use this hardware and software effectively or are not teaching them how to integrate technology creatively into their classes by giving low-level instructions, then we are not making efficient use of our resources.

So the lack of Efficiency can occur in many ways. Another term that we sometimes use is return on investment. In the field of business, return on investment means the rate of profit as a function of your capital investment. However, in the education system, it's sometimes not so easy to make these calculations; in fact, it's often extremely difficult. But we can do it in certain indirect ways; for example, we can say that in the United States, the current statistics are that overall, only 60% to 70% of the students who enter the ninth grade (the first grade of high school) actually graduate from high school within six or seven years. This would mean a very low return on our investments. So we consider the investments on students up to the ninth grade and then maybe one or two years into high school. If the students still do not achieve a high school diploma, then we know that we're not getting a very good return on investment from the societal point of view as well as in terms of the amount of time. This low return on investment plays out in their very poor economic prospects. In the United States, students or individuals without high school diplomas will have much lower chances of landing a job, earning a decent wage, etc. So there is a very poor return on investment from both societal and individual points of view. All of these are different examples of what we would mean by an efficiency goal or efficiency target within an education policy framework.

Equity

Equity means different things to different people, but it basically has to do with fairness. In English, equity has the connotation of fairness. For example, we would argue that an important goal for any nation is to make sure that

the opportunities for access and quality education are fairly distributed across all relevant subgroups.

In the United States, children who live in very poor neighborhoods or poor districts do not have equal opportunities because, typically, the resources that are available in the schools that they attend are much lower than the resources available to children attending schools in middle-class or upper-class districts. This is because in the United States, in most cases, school funding comes from the property taxes of the local neighborhood. So if you come from a poor community, you will have a lower tax base. Therefore, there are fewer funds available for education. If you live in a wealthy suburb, the taxes are higher, and so more money can be allocated to education. So there are very substantial differences in the learning opportunities because the available resources differ even across districts within the same state.

Further, Equity not only means the fair distribution of opportunities but also active efforts to reduce what I would call structural inequities. So differences in financing are structural inequities. Other structural inequities, at least in the United States, come from the fact that union contracts allow teachers and senior teachers to decide what kinds of schools they want to go to. So senior teachers who maybe very good typically do not choose to go to poor schools. Therefore, the students who are most in need of the best teachers are least likely to get them. This is a structural inequity that has to be dealt with directly.

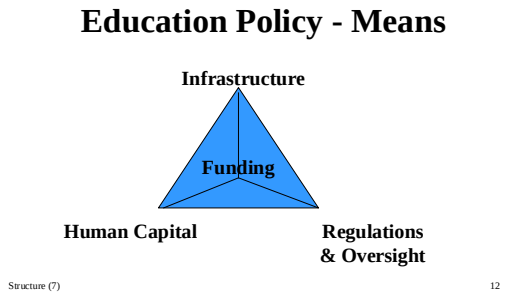
From a purely logical point of view, I think you could argue that Equity is an aspect of Quality. But it's such an important societal issue in itscultural and political aspects, that it deserves its own place as a vertex in our framework. Another reason is that if Equity doesn't get its own space, it is often overlooked. So I argue that it's very important to have Equity as a goal equal to Quality and Access.

Strategy

In preparing this lecture, I have felt that Strategy tends to be country specific. So I'm not going to say much about Strategy because I think that it tends to be very much conditioned by the political and cultural conditions, which of course change over time even in a single country and can differ very much across countries. So in Japan, where you have a very centralized system, I think that a lot of the strategy really comes from the central government and is played out through the different prefectures. In the United States, it's quite different because under the U.S. constitution, education is under the legal purview of the states. The federal government plays a very limited role in education. So, different states have very different ways of playing out education strategies. Another difference is that the federal government tries to influence all the states to move in a particular direction. Thus, the whole strategic issue and the way in which strategy is developed are very contextual, so I don't want to say too much about them here.

Means -Funding, Infrastructure, Human Capital, and Regulations & Oversight

Remember that Means are the tools, the decisions, and the activities that the government undertakes to achieve the goals, either directly or through its actions on other stakeholders. Again, I'm trying to restrict myself and I want to talk about four different kinds of Means: Funding, Infrastructure, Human Capital, and Regulations and Oversight.



Infrastructure and Human Capital

Infrastructure and Human Capital represent what we would typically consider as the basic aspect of governments' investments in education to achieve the goals. So they might be the physical plant—physical plant means the buildings and other hard objects and real estate—that is dedicated to education. They also represent the resources, textbooks, computers, and other supplies that we need as well as the teaching staff. When I refer to human capital, I'm talking about the people who are

playing some role within the education system, such as teachers, principals, school leaders, leaders at the prefecture level, leaders at the federal ministry level, and also the private sector. Both in Japan and the United States, the private sector plays an important role in education, for example, the juku schools and companies such as text publishers, test publishers, and curriculum publishers. So these are all part of what I call infrastructure and human capital and are Means to achieve goals.

Funding

Of course, all this takes money, and that's what Funding is all about. Funding itself has a number of different aspects, namely, its different sources. Again, I am not sure about how it works in Japan, but in the United States, we have multiple sources of funding—the federal government, state governments, and local governments. But above all this you have private or philanthropic foundations that invest literally billions of dollars a year in education. So they can also have an important influence in terms of funding.

We also have the distribution of funds, and how that plays out across different districts and at the school and student level within a district. In the United States, the federal government allocates several billion dollars a year, which is supposed to be focused primarily, but not entirely, on students who are very poor or who attend schools with large numbers of poor children. This funding is called "Title One." So there is a very specific funding stream that directs literally billions of dollars to that particular segment of the student population. This is why it is important to know about the funding, its sources, the amount, and the distribution.

I think it is equally important to know the time horizon, that is, whether the funding commitments are short term or long term. You know that serious improvement in the educational systems of the United States and Japan cannot be achieved in a year or two. Educational systems are extremely complex. They have a lot of inertia and don't change easily. Therefore, if you want to make a change, you have to allocate resources over a long period of time. Too often, at least from our experience in the United States, the commitments are really very politically driven. I mean that when the government changes, the commitments are changed and the reform effort becomes considerably weaker. So the time horizon—how far out the commitment is—is also a very important aspect of funding.

Regulations

Regulations are like laws, but they are typically more bureaucratic and administrative. When the state government or the federal government passes a new law, it suggests how things should take place, but the suggestions are never specific enough. As a result, these laws are interpreted by the appropriate Department of Education or the Ministry of Education. This results in thousands of pages of regulations that spell out very specific and detailed procedures. As an example, one issue concerns how each component of the system operates. For example, what do we expect from early childhood education, elementary education, junior high school, senior high school, etc.? What about non-public operators of schools? In the United States, many of the preschools are operated not by the government or the local school district but by private providers. What are the regulations governing these institutions? Who can provide such education? What new rules should be passed? What kind of staff do they need to have? In the United States, of course, we have many private schools, some are nonsectarian, nonreligious. We have religious private schools and charter schools. So there are many different kinds of schools and a whole body of regulations that govern their operation.

We also have regulations that govern the relationships among different components and levels. For example, how does a student pass from one level of the education system to another? What kind of coordination is needed between the different levels? Do regulations also govern the credentials and requirements of the school staff? What kind of qualifications does a teacher need to have? Do teachers need to reapply for their certification after so many years? In many states in our country, teachers do not receive permanent licenses and have to recertify. They need to be more professionally trained. Every five or ten years, they have to take courses in order to qualify

for their certificates. Principals also have credentialling requirements.

Oversight

Then what are the rights and obligations of each group—students, parents, teachers, administrators, etc.? What are the requirements for the curriculum and for assessment? Will the regulations detail the kind of curriculum you must specify for each grade and the level of detail, the kinds of tests that should be provided, how those tests should be related to the curriculum, etc.? All of these are governed by regulations, and oversight is very closely related to regulations. However, it has a slightly different connotation. It really has to do more with monitoring. In other words, are the different parts of the system actually following those regulations, and if not, why? And if they are, are those regulations having the desired effect in terms of the Goals of Quality, Access, Efficiency, and Equity? So oversight has to do with how we monitor the functioning of the system. Sometimes, it has to do with the determination of accountability, which means we have to decide if things are working well, whether we should reward some of the individuals within the system, and what we should do if things are not working well -- Should we fire the teachers, fire the principal, force them to change schools—all this is part of oversight.

Another very important part of oversight, although it's typically not a formal regulation, is how the results of this monitoring and accountability are reported to authorities in different levels and to the public at large. This is because oversight has its greatest impact when the results of the monitoring are reported to the public in ways that the public can understand and react to.

The Role of Testing

As you can see from this discussion, this framework doesn't necessarily call on testing. So we can establish the whole framework of our policy without thinking about testing. In principle, one can imagine a whole system that involves absolutely no testing. This doesn't really exist in too many countries and is probably less likely to exist in the future.

While the goals are set by society, the testing is an instrument of policy; it is just a means of achieving policy and typically operates through regulations and oversight or supervision. The role of testing is to provide tools for regulation and oversight, which are in turn means to achieve the goals. So typically, testing is not directly related to goals, but is part of a tool system that we use, which we have developed in order to develop better means for achieving the goals.

To achieve the Goal of "Access"

How does this play out? Well, we have test results that are obviously used for admission to an education program, for job selection, or for a professional license. This is an example of using testing as a means to achieve certain goals of access. So for selection, we use testing as a way of screening potential candidates to decide which students are most qualified to enter. In other cases, for example in e-learning, we use testing not to exclude students but to put them on the right path so that they can take the most appropriate course to achieve the required level or a combination of their level of achievement and their particular goals and targets for education and training.

Test results can also play a role in access because sometimes, in the absence of formal credentials, we can use test results to allow individuals to demonstrate their mastery. For example, in the United States students can sometimes avoid taking some courses if they do well enough in a certain test; they can move on to a next-level course. This is another example of how testing can influence access. But again, the test is not the goal. It is simply a means of achieving the goal.

To achieve the Goal of "Quality"

What about quality? Here things get a little trickier because—and I think this is one of the problems for testing—we sometimes use the test to define quality. I think when we do this, we tread on dangerous ground because we then confuse the test itself with the education goal, which become synonyms in the mind of the public. Yet, from a theoretical point of view, an education goal should never really be defined principally in terms of a test. However, , schools often define quality by saying that they want 80% of the students to pass the state exam

at the end of the tenth or eleventh grade. So we use the test results directly to define a quality goal. The problem, as you will see later, is that when the test is not very good, using it to define a goal can often lead to educationally unproductive situations; in such a case, we're better off trying to keep a distance between test results and education goals. In fact, in America, many test critics often protest that we define quality very narrowly in terms of test results, particularly because many of the tests are not that good. According to them, education should have a much broader definition. I'll speak a little more about this later.

On the positive side, particularly if the tests are good, test results can highlight where quality is lacking and indicate possible problems. For example, we might learn from a test that students in a particular school, in a particular district, are having trouble with algebra, which becomes a signal to the district or school administrator that the quality of the teaching in algebra needs to be upgraded or that there may be some other issues. So testing can provide a kind of warning light that can be very useful.

To achieve the Goal of "Efficiency"

Through monitoring, tests play a role in efficiency as well because they can help us understand the extent to which we have achieved certain levels of efficiency within the education system. For example, a national test or large-scale national surveys can be used to make comparisons of how the system is functioning among different districts or region. In the United States, for example, we have a survey called the National Assessment of Educational Progress (NAEP), which administers a test every two years in reading and mathematics for Grades4 and 8, as well as in other subjects on an irregular basis. This is given to a very large sample of students in every state. So the states use the NAEP results to compare themselves with other states and perhaps to make some judgments about whether their resources and educational investments are being used efficiently. They do so by comparing how their students are performing relative the students of other states. Obviously, since each state has its own testing system, we can't directly make comparisons from state to state, but we can do so through the NAEP.

Sometimes, specialized analyses of aggregate test results can be used to evaluate the contributions of individual units, like districts, schools, and teachers. That is, we can use the test results of groups of students within a classroom, within a school, or within a district to learn something about the effectiveness of that unit. It's not a simple issue, but it still is a potential use of test results that play a role in monitoring the system.

Further, we have international surveys like TIMSS or PISA that can provide inter-country comparisons. Apparently, in the last PISA results, both the United States and Japan didn't do so well. There is a great deal of discussion about how each country will improve its performance in comparison to other countries.

So I think testing can play the sort of warning role that may suggest that you are very happy with what you see internally by making internal comparisons. But now all of sudden, you find that compared with Finland or Singapore or South Korea, you're not doing so well.

To achieve the Goal of "Equity"

Finally, test results can also monitor equity. For example, if we see continuing systematic differences in levels of accomplishment between students from very rural communities and those from suburban or urban communities, then we see that there is an equity problem. Test results can therefore help us monitor lack of equity, and sometimes supplementary testing can be used to provide support to needy students to try to improve equity. So testing can play a very constructive role in moving towards this goal of equity.

Problems with Testing

The question we have to ask here is that if testing can play so many important roles in education policy, why is it so widely criticized? There are a couple of answers.

Answer (1) The quality and suitability of the tests

The first answer, which I have already suggested a few minutes ago, is the success of the testing in carrying out its roles. Again, the role is a means to achieving the goals; it's success depends on the quality and the suitability of the test for those goals, which are often inadequate for the

intended purposes. We know that many of the tests that we are using are not very good. This may be so because they've not been very professionally developed, or that they've been professionally developed but -- because of economic constraints -- they test very simple material rather than the more complex material. So students or teachers can play games more easily with these test results. Students can take the test and do well in it, but they are not in fact progressing educationally. So that's the first answer: Part of the problem is the quality of the tests themselves.

Answer (2) Campbell's Law

The second answer is something that we call Campbell's Law. Dr. Donald Campbell was a very famous social scientist in the United States. His law is as follows: "The more that any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures, and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

Campbell was saying that for any quantitative indicator (something that can tell us how a system is working, whether it's an education system or business), the more power it has (e.g. the higher the consequences attached to its results, then it will be subjected to more pressures—what we call corruption pressures—by people trying to fool the system. That is, they want to do well with respect to the indicator, but who are actually not doing well with respect to the process that the indicator is trying to track.

Let me give you an example from the business world. If we judge the management team of a business in terms of the company's stock price, then we know that the management team can manipulate the stock price in different ways to make it appear as if the business is doing very well, even if this not exactly true. Both the United States and Japan may have a similar system. In the United States, we had the Enron crisis, the WorldCom crisis, the Tyco case. So the social indicator, which in this case is an economic indicator (stock price), was subject to corruption because so much depended on it.

Well, it's not too hard to apply it to the education system. If the success of the principal and the teachers depends on how well the students do on the test, then certainly the hope is that the principals and the teachers will work hard to teach the children. But sometimes, they find shortcuts by which they can improve the test scores of the students directly, without necessarily educating them; this is a type of corruption. Campbell says that this should not be seen as an unusual event. He argues that this is a natural part of the way in which social and political systems work.

The risk of High-stakes testing

- Narrowing of the curriculum

So just to be more specific, when we talk about high-stakes testing—tests that have important consequences for the students, teachers, or the principal, we get a narrowing of the curriculum. That is, people are focusing just on what is being tested, and not necessarily on the whole curriculum. And again, if the test is not a very good test—if it contains, for example, just certain multiple-choice questions that focus on certain aspects of the curriculum—then school systems will tend to focus only on those aspects that are tested even though other aspects may be equal or even more important but difficult to test. An obvious example is that it's easy to test for factual information and simple logical skills, but much harder to test for creativity, problem-solving, etc. Further, tests that are devised to draw that kind of information are very expensive. So if we have economic constraints, as we usually do, they tend to drive out the more complex kinds of tasks. Hence, we end up with tests that lead to a narrowing of the curriculum.

- Inequitable allocation of teacher resources

We can also have inequitable allocation of teacher resources. What we're seeing now in the United States with this new accountability law, which focuses on the number of children in a particular grade or a particular class that are moving above a certain level, is that many teachers are focusing all their energy on the children just below the required level because they are the ones most likely to benefit from intensive instruction to move to the next level. Why should we bother about students who are very good? They are already above the level, and students who are too far behind can never catch up, so why bother? An unproductive, inequitable allocation of teacher

resources is the result, and many studies have shown that this is actually happening.

- Unfair treatment of students

It may also result in unfair treatment of students. If schools and school principals are held accountable for the proportion of students who meet a certain standard, sometimes students are encouraged to leave school or, because many of them have discipline problems, are suspended just before the testing takes place. Since they are not in school, the statistics of the school are improved. We would not call these as educationally productive actions.

- Widespread cheating

Finally, it leads to cheating either by the students or by the teachers. These are just different examples of Campbell's Law playing out in the real world of school accountability.

Answer (3) Poor test quality causes more serious problems

My answer number three is really a combination of one and two, which are already described. It mainly states that if Campbell's Law says that many of the problems of testing—in fact, maybe most of the problems of testing—are inherent or unavoidable because of its roles with respect to regulations and oversight, then poor test quality, in a particular context, makes it more likely that the consequences of Campbell's Law will occur. In other words, I'm saying that although Campbell's Law is pretty well shown to operate in most social contexts, the better the test, the less likely it is for those taking the test to engage in corruption.

For example, if you want to become a pilot of a plane, you have two kinds of tests—a written test and a flying test. You can have some pretty good written tests, but maybe it's possible to pass the test by studying in a certain way without really knowing the material. However, there is only one way of succeeding in flying a plane—you have to fly the plane, and if you don't, the consequences are very grave for you and whoever else is in the plane with you. So the authentic tests, the tests that are very close to what you're really trying to do educationally or in the training session, are harder to manipulate. It's harder to corrupt such an indicator because it's so close to what you're actually doing.

In fact, one of the sayings that's quite popular now in the United States is "Let's build tests that are worth teaching to." In other words, "Let's build tests that are so good, so clearly articulated with our education goals, that teaching students how to pass them is, in fact, teaching them the curriculum". The greater the distance between the test and the curriculum, the more likely it is for the predictions of Campbell's Law to come true. I think that a lot of the criticisms of testing in society have to do with Campbell's Law. People are reacting to the operation of Campbell's Law in a context where the tests are really not as good as they should be and where, in particular, we're defining education goals in terms of the test and putting too much pressure on the system to produce certain kinds of results, sometimes without the appropriate capacity.

Systemic Validity

So what's the answer? Well, there's no simple answer, but I'd like to propose here that we think about something that I and others call "Systemic Validity." Systemic, which is an adjective, means having to do with the system, and validity means that it is valuable and appropriate for the system. I'm going to define this in more detail on the next slide.

If we're going to strengthen and improve the contributions of testing to achieving goals, there are roughly two directions that we can take. One is a "top-down approach", using the idea of systemic validity in the way in which we implement education policy and the way in which we play out the means of accomplishing it. We use testing to accomplish the goals, and I'll say something about that. The second suggestion is more of a "bottom-up approach", using testing to enhance the infrastructure, which involves building the system's capacity to do better and to achieve the goals. I think both of these are viable directions; in fact, in order to take the best advantage of testing in an education system, I think we need to use both the top-down and bottom-up approaches.

Anil Kanjee, a researcher in South Africa, and I gave the following definition in a paper that we wrote:

"Assessment practices are systematically valid if they generate useful information that supports the improvement in one or more aspects of Access, Quality, Equity and Efficiency without causing undue deterioration in other aspects or levels." What we're trying to get at here is that too often when we develop policies, we focus on one of the goals and say that this policy will improve this goal. But we either ignore or overlook the fact that by improving this goal, we may in fact be creating another problem with respect to another goal that's even more serious. So overall, we're not making any improvement.

Let me give you some examples. In the United States, there has been a move, probably from the last 15 years, to reduce class size. That is, we're saying our classes are too big; 25–30 students in a class is too many and it would be better to reduce class size to a level of maybe 20 students per teacher. There is some research to show that this is reasonable. So sometime in the mid- or late-90s in California, they said, "Yes, we're going to improve our system, so we're going to issue regulations that say that in the elementary grades, no class can have more than 20 students." This sounded good. But then, they had to think about how they would actually implement this law. First, they needed many more classrooms, but there was not enough money for those classrooms. So they brought in what were called trailers, which were parked in the playgrounds. Extra classrooms could actually be made in these trailers, which, of course, was not an ideal setting. Then there was another problem—where would they get all the teachers? Well, it turned out there were not enough qualified teachers. For example, if they increased the number of classes by 35%, they needed 35% more teachers, which didn't exist. They then had to allow unqualified or uncertified teachers or bring about what they called emergency certification; as a result, the teacher quality decreased. So we can ask whether we are really better off having smaller classes but with more students being taught by unqualified teachers. We have to ask whether this law, which sounds very nice, was able to accomplish its goal, which is quality education—that students are learning more and in greater depth to prepare themselves to move to the next level. I think the answer is probably no.

There is another aspect explained by another example in the 80s in New York City. New York City has its own college systems, called City University of New York, which has both four-year colleges and two-year colleges. Many years ago, this was actually a very high-quality system. They had very stringent entrance requirements, very good faculty, and in fact graduated many Nobel Prize winners. But around the 80s or 90s, different immigrant groups came into New York City—particularly from Mexico and Spanish-speaking countries in Latin America—who were having much trouble with the English language and weren't able to meet the entrance requirements.

Since everyone pays taxes to the system, there was political pressure for everybody to have access to the system. So they changed the system and said, "We're going to have open admissions, so anyone who has a high school diploma can enter the system." Unfortunately, in the United States, having a high school diploma is no guarantee of your readiness for tertiary education. So they have this enormous influx of students with access, but the quality and efficiency decreased because more and more of the university system's resources were being devoted to providing remedial courses on elementary reading, writing, and arithmetic to students who were nominally in college. This led to what I would call a lack of systemic validity by trying to focus on or by focusing on a single goal—access—and ignoring how all this would play out with respect to other goals. We ended up with something that turned out to be educationally unproductive, not only for the system as a whole but for the students themselves, because the students who were ready for college were not really benefiting from it. The system has now gone back to a more selective access policy, but for a long time, it had almost destroyed the university.

Systemic Validity (Extended)

So our argument is that we're now talking about having more general education policies, and I think we can use the same idea. Education policies are systematically valid if they result in decisions and actions that lead to progress for one or more of these intended goals—access, quality, equity, and efficiency—without causing regression with respect to other goals. So this requires us to say, "always

have a broader view, even if your intended policy is focusing on a particular goal." Think about how this is going to play out with respect to your other goals. Make sure that your system has the capacity to achieve this goal without causing regression in other goals.

As I said, people have used the term "systemic validity" in a number of different directions, but typically more in a pure testing setting. Heyneman(1987), Heyneman and Ransom(1990), and Frederiksen and Collins(1989) have all have used this term with respect to testing, particularly with what they call the "backwash effects" of a test. That is, if you have a very good test, then it can be systemically valid in the sense that it can encourage teachers to teach to a higher level in order to meet the test requirements. This would be a positive backwash effect and would therefore, in some sense, contribute to or be an example of systemic validity.

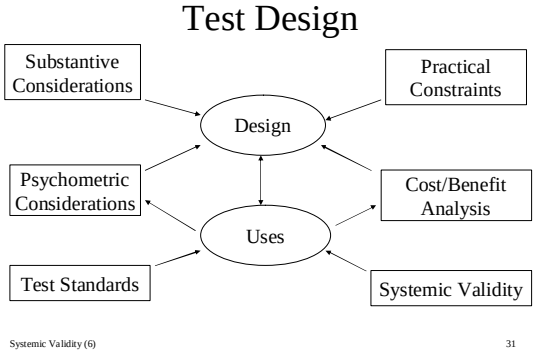
Another related concept is due to Samuel Messick(1989), my late colleague at ETS, who developed the criterion of "consequential validly". According to him, when we think about the validity of a test, we don't have to restrict ourselves to construct validity, which is the theoretical or empirical justification for the test -- we have to also think about how the test is played out and how it is used in society. Consequential validity is how the test results are interpreted and the consequences for individual institutions. He argued that it should be part of the overall framework for test validity and it is clearly closely related to the idea of systemic validity.

How can we enhance Systemic Validity?

So how do we enhance systemic validity? How do we move toward it from the point of view of testing? Well, the two most obvious ways are to improve our test design; that is, pay more attention to the quality of the test. This is based on my previous argument that many of the corrupting influences that Campbell's Law predicts come from the fact that our tests are not really as good and as appropriate as they need to be. We also need to follow proper test use guidelines that say, for example, that we should never use a single test to make high-stakes decisions about students or teachers. You should always use a broader range of information, although in fact we too often use just a single test. Of course, we're therefore putting too much pressure on that test, which leads to pressure to corrupt that indicator.

Also, my third point says that there ought to be more realistic scenarios of planning, and that before we actually develop these laws and regulations, we really ought to think about how they're going to play out over time, given realistic expectations of the capacity of the system. So going back to the example in California, it probably wouldn't have taken a lot of thought for people to realize that if the number of classes is going to increase by a third, then they would need a third more space and a third more teachers. Where would all this come from? But somehow, incredibly, nobody seems to have thought this through!

Test Design



Now what about test design? What I have here is a kind of a schematic of the influences of test design. This was meant to try to show you how notions of systemic validity can influence design. So we have typically substantive considerations like subject matter, etc. and practical constraints, such as the cost of the tests, the way in which they're going to be administered, etc. All of these influence design. Psychometric considerations of validity, reliability, etc. influence the design and cost-benefit analysis, of course.

At the same time, it is clear that the test design should be influenced by the use of the test. In fact, some people argue that test design should be a process that works backward, from the use to the design. But we talk about test standards and systemic validity only when we think about test uses. So I think that if we are more serious

about thinking through systemic validity with respect to the different uses of tests, it will eventually influence test design in the same way that psychometric consideration now do.

Example: Testing for Accountability

Let's talk about an example. Again, if we're testing for accountability where we're going to use the test results to evaluate districts or schools or teachers, the question is—Are they doing a satisfactory or competent job in education with the evaluation leading to a wide range of consequences from rewards to sanctions, punishments, or dismissal?

The simple-minded rationale for this is that high-stakes testing will stimulate people to work harder. It will also cause a productive reallocation of the existing resources. In other words, the governments are saying, "Look, the education system is very inefficient. We're spending money and we're not getting very much for it because people are really not paying attention to what's important." So if we introduce high-stakes testing, it will be a means to achieve equality because what's going to happen is that people are going to realize that we're really very serious and that we're really watching what they're doing, and they are going to then reallocate their efforts in a more productive way. That's what, in policy, is called the theory of action. How is high-stakes testing going to actually improve things?

What we're going to be talking about therefore is test quality and construct validity. Construct validity involves checking the theoretical basis for saying that this test will accomplish its targets. For example, being able to identify who has mastered this or who hasn't. The two threats to construct validity are construct under representation—when the test is narrower than the curriculum—and construct-irrelevant variance—if there are sources of noise in the testing system that are irrelevant to what you are trying to test. For example, we use the term "standardized testing" because the original meaning of standardized testing is that tests are given under standard conditions, that is, everybody takes a test under the same kinds of conditions and with the same amount of time. This is a fairness issue so that differences in the way tests are administered do not contribute to the variance in the results. We want the variance in the results to reflect mostly variance in the construct, as opposed to variance in the difference of the means of testing.

We also have to ask questions about the administration. Are we going to use multiple forms? If you use the same test repeatedly, and it's a high-stakes test, then people are going to memorize the items; they're going to pass them. This is what we have found with high-stakes tests all over the world—no item is secure. Therefore, as long as we're going to make decisions on a single test, we're going to have this kind of a problem. If you're using multiple test forms, are they equated properly? If they are not equated properly, then we have a fairness problem because some students taking one form will be disadvantaged relative to other students taking another form. And also of course, with administration protocols that permit various kinds of accommodations, psychometric questions about comparability arise.

So when we talk about test use in context, we're asking about whether this is going to be the sole basis for evaluation. To the extent that the test is carrying the burden of accountability, Campbell's Law is going to play itself out.

What are the standards for the evaluation? How are we going to make judgments about the individuals on the basis of the test? Is there any appreciation for the fact that any test result is subject to uncertainty, and how do we take that into account? And finally, what's the relationship of this test result with the larger accountability system? All of these things have to be thought through in systemic validity.

Perspective of Systemic Validity

What I'm arguing here is that from the policy point of view, from a top-down perspective, if we are proposing an accountability system that involves testing, we need to be more specific about the theory of action. For example, how will this system accomplish its goals? What is the empirical evidence to support this theory of action, and is there empirical evidence to the contrary?

Let me give you an example. One of the continuing controversies in public education in the United States is whether we should hold students back if they don't do well in a particular grade. For example, tests are given at the end of Grade 4, and if the students don't meet a certain standard, they must repeat the grade.

In many states, we have what's called social promotion; that is, even if students don't meet the standard, they can move ahead to Grade 5. But in our current accountability system, many school districts are implementing these more rigorous standards for evaluation, and they're saying, "No, if the students don't meet a certain standard, they must repeat the grade."

But the research shows that having students repeat the grade doesn't help; they don't do better, they just fall behind. Then in later grades, they are encouraged to just drop out. So if you have a 16-year-old child who is studying in the ninth grade and has failed two or three times, he is quite likely to drop out of school altogether.

Perhaps, in principle or in theory, some would argue that this is a good thing. A student who doesn't meet the standard should have another chance to achieve it, but if the empirical evidence says it's not working, then that is or should be a stimulant to say, "Well, what's the alternative?" Maybe the alternative is, "Let's pass him on to the fifth grade, but let's give him extra instruction time and special tutoring so that he has an opportunity to catch up while he continues to move along with his age cohort." Maybe this would be a better strategy than promoting such students without any extra support or holding them back and creating other problems.

While it involves more work, it seems sensible to me because I think that without these kinds of systemic thinking, we end up with very poor efficiency in the system one way or another. So for systemic validity, I think we need to develop our plausible alternative scenarios, incorporating both the positive and negative unintended consequences and talking about the probabilities and costs associated with those unintended consequences.

This is well established in the business world. The Shell Oil Company is one of the pioneers in this sort of scenario planning and, to my mind, there is no reason why we shouldn't be applying this in general social policy and education policy. In fact, I would argue that we don't need very sophisticated policy analyses or scenario analyses in order to see problems with some of the very simple suggestions that come up in the education debates in the political circles -- at least in the United States.

I would like to see a more interactive policy design, now focusing on policies that involve testing. As we play out the different scenarios, we recognize that our initial idea was not so good, and then we refine that initial plan, which is not just a test itself but the system in which the test is embedded. This is in order to mitigate or make small some of those negative unintended consequences and improve the probability of getting what we expect. These design modifications could include improving the construct validity. We could say that given the immense importance of this test, we really need to spend more money, get more open-ended items, and get a broader test; otherwise, we're going to end up with educationally unproductive activities. We have evidence that this happens over and over again. We may need to have more than one test to support the evaluation. We need to set realistic standards; by this I mean that if we are setting very unrealistic standards, and if the people in the system recognize that they don't have the capacity to meet those standards and are yet going to be held accountable, then they will find other ways to get around the system.

We want to make sure, both substantively and from the point of view of appearance, that what we have is a fair system. So if we're setting standards and are using the test in a way that appears to be unfair, then we are going to harm the systemic validity. We want to make sure that we're also thinking about the appropriate allocation of resources. If we're introducing a new curriculum, new tests, and new standards for those tests, but are unwilling to invest in preparing the teachers to teach the new curriculum, then we're going to have problems. So the lack of allocation of resources has to be thought through in advance, not after it has had its effects.

Often, particularly when we're talking about radical changes in policy, we need to think about phasing it in, rather than making it all happen in one year. We may need to phase it out over a period of years so that the system can adjust appropriately. Trying to do things too quickly in very complex systems often leads to negative, unintended consequences.

I think these are just examples of how systemic thinking can help us develop more productive and more intelligent policies. I have to say that at least in the United States, many of the policies that we've been working with lately have not really been subject to good systemic validity thinking. Therefore, we are getting many negative, unintended consequences from them.

As I said earlier, this is looking at systemic validity as a top-down approach. In other words, this is something that comes from the policy level down through the system. When policy makers use tests to achieve their goals, then in my view, systemic thinking needs to take place so that we can make the best use of the test. If we don't do this, we shouldn't be surprised if testing becomes a very unpopular part of the system and actually becomes a basis for attacking the whole education system.

Building Productive Capacity

I think there is also a more constructive way to use testing: to build productive capacity, that is, to build the capacity of the system to achieve positive education goals. Data from the monitoring test can be used to build system capacity so that individuals and units—whether it's an individual teacher or a school that needs training and support—can make good use of this data, and the quality of that data will therefore affect its usefulness. I believe that to build capacity, you must invest in capacity. Testing is actually a very cheap way of becoming productive—of moving toward productive capacity.

Let me give you some examples. Many schools in the United States are engaging in what we call "data-driven" decision-making, where they are using data from their state tests to identify weaknesses—either at the classroom level or by subject—and then to allocate their resources and attention to those areas. So if a school teacher or school principal sees that her children in the elementary grades are not doing very well in reading, then she knows that she has to allocate more resources to improve the quality of teaching of reading. In those early grades, they need to do more monitoring of students and identify students who are having problems. These can be health problems, need for eyeglasses, etc. In poor districts in the United States, many of the children's learning problems often come from lack of access to proper healthcare. They don't have good nutrition, etc., and they may have also various kinds of disabilities that haven't been identified. So there are many potential causes of this problem. But without the test results, the school principal often doesn't know that these problems exist until it's too late. So this is a kind of simple example of how we can use tests to build productive capacity within the system.

Also, as I had said earlier, many school districts are using state accountability test data to identify teachers that appear less effective and to provide them with professional development to improve their quality. This is another example of productive capacity. As another example, the TIMSS—which is the international survey for mathematics and science assessment—often includes not only the tests themselves but very extensive pedagogical surveys, which look at the ways in which these subjects—mathematics and science—are taught in different countries.

Many countries are using the very rich database of the TIMSS—not just the results themselves but the database around the pedagogy, curriculum, and instruction—to rethink their own approaches and to suggest, for example, that they need to train teachers in a new way. Sometimes in the United States, we often hear that in Japan, mathematics is usually taught in a more conceptual way, and that while the number of topics that are covered here is much smaller than that covered in America, the topics are covered in more depth and students get a deeper understanding. I don't know if this is true, but this is what American researchers are saying, and they are trying to use that research to influence the teaching of mathematics in America.

Developing countries, countries in Africa or Latin America, are using the TIMSS resources to reform their own curriculums. That is, they are using the TIMSS assessments, the scoring guides or the open-ended assessments, to help them think through the changes in their own curriculums and in the way their teachers are being trained. In fact, they are also disseminating assessment results through their teachers or training institutes so that over time, the teachers will begin to move toward teaching to an "international standard" rather than a local standard. .

Formative Assessment/ Assessment for Learning

What's nice about these examples is that these resources are already there; that is, the amount of additional investment that a country needs—whether it's a developed country like the United States or a developing country like South Africa—is relatively small relative to the investment that's already been made in this international survey. So it's a very cost-effective way of building capacity. So I think that testing can also be used to build productive capacity through this bottom-up approach rather than a top-down one. In fact, the most powerful way is through what we call "Formative Assessment" or "Assessment for Learning."

Assessment for learning is any assessment whose first priority in its design and practice is to serve the purpose of promoting pupils' learning. That is, it's not to hold students accountable or to give them a grade, but to help them understand their strengths and weaknesses. An assessment activity can help learning if it provides information that teachers and pupils can use as feedback in assessing themselves and each other. Such assessment becomes formative assessment when the evidence is actually used to adapt the teaching work to meet the learning needs. This comes from a 2002 book by Black and William called "Inside the Black Box."

The idea is that instead of thinking of assessment simply as a regulatory tool or as an oversight tool, it should be regarded as a more integral part of instruction. While teachers always test their students as a matter of course, it's often done in a very informal way, and it's not always clear how the teachers use the assessment results to adapt their teaching practices. The idea behind formative assessment is that it does not actually become more formal, but the use of assessment data becomes an integral part of the way in which the teacher adapts his or her approach, either to the class as a whole or to individual students. Formative assessment becomes a day-to-day monitoring activity within the classroom. What's very interesting is that there have been a number of research studies both in the United States and in England where it's been shown that with strong formative assessment, you can have a substantial impact on student learning with effect sizes as much as 0.5 or more. That is, when we compare classes in which teachers have been trained to use formative assessment and to use it effectively, then often, their students can improve by half a standard deviation or more relative to comparable students and schools where the teachers are not using formative assessment in this more integrated fashion..

But to achieve that kind of effect size—and 0.5 is certainly very substantial in education—requires considerable investment in teacher development over two to three years. Teachers need time to become more comfortable with this and they need training to look at student material. It may be hard to believe, but at least in the United States, teachers are not taught very much about testing. If they take a course on testing, it's about how to interpret test results. They are not given courses on how to design tests or how to interpret test results in a way that really influences their instruction in a meaningful way; and that, in effect, is what formative assessment training is all about. It's really giving teachers the kind of professional development that they should have had during their initial training. Unfortunately, in the United States, they don't get such training.

Formative assessment can contribute to achieving the goals of quality and efficiency by improving both teachers' knowledge and skills. It does so by helping the teacher to regulate the learning environment, by which I mean that she is able to differentiate among students in terms of their needs and to adjust instruction appropriately. It increases student engagement because by doing this, you are identifying and helping the students who need that

help the most; therefore, it eventually enhances their understanding. That's why it is possible to get effect sizes of around 0.5 from properly implemented formative assessment.

Conclusion

In conclusion, we all know that the education system in a country like Japan or the United States is a very complex, multi-faceted enterprise and has strong traditions. It's hard to change. It's also influenced by many different political and professional interests—by teacher unions or principal unions, politicians who are trying to make statements, businesses, etc. Real lasting change is very difficult to achieve. While testing has usually played an important role in regulation and oversight, I think that it can do much more. By this creative infusion of testing into the learning process, I think we can substantially enhance quality and efficiency, and by rigorous application of systemic validity, I think we can improve the policy uses of testing. We really need a better balance between assessment for learning, which involves more formative approaches, and assessment of learning, which involves more summative assessment for accountability.

Right now, there is an imbalance in most countries. That is, we're doing most of our testing as assessment of learning for regulation and oversight. There's relatively little assessment for learning, and what's being done is not being done very well. We need to allocate more resources to strengthening assessment for learning, and we need to start today. I think this is an area in which technology can play an important role in supporting both teacher professional development and formative assessments. For example, web-based delivery can offer students immediate and continuous access to formative or diagnostic assessments in order to accelerate their learning in an individualized way.

To realize these possibilities, we need more disciplined approaches to the formulation of education policies. This can be achieved through careful consideration of systemic validity and return on investment. We can do a lot more through creative use of existing resources. A good example is the resources that exist around the large international surveys, which are sadly underutilized. Further, more involvement of measurement specialists in the political process is needed. As scientists, we tend to avoid political issues because we think that they're really somebody else's job. Though of course in this and other countries, scientists—in this case measurement scientists—are sometimes called in to provide advice on the political process. But often, they are called in simply to affirm or confirm what the politicians or the policy makers have already decided. I think we need to become more involved in those political processes because those processes or policy-making activities often do more to shape the role of testing than what we do in our classrooms and research laboratories. Unless we play a more active role, the issues of systemic validity and productive capacity are not going to get the attention that they deserve and that the education system needs. As a result, we will continue to invest resources unproductively and will always be complaining about lagging behind.

Thank you very much.

Questions & Answers

Q1: Could you give us a very good working example of "assessment for learning"?

A1: There are lots of examples in a particular area. For example, in mathematics, people like Black and Wiliam are working on a set of questions that are designed to pinpoint where someone is. Let me give you an example. There is a computer-based system for mathematics assessment called ALEKS (http://www.k12.aleks.com/). It was developed by a cognitive scientist Jean-Claude Falmagne, who is at UC Irvine. It's a very sophisticated, well-developed test system. It's an adaptive test that covers the content from algebra to pre-calculus. It yields a profile of that student in terms of strengths and weaknesses. On that basis, a student entering a class meets a teacher who understands very clearly where he or she needs some extra help. No grades are given, and it doesn't play a role in the student's promotion to the next year. It's really a way of informing the students and teachers about how to allocate their effors. So this would be a real-world example of assessment for learning.

Q2: Are the diagnostic functions included in those assessments?

A2: Yes. Dylan Wiliam is one of the leaders in this area. He joined ETS about three years ago and likes to make a distinction between diagnostic and formative, and it's a subtle distinction. But he would say diagnostic tests are not informative unless the teacher knows how to use the results productively to address the problem areas. So you can have a wonderful diagnosis, but if nobody does anything with it, it's not formative assessment. It's similar to a doctor doing a large set of medical tests. He may get a very good diagnosis, but if he doesn't treat the patient then he hasn't helped him. Of course, we would like to think that any diagnostic tests produce results, which then get used.

Q3: Do you think that we can use a large-scale test that students all over the country will take as a diagnostic test?

A3: That's a good question. There's an ongoing debate about that. I think most people believe that large-scale tests that are primarily designed for "summative assessment", or "assessment of learning", cannot function equally well in "assessment for learning". This, of course, is part of the problem, and I've given you the example of ALEKS, which is not a large-scale test and is adaptive. So the item sequence is adjusted according to students' requirements.. It's very hard for one single broad assessment to provide useful formative information for all students, and it's usually not part of the design. We sometimes try to do the best we can; for example, there are large-scale assessments in mathematics with different subscales. For example, in middle school mathematics—where you might be doing numbers and operations, measurements, data analysis, or geometry—you could have several subscales. Some people argue that providing results at the subscale level can help make such a summative assessment if not formative, then at least diagnostic. The problem is that typically, the results of these different scales are not very reliable and highly correlated, once you make an adjustment for the unreliability of the measures. It is not clear how much distinct useful information is being provided by the different scales.

Let me give you an example. The SAT, the Scholastic Achievement Test of the College Board, is one of the major college entrance exams (built by ETS) in the United States. Until a few years ago, the SAT score report would simply give you're the score in verbal reasoning, the score in mathematical reasoning, as well as the corresponding percentiles. But now it also provides a diagnostic profile based on a version of the rule-space model. But there is a lot of criticism Because the SAT is designed to measure individuals' verbal reasoning or mathematical reasoning, each along a unidimensional scale. There is a limit to the amount of divergent information it can hold about subscales or deviations from the unidimensional model I think there is actually some useful information that's coming out of the rule-space model, and I encouraged the College Board to move in this direction as an interim measure until we could get better diagnostic assessments. But I certainly admit that no fixed large-scale assessment can really function equally well as a formative assessment because the design principles are so different.

If you want to do formative assessment, you are going to have to approach it from a different perspective. If you think about designing from the end back to the beginning, you have to ask yourself these questions: What are my goals for this assessment? What kinds of information do I want to get or what kinds of decisions do I want to make? What kind of evidence do I need for those decisions? What kinds of tasks do I need to get that evidence? What kind of assessment can I have to put these tests together? You end up with a very different instrument if you simply want to determine the ranks of students on their verbal reasoning. It's a very different set of goals and you end up with a very different assessment. There is an analogy with cars that can go in water. They can work both on land and in water, but they don't work very well in either setting!

Q4: I would like to comment on efficiency. I think ever since the Coleman report, there has been a long history of using summative tests as some kind of guidelines for the more efficient use of resources on education, and

researches have constantly been carried out on this issue. For example, I remember that there is some evidence that shows that the investment on the number of teachers is less efficient than that on the quality of teachers. I think it's a problem of how to use summative tests as a basis for policies.

A4: One of the problems in using summative assessments as a basis for making preliminary judgments about system efficiency is that summative assessments reflect the entire course of the students' education careers up to that point; therefore, it's not a very good basis for determining, for example, the efficiency level within say the last two or three years. This is one of the reasons why the schools where it has been applied are trying to understand how much the students have grown and what has been their growth over the last three years. They use comparisons based on growth to make judgments about efficiency or effectiveness in a more focused – and fairer -- way.

The Coleman study is kind of interesting. The original Coleman report was published in 1966. His argument was that family background was more predictive of students' attainment than schools. But I think part of the problem was that, again, he was looking particularly at those students who didn't have really good test scores, and there was more variation in the student background than in the school. So a number of studies have suggested that there was some sort of fundamental flaw in Coleman's analysis. I don't know whether this is part of the curriculum here, but when we try to understand the efficiency or return on investment with respect to educational programs, there is a very poor record of being able to do that very well. There is a new book, the second edition of Levin and McEwan, that does a very nice analysis. Henry Levin is a Professor of Economics and Education at Teachers College. They do a very nice analysis of this whole issue of trying to get estimates of return on investment in the educational sector at both the school and program levels. Basically, he says that there is not a single good example of that. So it's really an open question and a very important one because policy makers and the public are always asking this question. I think it is the best that we can do with these kinds of very gross comparisons, but unless we follow up with more in-depth humanistic investigations, we are as likely to be misled by those quantitative indicators as we are to be properly informed.

Q5: How do you think about the idea of "educational production function." I can say that it sometimes lacks attention to technology and just focuses on the physical aspects.
Let me mention one example. The public class size in Japan is much bigger than that in the U.S.; nonetheless, the achievement level is just the opposite. So there may be some differences in the technology or financial conditions.

A5: There was a famous social experiment in the United States in the State of Tennessee: The Tennessee class size experiments in the mid- to late-80s, where they randomly allocated students and teachers into different treatment arms. In one treatment arm, the class size had to be less than about 18 or 19; in another, it would be about 23 to 25 (somewhat similar to regular classes); and in another, it was regular classes with a class aide to help the teacher. They wanted to compare the performances of the students in these different settings. In fact, they found that the students in the smallest classes, with 15 or 17 students, did learn more; but they also found that the impact was the greatest for poor minority children. So the children who came from very poor and impoverished backgrounds without very much preparation for school were able to receive more individualized attention in the smaller classes, which really made a difference. But for most other students it wasn't very important. So I think the argument is that we're talking about empirical evidence. We're saying that if we're going to make an investment in class size, which will be enormous, then it should be much more targeted than simply making statements as in California: "Let's put everybody in small classes" and raising the possibility of unintended consequences.

There is a book by Richard Rothstein called "Class and Schools." Part of Richard Rothstein's argument is that in the United States, testing alone is not going to improve education, and high-stakes testing alone isn't going to make American education the first in the world. This is pretty obvious to us over here. In fact, he argues that not

only the achievement gap between the United States and other countries but also that within the United States is a function of inequities with respect to nutrition, housing, health, etc., and these inequities are simply reflected later on in educational differences. He is an economist by training.

He considers the cost of reducing class size in elementary grades for the whole country. He comes up with a very large number, in the range of billions of dollars. He then arrives at a number of the same order of magnitude by estimating the cost of providing every child in America has a decent breakfast, a decent supper, enough healthcare, glasses or hearing aids if needed, etc. and that their parents have enough money and a permanent home so that their children are not in foster care. He actually argues that it would be more cost effective to focus energy and resources on the context in which the children and their parents live than to invest them indirectly in the education system, like in reducing class size. I don't believe that everybody would accept Rothstein's argument. I'm just saying that this is the beginning of a long argument that if you really want to improve things, this is how it should be done. This is not just a matter of means but also of strategy, and that's why I say that discussions on strategy are really so contextual because you have different issues of equity in terms of access to health and food in Japan than we have in our country. I don't think you have the same gaps as we do.

In fact, one of my colleagues at ETS, Paul Holland, was on a national panel that was trying to estimate food insecurity. That is, they try to identify different categories of families and children at different levels of food insecurity. I have no food insecurity problems. In the last year, I've always been able to find enough food to eat; in fact, I had too much food to eat! But their study took into consideration the proportion of families that in the last six or twelve months had at least one month of continued problems in terms of getting enough food. The number of school-age children in families with high levels of food insecurity is enormous, and we should ask whether children who come to school hungry can really learn well. We have all sorts of school lunch programs. But school breakfast programs are provided only at local levels, if at all.

A study was carried out in Chicago. It was found that of the 5,000 children in elementary grades who were performing well below the grade level, some 40% or 50% had health problems -- they needed eyeglasses or hearing aids. They also needed other things. They were suffering from asthma and other such diseases, which were really interfering with their learning. This may not be an issue in Japan, but it's a very big issue for us, particularly because the federal government doesn't recognize or chooses not to recognize the fact that we have these enormous inequities in the resources available to students, quite apart from the educational issues. So it's an ongoing discussion. For these reasons I find many discussions on education production functions quite sterile and not particularly helpful.

Q6: You mentioned about the diagnostic tests, and I totally agree with their importance. But they can be very costly and I am not sure about the level of effectiveness of the present tests. How many batches of such tests are going around in the U.S. and who is producing them? Who are using these tests in reality and how do they choose them?

A6: I haven't done a survey in the market. I discussed ALEKS because it has been around for quite a while and is getting good publicity. I do know of some "local initiatives." For example, Uri Triesman at the Dana Center of the University of Texas in Austin has been working in the area of formative assessment. I think some work is being done at the University of California, Berkeley, and UCLA, CRESST. So there are different types of local initiatives. There are also these formal programs like Success for All and Reading First type of initiatives, which have very rigid structures but also would in some sense fit under the umbrella of formative assessment.

We tend to think of the test as providing good formative assessment. But unless there is a teacher who can utilize the information, you won't really get much leverage. Unfortunately, teacher training requires a much bigger

budget than testing. This is because once the system is in place, for example, the ALEKS system, the cost per student is quite small This is not the case with training teachers to take advantage of ALEKS. I think one of the failures of our system is the inability to recognize that having teachers grade these kinds of assessments, taking more time to work with other teachers, and to score student material contributes enormous value to professional development..

In the advanced placement program in the United States, in which high school students can take college courses, the tests that are given at the end of the year are accountability tests. Most of the assessments have both multiple-choice questions and constructed responses. Students have to write essays or solve long problems. Then many teachers are brought together from different parts of the country to do the grading, and they have a chance to discuss the student work with other teachers. Teachers are not so well treated, they have to live in dormitories and they hardly get paid. But they love it because they find it to be a very useful and stimulating experience. This is the reason why we had them take this model and used it more generally. In fact, Dylan William is doing that right now in the United States. He is trying to help schools recognize that a formative assessment program tied to the professional development of teaching is the way to really build capacity in our schools.

(end)