第1回日本テスト学会賞授賞記念講演

比較の測定から変化の測定へ

池田　央
日本テスト学会理事長
東京大学教育学部156番教室
2007/12/01

はじめに

- この夏、思いがけず、第1回の日本テスト学会賞受賞の栄誉に預かりました。これもひとえに事務局はじめ、学会成立から今日まで支えていただいた多くの方々のご努力の賜物と深く感謝申し上げます。
- 何か記念講演をということですが、改まって申し上げる話題も思いつきませんので、本学会事業の一つとして完成いたしました「テストスタンダード」などを素材に、学会とテスト研究者と社会とのこれからの結びつきで日ごろ大事だと感じているようなことを中心にお話してみたいと思います。

本学会の特色

- **会員構成**：何らかの形でテストに深くかかわりを持つ人の集まりです。

  (専門領域) 心理・教育測定、情報通信、言語、医学、など多岐に、また
  (業務) 教育、研究、開発、臨床、行政、実務、出版、など多岐にわたります。
  特色：学際的、テストの理論と応用、新テストの開発などに関心。
- **目的**：学会でなければ難しい課題の解決や支援に役立つこと、そのため
- 科学的基礎に裏付けられたテスト法の社会への定着、普及、
- テスト諸業務の無駄をなくし、規格化、標準化、共有化への提案、協力、
- 情報通信の先端技術 (ICT)を駆使したテスト法の研究開発と実用化、
- ICT時代に即応したテスト測定技術分野の研究者・利用者の育成、
- 海外との情報交換、国際的にも通用するアセスメント技術の研究開発など。

まずはテストへの共通認識をたかめるため「テスト・スタンダード」を作りました。

本日のテーマは

- われわれがテストといえばすぐにイメージするのは**選抜**のためのテストである。
- それは、テストの点数を人と**比較**して判断するのに使われている。
- テストの点数は選抜だけでなく、人がどれだけ学習し、成長し、**変化した**かを知る道しるべ（milestone）として役立つものにしたい。（それは思いのほか難しいが）
- それに近づけるには何が必要か。そのために**テスト学会**が貢献できるとすればなにか考えてみたい。

テストの点数（得点）が持つ意味

次のステートメントに示すテストの点数の解釈は正しいでしょうか？（「テスト・スタンダード」（Q&A22）より）

（例1）あるテストで70点を取った人と60点を取った人とでは、このテストで測られる能力は70点取った人の方が高い。

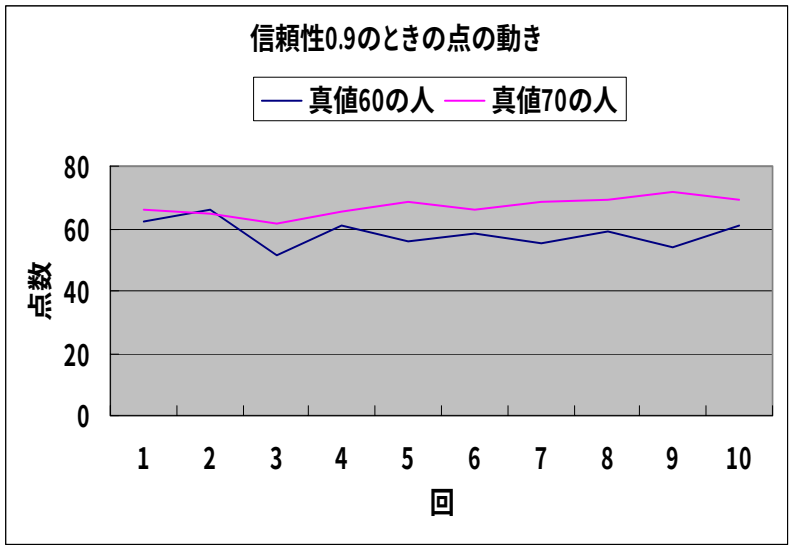
（例2）国語のテストで70点取った人と算数のテストで60点取った人は国語の能力の方が高い。

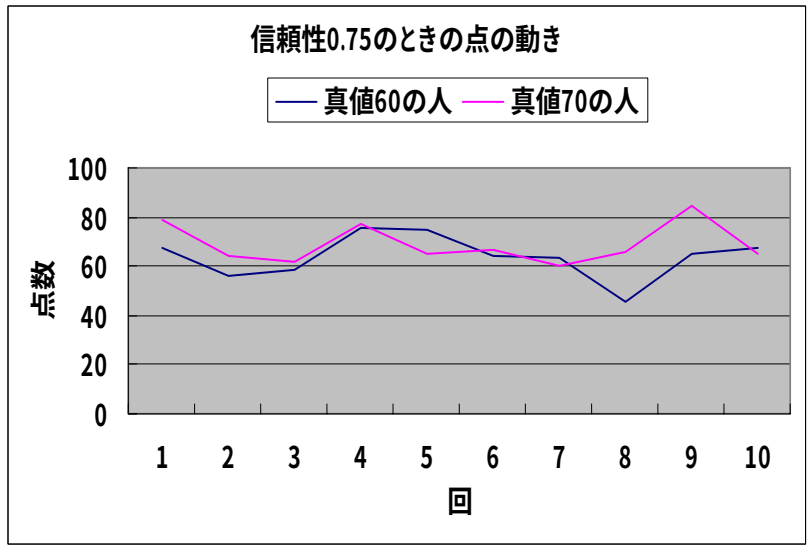
（例3）ある科目のテストで1学期の点が60点、2学期の点が70点取った子はその科目の学力が向上したといえる。

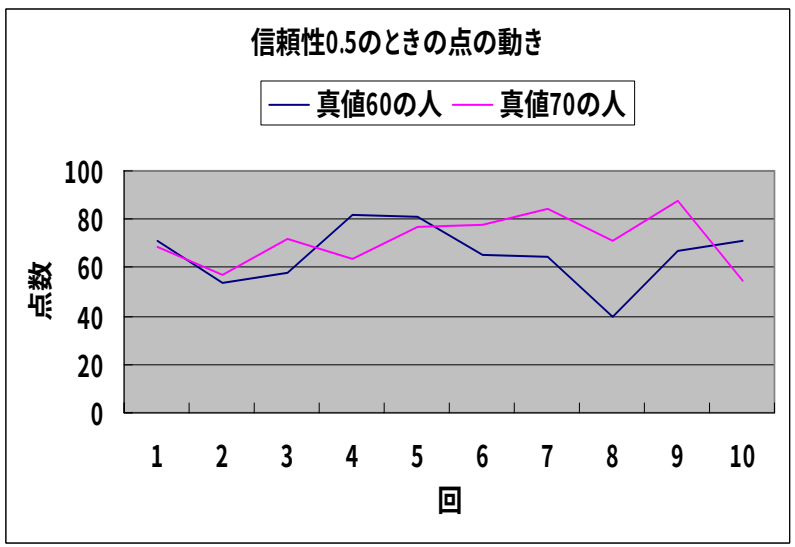
（例4）学期初めと学期末に同じテストを使って学力の伸びを調べた。学期初めが30点で学期末が50点の子と学期初めが60点で学期末が70点の子では、前の子のほうが学力の伸びが大きい。

答：以上のいずれもNo

(例1) テストが同じでもそのままの点数での比較は難しい







信頼性が異なる3つのケースで真値が60,70点の人の観測値の動き (シミュレーション)

100回のシミュレーションで、真値の差が10点あるのに観測値では下回ったケースの数

信頼性係数	0.9	0.75	0.5
X60>X70	4	15	20

⑵例2⑵テストが違うと比較は一層困難になる

- 例えば大学入試センター試験の主要科目で見ても、平成9年～18年の10年間で一つの科目の平均点の変動が少ない科目でも9.3点、大きい科目では27.5点もあり、3倍近くの開きがある。
- また同じ年で見ると、いちばん平均点の高かった科目と低かった科目との開きは、小さい年で11.5点、また大きい年では35.9点と、これも年によって3倍くらいの違いがあり一定しない。
- このように、点数の大小をそのまま比較しただけでは、それが能力の差を表しているかどうか判断するのが難しい。

一方、社会における評価への期待も変化してきている

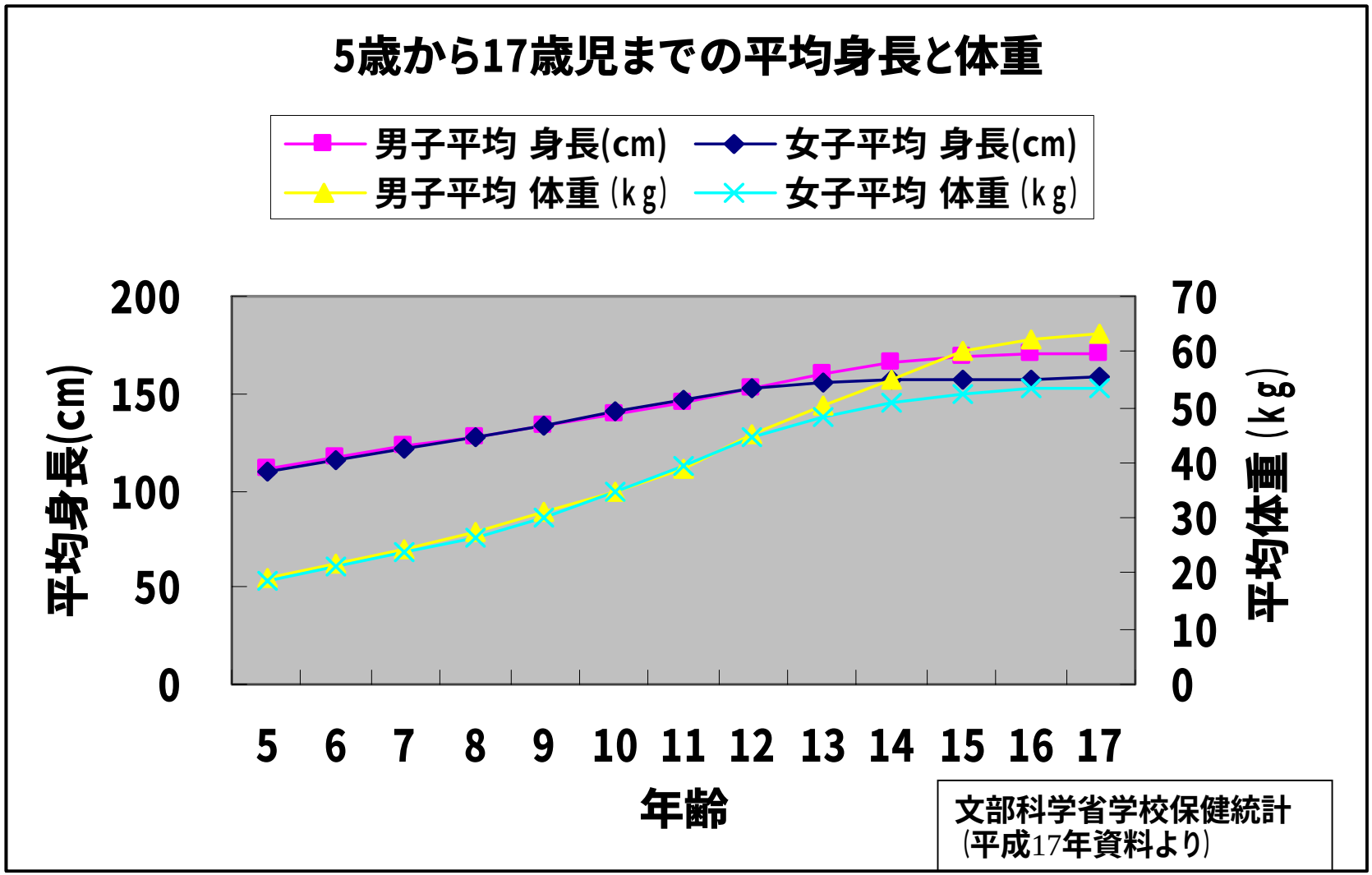
- 例えば、相対評価から絶対評価への重視、
- 従来の5段階評価をやめ、偏差値表示も避ける。
- 総括的評価から到達度評価、そして形成的評価へ。
- さらに、学習への意欲、関心、態度の重視。
- ほしいのは現状を知り、成長・発達、学習成果を表す情報。
- つまり学習経過が分かる情報、それには連続測定が必要。
- これは個人間の比較から個人内変化の重視を意味する。
- しかし、いずれもうまく行ってはいない。
- その測定に必要な技術と環境が伴っていないからといえる。

数値で変化がわかる情報の例は

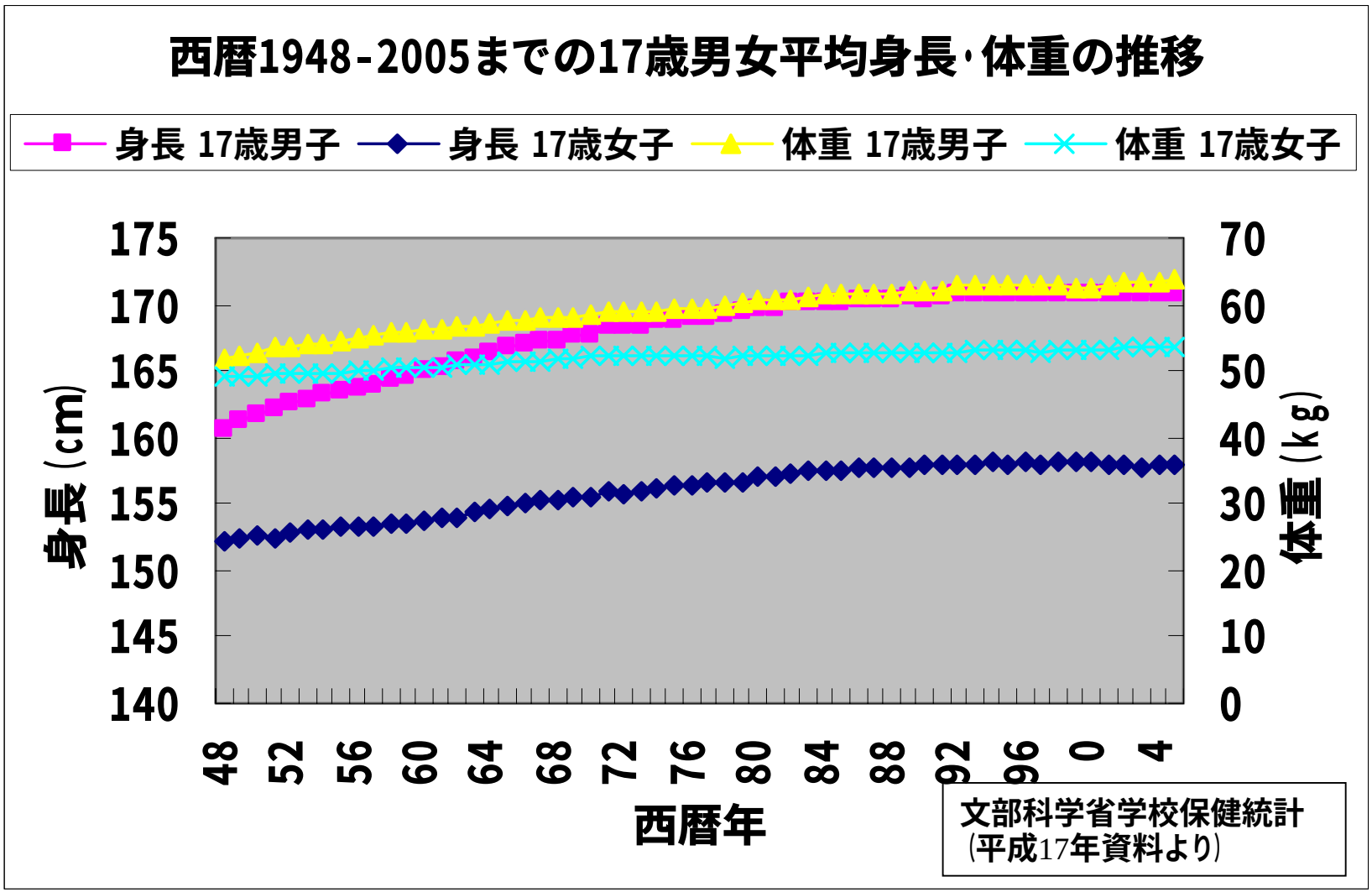
- 5年生の男の子が1学期終わりに身長を測ったら142cm、2学期終わりに測ったら147cmだった。その子の身長は伸びたか？これは（例3）の問を身長に置き換えたものである。
- ある子どもの体重は30kgあったのが、50kgに増えた。一方別の大人は同じ期間に体重が60kgから70kgに増えた。子供の体重の増え方の方が大きかったといえるか。これは（例4）の問を体重に置き換えたものである。

- これなら、前者では5cm伸びたといってよいし、後者では、子どもは20ｋｇ、大人は10kg増えたので、子どもの方が大人の2倍増えたといっても不自然ではない。変化が分かる。
- 一層詳しく次の例を見よう。

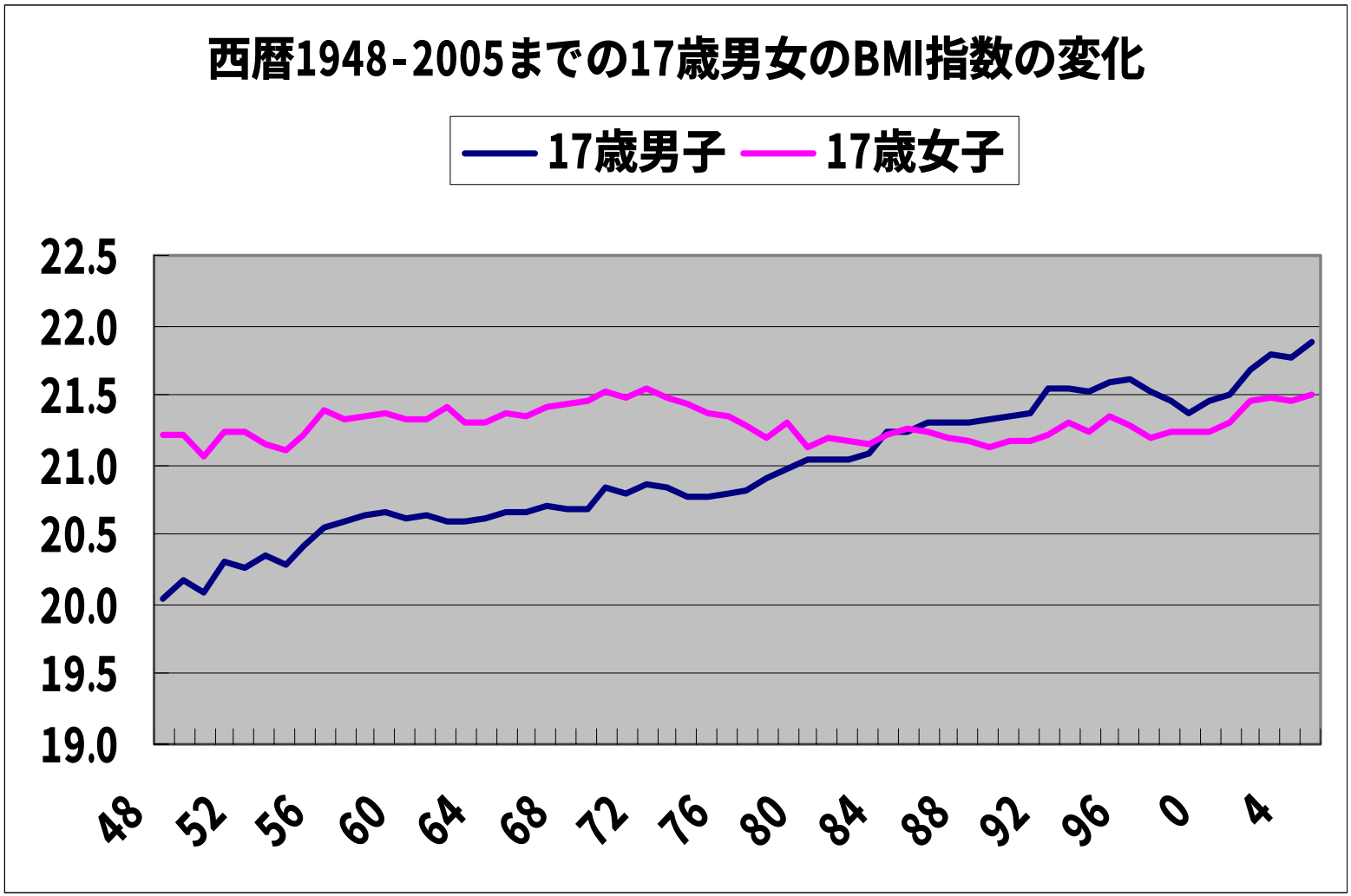
年齢とともに伸びる


5歳から17歳児までの平均身長と体重
男子平均 身長(cm)
女子平均 身長(cm)
男子平均 体重（kg）
女子平均 体重（kg）
平均身長(cm)
200
150
100
50
0
70
60
50
40
30
20
10
0
平均体重（kg）
5 6 7 8 9 10 11 12 13 14 15 16 17
年齢
文部科学省学校保健統計
（平成17年資料より）

時代とともに変わる17歳の平均身長と体重



時代とともに変わる17歳の平均体重kg/平均身長m^2



©Hiroshi Ikeda, 2006/9/17

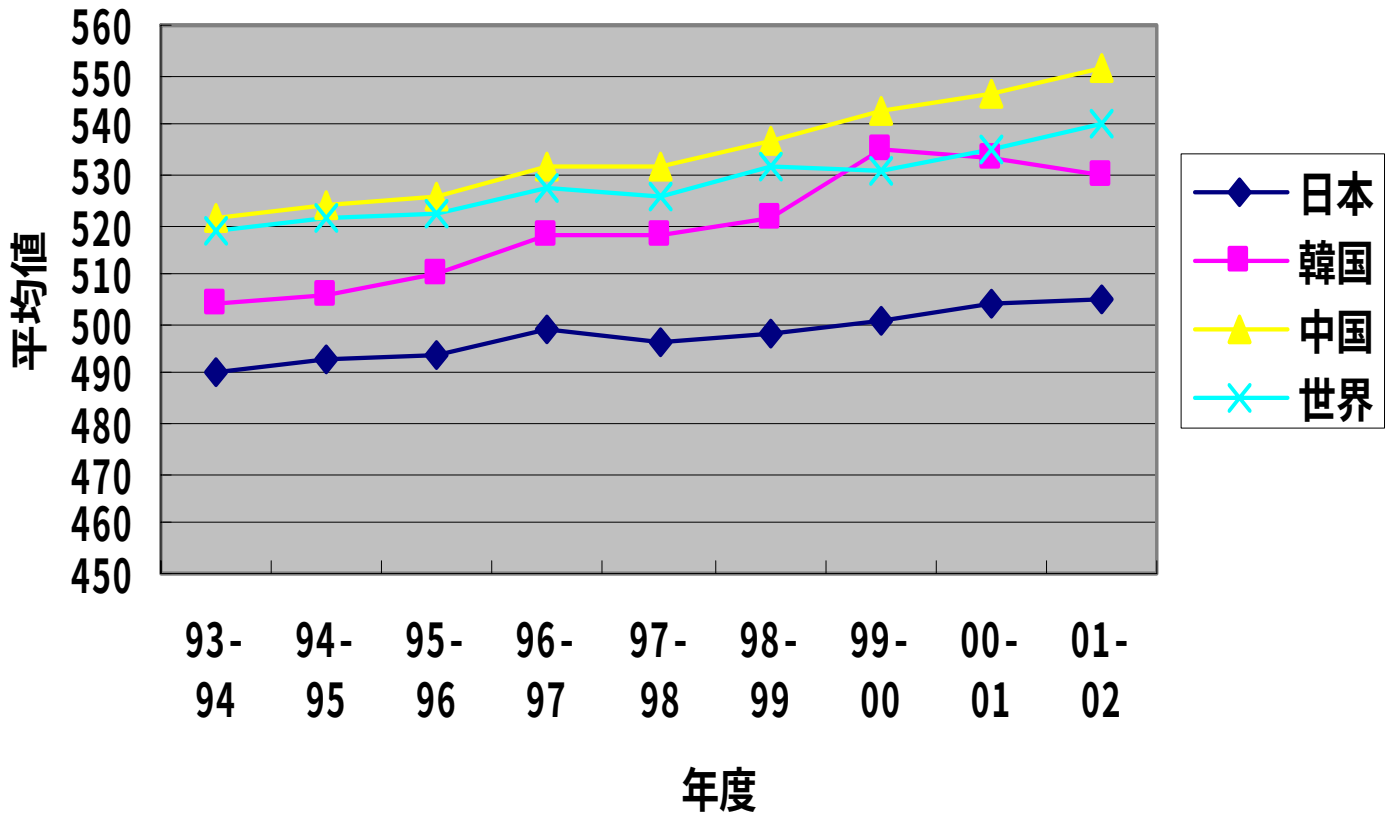
これらの変化の図を見ていえることは

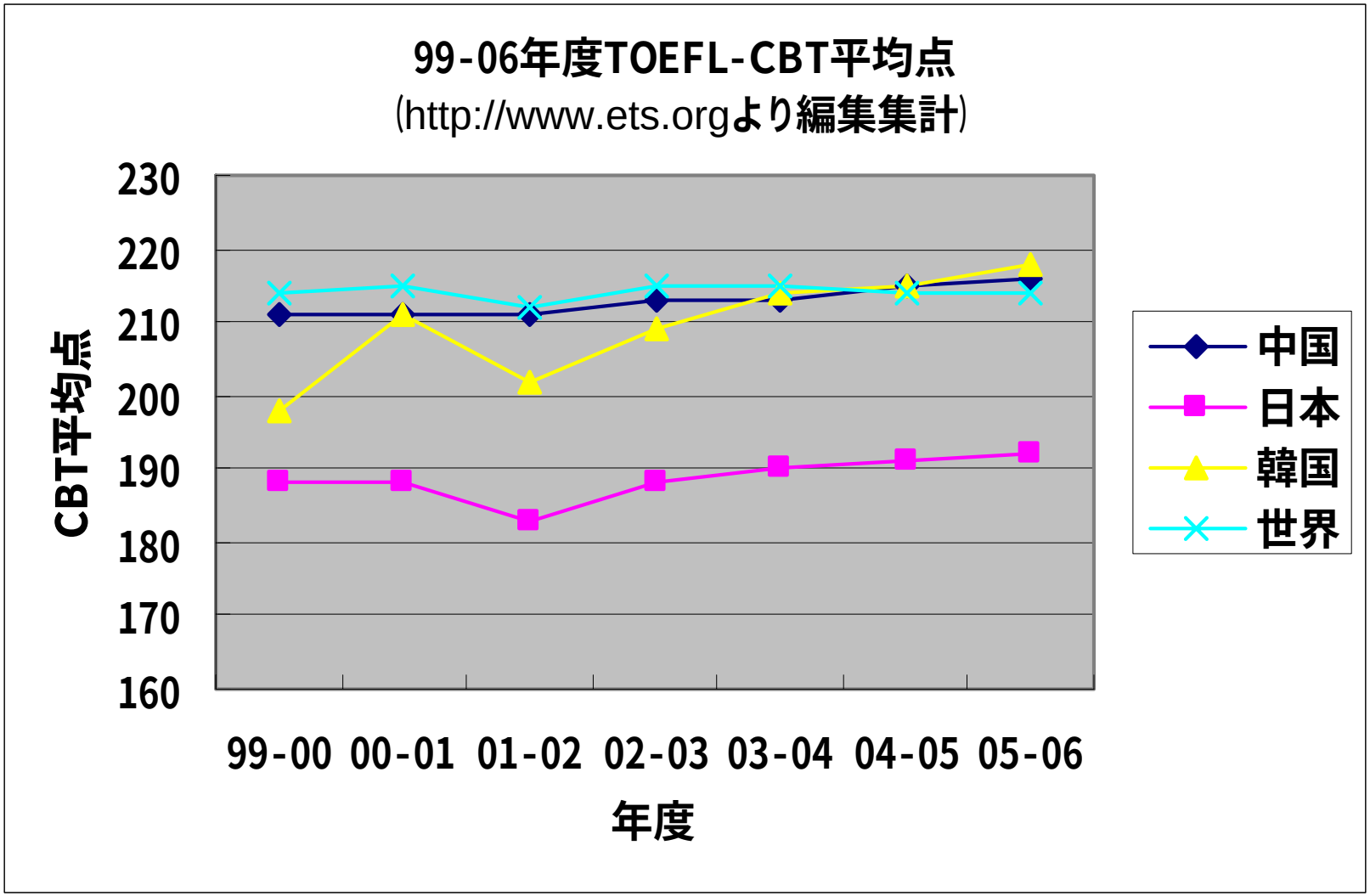
身長や体重の年齢および時代的変化を見ていえる特徴は、

- きわめて連続的で**スムーズ**な変化を示し、平均値が急激に変化したり、不規則に大きな変動をすることはない。
- それは常に**同じ尺度**（物差し）で計測しているからで、尺度が変わっては変化の様子は分からなくなる。
- では能力などを測るテストではどうだろうか。尺度の統一化という作業（専門用語で**等化**という）をすればどうなるか、
- 外国語としての英語能力テスト（TOEFL）（ペーパーテスト（PBT93-02）とコンピュータテスト（CBT99-06））の例やOECDの15歳児学習到達度調査（PISA2000,2003)などの例をみてみよう。

TOEFL-PBT総点平均値の国別年次変化

(http://www.ets.orgより編集集計)




99-06年度TOEFL-CBT平均点
(http://www.ets.orgより編集集計)
CBT平均点
230
220
210
200
190
180
170
160
99-00
00-01
01-02
02-03
03-04
04-05
05-06
年度
中国
日本
韓国
世界

PISA（読解力テスト）が示す国際比較

順位	国名	2000年	国名	2003年
1	フィンランド	546	フィンランド	543
2	カナダ	534	韓国	534
3	ニュージーランド	529	カナダ	528
4	オーストラリア	528	オーストラリア	525
5	アイルランド	527	リヒテンシュタイン	525
6	韓国	525	ニュージーランド	522
7	イギリス	523	アイルランド	515
8	日本	522	スウェーデン	514
9	スウェーデン	516	オランダ	513
10	オーストリア	507	香港	510
11	ベルギー	507	ベルギー	507
12	アイスランド	507	ノルウェー	500
13	ノルウェー	505	スイス	499
14	フランス	505	日本	498

http://www.mext.go.jpより編集

これらのグラフから分かること

- TOEFL-PBTの 世界の平均値は徐々に増加、それに伴って、日韓中ともに増加、ただ国により増加模様は異なる。
- 各国共通した増加状況（96-97年度の上昇）と国で固有な状況（99-00年度の韓国）などから、それぞれの変化の特性が推察される。
- TOEFL-CBTの99-02の日韓の乱れはPBTからCBTへの移行期における混乱（世界平均は安定）による。CBTが本格的に始まった03以降は安定化に向かう。中国はPBTがまだ主流。そうした変化の様子、国の特殊事情などが平均点の動きから理解できる。
- 米国の多くの大学で、入学にはPPTで550（あるいは600）以上、CBTなら213（あるいは250）以上などと条件をつけて大きな不都合はないが、日本で仮に60%点以上を合格などと決めると年による変動が大きい。
- PISAもまだ歴史が浅いが、2000年と2003年を比較しても、第1位から14位まで比べて分かるように、国こそ入れ替わるものの平均点と順位は驚くほど一致している。　　こうした尺度なら変化の様子が比較できる。

成長･発達･学習成果がなぜ分からないか

厳密に変化を知る満足なテストは少ない。理由は、

- 適した問題がすぐには得られない（項目情報の不足）
- 解答の迅速な収集システムが未発達（集個技術）
- 採点の理論と技術の未完成（とくに実技、スキル）
- 解答分析の結果がすぐに出せない（即時処理）
- 誤差と変化を区別できる精度不足（規則･不規則変動）

プロセスの処理には時間が勝負、
　壁は作問過程と採点過程、CBT化はそのカギを握る。

変化の測定で考えねばならない問題

解決しなければならない問題にどんなものがあるか

- 何を使って**測定尺度**とするか（項目の確保、選択）
- **測定規模**はどのくらいか（受験者の確保）
  - ー**長期大規模**調査（long-term trend analysis）
    - 縦断的研究（longitudinal study）　集団選択の問題
    - 横断的研究（cross-sectional study）　サンプリングの問題
  - ー**短期変化**の分析（short-term variation analysis）
    - 小規模ローカル測定（教室サイズ、学校単位）
- **受験者人数**の問題（数十人、数百人、数千人、数万人）
- **時間間隔**の問題（時間単位、日、学期、1年、数年単位）

変化量は時間との関数。変化に対する尺度の**敏感性**が問われる（課題や項目の選定）。

測定規模を決める上での問題

- **大規模調査測定**（集団動向調査）
  縦断的研究（longitudinal study）　（コホート分析）
  －パネルの取得困難、特定集団による偏り、長期間調査の問題
  横断的研究（cross-sectional study）（国際比較に多い、）
  －横断的大規模調査、集団としての変化量、定期調査の必要性
- **小規模ローカル測定**（学習効果、教授法）
  －学校･教室単位、短期連続測定、事前事後測定、一般化された結論を出す難しさ、短期間に効果判定を出す困難さ

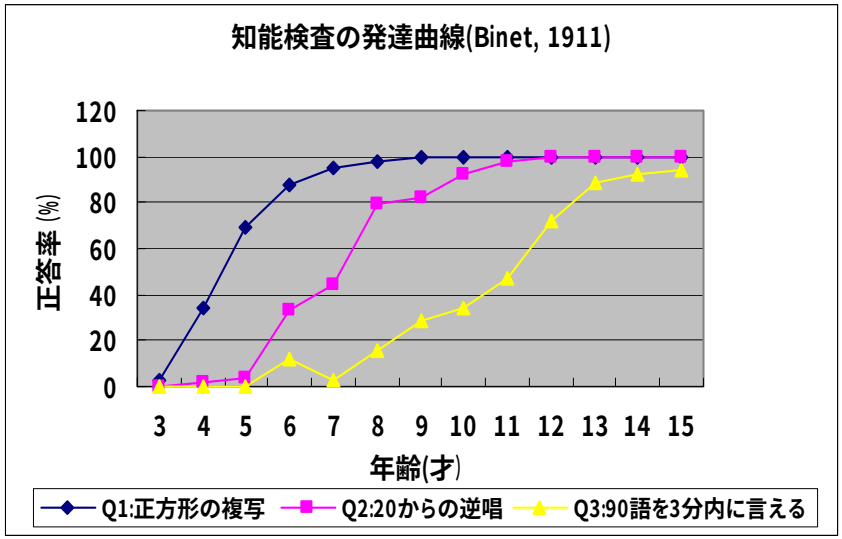
時間間隔のとり方の問題

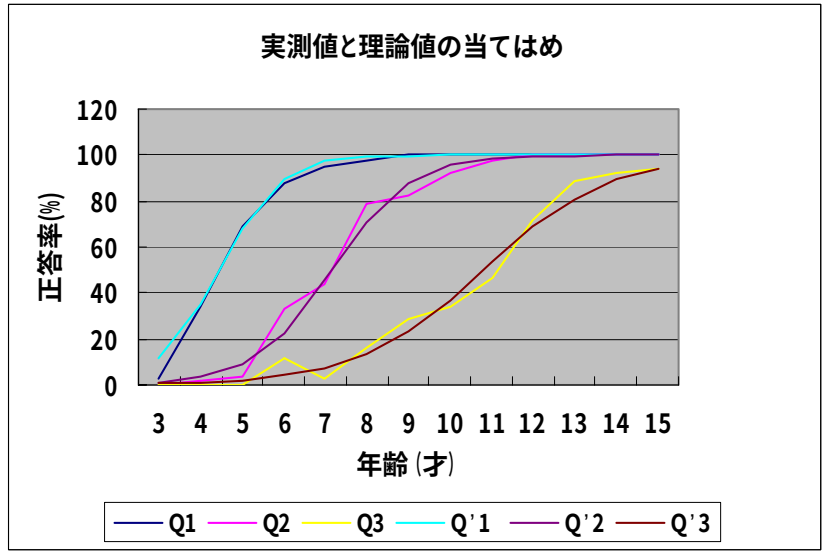
- **成長･発達**による変化の測定は時間との関数である。
  - ー変化の規則性をみるには時間軸を等間隔に。
  - ー成長発達期（幼児･生徒）は年齢との関数でみる。
  - ー間隔が長ければ変化は見やすいが、短いと誤差との区別が難しい。
- **長期変動調査**（long-term trend）　（NAEP, PISA, TIMSSなど）
  - ー「変化を測定するには、尺度を変えてはならない」（Beatonの言葉）
  - ー変化をみるには尺度の連続性、接続性が大事。異なる尺度は共通尺度化することが必要（vertical equating）。
- **短期変動調査**（short-term variation）
  - ー事前事後測定か、連続測定、継続測定か。効果が分かりにくい。
  - ー短期学習テストでは、開始からの時間、反復回数など計測可能な量との関係で見る（英単語量、漢字語彙量、単純計算の速さなど）
  - ー個人の変化は総括的評価テストによる全体変化と関連付けてみる。

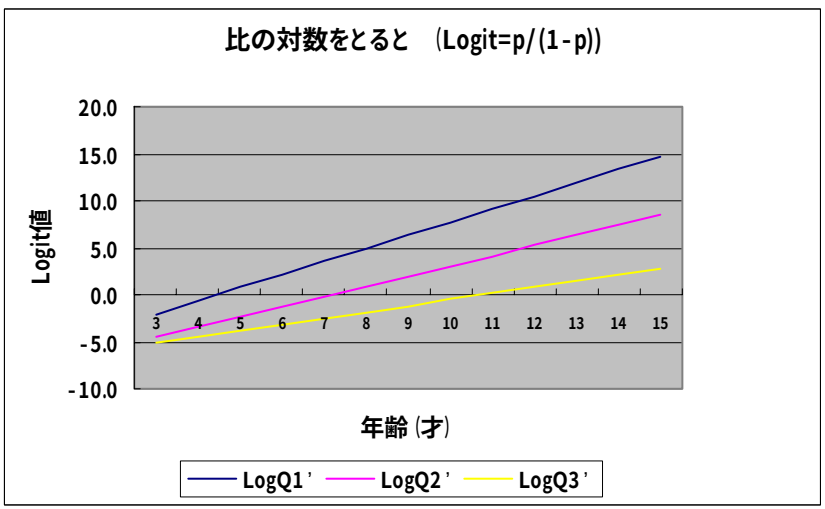
測定尺度(measure)のとり方の問題

いかにして、尺度の**同一性**を確保するか。いくつかのタイプ

- **問題固定**型：同じ問題を反復使用、発達検査、定期的見直しも必要。
  ～問題非公開。ビネー式知能検査、MMPIなど伝統的心理測定の方法。
- **集団固定**型：パネル方式、受験者を固定（規準集団）して、項目比較。
  ～一般化が集団に限定、作業の共通化、標準化で、メタ分析資料を容易に。
- **問題等化**型：共通問題、共通受験者の情報を利用する。
  ～予備テスト、アンカー問題の必要性、測定方法の設計が鍵となる
- **項目プール**型：豊富な同種問題を用意し、ランダム抽出で、等質性確保。
  ～パラメタ付項目プールの準備、完成までに時間がかかる。
- **同種反復**型：英単語、漢字語彙、簡単な反復計算など、単純作業が主。
  ～ここではレベルの分かった同種の素材と量を用意。連続測定できること。







キー項目の選定による発達尺度の構築例

Stanford-Binet **V**

Woodcock-Johnson **III** など

Raschモデルにより、等間隔平行項目を用意、編集：

変化に敏感な項目尺度を

立ちはだかる課題（SMEとICTの活用）

必要となる最大の課題を二つにまとめると、それは
作問技術と集個集約技術の開発といえる。

- 特性の分かった項目情報が不足している。
  - 一巷に問題は多いが、付随する項目情報（パラメタ値、正答率等）の分かっているものが少ない（Item Bank の充実）。
  - 一分野内容の専門家(SME：Subject-Matter Experts)との協力、
  - 一既存テストの再分析（IRTの観点から見直し、項目情報の付加）、
  - 一項目自動生成（AIG：Automated Item Generation）技術の開発。
- 解答者の反応を即座に集め、返す迅速技術が必要。
  - 一近年のICTの進歩は凄まじい。テストへの積極的実用化を。

作問技術の研究開発

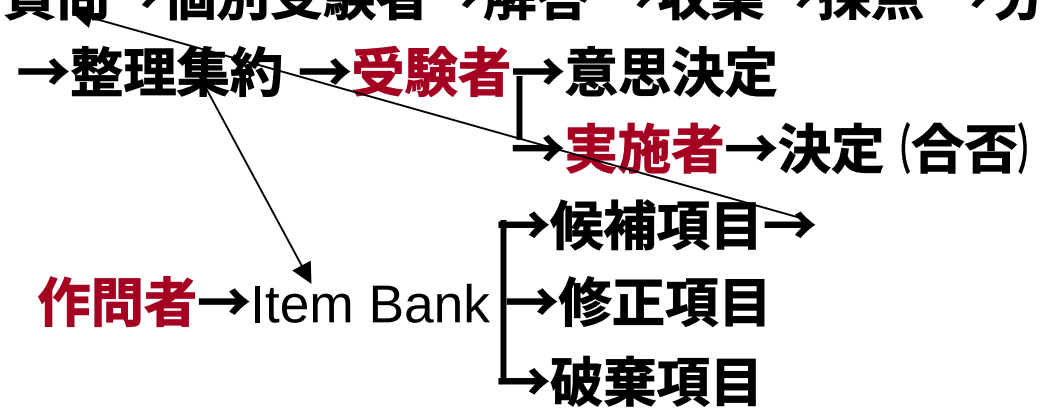
これはテスト研究の中でもいちばん遅れている分野

- 問題作成者の職人技を**製品化**へ（CBT化には不可欠）
- 測定**内容、認知対象、目標**の定義付け（分類法taxonomyの確立）
- **項目仕様、質問フォーマット**の規格化、標準化、普遍化
- **項目パラメタ**他、項目単位での情報整備　（項目の**部品化**）
- 項目の**分類・コード化**体系の構築　（**ラベル**作り）
- それにもとづく自動**検索編集**システムの開発
- 主に「テストスタンダード」第1章とくに「項目」の開発の部分
- **分野内容**の専門家（SME）と**心理・教育測定**専門家（PEM:Psycho-Educational Metricians）との協力体制が重要。

採点過程のシステム開発

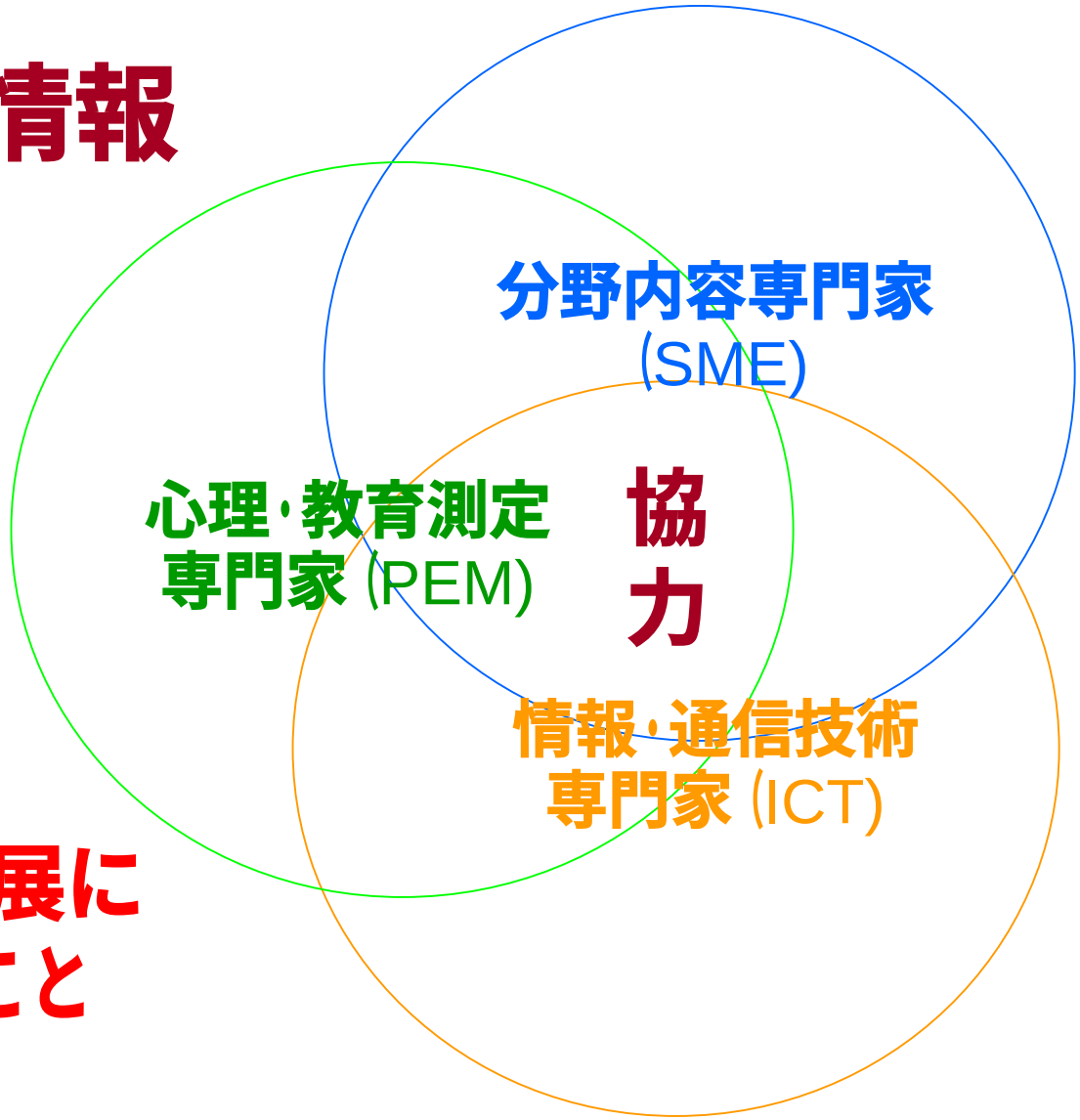
テスト･プロセス（情報の流れ）をスムースにするための研究開発

- 質問作成から実施、採点、分析、決定、情報管理に至る流れの一元化と自動化、迅速化が必要（CBT化）。　そこでは、
- 質問→個別受験者→解答 →収集→採点 →分析処理→（左へつづく）



- 解答方式、採点方式の規格化、標準化に対応できる設計。
- とくに人手が要るものPerformance Assessmentにおける評定システムの開発（Rubricの設定など）。
- ここではICT技術者と心理･教育測定家との共通理解と協力体制が大事。

テストは情報



テストの発展に必要なこと

学習者の自己理解を助け
持続的学習を維持するテスト

- 最後にひとこと：日本の学生は自ら学ぶことに慣れてない→テストはそれを**支援し育てる**テストでありたい。
- 強制練習（Drill）でなく**自発学習**（Self-study）を促すテストへ、
- しかし練習問題、質問、問題提起、課題解決機会は多くし、
- いかに続けさせるかが鍵→学習者に**役立つ**情報が必要、
- ナビゲータとしてのテストの役割→**目標が見える**形に、
  - ー**全体値（像）、現在値、目標値**の表示
  - ー**習得マップ**の作成
  - ーそれらは、いつでも**瞬時**に分かる状況でなくてはならない。
- 最初は**点**、それから**線**へ、そして**線**から**面**へと広げて行こう。

テストスコアには3つの情報が必要

それをイメージとして描くと～

1. **全体値** (マクロ情報)

Frameworking

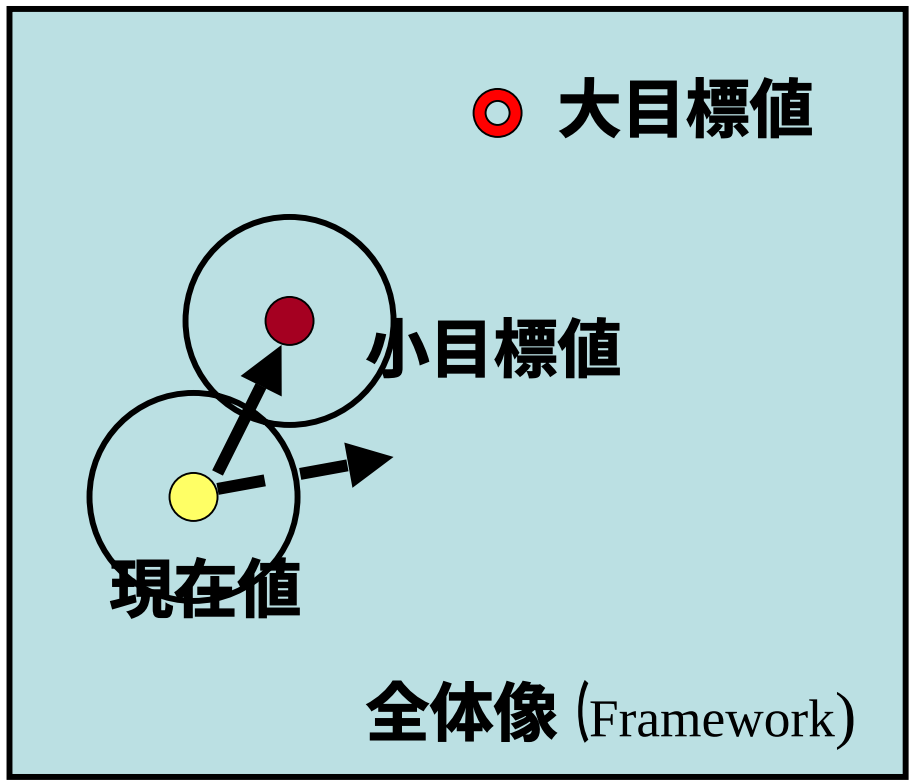
2. **現在値** (ミクロ情報)

Positioning

3. **目標値** (差分情報)

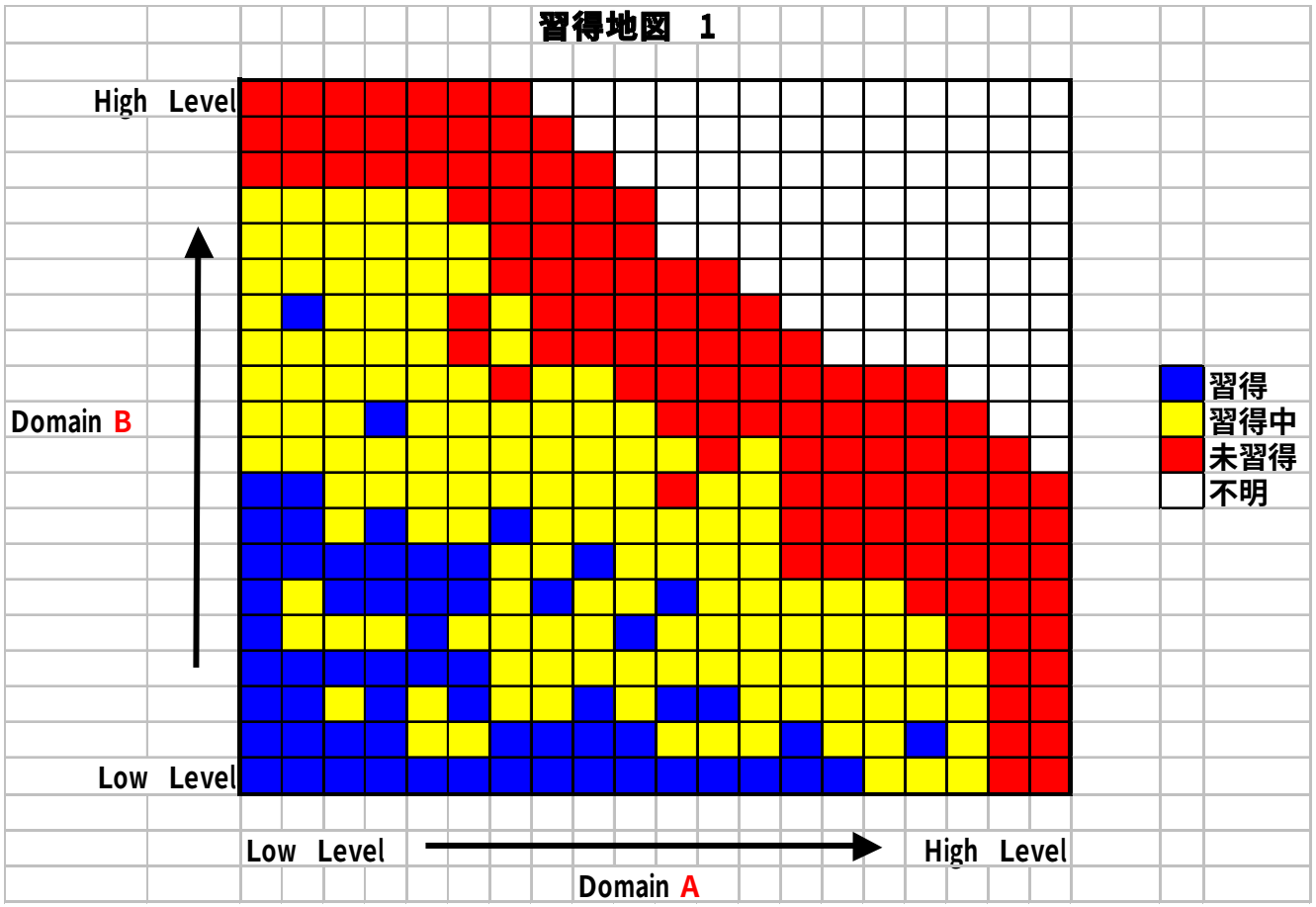
Goal setting

学習の動き (変化) が見える形で



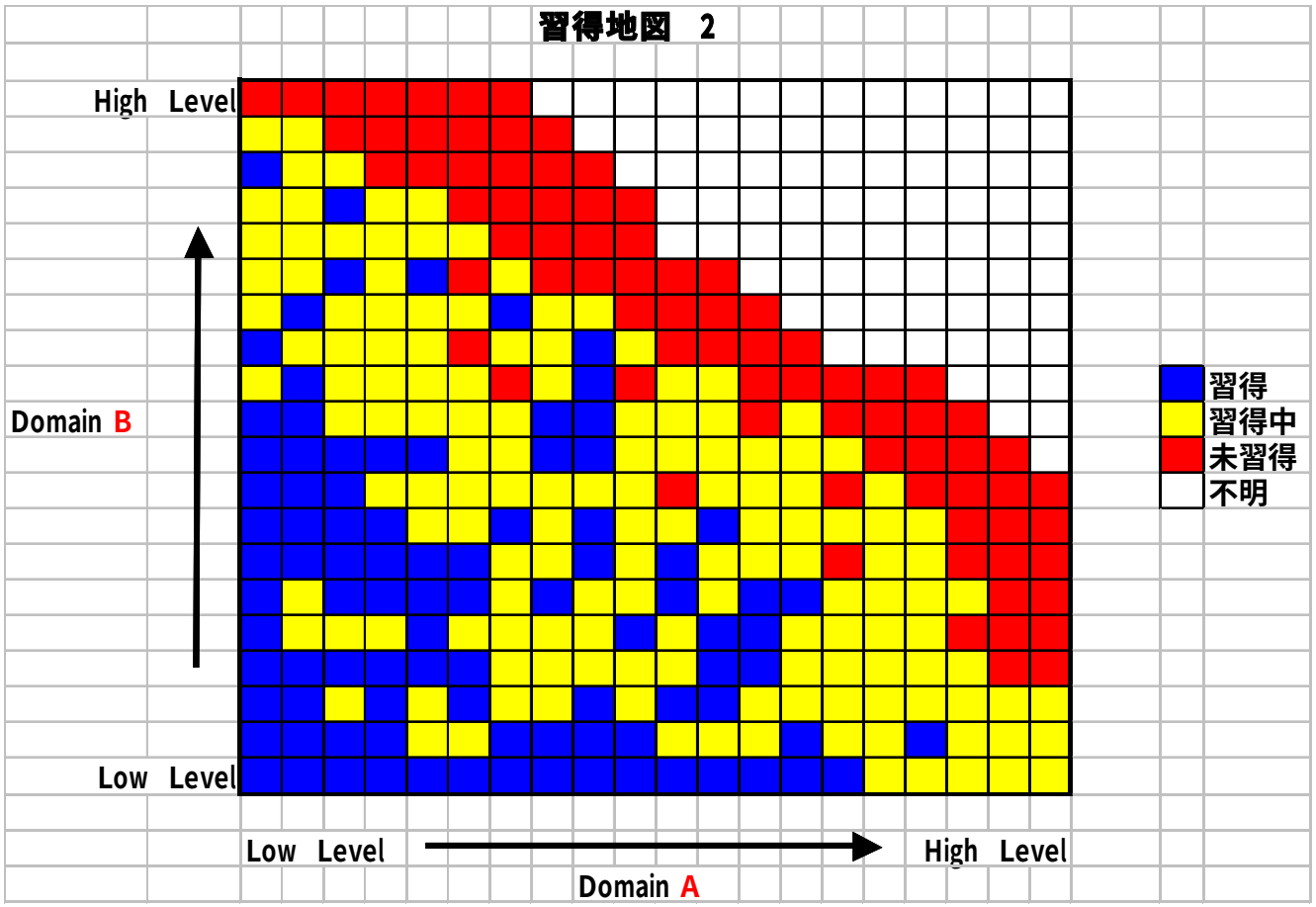
©Hiroshi Ikeda, 2006/9/17

Mastery Map の作成 1



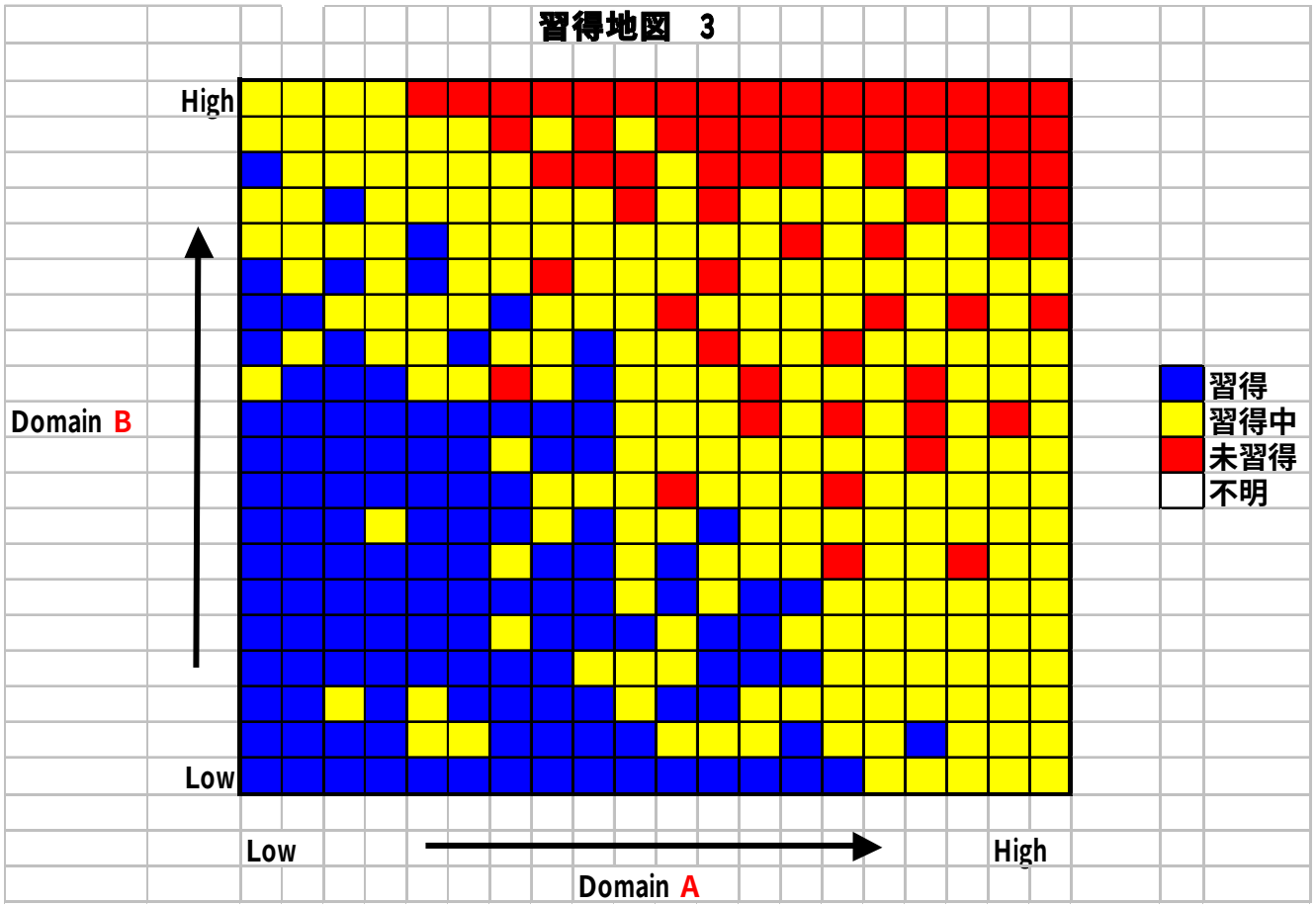
©Hiroshi Ikeda, 2006/9/17

Mastery Map の作成 2



©Hiroshi Ikeda, 2006/9/17

Mastery Map の作成 3



おわりに

- 今まで、テストの役割は選別をはじめとして、**他人と比較**することによって成り立つ数値を出すことに使われていました。
- しかも、それは**不安定**で、**不確か**な数値しか出すことが出来ず、それを少しでも確かなものにしようという努力も十分とはいえませんでした。
- われわれ、とくに学習者が欲しい情報は、日々の**成長、発達、学習**によってどれだけ**変化**が見られたか、それを一つの証拠として数字で示せるようなテスト情報といえます。
- **ＩＣＴ技術**や**テスト理論の**発達で、困難な問題は克服の路が見えてきています。それをもとに産官学の人たちが協力することで、**比較**のためのテストから**変化**を知るためのテストへと今後の**テスト事業**が展開していくことを願っている次第です。