What We Need for Research on Language Testing in Japan -A Psychometrician's View-

Hiroshi Ikeda

**The 21st Language Testing
Research Colloquium**

July 28, 1999

1

Ladies and gentlemen. Today, I have the great honor of speaking to you at the beginning of the 21st Language Testing Research Colloquium here in Tsukuba. I know this is the first colloquium which is held in Japan, and I would like to say, "Congratulations to all the people who have been waiting for today." This is truly the fruit of the long effort by the program organizers, especially the chair persons, Professors Ohtomo, Thrasher, Yuji Nakamura and Yoichi Nakamura, secretary general, in particular, and other working members. Thank you all for your hard work.

I would like to talk today about some modern test theories and the related technologies which have greatly progressed, in the United States in particular. I am neither an English teacher, nor a specialist of linguistics. So, my view may have some limitation from a psychometrician or a statistician who is working in the field of educational measurement. My topics may not be new for some participants who are familiar with them. My talk will be focused primarily on the Japanese audience who are unfamiliar with rapid changes in the testing world, because Japan is still an isolated country in terms of information on testing, the ideas of modern test theories, and research in particular. We must make an effort to make progress in these areas.

.. I sincerely apologize to the non-native attendants, for my later Japanese presentation, because I can't say in exact English what I intend to and I'm afraid of miscommunicating. So, I hope you will look at my abstract or the handouts on which I wrote the English headlines and some keywords which will be presented on the OHP. This may be an experiment of how well we can understand the ideas without oral explanation. Thank you. Now let's start.

Complexity of Language Testing

- ◆ Language ability cannot be measured by a single facet.
- ◆ Language ability cannot be measured by a single method.
- ◆ Language ability is assessed by an integration of multiple facets with multiple methods.
- ◆ Language testing requires advanced theories, research methods, technologies, data base, etc.

2

1. 言語能力を (verbal) **ability** という言葉で表すのが適切かどうか難しいのですが、ここではそうしておきます。しかし、それは一つのはっきりした概念を指すのではなく、たくさんの顔を持っています。ときには、

Competence, Proficiency, Performance, etc

といってよいものも含みます。

2. したがってそれは、ペーパーテストのような、**単一の方法**で測ることは出来ません。

3. 言語能力はたくさんの側面をたくさんの方法で**統合して**、**assess**する必要があります。

4. そのためには**高度な測定評価理論**、研究方法、情報技術(IT)、データベースなどの整備充実が必要です。それが故に、言語テストは**測定論**の立場から見ても興味ある研究対象なのです。

5. そこで今日は以下のことについて日ごろ考えていることの一部をお話してみたいと思います。

The Outline of Topics

1. More use of modern technologies
2. Application of modern test theories
3. Building test item bank
4. Need of validation studies
5. Close network between teachers and researchers

3

ここでお話したい話題の概要

1. 現代は**技術**の時代です。ゲームの世界を見てもお分かりのように、テストも **virtual reality** の時代になりました。
2. テスト**理論**もいまは高度に発達しております。それは技術の陰に隠れていますが、目に見えない背後にある**理論**が大事です。日本は戦後、背後の理論にあまり関心を払うことなく、ペーパーテストを導入しました。いまのコンピュータ化に際しても、それと同じ轍を踏まないか心配です。
3. 同時にそれを利用できる**データ**が必要です。データは**知的財産**です。使っては捨てでは、永久に振りだし双六でゴールには達しません。
4. そのためにもっと**研究**が行われなければなりません。研究はつまるところ**データの比較作業**です。その統計結果は**confidential**にすべきではありません。
5. そのためには、**みんなの協力**が必要です。手分けして、データを集め、それを共有する**ネットワーク**は作れないでしょうか。

1. Use of Modern Technologies

- ◆ Rapid development of computer-based testing (CBT)
 - e.g. GRE, GMAT, TOEFL, ASVAB, ARE, USMLE, NCLEX
- ◆ Change of question-answer types in testing
 - Audio-visual presentation and keyboard answering
- ◆ Change of measurable constructs
 - Constructs not measured by paper-and-pencil tests

4

最近急速に Computerized Tests が盛んになりました。

主な **Computerized Tests** をあげます。

GRE: Graduate Record Examination (Educational Testing Service)

大学院入試用です。

GMAT: Graduate Management Admission Test (Educational Testing Service)

経営大学院、ビジネススクール入試用です。

TOEFL: Test of English as a Foreign Language (Educational Testing Service)

外国人留学生用英語能力試験

CAT-ASVAB: Armed Services Vocational Aptitude Battery (US Department of Defense)

軍隊式職業適性検査 (18-23歳志願者230万人が受験)

ARE: The Architect Registration Examination (National Council of Architectural Review Boards: NCARB)

建築家登録試験

USMLE: United States Medical Licensing Examination (National Boards of Medical Examiners: NBME)

医師免許試験

NCLEX: National Council Licensure Examination (National Council of State Boards of Nursing: NCSBN)

看護婦免許試験

その他知るだけで30～40は優に越えます。

Major Differences between Traditional and Modern Testing Methods

	<u>Traditional</u>	<u>New</u>
Questions	Printed	Multimedia devices
Ans. Form	Mark sheets	Mouse, Keyboard
Present Order	Fixed	Flexible
Resp. Time	No count	Countable
Administration	Group	Individual

5

CBT の利点

- ・個人ごとに受験できる。
- ・いつでも受けられる。
- ・近くで受けられる。
- ・すぐに結果がわかる。
- ・障害者にも不公平さが少ない。
- ・いろいろな種類の問題が出せる。(文字、音、画像、映像)
- ・いろいろな種類の解答が求められる。(文章、数式、図表、音声、絵、運動)
- ・検査時間も測ることができる。

さらにCATでは.

- ・少ない数の問題で済む。
- ・短い時間で済む。

Classification of Test Types

6

いまあるテストをタイプ分けしてみますと

PPT: Paper and Pencil Test	ペーパー・テスト
CBT: Computer-Based Testing	コンピュータ・テスト
CAT: Computerized Adaptive Testing	コンピュータ適応型テスト
MCQ: Multiple-Choice Question	多肢選択式質問
DDQ: Drag and Drop Question	ドラッグ&ドロップ式質問

Authentic Test	真生テスト(本物テスト)
Mulidays-Multisites Testing	(いつでもどこでも受けられる)テスト
On-line Testing	オンライン・テスト
Performance Testing	パフォーマンス・テスト
Remote Testing	遠隔テスト
Simulation Test	シミュレーション・テスト
Test on Demand	オンデマンド・テスト
Virtual Reality Test	仮想現実型テスト

2. Modern Test Theories

- ◆ Item Response Theory (IRT)
 - Dichotomous model (1-, 2-, 3PL)
 - Polytomous model (nominal, graded, partial credit)
 - Multidimensional IRT, Nonparametric IRT

Application → Adaptive Testing (CAT)
Classification Testing (CCT)
- ◆ Generalizability Theory (GT)

Application → Performance Assessment (PA)

7

現代の代表的テスト理論を2つあげると、

1. 項目応答(反応)理論(IRT)

Dichotomous model ; One- Two- Three Parameter Logistic Model

(2値型-正誤解答) (1-, 2-, 3-パラメタ・ロジスティック・モデル)

Polytomous model : Nominal Response, Graded Response, and Partial Credit

(多値型) (名義反応) (段階反応) (部分採点)

その他: 多次元IRT, ノンパラメトリックIRT、などに分かれます。

その応用として、適応型テスト(Computerized Adaptive Testing: CAT)

Computer Classification Testing : Sequential Probability Ratio Test (SPRT)

2. 一般化可能性理論(GT)

論述テストは同じ受験者に対し、評定者によって、課題によって、評点が変わります。それをどう扱うか。それらの評点はどのくらい一般化可能か。パフォーマンス・アセスメントの評点などに関連して重要視されるようになりました。

Classical vs. Modern Test Scoring

- ◆ Conventional Test Scoring
 - Number Right or Formula Scoring
 - Scaled Scoring (Standard or T-scale, etc.)
- ◆ Adaptive Test Scoring (CAT)
 - Maximum Likelihood (or Bayesian) Estimation with the least number of items
- ◆ Computerized Classification Testing (CCT)
 - Sequential Probability Ratio Test (SPRT)

8

1. 現在行われている主な採点法には、次のようなものがあります。

素点法: 出来た正答に重み(配点)を掛けて合計します。

公式法: 誤答に対して正答から(選択肢数-1)分の1の割合で減点します。

偏差値法: (各素点-平均値) / 標準偏差 × 10 + 50 の公式で出します。異なるテストと得点を共通尺度化するためにいまもよく使われています。

等分位法: 同じ点は同じ順位を意味するように素点を変換した得点です。

2. **CAT:** 予め問題に定められた項目情報を使って高度な数学予測式を使い能力を推定します(最尤推定法、ベイズ推定法など)。予め難しさなどが分かっている問題を使いますから、一人一人出される問題が違っても能力レベルを決めることができます。今多くのアメリカのテストで実用化されています。

3. **SPRT(Sequential Probability Ratio Test):** 逐次確率比テスト。CCTで使われる。第2次世界大戦中A.Wald(1947)によって考え出された品質管理の手法を、資格・検定試験に発展させたもの。(Cronbach & Gleser, 1965, Ferguson (1969), Reckcase(1983)) アメリカではすでに医師免許資格試験、建築家登録試験などにこの考え方が使われ実用に達しています。

(帰無仮説0のもとでの尤度をL0, 対立仮説1のもとでの尤度をL1とし、

$\beta / (1 - \alpha) < L1/L0 < (1 - \beta) / \alpha$ の範囲を越えたら合否が決まる。それまでテストを続けます。 α 、 β はそれぞれ第1種、2種の危険率。)

従来の Paper & Pencil Test の採点法

1. たくさんある問題項目から、ある一組のテスト問題を選び、それを全受験者に同じ時間だけ与え、解答してもらいます。
2. そのうちの正解した問題数(量)で得点を決めます。
3. しかし、取り出した問題の組み合わせ方によって得点と同じでも、能力が同じことを意味しません。
4. そこで、問題の組み合わせが違うテスト得点の間を比較するために、平均値からのずれの大きさを標準偏差でわって標準化します(偏差値)。
5. (特徴)あらかじめ問題の難易度が分からなくても、同じ受験者群なら受験者の相対的位置が決められます。

新しい適応型テストの採点法

1. それぞれの測定したい能力領域で、やさしい問題からむずかしい問題まで**難度**の分かった問題をたくさん用意します。
2. 能力水準の分からない受験者に対し中度の難しさの問題を与えます。できれば次にもっとむずかしい問題を、できなければ次はもっとやさしい問題を与えます。
3. 以上の過程を難度が安定するまで続けます。言い換えれば解答の正誤が半々の確率になるように問題を選んで与えるわけです。(コンピュータがそれを行うのをCATといいます。実際はそれがなければ出来ないでしょう。)
4. 収束した問題の難易度水準をみれば、その人の能力水準が分かります。
5. (特徴) **少ない問題数**(通常の半分から3分の1で済むといわれます)で能力水準を決めることができます。

3. Building Test Item Bank

- ◆ The first test item is not necessarily a good item.
- ◆ Good word usage could be reinforced by repeat testing.
- ◆ Item parameters in test question have priceless property of value-added information.
- ◆ Item parameters must be kept up-to-date with continuous checking.

11

項目パラメータ作りの必要性

1. どの製品も同じですが、テストも使い始めの試作問題 (**virgin item**) が良い問題とは限りません。問題は精選して出すべきです。ふつうの製品では工場で作った試作品をいきなり、市場に出すことはありません。
2. よい表現(とくに英語の初期)は繰り返し口に出して、定着させるべきです。昔は**recitation**といって、基本文を何回も反復して覚えるまで言わしたそうです。問題の事前予告をして、練習に励むことがあっていいとすら考えます。
3. そのためには、各問題項目ごとにその問題の特性を示す、付加情報の蓄積が必要で、その情報は何事にも代えがたい一大資産と考えます。
4. そしてその情報は常に現在どうなっているか **update** にしておく必要があります。(On-line check)。30年前のデータが分かっても意味がありません。)
5. 問題はそうしたことをどうやって可能にするかです。(野球選手の人気投票率は何百万もあって分かるのに)。統計的データの値は数が大きくなれば安定してきて、少々のことでは動きません。それだけに多くの数を集めることが重要で、蓄積された結果は大きな財産となります。

4. Need of Validation Studies

- ◆ Setting reliability standard
 - Check of MCQ is easy by simulation study.
 - Contamination of signal and noise in test scores
- ◆ Accumulation of validation evidences
 - Correlational vs. experimental approach
 - Distinction of method and trait factors
 - Distinction of rater and ratee differences

12

テストでは**信頼性**と**妥当性**が重視されます。

1. 信頼性の意味はシミュレーション実験してみるとよく分かります。
2. 実際のテストスコアは情報(信号)と雑音とが混ざったもので、見ただけではどれだけの割合が信号で、どれだけの割合が雑音か分からないのが実情です(Table 1、Figure 1)。信頼性はその情報を与えてくれます。
3. 妥当性研究には**相関的研究法**と**実験的研究法**とがあります。
4. 妥当性データをみるには**方法要因**と**特性要因**とを区別すべきです。
5. 同様に客観テストでない評価データは、**受験者要因**と**評定者要因**の違いを区別すべきです。

Meaningful and Nonsense Tests - A Simulation Study -

- ◆ Table 1a, 1b

Real response and random response data

- ◆ Figure 1a, 1b

Distribution of test scores

13

表1aはコンピュータに対する予備知識を449名の学生に5段階評定法(スライド18 Case A 参照)で訊ねた**実測データ**の一部です。

表1bは各解答選択肢の選択率が同じになるように**乱数**を発生させて作った擬似データです。これは雑音だけの分布といえます。

各行の横の合計が各質問に答えた点の合計値(全部で10問)です。見た目では実測データと乱数データの違いは分かりません。

図1aの2本のグラフはその**得点分布**を重ね合わせて描いたものです。両者の平均値は各選択肢選択率が同じですから同じになります。

どこが違うでしょうか。(標準偏差が実測値の方が大きくなります。なぜでしょうか? 実測値は信号と雑音の両方を含んでいますが、乱数は雑音しか含んでいないため。)

なお、図1bはコンピュータへの学習意欲を訊ねる10個の質問について、同じように試したものです(スライド18 Case B 参照)。

Table 1a. Real Response Data

ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Real
1	2	3	4	4	3	4	3	4	4	5	36
2	3	3	4	4	1	3	1	1	4	4	28
3	1	1	1	1	1	1	1	1	3	3	14
4	1	1	1	1	1	3	1	1	3	2	15
5	1	1	1	1	1	1	1	1	3	3	14
6	1	1	1	1	1	2	1	1	2	1	12
7	1	1	1	4	1	3	1	1	3	4	20
8	1	1	1	1	1	3	1	1	3	3	16
9	4	1	4	1	1	1	1	1	1	1	16
10	1	1	1	5	1	5	1	1	3	5	24
11	4	4	4	5	5	5	4	5	4	4	44
12	2	3	2	3	1	3	2	2	3	3	24
13	1	1	1	2	2	3	1	1	3	1	16
14	4	2	1	3	1	3	1	1	3	4	23
15	3	3	3	3	3	1	1	3	3	4	27
16	4	4	4	4	4	4	4	4	4	4	40
17	3	1	2	4	1	4	1	1	3	4	24
18	1	1	1	1	1	1	1	1	1	1	10
19	3	3	4	4	3	2	3	1	1	2	26
20	3	2	4	4	4	3	4	5	2	3	34
21	2	2	3	4	2	2	3	4	3	3	28
22	3	1	3	3	1	1	1	1	3	1	18
23	2	1	1	3	1	1	1	2	2	2	16
24	1	1	1	4	2	2	1	1	1	2	16
25	2	3	3	3	3	3	1	1	3	3	25
26	1	1	1	1	1	1	1	1	1	1	10
27	3	1	3	4	5	4	1	5	3	4	33
28	2	1	3	4	1	3	1	5	3	1	24
29	1	1	1	3	1	3	4	4	4	4	26
30	1	1	1	5	1	4	2	2	3	4	24

Table 1b. Random Response Data

ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Rnd
1	2	3	1	2	1	2	5	1	3	1	21
2	1	1	3	3	1	3	1	1	2	3	19
3	2	3	1	1	3	3	4	2	3	5	27
4	1	1	1	4	3	1	2	5	5	4	27
5	3	1	1	1	2	3	1	5	1	3	21
6	1	2	3	4	1	3	4	5	3	3	29
7	5	4	2	4	1	1	1	1	4	4	27
8	2	1	2	3	1	5	4	1	1	3	23
9	4	1	1	4	3	3	5	5	3	3	32
10	1	4	3	3	1	3	3	1	3	3	25
11	1	1	1	1	1	2	1	1	3	3	15
12	1	1	1	2	3	2	3	1	3	3	20
13	3	4	3	1	4	3	1	1	3	2	25
14	1	3	1	1	1	1	1	1	2	3	15
15	1	1	4	3	1	3	1	1	3	1	19
16	3	1	1	4	1	3	1	1	3	4	22
17	1	1	2	2	1	1	2	1	3	5	19
18	4	1	1	1	5	3	3	1	1	2	22
19	3	2	2	3	1	3	1	1	3	3	22
20	3	2	1	3	3	4	1	1	4	4	26
21	3	3	1	4	1	1	4	1	3	2	23
22	1	3	3	3	1	3	1	1	3	4	23
23	1	1	3	3	2	2	1	5	5	1	24
24	1	1	1	2	3	1	1	1	1	3	15
25	2	3	4	3	2	3	2	1	3	5	28
26	5	4	1	4	1	2	2	1	4	2	26
27	1	1	2	3	3	2	1	5	3	2	23
28	3	1	1	1	1	3	1	4	3	1	19
29	1	1	1	3	1	2	1	5	1	3	19
30	2	3	3	3	1	5	1	1	3	4	26

Fig.1a Score distributions (CaseA)

Fig.1b Score distributions (CaseB)

Full Information of Variance-Covariance Item Matrix

	Item 1	Item 2	Item 3	Item 4	Sum
Item 1	C	A	A	A	B
Item 2	A	C	A	A	B
Item 3	A	A	C	A	B
Item 4	A	A	A	C	B
Sum	B	B	B	B	D

14

テストの分析には項目間どうし、項目と合計点どうしを含めて、**分散と共分散**の一覧表を作ります。

そこには図のように、A, B, C, D, 4種の情報が含まれます。

項目間分散-共分散行列の情報要素

A = Item Covariances (項目間共分散)

B = Item-Total Covariances (項目-全体共分散、周辺共分散、点双列相関)

C = Item Variances (項目分散)

D = Total Variance (全分散)

Required Information for Obtaining Reliability Estimates

	Covariance		Variance	
	Item	Marginal	Item	Total
Matrix elements	A	B	C	D
Kuder-Richardson 21	No	No	No	Yes & Mean
Cronbach α	No	No	Yes	Yes
Congeneric	No	Yes	Yes	Yes
Principal Component	Yes	Yes	Yes	Yes

15

これら4つの情報を使って、信頼性を計算することができます。

信頼性計算に必要な情報

1. 全平均値と全分散のみ (n = 項目数)

$$\text{Kuder-Richardsonの公式21} = n/(n-1) [1 - \mu_X (1 - \mu_X/n) / \sigma_X^2]$$

2. 全分散と項目分散

$$\text{Kuder-Richardsonの公式20} = n/(n-1) [1 - \sum p_i q_i / \sigma_X^2]$$

$$\text{Cronbachの } \alpha = n/(n-1) [1 - \sum \sigma_i^2 / \sigma_X^2]$$

3. 全分散と項目分散と周辺分散

$$\text{(強) 同族信頼性} = (1 - \sum \sigma_i^2 / \sigma_X^2) / (1 - \sum (\sigma_{iX})^2 / \sigma_X^2)$$

$$\text{池田の公式}(n=2 \text{ の場合}) = r_{12} / r_{1X} r_{2X} \quad (X = X_1 + X_2)$$

4. 全分散と項目間分散共分散 (第1主成分、 λ_1 = 第1固有値)

$$\text{(弱) 同族信頼性} = (1 - \sum \sigma_i^2 / \sigma_X^2) / (1 - \sum a_i^2 / \lambda_1)$$

たくさんの情報を利用するほど、正確な信頼性を推定することができます。

Concept of Reliability

- An Illustration by Simulation -

- ◆ Test score does not represent a true ability.
 - Test score consists of information(signal) and noise.
- ◆ Reliability Coefficient
 - How much meaningful information is contained in test scores ?
- ◆ Standard Error of Measurement (SEM)
 - A confidence band of the given score

16

テスト得点は2つの要素に分けて考えられます。

1. テスト得点 = 信号 + 雑音 (以下いろいろな呼び名でいわれます)
 - = 真の得点 + 誤差
 - = 関係情報 + 無関係情報
 - = 説明情報 + 残差

2. 誤差の中には受験者要因、採点者要因、それに課題差要因が加わります。そのうち採点者要因を固定化したのが客観テストですが、課題差要因は残ります。

3. 課題差要因はテスト得点の中では雑音として働き、数ある課題の中の個人的傾向(平均とトレンド)が信号として働きます。それは経済指標を類推してみるとよく分かります。個人の能力推定が難しいのは、この課題差要因が大きいことによります。課題要因を決める項目パラメータ値が分かっているならば、個人の能力は偏差値による相対評価をしなくても決められます。

4. 信頼性係数は標本誤差における信頼係数(confidence coefficient)とは異なります。類似語として、信頼性指数、信号雑音比、などがあります。またその出し方にも、手に入るデータの多寡によっていくつかの方法があります。

5. 測定の標準誤差とは、測定誤差の標準偏差で、標本誤差の標準偏差とは異なります。標本誤差はサンプル数(受験者数)を多くすることで小さく出来ますが、テスト得点の測定誤差は問題数を多くすることで小さく出来ます。しかし、それは結構難しいことです。

Four Levels of Reliability Standard -A Rule of Thumb-

Standard	Reliability	SEM<aSD
◆ The highest (Ideal)	$R > 0.9$	$a < 0.3$
◆ Permissible (Desirable)	$R > 0.75$	$a < 0.5$
◆ Suspended (Revise)	$R > 0.5$	$a < 0.7$
◆ Prohibitive (Harmful)	$R \leq 0.5$	$a \geq 0.7$

17

信頼性のあるテストとそうでないテストの品質の違いを4つの水準に分けて考えたいと思います。上の基準は別の言葉でいえば、次のようにいうことが出来ます。

1. 理想水準 誤差のSDが得点のSDの3割(真値の1/3)より小さくなるように。
2. 許容水準 誤差のSDが得点のSDの半分よりは小さくなるように。
3. 留保水準 誤差のSDが得点のSDの7割もある。(このままではダメ、要改訂)
4. 破棄水準 誤差のSDが真値のSDよりも大きくなってしまう。(要破棄)
(SEM: 誤差の標準偏差、SD: 標準偏差)

Figure 2a. Mean Scores of Real, Random, and Revised Tests

Figure 2b. Reliability of Real, Random, and Revised Tests

Figure 2c. SDs and SEMs of Real, Random and Revised Tests

Figure 3a. Discovery of Inappropriate Items (No items in random zone)

Figure 3b. Discovery of Inappropriate Items (Two items in random zone)

Comparison of Real and Random Response Tests

- ◆ Figure 2a Mean Scores
- ◆ 2b Reliabilities
- ◆ 2c SDs and SEMs
- ◆ Figure 3a, 3b Discovery of inappropriate
 items

18

項目分析 正解の正当性を探る(不良項目を見つけ、除去する方法)。

(例) コンピュータに対する意識調査(大学新入生449名)

Case A: Computer に関する予備知識。 5段階自己評定尺度10問構成

Scoreのつけ方: 5 = 人に内容を教えたり説明できる、4 = ちょっとだけなら内容を知っている、3 = 言葉だけなら知っている、2 = 以前に言葉を聞いたことがあるような気がする、1 = きょうが初耳である。○のつけた数値を合計。

Case B: Computer への学習意欲。 5段階自己評定尺度10問構成

Scoreのつけ方: 1 = 全然ない、2 = ややない、3 = ふつう、4 = ややある、5 = 大変ある。得点の高いものほど意欲がある。○のつけた数値を合計。

Case B': Case B にある2つの不良質問を除いて、採点し直したもの。

それによって信頼性が許容水準を満たすことになりました。

Figure 2a, 2b, 2c: 3つのケースについて 各尺度の平均値、信頼性係数、標準偏差と測定の標準誤差を比較したものです。左側が実測、右側が乱数データです。

Figure 3a, 3b: 上記Case A, B, について、横軸に項目平均値、縦軸に合計点と項目との点列相関(識別力にあたる)をとり、プロットしたものです。実測と乱数データでは落ちる位置が違いますが、Case Bで乱数項目と見分けのつかない実測項目が2項目あります。それを除いたのがCase B'です。

Correlational vs. Experimental Studies

- ❖ *Experimental* is primarily focused on group, while *Correlational* is on individuals.
- ❖ Avoid information loss, when either approach is possible.

19

1. L. J. Cronbachが心理学会会長就任講演でのべたことがきっかけです。
The two discipline of scientific psychology (1957).
American Psychologist, 12, 671-684.
2. その例として上げたのが、**適性処遇交互作用** (Aptitude-Treatment Interaction: **ATI**) の考え方であり、
3. テスト理論で発展させたのが、誤差の成分を特定するための、**一般化可能性理論** (Generalizability Theory: **GT**) による research design です。
4. どちらも交互作用 (Interaction) という概念が中心となり、それは実験計画法による**分散分析**の手法が大きな働きをします。
5. それらについては時間がありませんので、触れられませんが、私が言語研究の発表論文をみて、気になっていることを一つだけ上げさせていただくと、相関分析を行った方がよいと思うところを分散分析の手法で行っているケースが多いことです。
6. そのほかの統計的疑問については、最近出た本が役に立つと思います。
繁樹算男、柳井晴夫、森敏昭編著(1999)、心理学セミナーテキストライブラリー-3「Q&Aで知る統計データ解析、DOs and DON'Ts」、サイエンス社。
佐伯胖・松原望編著(2000)、「実践としての統計学」、東京大学出版会。

Quantitative Data Should be Treated as Quantitative

◆ Figure 4a

Analysis of Variance Approach

◆ Figure 4b

Regression Analysis Approach

20

1. 分散分析は本来グループの間の差に関心があり、個人の一つ一つの情報には関心がありません。この場合グループ分けは属性によってなされておらず、誤差を伴った確率変数です。したがって当然分類エラーも起こります。

Figure 4a の例は先のコンピュータの知識と学習意欲の関係を見るために Case A のスコアをもとに3つのグループに分け、それを学習意欲について分散分析したものです。図はグループごとの Case B の得点 (S_y) のばらつきを表わしています。グループは、Case A の得点 (S_x) によって

Low Group は $S_x = 10 \sim 18$ 、 $N=144$ 名

Middle Group は $S_x = 19 \sim 26$ 、 $N=168$ 名

High Group は $S_x = 27 \sim 48$ 、 $N=137$ 名 に分けたものです。

2. 回帰分析または相関分析の場合は、回帰直線の傾きが大きい(上昇傾向)、そうでないかに関心があります。またそのために個々のデータの情報を利用しています。そこでは単に相関があるかないかだけでなく、その大きさにも関心があるのが普通です(相関が0でないということが分かっても得る情報は少ない)。

Figure 4b の例は横軸に Case A (S_x) の得点を縦軸に Case B (S_y) の得点を取ったもので、回帰直線の傾きは0.192、相関係数は0.260 で、いずれも有意です。($\rho=0$ の仮説検定には、 $t = r\sqrt{(N-2)/\sqrt{(1-r^2)}}$ が使えます。)

Figure 4a Analysis of Variance Approach

Figure 4b Regression Analysis Approach

Distinction between Method and Trait Factors in Validity Data

- ◆ Knowledge and performance are different constructs.
- ◆ Knowledge and performance are measured by different methods (paper test and observation)
- ◆ Correlation between measured data is contaminated with both factors.

21

1. たとえば、**知識**と**技量**は違った構成概念です。
2. それぞれは違った方法で測られます。知識はペーパーテストで測られるかもしれませんが、技量は主に作業過程の観察か、成果作品 (output) の評価が中心です。
3. したがって、測定されたデータ (得点) 間の**相関**は、測定の方法要因と測る対象の特性要因とが混ざり合った形で得られます。従来のようなどちらも同じペーパーテストの場合、方法の違いの差は分からないか無視されがちでした。
4. そこでその両者を**区別**できる方法論が必要です。
5. ここでは、Campbell & Fiske (1959) の有名な
Convergent and discriminant validation by the multitrait-multimethod matrix.
Psychological Bulletin, 56, 81-105.
の考え方と、そのための分析法として最近よく使われる**共分散構造分析**による**構造方程式モデル** (Structural equation modeling) について触れたいと思います。

Measurement by MTMM Design

Bold: Convergent	Method 1	Method 2		
<i>Italic:</i> Discriminant	TraitA	TraitB	TraitA	TraitB
Method 1	Trait A	<u>0.870</u>		
	Trait B	<u>0.260</u>	0.713	
Method 2	Trait A	0.974	0.281	<u>0.847</u>
	Trait B	0.266	0.965	<u>0.288</u> 0.663

Trait A: Knowledge B: Learning Motivation
Method 1: 5-Point Scale 2: 3-Point Scale

22

この例で測ろうとしている尺度AとBの特性はそれぞれ

Trait A: Knowledge on computer (Self-report)

Trait B: Learning motivation for computer (Self-report)

また方法は5段階と3段階の評定尺度です。

Method 1: 5-point rating scale

Method 2: 3-point rating scale (ただし、ここでは両方法は独立でない)

そのとき相関の数値の違いを系統的に比較してみることによって、その違いが何の原因からくるものか識別できます。

Underline: Reliability coefficient (同一特性-同一方法による繰り返し測定)

Bold: Convergent validity (同一特性-異方法による測定)

Italic: Discriminant validity (異特性-同一方法による測定)

それを分析するのが、**多特性多方法分析(MTMM)**計画の考え方です。

それを図にモデル化したのが、上のパス図です。

この場合、4種の観測値は2つの特性要因と2つの方法要因および4つの誤差のいずれかの要素からなると考えられますから、その組み合わせは、次のように表わせます。(ただ、この例では、 e_1 と e_3 、 e_2 と e_4 との間は独立ではない)

$$X_{A1} = f_A + f_1 + e_1$$

$$X_{B1} = f_B + f_1 + e_2$$

$$X_{A2} = f_A + f_2 + e_3$$

$$X_{B2} = f_B + f_2 + e_4$$

主なソフトウェアとその発行所は、

AMOS/AMOSDraw: J. Arbuckle (Temple U) SPSS Japan.

CALIS: SAS Institute

EQS5: P. M. Bentler (UCLA) Multivariate Software

LISREL8/SIMPLIS, LISREL8/PRELIS2 K. Jöreskog & D. Sörbom
Scientific Software International, Inc.

LISCOMP: B. O. Muthén, Scientific Software International, Inc

Mx:Statistical Modeling: M. C. Neale 自分のところ

SEPATH: J. H. Steiger StatSoft

Measurement by Human Rating

- ◆ Subjects have individual differences.
- ◆ Raters have individual differences.
- ◆ Task contents also have different attributes.
- ◆ Errors of measurement are confounded with the interaction of different source of factors.
- ◆ Generalizability theory gives a research design for the systematic investigation of errors.

24

1. 人間の行う評定の分析は、**Performance Assessment** の需要とともに、主観評定を取り入れなければならない状況が増えてきました。
2. そこでは、少なくとも被験者、評定者、課題の3要因が得点差に入り込みます。ということは、3要因の3つの主効果のほかに4つの交互作用が誤差要因の中に加わります。
3. その誤差成分を組織的比較実験により評価推定し(**G研究**)、ある特定のテスト条件で評価される得点がどこまで一般化されるかたちで解釈できるか(**D研究**)を考えようとするのが、**一般化可能性理論(GT)**です。
4. 日本語で紹介された本としては私の「現代テスト理論」(朝倉書店)がありますが、ほかにわかりやすい原書をあげると、
Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer.*
Sage Publications.
Brennan, R. L. (1992). *Elements of generalizability theory.* The American College Testing Program.
でしょう。ソフトとしては **GENOVA** があります。

5. Building Network between Teachers and Researchers

We need:

- ◆ To exchange new test ideas through Special Interest Groups (SIG).
- ◆ More communication among teachers, researchers, IT engineers, test publishers, and citizens.
- ◆ To establishing data base center on language testing.
- ◆ To disclose educational reports and information.
- ◆ More academic activities like LTRC

25

1. このようにテストの問題は従来の枠組みを越えて、さまざまな方法が考えられようとしています。そのために関心のある多くの方がアイデアを持ちよって、思いつきでなく、**心理測定理論**の基礎に立って理論化する必要があります。MCQについては1世紀の研究でかなりのことが分かっていますが、**Constructive Type**や**Performance Assessment**についてはこれからです。

2. そのためには多分野にまたがる多くの専門家の協力が必要です。そして専門家(生産者)だけでなく、一般市民(利用者、消費者)の参加も忘れてはなりません。テストの本当のお客さんは消費者である利用者なのですから。

3. こうした研究が実を結ぶためには、同時に研究の基礎となる**データベース**の蓄積と情報の共有化が必要です。テスト問題を作っては捨てるの歴史が、わが国の発達を阻害してきました。残念ながらわが国の教育情報の多くは外国の研究からもたらされたものです。(わが国の教育論にはデータがない。)

(IEAの research report、TOEFLの research report)

4. しかも、どうも都合のよいデータは公開されますが、教育とか医療、金融とか人にかかわる情報や政府に都合の悪いデータは不完全だからとか、誤解されとかで、○秘扱いにされ、ほとんど一般には知らされてきませんでした。教育に関する議論はほとんど**データなし**の議論です。先の妥当性研究が日本で盛んでないのは、それを妨げる条件が多いからかも知れません。

5. 最後にこの**LTRC**のようなテストの研究学会もわが国には必要です。

Final Words

- ◆ Language competency cannot be improved without testing. It guides an achievement goal.
- ◆ Test construction is no longer an art but a technology. It is beyond a personal skill.
- ◆ So we must have more objective information on test data and the test evidence as public knowledge.
- ◆ The reform of these will surely improve our current stage of language education. Testing activity is a science.

26

結論

1. テスト開発は**総合科学**です。それは**教育プログラムの一貫**として組み込まなくてはなりません。教育プログラムは**一大イベント**で、しかも**長期のシナリオ**が必要です。(私の夢は電子教科書をつくること)

2. 教育プログラムには**内容提示**だけでなく、はっきりした**達成目標**と、そこに到達させるための**具体的手段**(興味を持続させるための方法、下位目標を明確にした評価方法、学習を支援するための道具立て)が緻密に用意されていなければなりません。

3. いまはアメリカではそれを達成するために**高度の情報技術(IT)**と**テスト理論**を駆使した電子教科書とか、電子図書館とか、電子テスト作りに、多くの心理測定技術者、認知心理学者、統計学者、情報工学者、コンピュータプログラマーが競って、この21世紀の**魅力ある教育テーマ**に取り組んでいます。

4. われわれに必要なのは**目の前の教育問題**だけでなく、また**教育についての従来の観念やイメージにとらわれる**のではなく、**効果が見えてくる20年後30年後の未来を見据えた教育改革**ではないかと思っています(本当に教育は100年の計)。テストの問題もその中にしっかりと位置付けて、これからの子孫に悔いを残さないようにしていかなければならないと思っています。

おわり