

第7回「日本テスト学会賞」記念講演

英語教育とテスト

第二言語習得における規準設定をめぐって

大友賢二(筑波大学名誉教授)

第7回

研究協力者: 法月 健(静岡産業大学教授)

成蹊大学 2013年12月7日

講演のあらまし

1. 英語教育と言語テストとの関わり

2. 第二言語習得研究の動向

- | | |
|------------------------------|------------------------------|
| 2. 1. Audio-Lingual Method | 2. 2. Input Hypothesis |
| 2. 3. Interaction Hypothesis | 2. 4. Corrective Feedback |
| 2. 5. Focus on Form | 2. 6. Task-based Instruction |

3. 規準設定の意味と方法

- 3. 1. 規準設定の意味
- 3. 2. 規準設定についての評価
- 3. 3. Bookmark Method とは何か
- 3. 4. PNO/TIN間の数値差を利用した分割点設定法
- 3. 5. Rasch ModelとLRTを併用した分割点設定法

4. 結び

1. 英語教育と言語テストとの関わり

- 1956 東北学院大学英文学科卒
- 1956－1960: 東小野田中・円田中学校教諭
- 1960－1962: 宮城学院高等学校教諭
- 1962－1974: ELEC主事・研修第一部長
- 1965－1966: Georgetown Univ. 留学
- 1974－1983: 神奈川大学助教授・教授
- 1979－1980: UCLA客員研究員
- 1983－1996: 筑波大学教授
- 1989－1990: Teachers College, Columbia Univ./
Simul M.A. Program in TESOL 講師
- 1996－2006: 常磐大学教授・学部長
- 2006 常磐大学退職

- 1965ー1966: Georgetown University : Robert Lado に師事
- 1976ー 現在: JACETテスト研究開発委員長・会員
- 1979ー1980: UCLA 客員研究員:E. Hatch, J. Popham に師事
- 1987ー1999: 国際交流基金日本語能力試験委員会委員
- 1991ー1995: 筑波大学外国語検定制度開発ワーキング・グループ委員長・検定運営委員会委員長
- 1992ー 現在: 国際言語テスト学会 (ILTA) 会員
- 1994ー2004: *Language Testing* Editorial Advisory Board
- 1996ー1999: 第21回言語テスト国際会議 (LTRC99) 実行委員会委員長
- 1996ー2004: 日本言語テスト学会 (JLTA) 会長
- 2003ー 現在: 日本テスト学会 (JART) 理事・会員
- 2004ー 現在: 日本言語テスト学会 (JLTA) 名誉会長
- 2006ー 現在: NCME(National Council on Measurement in Education) 会員

Washington, DC, USA,
1965



San Francisco, USA,
1980



Tsukuba, Japan,
1999



Bangkok, Thailand,
1970



Lancaster, UK,
1985



Kyoto, Japan,
2006



Los Angeles, USA,
1979



Lancaster, UK,
1985



Tokyo, Japan,
2013



The 21st Language Testing Research Colloquium , Tsukuba , 1999 (13か国, 198名参加)



The 21st Language Testing Research Colloquium
Tsukuba, Japan: July 28 - 31, 1999

2. 第二言語習得研究の動向

2.1. Audio-Lingual Method

1940~1950 年代に、アメリカ構造主義言語学と行動主義心理学とを理論的背景として提唱され、実践された教授法。

Charles C. Fries, Robert Lado (University of Michigan) が中心的存在。学習は stimulus, response, reinforcementによって成立すると考えられた。わが国の Oral Approachは、1956年創設された ELECによって普及された。

The Five Steps of Language Learning

1. recognition 2. imitation
3. repetition 4. variation
5. Selection

Teaching Procedure

- A. Review: choral reading, pattern practice, written test
- B. Presentation of New Material: defining sentences, mimicry-memorization
- C. Reading & Check of Understanding
- D. Consolidation

* * * (山家, 1972: 76, 281)

Charles C. Fries: University of Michigan

The word “oral” in the name of “oral approach” expresses what we want the pupil to be able to do.

It (approach) has been chosen in order to stress that we are concerned with a path or a goal---a path or a road that includes everything necessary to reach that goal. * * *
(Fries, 1958:14-15)

Peter H. Fries: Central Michigan University

My father called the approach to language teaching that he has developed “The Oral Approach”. This approach was often confused with a number of methods of teaching English such as the “audio-lingual method”, the “direct method”, and the “oral method”, etc. He constantly argued that his approach was very different in nature, though indeed, like those other methods, his oral approach used a great deal of oral practice. * * * (ELEC, 2013: 73)

わたくしの子供(3才)が話した
はじめての英語？

Don't touch.
That's mine.

アメリカ人(4才)と「おもちゃ」の取り合いになった時

2.2. Input Hypothesis

言語習得の必要十分条件は理解可能なインプットである。現在の段階を i とすると、それよりも一段階高いレベルの文法構造等を含んだインプット ($i + 1$) を学習者が理解することによって、言語習得は無意識的に進められる。
(input hypothesis)

さらに、意識的な「学習」は、自分の発話の正しさをチェックするモニターの役割しかない。
(monitor hypothesis) * * * Krashen (1985)

2.3. Interaction Hypothesis

言語は、インプットだけで習得できるのでしょうか？インプットだけで良いという仮説には、疑問が投げかけられている。

Interaction hypothesis によれば、学習者は、最初は理解困難と思われるインプットでも、**対話者と相互交流する中で**、理解可能なインプットへと変え、言語習得に到達するであろうということである。

*** Long (1996)

2.4. Corrective Feedback

学習者のおこす言語的誤りに対して、聞き手が誤りを修正する意図を持って与えるフィードバック。これには、(1)「明示的否定フィードバック」と(2)「暗示的否定フィードバック」がある。

何が間違っているかを学習者に明確に示す方法が(1)であり、誤りを明確に指摘し、正しい文法規則を説明して指導すること。間接的に誤りに注意を向けさせる方法が(2)であり、学習者の意識を言語項目に向けさせ、誤りを自己修正するように促すこと。

2.5. Focus on Form

Focus on form という指導理論において、もつとも重視されているのは、学習または、指導の中心が、「意味理解、意志の伝達などの第二言語によるコミュニケーション活動」に置かれるということである。

内容のある事柄に関して、第二言語学習者が目標言語を使用して、意味のある活動を行い、**その過程において文法習得を促すことが**、focus on form のねらいである。*** (村野井, 2006: 88), Long (1991)

< The Natural Order Hypothesis >

1980年代になって、第二言語習得においても、自然な習得順序があると結論づけた。

- (1) 進行形(-ing)・複数形(-s)・連結辞(be)
- (2) 助動詞(be)・冠詞(a, an, the)
- (3) 動詞不規則・過去形
- (4) 動詞規則過去形(-ed)・三単現(-s)・所有格(-'s)

* * * (鈴木・白畑, 2013: 146)

* * * Dulay, Burt & Krashen (1982)

< Focus on form の効果 >

- (1) Form-meaning-function の結びつきの理解を促すこの指導は、第二言語習得を促す。
- (2) 暗示的文法指導と明示的文法指導を融合したこの指導は、一定の条件下で第二言語習得を促す。
- (3) 意味交渉 (negotiation of meaning) を引き起こすこの指導は、一定の条件下で、特に学習者の心理言語的レディネスに合致している場合、第二言語習得を促す。
- (4) 学習者が中間言語と目標言語の「ギャップ」に気づくこと (noticing) を促す。*** (村野井, 2006: 109), Doughty & Williams (1998), Williams (2005)

2.6. Task-based language teaching: TBLT

- (1) the primary focus is on meaning
- (2) there is a need to communicate with an interlocutor
- (3) learners need to rely on their own linguistic and nonlinguistic resources to carry out the activity
- (4) there is a goal to the activity beyond just using the language itself (such as coming to a consensus or making a decision). * * * R. Ellis (2009)

< Task-based language assessment: TBLA >

Four main features characterize TBLA:

- (1) a formative assessment**
- (2) a performance-referenced assessment**
- (3) a direct assessment**
- (4) an authentic assessment**

*** * * (Shehadeh, 2012: 157)**

<言語能力の構成と測定>

- (1) Robert Lado (1961): Discrete point approach
- (2) John B. Carroll (1961): Integrative approach
- (3) David P. Harris (1969): Rate and general fluency
- (4) John W. Oller (1979): Unitary competence hypothesis
- (5) Bachman, J.F. & Palmer, A.S. (1996): Language knowledge & strategic competence
- (6) Norris, J.M. (ED.)(2002) : Task-based language assessment
- (7) Chalhoub-Deville, M. & Deville, G.(2006): ability-in-language user-in-context

3. 規準設定の意味と方法

3. 1. 規準設定の意味

Standard setting: Standard setting can be defined as a process by which a standard or cut score is established. * *
* (Cizek, 2006: 226)

基準と規準: 橋本(1983:28)では、**criterion** には「規準」を、**standard**には「基準」と述べているが、その後、皆見(2008)など、さまざまな議論があつたが、ここでは、池田(監訳)(2008:12)に準じて、**criterion**を「基準」、**standard**を「規準」とする。

3. 2. 規準設定についての評価

規準設定の方法:

- (1) Methods that involve review of test items and scoring rubrics
 - (2) Methods that involve review of candidates
 - (3) Methods that involve looking at candidate work
 - (4) Methods that involve panelist review of score profiles
- * * * (Hambleton & Pitoniak , 2006: 440)

主な規準設定法:

- (1) Angoff Method, Ebel Method, Nedelsky Method, Jaeger Method, Bookmark and Other Item Mapping Methods
- (2) Borderline-group Method, Contrasting-group Method
- (3) Item-by-item approaches, Holistic approaches
- (4) Judgmental policy capturing method, Dominant profile method

標準設定法に対する否定的見方

To summarize---there is no 'gold standard', there is no 'true' cut-off score, there is no best standard setting method, there is no perfect training, there is no flawless implementation of any standard setting method on any occasion and there is never sufficiently strong validity evidence. * * * (Kaftandjieva , 2004: 31)

Standard setting has been called the 'Achilles heel' of educational testing (Hambleton & Plake ,1998) largely because there is no clear consensus on the best choice among numerous methods and because the results of applying any method cannot easily be validated (Kane ,1994). * * * (Jaeger and Mills, 2001: 314)

規準設定法に対する中立的見方

According to Segal, 'A man with a watch knows what time it is. A man with two watches is never sure.' Because there is no equivalent of atomic clock in the field of standard setting, our recommendation is simply for practitioners to invest in a single watch of greatest quality given available resources. * * * (Cizek and Bunch, 2007: 320)

In this sense, all cutscores are subjective. Yet, once a cutscore has been set, the decisions based on it can be made objectively. Instead of a separate set of judgments for each test taker, you will have the same set of judgments applied to all test takers. Cutscores cannot be objectively determined, but they can be objectively applied. * * * (Zieky, Perie & Livingston, 2008: 197)

規準設定法に対する肯定的見方

Some writers in the measurement literature have been skeptical of the meaningfulness of achievement standards and described the standard-setting process as blatantly arbitrary. We argue that standard setting is more appropriately conceived of as a measurement process similar to student assessment. * * * (Nicholes, Twing, Mueller, & O'Malley, 2010: 14-24)

Findings suggest that Bookmark-based methods have comparable reliability, resulting cut scores, and panelist evaluations to Angoff. Given that Bookmark-methods are shorter in duration and less costly, Bookmark-based methods may be preferable to Angoff for NAEP standard setting. * * * (Peterson, Schulz & Engelhard Jr. , 2011: 3-14)

3. 3. Bookmark Method とは何か

<Bookmark Methodの誕生>

Lewis, D.M., Mitzel, H.C. and Green, D.R. (1996, June)
Standard Setting; A Bookmark Approach.

<Bookmark Methodの特徴>

- (1)「項目応答理論(IRT)」の活用
- (2)複数の分割点の設定
- (3)多肢選択形式テストでも記述式テストでも活用
- (4)審査員の作業は極度に簡素化
- (5)テスト項目の内容を反映した評価

< Core Steps in the Bookmark Standard Setting >

1. Define PLDs (performance level descriptors) and focus on minimal performance levels
2. Create an OIB (ordered item booklet) with RPV (response probability value)
3. Present the OIB and elicit a bookmark for each cut-off
4. Collect the judgments of each standard setter
5. Calculate the median judgment for each PLD cut-off.
* * * (Lissitz, 2013: 165)

3. 4. PNO/TIN の数値差を利用した方法

< ORDERED ITEM BOOKLET >

Item 22

Ability level required for .67 chance of answering correctly: 1.725

Passage = Yellowstone

Which of these subheadings most accurately reflects the information in paragraphs 1 and 2?

- A. Effects of the Yellowstone Fire
- B. Tourism Since the Yellowstone Fire
- * C. News Media Dramatically Reports Fire
- D. Biodiversity in Yellowstone Since the Fire

* * * (Cizek, Bunch, and Koons, 2004: 37)

< ORDERED BOOKLET ITEM PARAMETERS AND ASSOCIATED THETA VALUES >

PNO	TIN	DIFF	DISC	<u>THETA@RP=.67</u>
1	19	-3.395	0.493	-2.550
2	13	-2.770	0.997	-2.352
3	01	-2.757	1.441	-2.468
4	22	-2.409	0.461	-1.505
5	04	-2.282	0.527	-1.492
6	02	-2.203	0.607	-1.517

* * * (Cizek, Bunch and Koons, 2004: 39)

THETA@RP=.67: (この項目に67%の正答率が求められる能力水準)

2PLM: $P = 1 / (1 + \exp(-Da(\theta - b)))$: $\theta = \ln(P / (1 - P)) / (Da) + b$:
 $= \ln(.67 / (1 - .67)) / (1.7 * .493) + (-3.395) = -2.550$

＜審査員によるbookmarkの位置＞

分析結果: Bookmarkの置き場所をTIN=2としたものは7名, TIN=04としたものは3名, TIN=13としたものが2名であった。

Bookmarkを置くように指示したことは、「正答率が.67以下に下がると思われるOIBの最初の頁にbook-markを置くこと」(Cizek, 2006: 247)である。

この指示と審査員の意思決定の方法が、Bookmark Method は任意的決定法という批判を生んでいると考えられる。また、これを「古典的精神物理学(classical psychophysics)」と説明しているところに、問題がないだろうか？ 何か客観的解決手段はないのだろうか？

< PNO/TIN間の数値差を利用した方法 >

分割点決定のための客観性を高める方法として、さまざまなことが考えられるが、**PNO (page number in OIB)**や**TIN (test item number)**間の数値差を利用した方法では、審査員の恣意的判断は避けることができ、前述のデータでは、より明確な分割点の設定が可能であった。

前述のデータには受験者の回答データがないし、そのデータ数も十分ではないので、Wright and Stone (1979: 31) Table 2.3.1. Original Response of 35 persons 18 items on the KNOX CUBE TESTを利用して、さらなる検討を行った。

< KNOX CUBE TEST: Wright & Stone >

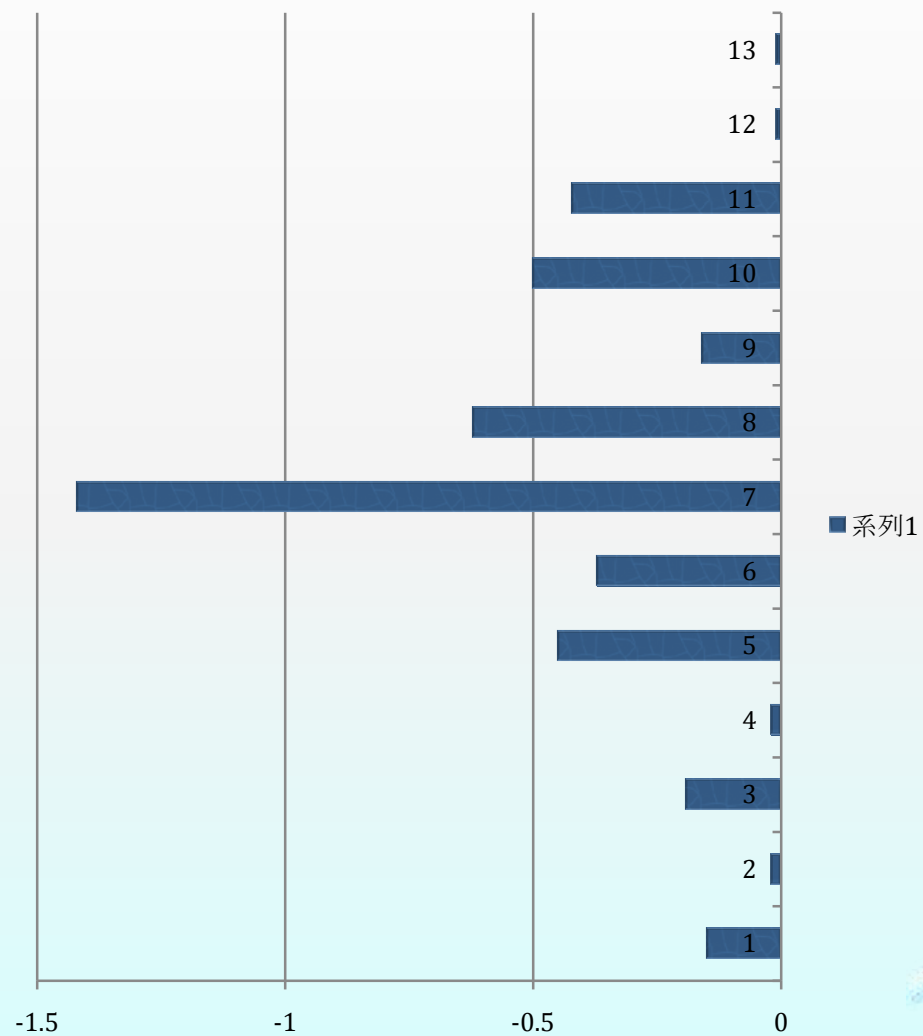
PNO	TIN	DIFF	DISC	THETA
1	4	-1.93	0.93	-1.482
2	7	-1.78	0.86	-1.296
3	5	-1.76	0.92	-1.307
4	6	-1.57	0.90	-1.107
5	9	-1.55	0.95	-1.112
6	8	-1.10	0.95	-0.662
7	10	-0.73	0.87	-0.251
8	11	0.69	0.85	1.180
9	13	1.31	0.93	1.758
10	12	1.47	0.85	1.960
11	14	1.97	0.85	2.460
12	17	2.39	0.95	2.828
13	16	2.40	0.95	2.838
14	15	2.41	0.94	2.853

< TIN 間の数値差を利用した推定法 >

DIFF (GDN(TIN-TIN)) DISC(GDN(TIN-TIN)) THET(GDN(TIN-TIN))

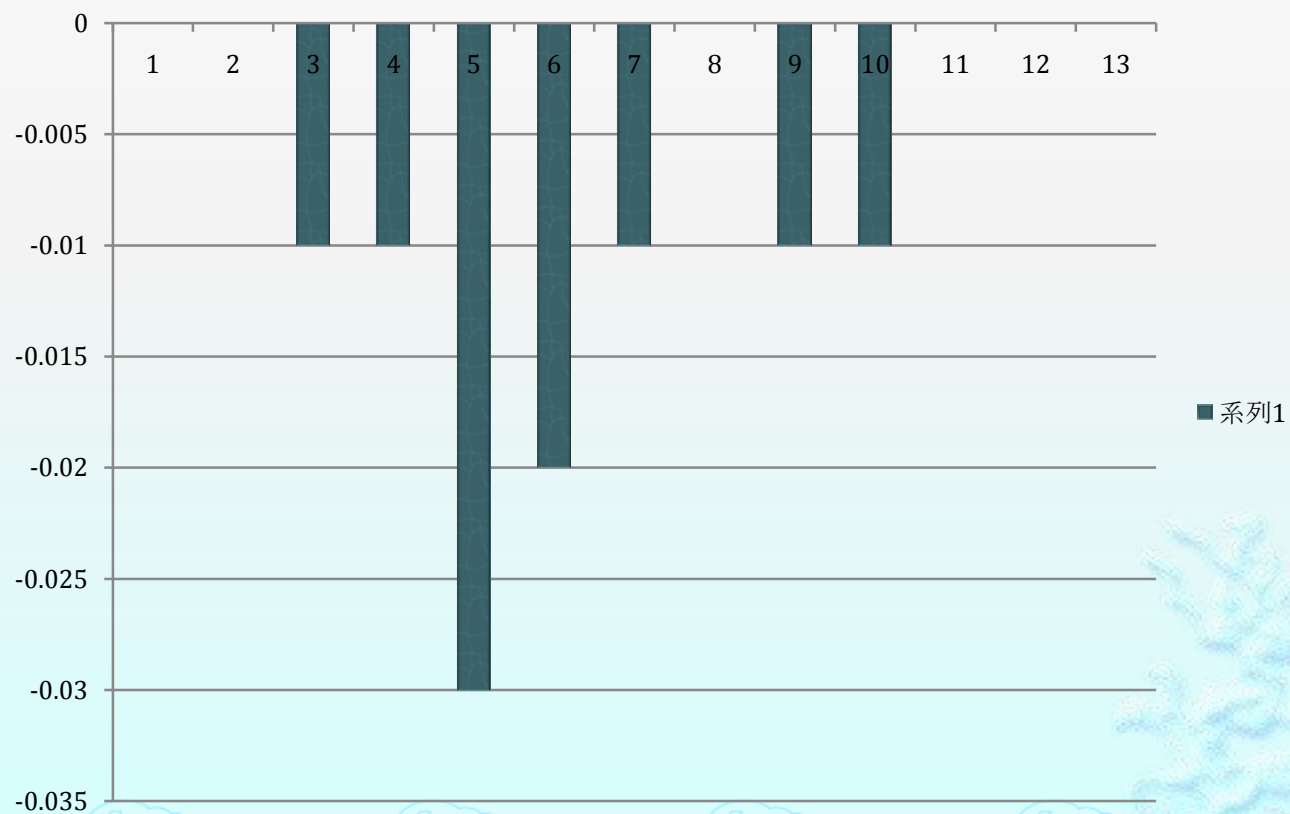
1. (04-07) -0.15	1. (11-12) 0.00	1. (04-05) -0.175
2. (07-05) -0.02	2. (12-14) 0.00	2. (05-07) -0.011
3. (05-06) -0.19	3. (14-07) -0.01	3. (07-09) -0.184
4. (06-09) -0.02	4. (07-10) -0.01	4. (09-06) -0.005
5. (09-08) -0.45	5. (10-06) -0.03	5. (06-08) -0.445
6. (08-10) -0.37	6. (06-05) -0.02	6. (08-10) -0.411
7. (10-11) -1.42	7. (05-04) -0.01	7. (10-11) -1.431
8. (11-13) -0.62	8. (04-13) 0.00	8. (11-13) -0.578
9. (13-12) -0.16	9. (13-15) -0.01	9. (13-12) -0.202
10. (12-14) -0.50	10. (15-09) -0.01	10. (12-14) -0.500
11. (14-17) -0.42	11. (09-08) 0.00	11. (14-17) -0.368
12. (17-16) -0.01	12. (08-17) 0.00	12. (17-16) -0.010
13. (16-15) -0.01	13. (17-16) 0.00	13. (16-15) -0.015
14. (15-	14. (16-	14. (15-

DIFFICULTY
(GDN(TIN-TIN))<7(10-11)>



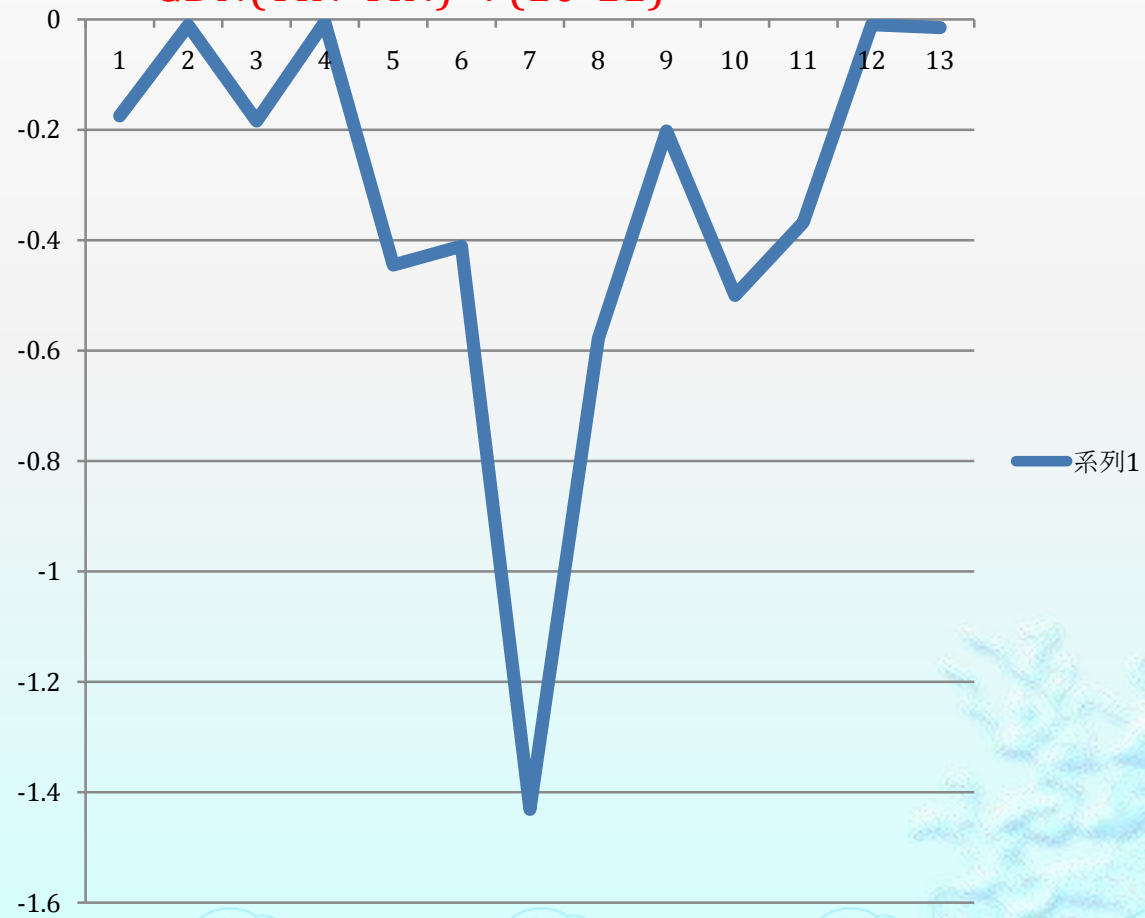
DISCRIMINATION

GDN(TIN-TIN)<5(10-6)>



THETA

GDN(TIN-TIN)<7(10-11)>



<PNO/TIN間の数値差を利用した推定法: 手順と結果>

1. テスト結果をIRTを用いて分析
2. RPを設定しTheta@RPを産出、OIBを作成
3. DIF, DIS, THE をそれぞれ低から高へ配列
4. PNO/TIN間の数値差を求め、GDNにそって表, グラフを作成
5. PNO/TIN間の数値差が最大のGDNとその前後のGDNを選定
6. 以上の2つのGDNに共通に, あるいは単独で、含まれるPNO/TINを選定
7. 以上のPNO/TINをbookmark の置き場所とする

<TIN=10>が分割点であることを, 審査員の主観的判断を必要とせず, 見つけ出すことが可能であった。

3. 5. Rasch ModelとLRTを併用した 分割点設定法

法月 健(静岡産業大学)



<なぜラッシュモデルと潜在ランク理論？>

1) ラッシュモデル (RM)

ラッシュモデルに基づくシステムは最善で恐らく唯一の規準維持の状況を説明する方法である (Bramley, 2010).

*** Bramley, T. (2010). *Locating objects on a latent trait using Rasch analysis of experts' judgments*. A paper presented at the conference “Probabilistic Models for Measurement in Education, Psychology, Social Science and Health,” Copenhagen, Denmark (June, 2010).

2) 潜在ランク理論 (LRT)

身長計や体重計が、ほとんど同じ2人の身長や体重の違いを見抜くことができる解像度(精度)の高い測定道具であるのに対し、テストは同じくらいの学力の違いを見分けるほど解像度が高い測定道具ではない. ⇒ 学力を段階評価するためのテスト理論 (荘島, 2010).

*** 荘島宏二郎 (2010)「ニューラルテスト理論」植野真臣・荘島宏二郎『学習評価の新潮流』, 東京: 朝倉書店.

<分析ツールと設定>

1) ラッシュモデル (RM)

Winsteps Ver. 3.80.1 (Linacre, 2013)

- * 項目難易度の平均(原点)を0に設定

- * 全員正解3項目、全問不正解1項目を除去

⇒全問不正解1名を除去(受験者34名, 14項目)

2) 潜在ランク理論 (LRT)

Exametrika Ver. 5.3 (荘島, 2011)

- * 自己組織化マップ (SOM), 2値データ,

潜在ランク数2, 一様分布に設定

<分析手順(1)>

S1:LRT (Exametrika) 分析ファイル(Excel) の
〈Examinee〉のシートに, RM (Winsteps) 分析
で得られた受験者能力(θ)と項目難易度(δ)
の値を挿入する.

S2:①受験者能力(θ)降順, ②潜在ランク降順,
③ランク・メンバーシップ・プロファイル(RMP)
の Rank 2 降順 $\Rightarrow$ 並べ替え

S3:①RMPランク2 $\rightarrow$ 1, ② θ の数値の変化,
③ δ の数値の変化 $\Rightarrow$ TINの分割点

<分析結果(1-1)>

A	B	C	D	E	F	G	H	I	J	K
ID	基本統計量				Rasch θ	潜在ランク 推定値	ランク・メンバーシップ・プロファイル		$\delta (\leq \theta) *$	Item No.
	回答数	回答率	正答数	正答率			Rank 1	Rank 2		
10007	14	1.000	11	0.786	3.73	2	0.000	1.000	3.37	14
10024	14	1.000	11	0.786	3.73	2	0.000	1.000	3.37	14
10015	14	1.000	10	0.714	2.85	2	0.000	1.000	2.24, 1.95	12, 13
10021	14	1.000	9	0.643	1.94	2	0.004	0.996	0.79	11
10022	14	1.000	9	0.643	1.94	2	0.005	0.995	0.79	11
10023	14	1.000	9	0.643	1.94	2	0.010	0.990	0.79	11
10034	14	1.000	9	0.643	1.94	2	0.010	0.990	0.79	11
10010	14	1.000	8	0.571	0.92	2	0.072	0.928	0.79	11
10014	14	1.000	8	0.571	0.92	2	0.072	0.928	0.79	11
10032	14	1.000	8	0.571	0.92	2	0.072	0.928	0.79	11
10018	14	1.000	8	0.571	0.92	2	0.118	0.882	0.79	11
10020	14	1.000	8	0.571	0.92	2	0.118	0.882	0.79	11
10013	14	1.000	7	0.500	-0.26	2	0.007	0.993	-1.57	10
10029	14	1.000	7	0.500	-0.26	2	0.017	0.983	-1.57	10
10016	14	1.000	7	0.500	-0.26	2	0.172	0.828	-1.57	10
10028	14	1.000	7	0.500	-0.26	2	0.172	0.828	-1.57	10
10003	14	1.000	7	0.500	-0.26	2	0.278	0.722	-1.57	10
10002	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10005	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10006	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10008	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10009	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10026	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10031	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10019	14	1.000	6	0.429	-1.37	1	0.866	0.134	-1.57	10
10030	14	1.000	6	0.429	-1.37	1	0.866	0.134	-1.57	10
10017	14	1.000	6	0.429	-1.37	1	0.970	0.030	-1.57	10
10012	14	1.000	5	0.357	-2.23	1	0.973	0.027	-2.35	8
10011	14	1.000	5	0.357	-2.23	1	0.988	0.012	-2.35	8
10001	14	1.000	4	0.286	-2.94	1	0.998	0.002	-3.38	6
10027	14	1.000	4	0.286	-2.94	1	0.998	0.002	-3.38	6
10004	14	1.000	3	0.214	-3.61	1	0.999	0.001	-4.4	4
10033	14	1.000	3	0.214	-3.61	1	1.000	0.000	-4.4	4
10025	14	1.000	2	0.143	-4.32	1	1.000	0.000	n/a	n/a

<分析結果(1-2)>

A	B	C	D	E	F	G	H	I	J	K
ID	基本統計量					潜在ランク 推定値	ランク・コンバージョン・プロファイル		$\delta (\leq \theta) *$	Item No.
	回答数	回答率	正答数	正答率	Rasch θ		Rank 1	Rank 2		
10007	14	1.000	11	0.786	3.73	2	0.000	1.000	3.37	14
10024	14	1.000	11	0.786	3.73	2	0.000	1.000	3.37	14
10015	14	1.000	10	0.714	2.85	2	0.000	1.000	2.24, 1.95	12, 13
10021	14	1.000	9	0.643	1.94	2	0.004	0.996	0.79	11
10022	14	1.000	9	0.643	1.94	2	0.005	0.995	0.79	11
10023	14	1.000	9	0.643	1.94	2	0.010	0.990	0.79	11
10034	14	1.000	9	0.643	1.94	2	0.010	0.990	0.79	11
10010	14	1.000	8	0.571	0.92	2	0.072	0.928	0.79	11
10014	14	1.000	8	0.571	0.92	2	0.072	0.928	0.79	11
10032	14	1.000	8	0.571	0.92	2	0.072	0.928	0.79	11
10018	14	1.000	8	0.571	0.92	2	0.118	0.882	0.79	11
10020	14	1.000	8	0.571	0.92	2	0.118	0.882	0.79	11
10013	14	1.000	7	0.500	-0.26	2	0.007	0.993	-1.57	10
10029	14	1.000	7	0.500	-0.26	2	0.017	0.983	-1.57	10
10016	14	1.000	7	0.500	-0.26	2	0.172	0.828	-1.57	10
10028	14	1.000	7	0.500	-0.26	2	0.172	0.828	-1.57	10
10003	14	1.000	7	0.500	-0.26	2	0.278	0.722	-1.57	10
10002	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10005	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10006	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10008	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10009	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10026	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10031	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10019	14	1.000	6	0.429	-1.37	1	0.866	0.134	-1.57	10
10030	14	1.000	6	0.429	-1.37	1	0.866	0.134	-1.57	10
10017	14	1.000	6	0.429	-1.37	1	0.970	0.030	-1.57	10
10012	14	1.000	5	0.357	-2.23	1	0.973	0.027	-2.35	8
10011	14	1.000	5	0.357	-2.23	1	0.988	0.012	-2.35	8
10001	14	1.000	4	0.286	-2.94	1	0.998	0.002	-3.38	6
10027	14	1.000	4	0.286	-2.94	1	0.998	0.002	-3.38	6
10004	14	1.000	3	0.214	-3.61	1	0.999	0.001	-4.4	4
10033	14	1.000	3	0.214	-3.61	1	1.000	0.000	-4.4	4
10025	14	1.000	2	0.143	-4.32	1	1.000	0.000	n/a	n/a

<分析結果(1-3)>

A	B	C	D	E	F	G	H	I	J	K
	基本統計量					潜在ランク	ラング・メンバーシップ・プロフィール			
ID	回答数	回答率	正答数	正答率	Rasch θ	推定値	Rank 1	Rank 2	$\delta(\leq \theta)*$	Item No.
10007	14	1.000	11	0.786	3.73	2	0.000	1.000	3.37	14
10024	14	1.000	11	0.786	3.73	2	0.000	1.000	3.37	14
10015	14	1.000	10	0.714	2.85	2	0.000	1.000	2.24, 1.95	12, 13
10021	14	1.000	9	0.643	1.94	2	0.004	0.996	0.79	11
10022	14	1.000	9	0.643	1.94	2	0.005	0.995	0.79	11
10023	14	1.000	9	0.643	1.94	2	0.010	0.990	0.79	11
10034	14	1.000	9	0.643	1.94	2	0.010	0.990	0.79	11
10010	14	1.000	8	0.571	0.92	2	0.072	0.928	0.79	11
10014	14	1.000	8	0.571	0.92	2	0.072	0.928	0.79	11
10032	14	1.000	8	0.571	0.92	2	0.072	0.928	0.79	11
10018	14	1.000	8	0.571	0.92	2	0.118	0.882	0.79	11
10020	14	1.000	8	0.571	0.92	2	0.118	0.882	0.79	11
10013	14	1.000	7	0.500	-0.26	2	0.007	0.993	-1.57	10
10029	14	1.000	7	0.500	-0.26	2	0.017	0.983	-1.57	10
10016	14	1.000	7	0.500	-0.26	2	0.172	0.828	-1.57	10
10028	14	1.000	7	0.500	-0.26	2	0.172	0.828	-1.57	10
10003	14	1.000	7	0.500	-0.26	2	0.278	0.722	-1.57	10
10002	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10

- 10003番の受験者⇒ ランク2で θ が最も低い受験者のうち
 ランク2に所属する確率が最も低い受験者
- $\delta(\leq \theta)* \Rightarrow \theta$ の値以下でそれに 近接する項目難易度の値
 (右隣はその項目)

<分析結果(1-4)>

A	B	C	D	E	F	G	H	I	J	K
	基本統計量					潜在ランク	ランク・メンバーシップ・プロフィール			
ID	回答数	回答率	正答数	正答率	Rasch θ	推定値	Rank 1	Rank 2	$\delta(\leq \theta)^*$	Item No.
10003	14	1.000	7	0.500	-0.26	2	0.278	0.722	-1.57	10
10002	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10005	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10006	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10008	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10009	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10026	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10031	14	1.000	7	0.500	-0.26	1	0.708	0.292	-1.57	10
10019	14	1.000	6	0.429	-1.37	1	0.866	0.134	-1.57	10
10030	14	1.000	6	0.429	-1.37	1	0.866	0.134	-1.57	10
10017	14	1.000	6	0.429	-1.37	1	0.970	0.030	-1.57	10
10012	14	1.000	5	0.357	-2.23	1	0.973	0.027	-2.35	8
10011	14	1.000	5	0.357	-2.23	1	0.988	0.012	-2.35	8
10001	14	1.000	4	0.286	-2.94	1	0.998	0.002	-3.38	6
10027	14	1.000	4	0.286	-2.94	1	0.998	0.002	-3.38	6
10004	14	1.000	3	0.214	-3.61	1	0.999	0.001	-4.4	4
10033	14	1.000	3	0.214	-3.61	1	1.000	0.000	-4.4	4
10025	14	1.000	2	0.143	-4.32	1	1.000	0.000	n/a	n/a

- 10031番の受験者 $\Rightarrow \theta = -0.26$ の終点, これ未満の $\theta = -1.37$ (3名) もTIN10の $\delta = -1.57$ よりも高い値
- 10017番の受験者 $\Rightarrow \theta = -1.37$ の終点, これ未満の $\theta = -2.23$ はTIN10の $\delta = -1.57$ よりも低い値

*** TIN10が分割点の有力候補**

<分析手順(2)>

*LRT分析ファイルの <Item>シート

①項目参照プロファイル(IRP) の Rank 1昇順、

②IRP指標の Beta 降順

⇒ 並べ替え

分析結果(2)

	統計量				項目参照プロフィール(IRP)		IRP指標					
項目	正答数	正答率	標準偏差(σ)	項目得点μ	Rank 1	Rank 2	Alpha	A	Beta	B	Gamma	C
問15	1	0.029	0.171	0.339	0.005	0.054	1	0.049	2	0.054	0.000	0.000
問16	1	0.029	0.171	0.339	0.005	0.055	1	0.050	2	0.055	0.000	0.000
問17	1	0.029	0.171	0.339	0.005	0.055	1	0.050	2	0.055	0.000	0.000
問14	3	0.088	0.288	0.208	0.014	0.161	1	0.147	2	0.161	0.000	0.000
問12	6	0.176	0.387	0.420	0.028	0.324	1	0.296	2	0.324	0.000	0.000
問13	7	0.206	0.410	0.607	0.033	0.380	1	0.347	2	0.380	0.000	0.000
問11	12	0.353	0.485	0.552	0.056	0.647	1	0.592	2	0.647	0.000	0.000
問10	24	0.706	0.462	0.570	0.607	0.804	1	0.197	1	0.607	0.000	0.000
問8	27	0.794	0.410	0.718	0.670	0.918	1	0.248	1	0.670	0.000	0.000
問9	30	0.882	0.327	0.646	0.783	0.981	1	0.197	1	0.783	0.000	0.000
問6	30	0.882	0.327	0.427	0.833	0.931	1	0.098	1	0.833	0.000	0.000
問5	31	0.912	0.288	0.439	0.837	0.986	1	0.149	1	0.837	0.000	0.000
問4	32	0.941	0.239	0.413	0.891	0.990	1	0.099	1	0.891	0.000	0.000
問7	31	0.912	0.288	0.240	0.936	0.887	1	0.000	2	0.887	1.000	-0.049

- Beta =2の場合は、ランク2の受験者の正答率のほうが、ランク1の受験者の正答率よりも、50%に近接
- Beta =1の項目中、**TIN10** はランク1,2受験者の正答率が一番低く、Beta =2下限の **TIN11**よりもランク間の正答率が均衡
- ランク1受験者の正答率がランク2の正答率を上回り、Beta=2を示した**TIN7**は例外（弁別力低、RMでもミスフィット）

*** TIN10は分割点領域に位置する**

4. むすび

教育再生実行会議なども含めて、これからの大学教育等の在り方をめぐって、議論が戦わされている昨今である。大学入試・卒業にTOEFL等を導入。一点刻みを改め、段階評価。「英語教育, 迫り来る破綻」。「点数不足ならともかく、人物本位での落第となると衝撃はいや増す」。CAN-DO リスト。PISAは過去最高か？しかし、もうこの「苦しみの連鎖」から脱出しませんか？どんなことから、どんな順序で？

近頃、language assessment literacy という言葉によく巡り会う。その意味は、an understanding of the principles of sound assessment ということである。その適切で妥当な知見を可能な限り広めることが、我々に与えられた任務のひとつではなかろうか。

謝辞:「3. 規準設定の意味と方法」の一部は、公益財団法人日本英語検定協会:英語教育研究センター委託研究のための助成を受けて行った研究である。

参考文献

- Bachman, L.F. and Palmer, A.S. (1996). *Language Testing in Practice*. OUP.
- Carroll, J.B.(1961). 'Fundamental consideration in testing English proficiency of foreign students'. In *Testing the English Proficiency of Foreign Students* .(pp-30-40). Center for Applied Linguistics.
- Chalhoub-Deville, M. and Deville, C. (2006). 'Old, Borrowed, and New Thoughts in Second Language Testing'. In Brennan, R.L. (ed.). *Educational Measurement (Fourth Edition)*. American Council on Education and Praeger Publishers
- Cizek, G.J. and Bunch, M.B. (2007). *Standard Setting, A Guide to Establishing and Evaluating Performance Standards on Tests*.(p.320). Sage.
- Cizek, G.J.(2006). 'Standard Setting'. In S.M. Downing & T.M. Haladyna (Eds.) *Handbook of Test Development*, (p.226, p.247). Lawrence Erlbaum Associates, Publishers.
- Cizek, G.J., Bunch,M.B., and Koons, H. (2004). Setting Performance Standards: Contemporary Methods, *Educational Measurements: Issues and Practice*, 23, (4).31-50.
- Doughty, C. & Williams, J. (1998). 'Pedagogical choices in focus on form' . In C. Doughty & J. Williams (Eds.) *Focus on form in classroom second language acquisition* (pp. 197-261). Cambridge University Press.
- Dulay, H., Burt, M. and Krashen, S. (1982). *Language Two*. Oxford University Press.

- ELEC (2013).『日本の英語教育とELEC』英語教育協議会 (p. 73).
- Ellis, R. (2009). 'Task-Based Language Teaching: Sorting Out the Misunderstandings' *International Journal of Applied Linguistics*. 19, 3 : 221-246.
- Fries, C.C. (1958). On the Oral Approach. *Lectures by C.C. Fries and W.F. Twaddell*. 研究社出版株式会社. (pp. 14-15).
- Fulcher, G. (2010). *Practical Language Testing*. (p.323) Hodder Education.
- Hambleton, R. K. & Pitoniak, M. J. (2006). Setting Performance Standards. In Brennan, R.L. (ed.) *Educational Measurement* (Fourth Edition).(p.440) ACE
- Harris, D.P. (1969). *Testing English as a Second Language*. McGraw-Hill, Inc.
- Jaeger, R.M. and Mills, C.N. (2001). An Integrated Judgment Procedure for Setting Standards on Complex, Large-scale Assessments. In Cizek, G.T. (ed.) *Setting Performance Standards*, (p.314) Lawrence Erlbaum Associates, Publishers
- Kaftandjieva, F.(2004). Section B: Standard Setting., *Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR*, (p.31).Council of Europe.
- Krashen, S. (1985). *The input hypothesis: Issues and implications*. Longman.
- Lado, L. (1961). *Language Testing*. MacGraw-Hill, Inc.
- Lewis, D.M., Mitzel, H.C., and Green , D.R. (1996, June). Standard Setting: A Bookmark Approach. In Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*, Symposium conducted at Council of Chief State School Officers National Conference on Large-Scale Assessment , Phoenix, AZ.

- Lissitz, R.W. (2013). Standard Setting: Past, Present, and Perhaps Future. In Simon, Ercikan, and Rousseau (Eds.). *Improving Large-Scale Assessment in Education*. (p.165). Taylor & Francis.
- Long, M. (1991). 'Focus on form: a design feature in language teaching methodology' In K.R. de Bot, Ginsberg, and C. Kramsch (eds.): *Foreign Language Research in Cross-Cultural Perspective*. John Benjamins.
- Long, M. (1996). 'The role of the linguistic environment in second language acquisition'. In W. Ritchie and T. Bhatia (eds.): *Handbook of Second Language Acquisition*. Academic Press. (pp.413-468).
- Lyster, R. & Ranta, L. (1997). 'Corrective feedback and learner uptake: Negotiation of form in communicative classrooms'. *Studies in Second Language Acquisition*, 19, 37-66.
- Nicholes, P., Twing, J., Mueller, C.D. and O'Malley, K. (2010). Standard-Setting Methods as Measurement Process, *Educational Measurement: Issues and Practice*, 29 (1), 14-24.
- Norris, J.M. (2002). Special Issue: Task-based language assessment. *Language Testing*, Vol.19, Issue 4.
- Oller, J.W. (1979). *Language Tests at School*. Longman Group Ltd.
- Peterson, C.H., Schulz, E.M. and Engelhard Jr., G. (2011). Reliability and Validity of Bookmark-based Methods for Standard Setting, *Educational Measurement: Issues and Practice*. 30 (2), 3-14.

- Shehadeh, A. (2012). 'Task-Based Language Assessment: Components, Development, and Implementation,' In Coombe, Davidson, O'Sullivan & Stoyloff (eds.). *The Cambridge Guide to Second Language Assessment*, (pp. 156-163). Cambridge University Press.
- Williams, J. (2005). Form-focused instruction. In E. Hinkel (ed.) *Handbook of research in second language teaching and learning*. (pp. 671-691). Lawrence Erlbaum.
- Wright, B.D. and Stone, M.H. (1979). *Best Test Design* (p.31), MESA.
- Zieky, M.J. , Perie, M., and Livingston, S.A. (2008). *Cutscore: A Manual for Setting Standards of Performance on Educational and Occupational Tests*, (p.197). ETS
- 皆見英代 (2008).「規準」と「基準」: criterion と standard の区別と英和照合——教育評価の専門用語和訳に戸惑う.『国立教育政策研究所紀要137』
- 橋本重治 (1983).『続・到達度評価の研究——到達基準設定の方法』(p.28). 日本図書文化協会.
- 山家 保 (1972).『実践英語教育』ELEC . (p. 76, 281).
- 村野井仁(2006).『第二言語習得研究から見た効果的な英語学習法・指導法』. 大修館書店. (p.88, 109).
- 池田 央:日本語版監訳(2008).『テスト作成ハンドブック』 (p.12). 教育測定研究所.
- 鈴木孝明・白畑知彦 (2013).『ことばの習得:母語獲得と第二言語習得』くろしお出版 (p.146).
- JACET SLA研究会 (2013).『第二言語習得と英語教育法』. 開拓社. (pp.41-45).