英語教師のための
言語テストに関する基礎知識[1)]

島　田　勝　正

はじめに

テストは人生の意思決定に重要な役割を果たす。したがって，それは「よい」テストでなければならない。よいテストの条件として妥当性，信頼性，実用性が挙げられる。本稿では妥当性，信頼性，実用性について解説し，テストを改善するための方策についてテスト全体とテスト項目一つひとつから検討する。また，テスト得点の解釈やスコアの算出のための基本的な統計的知識も併せて示す。

1. テストの目的

1.1　指導・学習・評価

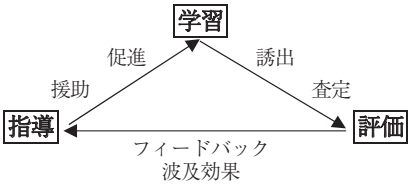


図 1：指導・学習・評価の関係（島田，2000 に基づく）

キーワード：言語テスト，英語教師，「よい」テスト

テストは評価のためのデータを得る重要な手段である。テストの役割を，指導（teaching）と学習（learning）と評価（testing）の相互関係という観点から考えてみると，その役割がよくわかる。図１が示すように，指導は学習を促進し（facilitate），援助（assist）する。一方，テストは学習の成果を誘出し（elicit），その成果をある規準･基準に基づいて査定（assess）する。そして，その情報を指導や学習に役立つように提供する（feedback）。

テストのもう一つの機能として，指導や学習への影響がある。例えば，2006 年度から大学入試センターにリスニングテストが導入された結果，高等学校の授業ではリスニングやスピーキングが重視されるようになり，高校生もその対策をするようになった。このように，テストが先導して指導や学習に影響を与えることをテストの波及効果（washback effect）という。

1. 2　テストの用途

生徒の能力を測定するために教育プログラムではテストがさまざまな目的で使用される。Harris（1969）は，テストの目的（機能）として次の 6 点を挙げている（pp. 2–3）。

（1）教育プログラムに入るための準備（readiness）ができている者といない者を選別する（例：留学テスト（TOEFL, IELTS[2)]），入学試験）
（2）能力レベルに応じて適したクラスに割り振る（例：クラス分けテスト（placement test））
（3）個々の生徒の得意な，または不得意な技能・分野を診断する（例：診断テスト（diagnostic test））
（4）将来の成績を予測するために学習に対する適性（aptitude）を測定する（例：適性テスト）
（5）教育目標に対して生徒がどの程度到達（achievement）できたかを測定する（例：中間テスト・期末テスト）

(6) 学習者が現時点でどの程度の言語能力（proficiency）を持っているかその熟達度を測定する[3)]（例：実用英語検定）

(7) 教育プログラムの効果を測定する

また, Hughes（1989）は, テストをその使用目的に応じて, 熟達度テスト, 到達度テスト, 診断テスト, クラス分けテストの４つにタイプに分類している（p. 7）。このように, テストの用途は実に多岐にわたる。

2. 妥当性

2. 1　妥当性の定義

「妥当性（validity）」に関して, *Lougman Dictionary of Language Teaching & Applied Linguistics* は “the degree to which a test measures what it is to supposed to measure, or can be used successfully for the purposes for which it is intended”（Richards & Schmidt, 2010, p. 622）と定義している。簡潔に言えば, 妥当性とは, そのテストが測ろうとしているものを正確に測っているかどうかということである。

テストの基本的な機能は, 測定（measurement）である。「測られるもの」として「能力」があり, それを「測るもの」としての「道具」がテストである。したがって,「測られるもの」と「測るもの」のマッチングが重要になってくる。測られるもの（能力）を測るもの（道具）が正確に測っていないとすれば, 測られる方は堪ったものではない。

テストの妥当性という概念にはいくつかの側面があり, それぞれの側面の妥当性を調べるために, 内容的妥当性（content validity）, 構成概念妥当性（construct validity）, 併存的妥当性（concurrent validity）, 予測的妥当性（predictive validity）の方略がある。

内容的妥当性とは, テストの内容が測定しようとしている知識・技能を代表しているか, 総括的であるかどうかをいう。構成概念妥当性とは, テストがそ

の構成概念（言語理論の枠組みで想定された，基底となる能力）を反映したものになっているかどうかをいう。併存的妥当性とは，新しく開発されたテストが，既存の標準化されたテストとどの程度相関するかをいう。そして，予測的妥当性とは，当該テストが外部規準での成績をどの程度予測できるかをいう。

2.2　テストの進化と妥当性

2.2.1　TOEFL

TOEFL（Test of English as a Foreign Language）の変遷は，留学テストとしての妥当性の追求であったといってもよい。第1世代の筆記試験（Paper-Based Test; PBT）はリスニング，リーディングと文法･語法から構成されていた。しかし，この筆記試験で高得点をとった学生が，はたして，北米の大学に留学して英語で行われる授業に問題なくついていけるかどうか，例えば，与えられた課題について英語でレポートを書いたり，講師の先生に英語で質問したりすることができるかどうかは，妥当性の観点からみると甚だ疑問である。

そして，コンピューターの発達にともない，コンピューター・ベース（Computer-Based Test; CBT）の第2世代のTOEFLにはエッセイライティングが加わることになった。また，CBTの導入により，項目応答理論（item response theory; IRT）に基づいて，受験者の能力に応じて，提示するテスト項目の困難度を変えることが可能になった。このテストは日本では2000年から2006年まで実施された。

さらに，インターネットの普及にともない，コンピューター・ベースのTOEFLは第3世代のインターネット・ベース（Internet-Based Test; IBT）に取って代わった。そして，スピーキングが加わって4技能をすべてカバーすることになったのである。

TOEFLが進化するにつれて，CBTではライティング，そして，IBT

ではスピーキングと，テストされる技能が増えていった。このようなTOEFLの変遷から，留学テストとしては，ライティングやスピーキングの技能も測定しないと，その妥当性が低いと解釈することができる。さらに，IBTでは，複数の技能を統合的に（integrated）測定するようになっている。留学のさまざまな場面に対応するために，より真正性（authenticity）が向上したといえる。

2.2.2　大学入学試験問題

大学入試センターが実施した英語のテストでは，筆記テスト形式で発音やアクセントを問う問題が長く続いた。以下はその例である。

下線部の発音が，他の3つの場合と異なるものを，それぞれA.～D.のうちから一つずつ選びなさい。

問1	A. b<u>oo</u>t	B. g<u>oo</u>se	C. pr<u>oo</u>f	D. w<u>oo</u>l
問2	A. bread<u>th</u>	B. fai<u>th</u>	C. leng<u>th</u>	D. smoo<u>th</u>
問3	A. <u>ear</u>n	B. h<u>ear</u>t	C. p<u>ear</u>l	D. s<u>ear</u>ch
問4	A. lei<u>s</u>ure	B. mea<u>s</u>ure	C. phy<u>s</u>ics	D. vi<u>s</u>ion

与えられた語と第1アクセント（第1強勢）の位置が同じ語を，それぞれA.～D.のうちから一つずつ選びなさい。

問1　damage

A. convince　B. effort　C. prefer　D. throughout

問2　recommend

A. guarantee　B. museum　C. objective　D. satisfy

問3　fortunately

A. appreciate　B. elevator　C. manufacture　D. sympathetic

このような筆記テストで高得点をとった受験者が，実技テストにおいて実際にその単語を正しく発音できるかどうかは疑問である。大学入試センター試験において発音やアクセントを単独で問う問題は，2021 年度から実施された大学入学共通テストでは出題されなくなった。このような問題作成方針の転換はテストの妥当性で説明できる。

3. 相関

3.1　相関の概念

前節の妥当性や後節の信頼性について理解するためには「相関（correlation）」という概念について知っておく必要がある。相関とは，2 つの変数（x，y）の変動がどの程度似ているかというその度合い（関係）のことをいう。

表 1 は，テスト A とテスト B の 2 つのテストを実施して得られた得点を示している。

表 1 のデータを，x 軸に Test A，y 軸に Test B をとり，その 2 つの得点が交差するところに点を打つと，図 2 のグラフができる。このグラフを散布図（scattergram）という。散布図には 2 つの変数の関係を視覚的に把握できるという効果がある。

仮にテスト A の得点とテスト B の得点が同一であるとすると，この点をつなぐと一直線になる。一直線になった場合，相関係数（r）は 1.000 と

表 1：テスト A とテスト B の得点

ID No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
OIT	24	29	22	22	28	32	26	26	25	22	18	28	22	11	25	25	18	15	24	28	30	27	22	32	34	25	28	16	28	19
OTT	27	30	26	18	31	32	31	25	22	24	20	31	18	11	31	29	27	19	27	30	33	24	22	29	29	27	31	31	33	25

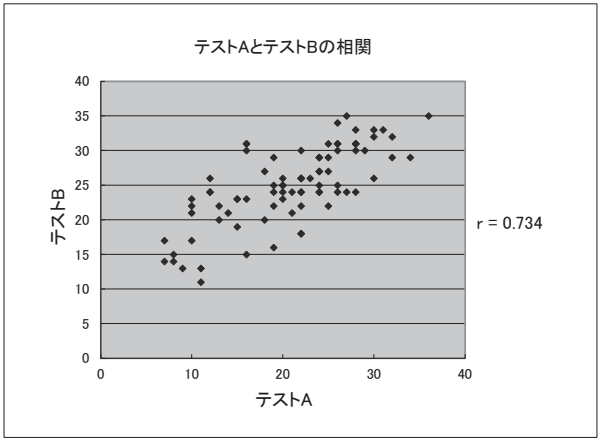


図 2：散布図

なり，この 2 つのテスト間には非常に強い相関があるといえる。直線の向きが右上がりの場合を正の相関といい，右下がりの場合を負の相関という。相関係数は +1 〜−1 の間を変異する。一般的に r = 0.7 以上を強い相関，r = 0.3 以下を弱い相関とよんでいる。

3. 2　相関と妥当性

もしも，前述した大学入試センターの発音・アクセント問題に関して，その「筆記テスト」と「実技テスト」の相関が高ければ，たとえば，相関係数が 0.7 であれば，その決定係数は $(0.7)^2 = 0.7 \times 0.7 = 0.49$ となり，2 つのテストはその重複部分となる約半分（49%）が同じ能力を測定していることになる。

さらに，相関係数が 0.9 であれば，その決定係数は $(0.9)^2 = 0.9 \times 0.9 = 0.81$ となり，両者は，5 分の 4 以上（81%）の同じ能力を測定していることになる（図 3）。よって，「筆記テスト」で実際に正しい発音・正しいアクセントで単語を発音する能力をかなりの度合いで測定することができるとい

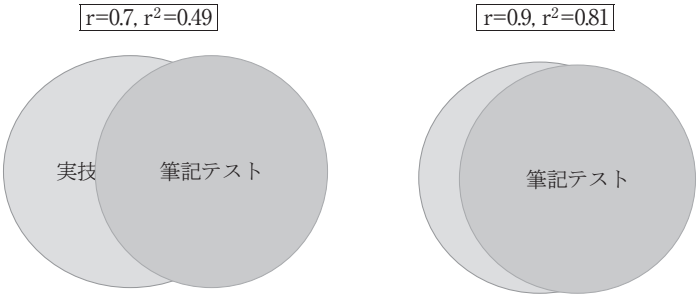


図 3：相関係数の解釈

うことになる。つまり，そのような結果になれば，「筆記テスト」が「実技能力」を測る妥当性はかなり高くて，「筆記テスト」が「実技テスト」の代替となる可能性があるということになるが，実際にはその可能性は低いだろう。

4. 信頼性

4. 1　信頼性の定義

信頼性（reliability）とは *Longman Dictionary of Language Teaching and Applied Linguistics* によれば，“A measure of the degree to which a test gives consistent results. A test is said to be reliable if it gives the same results when it is given on different occasions or when it is used by different people”（Richards & Schmidt, 2010, p. 495）と定義されているように，異なった状況や人々に対して繰り返し実施しても同じ一貫した結果を示す程度のことをいう。

また，*Dictionary of Language Testing* によれば，信頼性とは “The actual level of agreement between the results of one test with itself or with another test. Such agreement, ideally, would be the same if there were no measurement error”（Davies et al., 1999, p. 168）と定義されて

おり，測定誤差（measurement error）がないと一致度は同じになる。

信頼性とは観測得点（observed score）の中に真の得点（true score）が占める割合をいう。観測得点とは，その受験者がテストを受けて実際にとった点数のことである。しかし，テストの観測得点が受験者の能力を必ずしもすべて反映しているとは限らない。観測得点には真の得点だけでなく誤差（error）も含まれる。

$$\text{信頼性} = \frac{\text{真の得点}}{\text{観測得点}} = \frac{\text{真の得点}}{\text{真の得点} + \text{誤差}}$$

4.2　信頼性係数

信頼性は信頼性係数として数値化されるが，その方法には安定性（stability）を判断するものと，内的一貫性（internal consistency）を判断するものの2種類がある。

安定性は，再テスト法（test-retest method）や代替テスト法（alternative forms method）により測定される。再テスト法とは，同じテストを同じ受験者集団に2回実施しその相関係数を算出する方法である。1回目と2回目の間に高い相関が得られたならば，そのテストは安定している（stable）ということができる。しかしながら，2回も同じテストを受けなければならないという実用性の低さや1回目に受けたことが練習効果として2回目に影響するなどの問題点が指摘される。そこで同じテストの異なったフォームを実施する代替テスト法という方法が考えられた。

安定性を測るにはテストを2回実施しなければならないという問題点がある。そこで，テストは実際は1回しか実施しないが，全項目を折半（split-half）して2つのテストを実施したとみなす方法が考案された。この方法では，項目のその内部一貫性，すなわち各項目が同質（homogeneous）であるかをみるために，1つのテストを折半してその半数の項目からなる

2セットの得点間の相関係数を算出する。そして，スピアマン・ブラウンの予測公式（Speaman-Brown prophecy formula）で信頼性を推定する。全項目を折半する方法にはいくつかあるが，奇数番号の項目群と偶数番号の項目群に分ける方法が一般的である。

折半法では，テストを2つに折半する方法により，信頼性係数が変異する。そこで考え出されたのが，このすべての組み合わせの相関係数の平均値をとるもので，正解と不正解からなる1, 0データでは，キューダー・リチャードソンの公式（Kuder-Richardson（KR）formulas），項目ごとに配点を変える重みづけされた連続データでは，クロンバック α（Cronbach's alpha）がある（Brown, 1996, p. 202）。以下に精度がもっとも高いといわれるKR20の公式を示す。

$$\mathrm{KR20} = \frac{k}{k-1}\left(1 - \frac{\Sigma\,\sigma^2 i}{\sigma^2 x}\right)$$

ただし，k は項目数，$\sigma^2 i$ はテスト項目の分散，$\sigma^2 x$ はテスト総得点の分散である。この公式から受験者の能力の分散が大きいほど信頼性係数が高くなることがわかる。

4. 3　誤差

前述したように，観測得点の中には誤差が含まれている。つまり，観測得点は真の得点と誤差から成り立っている。例えば多肢選択式テストで正解がわからず，鉛筆を転がして回答したら，それがたまたま正解だったということがある。このようなまぐれあたりのことを，当て推量（guessing）という。当て推量によって正解を得たとしても，それはその受験者の実力ではない。したがって，本当の実力を知りたい場合は，観測得点から誤差を差し引かなければならない。仮に，その受験者が10点をとったとしても，

表 2：TOEFL（PBT）における誤差

	ITP Pre-TOEFL		ITP TOEFL
得点範囲	200–500		310–677
問題数	95	<	140
テスト時間	70		115
信頼性係数	0.89	<	0.95
標準測定誤差	16.4	>	13.7

そのうちの 2 点はまぐれあたりだったのかもしれない。そうすると，この受験者の実力は 8 点ということになる。このように，テストにはつねに誤差がつきまとう。

上の表は 2 種類の PBT–TOEFL のテストのデータを比較したものである。筆記試験（PBT）タイプの TOEFL は，今でも団体受験（Institutional Testing Program; ITP）用に使われている。ITP には，通常の TOEFL（677 点満点）と，易しい項目で構成された簡易版の Pre-TOEFL（500 点満点）がある。

表 2 に示す標準測定誤差（Standard Error of Measurement; SEM）とは，受験者の観測得点が，真の得点からどれだけずれているかを表す数値である。SEM の算出式を下記に示す。

標準測定誤差＝標準偏差（Standard Deviation; SD）×（1－信頼性係数）

標準測定誤差から，個人の真の得点が変異する範囲を予測することができる。例えば，A 君が ITP Pre-TOEFL で 400 点をとったとする。標準測定誤差は ± 16.4 であるから，

$$400 - 16.4 = 383.6 \qquad 400 + 16.4 = 416.4$$

A 君の実力，つまり，真の得点は，383.6 ～ 416.4 の間に入る確率が 68.3%（正規分布曲線下の平均点 ± 1 標準偏差が占める面積）あるということになる。

5. 実用性

実用性（practicality）とは，*Dictionary of Language Testing* によれば，“The term practicality covers a range of issues, such as the cost of development and maintenance, test length, ease of marking, time required to administer the test, ease of administration, and equipment required”（Davies et al, 1999, p. 148）と定義されている。いくらそのテストが妥当性や信頼性が高くても実用性が低ければ実際に使えない。実用性には，費用，テストの長さ，採点，テスト時間，設備等，さまざまな要因が関係する。

例えば，表 2 におけるテスト時間をみると，Pre-TOEFL は TOEFL に比べて，項目数が 140 から 95 に 45 項目も減った分，テスト時間が 115 分から 70 分へと 45 分も縮小されている。したがって，Pre-TOEFL は短時間で実施でき，TOEFL よりも実用性はより高いといえる。

信頼性を上げるためには，項目数を増やさなければならないが，項目数を増やせば，それを解くためのテスト時間が増える。したがって，信頼性と実用性は，一種の相殺関係（trade-off）にあり，どこかで妥協が必要となる。

また，スピーキング能力を測ろうとすればそれを採点する採点者（rater）の確保，公正平等な採点のための採点者の訓練（rater training）が必須となる。近年，PC 等の機械を使ったテストにおいてはその設備の確保はいうまでもない。英語の 4 技能を測る民間テストの活用がいったんは決まったのに 2019 年に見送りとなったのは，公平な採点や地域格差・経済格差などの実用性の問題が関係している。

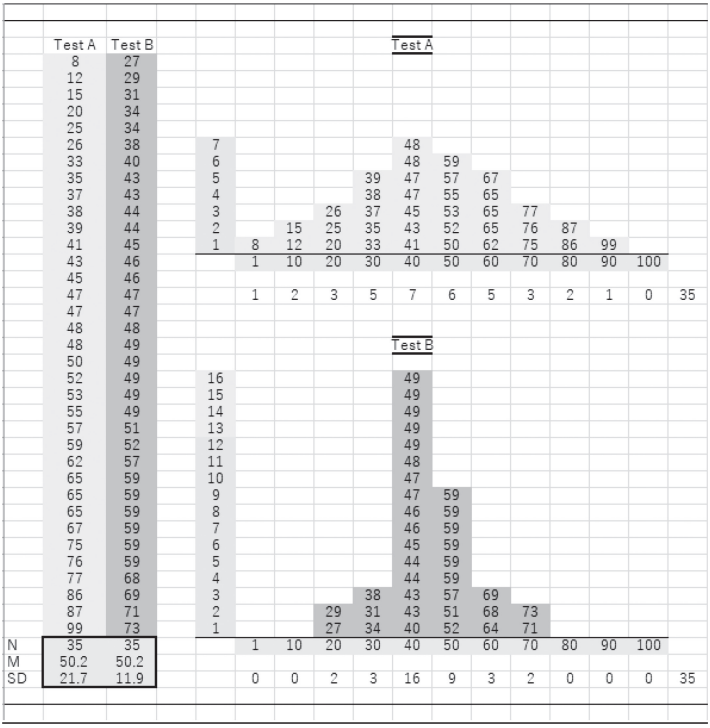
6. テスト得点の解釈

テストの結果が出たときに教師も生徒もともにもっとも関心を示すのが平均点（mean）である。しかし，2つのテストの得点の分布（distribution）は平均点だけでは比較できない。ここでは分布の1つの指標である標準偏差（standard deviation; SD）について解説する。図4の左側の数値は，生徒35名のテストAおよびテストBの得点である。2つのテストの平均点は50.2と同一である。そして，図4の右側の柱状グラフはその得点を得た人数（頻度）を10点刻みで表したものである[4)]。テストAは平均点を中心に受験者が万遍なく分散しているが，テストBは平均点周辺に集中している。つまり，テストAの方がテストBよりも得点のばらつきが大きいことがわかる。このばらつきの違いを表す数値が標準偏差である。テストAの標準偏差は21.7，テストBの標準偏差は11.9で，両者には2倍ほどの差がある。

偏差とは個々の得点と平均点の距離を意味する。そして，標準偏差とはそれらの平均をとったものである。標準偏差の算出式は以下の通りである。

$$\text{標準偏差} = \sqrt{\frac{\Sigma(\bar{x}-x)^2}{n}}$$

ただし，xは受験者個々の得点，$\bar{x}$は平均点，nは受験者数を表す。個々の得点と平均点との距離は正（+）の場合と負（−）の場合があり，両者が相殺されて0にならないように，2乗して正の数にしている。なお，標準偏差を2乗したものを分散（variance）という。



	Test A	Test B
	8	27
	12	29
	15	31
	20	34
	25	34
	26	38
	33	40
	35	43
	37	43
	38	44
	39	44
	41	45
	43	46
	45	46
	47	47
	47	47
	48	48
	48	49
	50	49
	52	49
	53	49
	55	49
	57	51
	59	52
	62	57
	65	59
	65	59
	65	59
	67	59
	75	59
	76	59
	77	68
	86	69
	87	71
	99	73
N	35	35
M	50.2	50.2
SD	21.7	11.9

Test A

7						48							
6						48	59						
5					39	47	57	67					
4					38	47	55	65					
3				26	37	45	53	65	77				
2			15	25	35	43	52	65	76	87			
1	8	12	20	33	41	50	62	75	86	99			
	1	10	20	30	40	50	60	70	80	90	100		
	1	2	3	5	7	6	5	3	2	1	0	35	

Test B

16					49								
15					49								
14					49								
13					49								
12					49								
11					48								
10					47								
9					47	59							
8					46	59							
7					46	59							
6					45	59							
5					44	59							
4					44	59							
3				38	43	57	69						
2			29	31	43	51	68	73					
1			27	34	40	52	64	71					
	1	10	20	30	40	50	60	70	80	90	100		
	0	0	2	3	16	9	3	2	0	0	0	35	

図 4：分布

7. テスト改善

7. 1　信頼性の向上

項目数が増えると信頼性は上がる。表 2 にしたように，Pre-TOEFL が 95 項目に対して，TOEFL では 140 項目と 45 項目も増えている。そして，項目数の増加にともない，標準測定誤差が 16.4 から 13.7 へと減少し，信頼性係数は 0.89 から 0.95 に向上している。

次に，増やした項目数に対して信頼性係数がどれだけ上がるかを予測するスピアマン・ブラウンの予測式を示す。

$$\text{予測される信頼性係数} = \frac{nr}{1 + (n - 1)\, r}$$

ただし，n は項目数を何倍に増やしたかを示す倍率であり，r は項目を増やす前の信頼性係数である。

例えば，20 項目のテストに 5 項目を追加した場合，n は（20 + 5）÷ 20 であり 1.25 倍に増やしたことになる（Henning, 1987, p. 85）。

前述した折半法では，一旦折半して半減した項目数から折半する前の項目数に戻して（すなわち 2 倍して）元の信頼性係数を予測する。例えば折半したテストセット間の信頼性係数（r）が 0.7 であるとき，折半する前の（つまり，項目数が 2 倍の）信頼性係数は，（2 × 0.7）÷（1 + 0.7）= 0.823 と予測できる。よって，この公式からも項目数が大きくなれば，信頼性係数が大きくなることがわかる。

7. 2　重みづけ

前節でみたように，項目数を増やせば信頼性は上がる。そこで，どの項目を増やせばいいのかという問題については，重みづけ（weighting）を考えてみる必要がある。

Ebel（1979）は，「ある到達度テストが 2 つの領域をカバーする場合に，一方がもう一方よりも 2 倍重要であれば，より重要な領域の項目を 2 倍にすべきである」と指摘している（cited in Alderson et al., 1995, p. 149）。つまり，領域の重要性に応じて 1 問を重く配点する傾斜配点を避けて，領域の重要度に応じて項目数を調整する必要がある。

表 3 に示すように，テスト 1 では，領域 A に 10 項目，領域 B に 20 項目出題している。一方，テスト 2 では，領域 A に 20 項目，領域 B に 10 項目出題している。いずれのテストも項目数合計は 30 である。領域 A に

表 3：重みづけ

	領域 A	領域 B	A ＋ B
重みづけ比	2	1	3
テスト 1			
項目数	10	20	30
配点	2 点× 10 項目 = 20 点	0.5 点× 20 項目 = 10 点	
テスト 2			
項目数	20	10	30
配点	1 点× 20 項目 = 20 点	1 点× 10 項目 = 10 点	

領域 B の 2 倍の重みづけをしたいわけだから，テスト 2 のように，すべての項目を 1 点とし，領域 A の項目数を 2 倍に増やす方がよい。

7. 3　項目分析

前節までは，テスト「全体」としてよいテストについて考察してきた。テスト全体がよいテストであるためには，その構成要素であるテスト項目一つひとつがよい項目でなければならない。本節では，項目困難度（item difficulty）と項目弁別度（item discriminability）を分析することにより，各テスト「項目」の良否について考察する。項目分析（item analysis）の目的は，よくない項目を特定して，それを修正すること，またはよい項目と置き換えることである。

表 4 は 10 項目からなるテストを 8 名の生徒が受けた結果を示したものである。1 は正解を 0 は不正解を意味する。項目困難度は，換言すれば正答率であり，全受験者のうち何名が正解したかを示す。項目困難度が高ければ易しい項目となり，低ければ難しい項目となる。理想的な項目困難度は 0.5 といわれているが, 現実的には 0 から 1 を 3 等分した中央の部分（0.33 ～ 0.67）が適正範囲と考えるべきである（Henning, 1987, pp. 49–50）[5)]。

項目弁別度とは，テスト項目が学力のある者と学力のない者を区別する能力をいう。問題が易しすぎると学力がある者も学力のない者も正解して

しまい，両者が区別できなくなる。また，問題が難しすぎると両者とも正解できなくなり，やはり，区別できなくなる。したがって，項目困難度が0.5 に近づくにつれて項目弁別度は上がる。項目弁別度は，表 4 では，標本分離（sample separation）と点双列相関係数（point-biserial correlation coefficient）の 2 つの指標で示されている。

標本分離は，上位群の項目困難度から下位群の項目困難度を引いた数値である。上位群は下位群よりも正答率が高いという前提に基づいている。上位群・下位群の配分の最適値には 28% が用いられ，標本分離による項目弁別度の受け入れられる最低値は 0.67 と設定される（Henning, 1987, pp. 51–52）[6]。しかしながら，この方法では中位群のデータが考慮されていない。そこでその代替となる指標が点双列相関係数である。これは，各項目（1, 0 データ）と合計点（連続データ）の相関係数である。ある項目が正解であればその受験者はテスト全体の得点も高く，正解でなければテスト全体の得点も低いことを前提にしている。Henning（1987）は適正値を 0.25 以上と設定している（p. 53）[7]。

表 4 における項目困難度および標本分離と点双列相関係数の 2 つの項目弁別度の数値をみると，このテストの項目 1，5，7 はよくない（poor/weak）項目と判断される。

表 4 が示すように，項目困難度や項目弁別度の低い項目（1，5，7）がそれぞれ削除されると，テストの信頼性係数は，削除前の信頼性係数（0.365）より上がることがわかる。例えば，項目 1 を削除した場合は 0.464 に，項目 5 を削除した場合は 0.516 に，項目 7 を削除した場合は 0.462 に信頼性係数は向上する。一般的には，項目数が減少すると信頼性係数は下がるが，このようによくない項目を削除すると，（たとえ項目数が減少したとしても）信頼性係数は上がる。

表 4：項目分析

受験者 / 項目	1	2	3	4	5	6	7	8	9	10	合計
5	1	1	1	1	1	1	0	1	1	1	9
1	1	1	1	1	0	1	0	1	1	0	7
2	0	1	1	1	0	1	0	0	1	1	6
7	1	1	0	0	1	1	0	1	0	1	6
10	1	1	1	0	0	1	0	0	1	1	6
6	1	0	0	1	1	0	1	0	1	0	5
3	0	0	0	1	1	1	0	0	0	1	4
9	1	1	0	0	0	0	0	0	1	1	4
4	1	0	0	0	1	1	0	0	0	0	3
8	1	1	0	0	0	0	0	0	1	0	3
平均	0.80	0.70	0.40	0.50	0.50	0.70	0.10	0.30	0.70	0.60	5.56
正解数	8	7	4	5	5	7	1	3	7	6	
不正解数	2	3	6	5	5	3	9	7	3	4	
項目困難度	0.80	0.70	0.40	0.50	0.50	0.70	0.10	0.30	0.70	0.60	
適正値	×	×				×	×	×	×		
上位群正解数	2	3	3	3	1	3	0	2	3	2	
下位群正解数	3	2	0	0	1	1	0	0	2	1	
差	−1	1	3	3	0	2	0	2	1	1	
標本分離	−0.33	0.33	1.00	1.00	0.00	0.67	0.00	0.67	0.33	0.33	
適正値	×				×		×				
点双列相関係数	0.084	0.475	0.775	0.502	0.056	0.475	−0.056	0.743	0.353	0.365	
適正値	×				×		×				
不良項目	×				×		×				
項目削除後	0.464	0.332	0.140	0.325	0.516	0.332	0.462	0.174	0.388	0.392	

項目削除前の信頼性係数 $\alpha = 0.395$

7. 4　テスト項目の改善

前述したように，テストを実施した後で個々のテスト項目の良否を判断するのが項目分析であるが，テストを実施する前に各項目に作成上の問題がないかチェックしておく必要がある[8)]。多肢選択肢式（multiple-choice）テストの作成上のガイドラインとして，次のチェック項目が挙げられる（Brown, 1996, p. 51）。

1. 正解は１つしかないか。
2. 問題文にヒントが隠れていないか。
3. 項目が１つのことをテストしているか。
4. 各々の選択肢を問題文（stem）に入れた場合に文法的に正しいか。
5. 各々の選択肢を問題文に入れた場合に意味的に正しいか。
6. 選択肢の長さはほぼ同じか。
7. 選択肢は概ね同じ難しさか。
8. 選択肢はすべて同じ分野・領域に関連しているか。

次に，多肢選択式の文法および語彙テストの具体例を挙げて，その問題点を指摘し，改善策を示すことにする。

(1) I was very pleased to hear that.

a. glad　　b. sad　　c. happy　　d. frightened

この項目には，正解が２つある（a. glad, c. happy）という問題が指摘される。これはガイドライン１に抵触する。a. c. のいずれかを，例えば depressed に置き換えればよい。

(2) Though his family was poor, Charles became extremely prosperous in later life.

a. lonely　　b. ill　　c. rich　　d. healthy

これは prosperous の意味を問う問題であるが，問題文に Though his family was poor というヒントが含まれているのでガイドライン２に抵触する。このヒント（Though）から poor の反対語は rich であるとわかる。したがって，このヒントとなる従属節を削除すればよい。

(3) After dinner we took a stroll in the park.

a. slow walk　　b. brief sleep　　c. light lunch　　d. short pause

ここでは問題文にある take（took）と共起（collocate）しない語が選択

肢に含まれていることが問題となる。British National Corpus（BNC）では have a sleep は確認できるが，take a sleep は見当たらない。したがって，この問題はガイドライン 4 に抵触することになる。ちなみに，lunch が take をとるよう例が 11 件みられるが，その中に take a lunch という例はない。さらに，take a pause という用例は 3 例しかみられない。そこで，選択肢の単語をすべて take と共起する単語（例えば take a brief nap, take a light meal, take a short break）に置き換えるという改善策が考えられる。

（4）The man in the portrait is a direct ancestor of mine.

a. break　　b. cost　　c. energy　　d. picture

a. break, b. cost, c. energy はいずれも問題文において man と意味的に共起しない。したがって，ガイドライン 5 に抵触する。例えば a. book, b. movie, c. paper と置き換えればいずれも共起する。

（5）Someone who designs houses is a .....

a. designer　　b. builder　　c. architect　　d. plumber

母音で始まる単語の前につける不定冠詞は，a ではなくて an である。 c. architect を問題文に入れると a architect となり，文法的に正しくなくなりガイドライン 4 に抵触することになる。問題文から不定冠詞 a を削除し，すべての選択肢に不定冠詞をつけて，a. a designer, b. a builder, c. an architect, d. a plumber とする。

（6）John soon returned to .....

a. work　　b. the prison　　c. home　　d. school

c. home を問題文に入れると returned to home となる。home は副詞であるので前置詞 to は不要である。これもガイドライン 4 に抵触する。例えば home に his をつけて his home と名詞句にすればよい。

（7）The officer dismissed the men after his brief speech.

a. sent away　b. rejected　c. called back　d. complained

問題文にある dismiss は他動詞で直接目的語をとるのに対して，選択肢の d. complained は自動詞で直接目的語をとらない。よって，ガイドライン 4 に抵触する。d. complained に about をつけて，complained about とする。さらに，a. sent away から away を削除して sent とし，1 つの単語の選択肢の数と 2 つの単語の選択肢の数を 2 つずつと同じにする。

(8) I never knew where .....

a. had Tom gone　b. Tom had gone　c. has Tom gone　d. Tom has gone

このテスト項目は語順と時制の両方をテストしているのでガイドライン 3 に抵触する。この項目は生徒が語順がわからなくて間違えたのか，それとも時制がわからなくて間違えたのかについての情報を提供しない。a. Tome goes, c. Tome will have gone に変えて a. Tom goes, b. Tom had gone, c. Tom will have gone, d. Tom has gone とすれば時制を問う問題になる。また，a. had Tom gone, b. Tom had gone, c. had gone Tom, d. Tom gone had とすれば，過去完了形の語順を問う問題となる。

(9) John tried to think deeply about the problem.

a. vex　b. poll　c. tug　d. contemplate

一目見て d. contemplate という単語だけが他の単語よりも長いので，ガイドライン 6 に抵触する。これを例えば ponder に置き換えれば他の単語とほぼ同じ長さになる。

(10) After five years, the old house was still vacant.

a. hideous　b. empty　c. shabby　d. sturdy

b. empty だけが頻度が高く他の選択肢より易しくなるのでガイドライン 7 に抵触する。例えば b. empty を uninhabited という低頻度の単語に置き換えれば他の単語との難しさはほぼ同じになる。また，他の単語を例えば a. broken, c. dirty, d. cheap と高い頻度の語に置き換えても，難しさは他

の単語とほぼ同じになる。

（11）John says he detests serious music.

a. doubts　　b. prefers　　c. hates　　d. tries

a. doubts, b. prefers, c. hates が感情を表す語であるのに対して try だけが感情語ではないので，ガイドライン 8 に抵触する。選択肢 d. tries を感情を表す単語，例えば fears に置き換えればよい。

8. まとめ

本稿では英語教師として備えておくべき言語テストに関する基礎知識について概説した。テストは生徒の将来の進路を決定するための貴重な判断材料となる。したがって，テストは「よい」テストでなければならない。よいテストはそれが測定したいと思っている能力を本当に測定しているという妥当性，繰り返し実施しても同じ結果が導かれるという信頼性，実際にテストが実施できるという実用性の 3 条件を満たしていなければならない。

これらの 3 条件に加えて，さらに，テスト結果を分析する際に不可欠な統計の考え方（標準偏差や相関係数），個々の項目の良否を分析する方法（項目分析），テスト改善のための多肢選択式の問題を作る際の留意事項も併せて示した。

注

1）本稿の第 1 節～第 4 節は，日本言語テスト学会（JLTA）Web Tutorial「「よい」テストの条件：妥当性，信頼性，実用性」（jllta2016.sakura.ne.jp/tutorial/1）%10intro/）をベースに，第 7 節は日本言語テスト学会（JLTA）第 46 回研究例会 / 第 4 回中部地区英語教育学会（CELES）近畿地区研究会ワークショップ「多肢選択式語彙・文法テスト作成上の留意点」（2017 年 10 月 28 日　桃山学院大学にて開催）をベースに加筆修正したものである。

2）TOEFL（Test of English as a Foreign Language）が主にアメリカおよびカナダ等での大学に留学を希望する，非英語母語話者のための熟達度テストであるのに対して，IELTS（International English Language Testing Systems）は主にイギリスおよびオーストラリア等での大学に留学を希望する，非英語母語話者のための熟達度テストである。

3）「到達度」テストが，出題範囲が特定のコースやシラバスに基づいているのに対して，「熟達度」テストは，出題範囲が特定のコースやシラバスに基づいていない。

4）この柱状グラフの個々の頻度の頂点をつないだ線が，平均点を中心に左右対称であれば正規分布曲線（normal distribution curb）という。その曲線は鐘に似ているので bell curb ともいわれる。

5）Heaton（1988）は，「多くのテスト作成者は困難度が 0.4 ～ 0.6 の項目を目指すが，実際は 0.3 ～ 0.7 を受け入れる用意がある」と述べている（p. 179）。また，Brown（1996）も同様に「0.3 ～ 0.7 が普通に受け入れられる」と考えている（p. 70）。

6）Ebel（1979）は，0.4 以上をとてもよい項目としている（Brown, 1996, p. 70）。

7）Hughes（1989）は，適正値を 0.3 以上と設定している（p. 160）。

8）Brown（1996）はこれを項目形式分析（item format analysis）とよんでいる。

引用文献

Alderson, C., Clapham, C. & Wall, D.（1995）. *Language test construction and evaluation*. Cambridge University Press.

Brown, J. D.（1996）. *Testing in language programs*. Prentice Hall Regents.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T.（1999）. *Dictionary of language testing*. Cambridge University Press.

Harris, D. P.（1969）. *Testing English as a second language*. McGraw-Hill Book Company.

Heaton, J. B.（1988）. *Writing English language tests*. Longman.

Henning, G.（1987）. *A guide to language testing: Development・evaluation・research*. Newbury House.

Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.

Richards, J. C., & Schmidt, R. (2010). *Dictionary of language teaching and applied linguistics. Fourth Edition*. Longman.

島田勝正（2000）.「「学習」中心の英語授業」『英語教育学論集―青木昭六先生古希記念論文集』 開隆堂　43-52 頁.

英語教師のための言語テストに関する基礎知識

Basic Knowledge in Language Testing for Teachers of English

SHIMADA Katsumasa

A schoolteacher's role involves not only teaching but also testing. This study outlines the basic knowledge in language testing that English teachers should possess. Tests provide students with valuable information for determining their future careers. Therefore, any "good" test must fulfill three conditions: validity, reliability, and practicality.

Validity refers to the degree to which a test measures what it is supposed to measure, whereas reliability indicates the stability of test scores; it measures the degree to which a test provides consistent results. A test is reliable if it provides the same results when administered to the same candidates on different occasions. Practicality shows whether the test is feasible.

In addition to these three conditions, we will introduce certain statistical concepts, including standard deviation and correlation, essential for analyzing test scores and item analysis to focus on the quality of individual items in terms of item difficulty and discrimination and discuss how to write a satisfactory test item for multiple-choice questions.