Flucher, G. (2010). *Practical Language Testing*. London: Hodder Education.

（担当：M.H.）

Chapter 2

7. Testing the test

■ Yerkes テストの作成者はテストを分析するために 2 つの方法を用いた (pp.45, Table2.2 参照)。

① 項目難易度：項目の正答数を産出すること。今日では、facility index と呼ばれる。

② グループ間比較：異なるグループの回答を比較すること。

8. Introducing reliability

■ 標準化テストで最も大切な要素が信頼性 (reliability)である。信頼性とは、テストを複数回実施したとき、その間に教育的介入がなければ、ほぼ同じスコアが得られるかという概念である。

■ 信頼性を損なう要因として、Lado (1960) は以下の 3 つをあげている。

(a) テストが実施される環境に関するもの (e.g., 時間や場所)

(b) テスト自体に関するもの (e.g., 問題の質: テスト項目は同質性を有する必要がある)

(c) 評価や採点に関するもの (e.g., ライティングやスピーキングでは採点結果が分散しやすい)

■ テストスコアは、

$$X = T + E$$

の式で表される (X=観測値、T=真値、E=誤差)。つまり、観測値とは、本来の能力値に誤差が加わったものとして理解できる。

9. Calculating reliability

以下では Lado が提案した信頼性の変動要因 3 つを中心に、信頼性の測定法を述べる。

(1) **Test administration**

■ 異なる場所や時間で実施されたテストのスコアが、一貫しているか否かを検証するには、ピアソンの積率相関によって、相関係数を産出する (see pp.48–49, Fig. 2.6. & Table2.3)。

$$Rxy = \frac{\sum xy}{N(sdx)(sdy)}$$

N is the number of scores

■ 相関性の強さは、-1 (スコアが完全な反比例関係)から 1 (完全に同じスコア)で表される。

■ それぞれのテストスコアがどのくらい重複しているのかを知るためには、相関係数を二乗する (see Fig. 2.7.)。

(2) **The test itself**

■ 信頼性の高いテストの項目は、同質的 (homogeneous)である必要がある。言い換えれば、テスト項目どうしが高い相関性を有していることが求められる。

■ テスト項目の相関性推移法として以下の 2 つをあげる。

① 折半法 (split-half method)

・1 つのテストを半分に分け、それぞれの合計点の相関をスピアマン・ブラウン公式によって産出する。

・分け方の例：問題 1 を A, 2 を B, 3 を A, 4 を B, 5 を A…

・公式

$$R = \frac{2rhh}{1+rhh}$$

rhh is the correction between two halves of the test

② クロンバックアルファ (Cronbach's alpha)

・全ての折半方法によって推定した信頼性の平均値を統計的に産出する手法。

・相関係数の産出において最も良く用いられる。

・公式

$$R = \frac{k}{k-1}\left\{1 - \frac{\sum pq}{s^2}\right\}$$

k is the number of items, s^2 is the test score variance, $\sum pq$ is the sum of the variances of individual items

(3) Marking or Rating

■ 評価の厳しさには個人差があり、特定のテストにおいて評価が高くなる採点者も存在するので、評価や採点は非常に複雑なものである。

■ 評価者間の信頼性を出す手法として、クロンバックアルファが用いられる。しかし、評価者は通常部分点 (e.g., 1~6 点)によって採点するので、テスト項目の場合の異なる公式を用いる。

$$\alpha = \frac{k}{k-1}\left\{1 - \frac{S_{r1}^2 + S_{r2}^2}{S_{r1+r2}^2}\right\}$$

k is the number of raters, S^2 is the variance of their scores, r1 and r2 stands for rater 1 and rater 2

10. Living with uncertainty

■ 標準化テストの最も重要なツールの 1 つが､測定の標準誤差 (standard error of measurement)である。

■ 標準誤差は観測値が真値からどのくらい離れているかを示すものなので、より実用的な示唆を与える。

$$Se = sd\sqrt{1-R}$$

■ 実際に (1)linguality test と (2) writing test の標準誤差を用いて、観測値の信頼区間を計算し、真値の位置を推定する (母集団が正規分布していると過程)。

(1) linguality test

・標準誤差 = 1.7

・95%信頼区間 = 1.7 × 1.96 (平均からの標準偏差) = 3.33

よって、95%の確率で真値は観測値の±3.33 点以内にあると推定できる。
(e.g., 14 点をとった場合、真値は 11 < 14 < 17)

(2) writing test
・標準誤差 = 1.15
・99%信頼区間 = 1.15 × 2.58 = 2.97
よって、99%の確率で真値は観測値の±2.97 点以内にあると推定できる。
(e.g., 6 点をとった場合、真値は 3 < 6 < 9)
■ この情報によって、テストが本来意図する目的のために、スコアを利用できるか否かを決定できる。そして、われわれはテスト作成者に常に標準誤差の情報を求めるべきである。

11. Reliability and test length
■ 信頼性は、項目数 (test length)によっても影響される。一般に項目数が多いほど、信頼性は高まる。
■ そのためには、特定の項目の回答は、他の項目から独立したものでなければならない (i.e., 受験者が A の問題に正解したのは、B の問題に正解できたからだ、というようなことがあってはならない)。
■ Lado (1961)は項目数と信頼性の関係を推定する公式を提案した。これによって、信頼係数を高めたいときに、どのくらい項目数を増やすべきなのかを知ることができる。

$$A = \frac{r_{AA}(1 - r_{11})}{r_{11}(1 - r_{AA})}$$

A is the proportion by which you would have to lengthen the test to get the desired reliability,
r_{AA} is the desired reliability and r_{11} is the reliability of the current test

■ だが、テストの時間的制約から、信頼性を高めるために項目数を増やすことは、現実的ではない。
→ 信頼性を向上させる最善の方法は、項目の質を高めることである。

12. Relationships with other measures
■ 同じ構成概念の、異なる 2 つの尺度による結果の相関性が高いことは、収束的妥当性 (convergent validity)の証しとなる (i.e., 既に妥当性の高さが確認されているテストとの相関性が高ければ、そのテストの妥当性も確認できる)。
■ Yerkes (1921)のグループテストは、難易度が高すぎたため最下位層を測定できなかった。そこで、妥当性に勝る個人テストとの相関性が測定された。
■ 結果、十分な相関性 (r = .79)が得られたので、個人テストが利用不可能なときは、グループテストを利用できることが示された。

13. Measurement
■ この章で議論した測定理論は、標準テストがどのように作られるのか、そしてどのように機能するかを理解するために、決定的に重要である。
■ さらに、言語テストとその評価の基礎も、測定理論に端を発しているということを留意しなければならない。

14. ディスカッションの内容

14.1 観測値、真値、誤差の関係

■ 英語の学力を例に取ると、

(1) 観測値：学力テスト等によって得られて得点

(2) 真値：英語の学力の本当の値。直接測定することは不可能なので、テストなどの媒体を利用して推定を試みる。

(3) 誤差：テストを受けたときのコンディションの違いなどから生じる、真値からのずれ。

と定義することができる。そして、古典的テスト理論 (classical test theory) では 3 者の関係は以下の式で表すことができるとされる。

$$観測値 = 真値 + 誤差$$

■ さらに、この式の基底には以下の仮定がある。

(1) 誤差はランダムに生じる (i.e., 誤差と真値との間に相関はない)

(2) 誤差はプラス、マイナスどちらにも発生し、その平均値はゼロである

(3) 誤差どうしの相関はゼロである。

■ さらに真値と誤差の間に相関がないという仮定から、観測値の分散は、真値の分散と誤差の分散の和によって表すことができる。

$$観測値の分散 = 真値の分散 + 誤差の分散$$

■ 古典的テスト理論における信頼性係数 (ρ)は、この式によって定義される。

$$信頼性係数\ (\rho) = \frac{真値の分散}{観測値の分散} = \frac{真値の分散}{真値の分散 + 誤差の分散}$$

つまり、誤差の分散が大きくなるほど、分母が信頼性係数は低くなる。

14.2 項目の同質性 (homogeneousity)とは？

■ 特定の能力を測定するテストにおいて、個々の項目が同一の能力や知識を測っている程度のこと。

■ 項目の同質性が満たされていないのならば、個々の項目が測定する能力は一貫せず、そのテストの信頼性は損なわれる。

■ 言い換えれば、信頼性の高いテストでは、それぞれの項目どうしの相関性が高くなければならない。これを確かめるために、折半法やクロンバック α などによって、テストの内的一貫性を推測する。

15. 参考文献

平井明代. (2010). 『テスト問題・教材再利用のすすめ：TEASY 理論編』, 東京：丸善プラネット.

三浦省五. (2004). 『英語教師のための教育データ分析入門 授業が変わる テスト・評価・研究』, 東京：大修館書店.