第 2 章　英語学力測定論　pp. 47-58

5. 信頼性

5.1　考え方

・信頼性(reliability)とは、**測定の一貫性(consistency)**あるいは**安定性(stability)**の**程度**である。

・テスト得点は、測定したい能力以外にもさまざまな要因によって影響を受け、変動する。

- 受験者の体調
- テスト実施手順の違い
- 時間の経過による受験者の能力の変化
- テストの版の違い
- 評定者の違い

などが**測定誤差(measurement error)**の原因となる。

→信頼あるテスト得点とは、測定したい能力を最大限反映し、測定誤差を極力含まない得点。

→測定誤差の原因を完全に除去することはできない。

・信頼性検証の理論として、**古典的テスト理論**（測定誤差を一括して扱う）と**一般化可能性理論**（誤差を要因に分けて扱う）がある。

5.2　古典的テスト理論に基づく信頼性推定

5.2.1.　信頼性推定のモデル

・古典的テスト理論では、測定誤差は一括してランダムに発生すると仮定する。

・テストの観測得点(observed score)= x、真の得点=x_t、誤差得点=x_eとすると、

$$\boldsymbol{x = x_t + x_e}$$

テスト観測得点の分散=σ_x^2、真の得点の分散=σ_t^2、ランダムな誤差分散=σ_e^2とすると、

$$\boldsymbol{\sigma_x^2 = \sigma_t^2 + \sigma_e^2}$$

の式が成り立つ。

・観測得点の分散に対して、真の得点の分散が大きければ大きいほど、また誤差分散が小さければ小さいほど、そのテストの観測得点はより信頼性が高いといえる。

・理論上の信頼性($r_{xx'}$)は、観測得点の分散にしめる真の得点の分散の割合、$\frac{\sigma_t^2}{(\sigma_t^2-\sigma_e^2)}$であると定義され、信頼性係数は $0 \leqq r \leqq 1$ の区間で示される。

→真の得点の分散や誤差分散の大きさは実際にはわからない。

5.2.2.　信頼性の推定方法

・古典的テスト理論に基づく信頼性の推定方法には、**(1)内部一貫性に基づく方法**と、**(2)安定性に基づく方法**があり、信頼性の推定は以下の3段階で行われる。

1. 測定誤差の原因を特定
2. 独立で平行な 2 つの得点を収集するための研究計画を立てる
3. 2 つの得点の適切な相関係数、またはテスト得点の分散に基づくα係数を算出する

(1) 内部一貫性の信頼性推定

・内部一貫性(internal consistency)の信頼性推定の方法として、**折半法(split-half method)**と**項目分散(item variance)を用いる方法**がある。

1) 折半法による信頼性推定

・テスト得点を 2 分割し、1 人の受験者に得られる 2 つの得点の相関係数を求め、信頼性の推定値とする方法。

・分割方法として、偶数番号と奇数番号の項目に分ける方法が一般的だが、内容や測定している能力の点から 2 分割するなど、何らかの基準を設けて 2 分割する場合もある。

・一般的には長いテスト(項目数が多い)ほうが短いテストよりも信頼性が高くなる。

・折半法ではテストを半分にしたので、この長さの短縮のため、得られた相関を修正する必要がある。このために用いられる公式が**スピアマン・ブラウンの修正公式**である。

$$\text{信頼性係数}=2\frac{\text{相関係数}}{1+\text{相関係数}}$$

2) 項目分散による信頼性推定

・個々の項目の分散に基づいて信頼性係数を計算する方法で、各項目の分散と、全体得点の分散を用いて内部一貫性の信頼性推定値を計算する。

・間隔尺度以上で部分点採点の場合、**α係数(coefficient alpha)**が用いられる。

$$\alpha=\frac{\text{項目数}(n)}{\text{項目数}-1}\times\frac{1-\text{項目分散の和}}{\text{合計点の分散}}$$

・正誤で採点される 2 値データの場合、**KR20 (Kuder-Richardson formula 20)**、**KR21** を用いる。

・項目数を増やしてテストを長くすると、ほかの条件が同じであれば信頼性は高くなる。

・テストを長くすることでどの程度の信頼性が得られるか推定する方法が**スピアマン・ブラウンの予想公式 (Spearman-Brown prophecy formula)**である。

・k=現在の項目数の倍数、$r_{tt'}$=望ましい信頼性の水準、$r_{xx'}$=現在の信頼性の水準としたとき、以下の公式で求められる。

$$k=\frac{r_{tt'}(1-r_{xx'})}{r_{xx'}(1-r_{tt'})}$$

(2) 安定性の信頼性推定

1) **再テスト信頼性推定値 (test-retest reliability estimates)**

・同じ受験者集団に同じテストを 2 回実施し、それらの相関を求めることで信頼性を推定する方法。

・2 つのテスト得点を平行測定として扱い、時間経過後のテスト得点の安定性を推定する。

・同じテストを 2 回実施するので誤差の原因は受験者にあると考えられる。

再テスト法の問題点

1. 練習効果(practice effect)が考えられること、
2. 2 回目のテスト実施までに受験者の能力値水準が変化すること、
3. 1 回目、2 回目ともにランダムな誤差が生じること

2) **等化性(平行測定)信頼性推定値(equivalence reliability estimates)**

・安定性のない測定はテストの版(forms)の違いから起こることもあり、テストの版の**等価性(equivalence)**を検討す

る必要がある。

・測定される能力と項目の内容や形式が等しいと考えられる 2 つのテストの版を作製し、同一集団に実施した後、2 つのテスト得点の相関係数を算出し、それを**等化性信頼性推定値**とする。

・順序効果の可能性を最小化するために、**釣合型計画(counterbalanced design)**を用い、集団ごとに受けるテストの版の順序を変えることが望ましい。

3） 評定者の一貫性による信頼性推定値

・一人の評定者が一貫性のない採点をする場合、複数の評定者間で最低の不一致が生ずる場合が考えられる。

評定者間信頼性(inter-rater reliability):

評定者が 2 名の場合、2 つの評定の相関係数、または α 係数を求め、信頼性の推定値とする。3 名以上の場合、評定者の評定を合計し、α 係数を求めることによって信頼性の推定値を求める。

評定者内信頼性(intra-rater reliability):

1 人の評定者に時期をあけて、2 度採点してもらい、同じ受験者について 2 通りの評定を得、評定群の相関係数、または α 係数を算出し、信頼性の推定と解釈する

5.2.3. 測定の標準誤差と信頼区間

・信頼性係数は、個人のテスト得点の正確さに関する情報は提供しない。

・個人のテスト得点の信頼性に関する情報を得るためには、**測定の標準誤差(standard error of measurement: SEM)**を求める必要がある。S_xは観測得点の標準偏差、$r_{xx'}$は信頼性係数である。

$$\mathbf{SEM} = S_x \sqrt{1 - r_{xx'}}$$

5.3 一般化可能性理論

5.3.1. 考え方

・古典的テスト理論では誤差成分が一まとめに扱われ、全ての誤差がランダムに発生すると考える。
→ランダムな測定誤差と系統的に発生する測定誤差とを区別することができない。

・**一般化可能性理論(generalizability theory: G-theory)**によって、テスト得点に影響を及ぼす様々な変動要因の大きさを推定することができる。

5.3.2. 一般化可能性研究（G研究）

・分散分析モデルを用いて、得点の変動がどのような要因によってどの程度発生しているのか検討する。

・例えば、ライティング能力を測定するために受験者にタスクを与え、複数の評定者が採点する場合では、測定対象は受験者個人のライティング能力であり、タスクと評定者が**相(facet)**となる。この場合、2 つの相があるため、**2 相計画(two-facet design)**と呼ばれる。

・考慮の対象となる全タスクと全評定者（母集団）を、それぞれ**許容観測母域(universe of admissible observations)**と呼ぶ。

・G 研究の目的は、様々な分散要因の相対的な割合を推定することで、推定値は分散成分(variance components)と呼ばれ、$\boldsymbol{\sigma^2}$で示される。

・テスト開発者は G 研究の第一段階として、テストに含める測定の相を特定する。

1)　　単相計画

・1つの相(facet)のみのテストの場合を単相計画(one-facet design)という。
・客観テストのようなケースでは変動する項目は受験者と出題項目だけであり、この場合の相は出題項目のみ。
・受験者がすべての項目を受ける場合、受験者と項目のデータが得られる。このことを**クロス計画(crossed design)**と呼ぶ。
・単相クロス計画では、3 つの変動要因、
　①測定の対象（受験者）p、②項目の相 i、③受験者と項目の交互作用 $p\times i$
　が考慮され、これ以外で得点に影響を与える成分は誤差の分散とする。
・誤差は交互作用の分散成分に含め、合計得点の分散σ_x^2はそれぞれ分散成分の和で説明される。
・分散成分の推定に加え、各項目の難易度の指標としてそれらの平均値が推定される。

2)　　2 相完全クロス計画

・ライティングのテストで、複数のタスクが受験者に与えられ、タスクに対する回答は複数の評定者によって採点され、また、全受験者がすべてのタスクに応答し、全回答が評定者全員によって採点されるような場合は、タスクと評定者の **2 相完全クロス計画(fully crossed design with two facets)**という。
　この場合、
　①受験者 p、②タスク t、③評定者 r、④受験者とタスクの交互作用 $p\times t$、
　⑤受験者と評定者の交互作用 $p\times r$、⑥評定者とタスクの交互作用 $r\times t$、
　⑦残差(residual)として受験者と評定者とタスクの交互作用と誤差 $p\times r\times t, e$
　を推定し、それらの合計得点の分散σ_x^2が、これらの 7 つの分散成分の合計で表される。
・項目相の平均値と評定者の相の平均値も得られ、それぞれ項目難易度および評定の厳しさの指標として解釈される。

3) 2 相枝分かれ計画

・すべての相がクロスされるのでなく、1 部の相が他の相と入れ子になっている場合、2 相枝分かれ計画(two-facets nested design)と呼ばれ、$p\times(i;t)$で表される。
・リーディングやリスニングのテストのように、1 つのテキストに複数の設問を伴う場合、設問（項目）がテキストに入れ子になっていると考え、$i{:}t$ で表す。
・2 相枝分かれ計画では
　①受験者 p、②テキスト t、③テキストの中の項目 $i{:}t$、④受験者とテキストの交互作用 $p\times t$、
　⑤残差（テキストの中の受験者と項目の交互作用と誤差）$p\times i:i, e$
　を推定し、合計得点の分散σ_x^2はこれら 5 つの分散成分の合計で説明される。
・各テキストの平均値が得られ、それらはテキストに含まれる項目によって測られたテキストの難易度の指標として解釈される。

5.3.3.　　決定研究（D研究）

・D 研究の目的は、G 研究で得られた分散成分の情報を用いて、測定誤差を最小化するような測定手続を計画することで、以下の情報が D 研究によって得られる。

① テスト得点に占める分散成分の相対的大きさの情報
② 観測得点が母得点の推定値にどの程度依存しているかという**信頼度(dependability)**の情報
・一般化可能性理論では合計得点の分散σ_x^2は、母得点分散σ_p^2と誤差分散σ_{error}^2の合計から構成される
・母得点分散は D 研究で推定されるので、信頼度推定のために、誤差分散を推定する。
・一般化可能性理論では測定誤差を、
① 相対的な測定誤差(relative measurement error):集団基準準拠テストに対応
② 絶対的な測定誤差(absolute measurement error):目標基準準拠テストに対応
の 2 つに区別する。
・相対的決定のための信頼度は**一般化可能性係数p^2(generalizability coefficient: G 係数)、**
・絶対的決定のための信頼度は**信頼度係数 ϕ (index of dependability: ϕ 係数)**と呼ばれる。
・一般可能性理論の分散成分や、信頼度推定値を算出するプログラムとして、**GENOVA** (Crick & Brennan, 1983)や **mGENOVA** (Brennan, 2001)などがある。

5.4 一般化可能性理論の利点と限界
・一般化可能性理論の古典的テスト理論に対する利点は、
① 複数の誤差要因の相対的な影響の大きさを推定することができる
② それぞれの測定誤差の要因の大きさを分散成分という形で推定することができる
③ 測定の各相における条件の数を調整することで、測定の信頼性を最適化できる
④ 相対的誤差と絶対的誤差とを区別し、NRT にも CRT にも対応可能な信頼度推定値を算出できる
・一方、一般化可能性理論の限界として、
① 2 つのテスト理論から得られる信頼性推定値は、特定の受験者集団に依存する
② どちらも測定誤差がどの能力値水準においても同じであることを前提としている
点が挙げられる。
→これらの問題点を克服するのが項目応答理論である。

6. 妥当性

6.1 考え方

・テストの得点を元に行われるこうした推測や決定は、テスト得点をそのように解釈し使用することの妥当性が担保されていることが前提となる。
・テスト得点の解釈や使用が妥当であることを保証するために様々な証拠を元に論理的説明をする必要があり、この過程を妥当性の**検討(validation)**という。
・妥当性は「測定していると主張する内容をどの程度測定しているか」に関する概念であるとされ、
① 内容妥当性(content validity)
② 基準関連妥当性(criterion-referenced validity)
③ 構成概念妥当性(construct validity)
の 3 種であるとされてきたが、現在では構成概念妥当性であらゆる妥当性を代表させるという考え方、妥当性は単一の概念(a unitary concept)であるという考えが主流となっている。
・Messick(1989)は妥当性を以下のように定義づけている。

「妥当性とは、テスト得点またはそれに類する他の評価法を下にして行う推論と行為の相応性ならびに適切性について、それを支持する経験的証拠と理論的理由づけの度合いを示す総合的な評価判断をいう」

・Millaer, Linn & Gronlund (2012: 72-73)は妥当性という用語を用いる注意点として以下の 4 点を挙げている

① 「テストの妥当性」は、より正確には「テスト結果をもとになされる解釈や得点使用の妥当性」

② 妥当性は程度問題であり、妥当性がまったくなかったり完全であったりということはない

③ 妥当性は、常に特定の受験者集団に対して、特定の黙亭での使用や解釈に適用されるもので、すべての目的に対して妥当であるような評価は存在しない

④ 妥当性は単一の概念であり、包括的な評価判断を含むものである

6.2 妥当性の検討

・テスト得点の特定の解釈が妥当であると考えられる説明的な論証を展開していく、論証に基づくアプローチ(argument-based approach)という考え方が、妥当性の検討の基本である。

・論証に基づくアプローチによる妥当性検討では、

① 解釈的論証(interpretive argument)…妥当性の論証を展開するための枠組みを提示

② 妥当性の論証(validity argument)…解釈的論証の全般的な評価

が行われる。

Kane(2006: 23-25)による論証に基づくアプローチ

・解釈的論証の段階では、①採点(scoring)、②一般化(generalization)、③外挿(extrapolation) 、④決定(decision)の4つの推論が与え、それぞれの推論に対し、p.57 のような仮説を考える。

・妥当性の論証の段階では、各推論における仮説が適切なものと判断できるか、適切な証拠を用いて評価する。採点の推論では、採点基準が適切か、採点の質管理が徹底して行われているか判断される。

・一般化の推論では、信頼性や一般化可能性研究、テストに含められる項目標本の代表制についての判断が求められる。

・外挿の推論では、テストが測る技能とコースで必要とされる技能の重なりの程度を判断したり、テスト得点とコースにおけるパフォーマンスの測定値（成績）との関係を実証分析したりする。

・決定の推論では、決定から得られた様々な種類の論拠には、各推論や支持する仮説に関連して、専門家の判断、実証研究、先行研究の結果、価値判断等が含まれる。

・Bachman(2005)や Bachman & Palmer(2010)は、kane(2002; 2006)の論証に基づく妥当性検討の枠組みを元に、言語テストの妥当性検討のために、**テスト使用に関する論証(Assessment Use Argument: AUA)**という考え方を展開した。

・AUA では、妥当性検討という用語の代わりに**正当化(justification)**を用いる。

・意図したテスト使用であることがどの程度正当化できるのか、という観点からテスト使用の正当化を行う。・

・AUA では各推論の段階で述べられる「主張(claim)」に対して、「理由付け(warrant)」と「反証(rebuttal)」が述べられる。

・それぞれに「裏づけ(backing)」となる証拠が示され、主張が支持されるか、却下されるかを検討していく。

【コメント】

Messick 以前、Messick 以後、で妥当性の考えかたに大きく違いが生まれたことは知っていたが、AUA の考え方についてこれまで知らなかった。また、古典的テスト理論についての知識は一通りあったものの、一般化可能性理論、項目応答理論に関する知識には穴が多すぎることを今回の発表を通じて自覚することができた。

一方、教育現場でテストの妥当性や信頼性の推定を行う際には、やはり深い専門的知識と労力が必要だということも再認識した。いかに実際にテストを行った際にそのテストの信頼性の推定を行い、妥当性の検証を行うか、そしてその方法論を現場の教師に広めることも重要なことだと感じた。テストの practicality と妥当性、信頼性の保持は反比例するようにこれまで感じてきたが、両者を保持し、学習者、教員どちらにも有益なテスト情報を与えるテストは何か考え続けなければならないと強く感じる。