Web テストにおける測定条件が信頼性に及ぼす影響†

＜あらまし＞　本研究では Web テストにおける測定条件，たとえば小問数，正答率，所要時間，入力方法などが信頼性に及ぼす影響について調査した．具体的には中学校数学の Web 教材に採点・送信機能を追加して各頁と他の頁の得点との相関を調べることにより小問数などの変化に応じて平均相関係数がどのように変わるかを調査した．本研究により，正答率と信頼性の関係など学説があるものについては実際のデータでの現れ方を確かめることができ，入力方式や１題当たり所要時間と信頼性の関係などについても一定の目安が得られた．

＜キーワード＞　教育測定，web 利用，テスト，信頼性，I-R 相関，データ解析

1. はじめに

テストの信頼性を調べるための指標としてクロンバックのα係数やSP分析におけるP注意係数が有名である．また，ある項目の得点と合計得点との相関（I-T 相関）や残りの項目の合計得点との相関（I-R 相関）を信頼性の指標とする方法もある．

この研究では Web 教材において不特定多数の読者が自分に必要な幾つかの項目だけに回答しているときに，I-R 相関の考え方を取り入れながら測定の信頼性を判断する方法を考えることによって，測定条件が信頼性に及ぼす影響を調べた．

2. I-R 相関と信頼性

I-T 相関（項目得点 item score と合計得点 total score の相関）が高ければその項目が測定している能力とテスト全体が測定している能力の整合性があると考えることができる．ただし，I-T 相関ではその項目自身の得点も合計に含まれるため，項目数が少ないときに実際以上に高くなる可能性がある．そこで，ある項目とその項目以外のすべての項目の合計得点との I-R 相関（項目得点 item score と残りの合計得点 remainder score の相関）を見るとその項目が測定している能力とテスト全体が測定している能力との整合性が分かり，その項目による測定の信頼性の指標とすることができる．

3. I-R 相関の代用

クロンバックのα係数や I-R 相関はすべての回答者がすべての項目に答えている場合には使い易いが，Web 教材のように不特定多数の読者が自分に必要な項目だけに回答しているようなデータには使いにくい．SP 分析を用いても同様である．

そこで，回答者が各々一部の問題だけに回答している答案から I-R 相関の趣旨に合う指標を求めるために次の方法を考える．

まず，同一人物が２つの項目の両方に回答している答案を抽出して２項目間の相関を求める．この作業をできる限り多くの組み合わせについて行う．具体的には回答数の多い 50 項目を選び ${}_{50}C_2$=1225 組を抽出して相関係数の計算を行う．

次に，各項目と他の項目との相関係数の平均値がその項目と残りの項目との整合性すなわち信頼性を表していると考える．

以上の作業から得られる相関係数の一覧表は２つの側面から見ることができる．一つは測定条件，たとえば小問数，正答率，所要時間，入力方法などが信頼性に影響を及ぼしているかどうかと見ることであり，もう一つは内容の相互関係がどうなっているかと見ることである．

この研究では，内容の相互関係を考える前提として測定条件と信頼性の関係について調査した結果をまとめた．

4. 調査の方法

筆者が公開している中学校数学の Web 教材について 2010 年 2 月 1 日から 2010 年 5 月 30 日までの期間に回答・返信のあった 20,722 件の答案から，学年が中学校１年生から３年生および卒業生であると答えた者のうち同一人物の答案を除いた 8,322 人の答案を抽出して集計した．

学年は自己申告であり他に確かめる方法はない．４月には秋以降に習う単元の問題に答えることができないことから分かるように，調査の時期によって各単元に回答する学年別人数構成は変化し，広い範囲に回答している者は「卒業生」と答えていることが多い．ただし，回答者を学年別に細分すると各項目のデータ数が少なくなり過ぎるので，この研究では回答者の学年別集計は行っていない．また，学年欄が無回答の答案は問題に対する解答も統計的に安定していないので集計に含めなかった．

各答案にはその頁（以下項目ともいう）の小項目の採点結果とともにそれ以前に行った頁の得点も自動的に記録されるものとした．実際には同一頁や同一小項目を何度も繰り返し行う者が多いが，同一頁についても同一小項目につ

いても第１回目の答案だけが記録されるようにした.

一般に行われているｅラーニングとは異なり，単なる Web ページに採点・返信機能を追加しただけのものであるので，厳密に個人を識別することは難しいが，IP アドレスの上位 16 ビットと回答パターンが一致するものは同一人物とみなした．学校のパソコン教室のように１教室分のパソコンが対外的には同一の IP アドレスとなって現れる場合，回答パターンが同じ答案はこの研究の集計には含まれない.

２つの項目の両方に回答している答案が 10 人未満のときはデータなしとした．また，無相関検定によって有意な相関があると認められる値のみを使用した.

5. 集計結果

縦横 50 項目（各項目は１頁に対応する）からなる表１のような一覧表が得られる.（表１には一部分のみ表示している．各項目の自分自身との相関は求めない．－－はデータなし，‥は有意な相関が認められない組）

表１

	p50	p51	p53	p54	p55	p57	p60
p18	‥	0.406	‥	‥	‥	‥	‥
p21	0.566	0.743	0.580	0.573	—	—	—
p22	‥	‥	‥	‥	—	0.390	0.592
p24	‥	‥	0.562	0.483	‥	0.486	0.484
p28	‥	‥	0.367	0.760	—	0.456	‥
p31	‥	‥	0.534	0.726	—	0.555	0.532

この表において各行または各列の平均が，各項目と他の項目との相関係数の平均値となりその項目における測定の整合性・信頼性を表すものと考える.

なお，中学校数学では３年間の教材が 20 弱の単元に分かれており，単元ごとに信頼性を判断する方がより合理的であるが，２・３年生の教材に対する答案が少ないためこの研究では３年間を通した枠組みの中で整合性・信頼性を求めた.

列ごとの平均値は次の表２（一部を表示）のようになる.

表２

p50	p51	p53	p54	p55	p57	p60
0.410	0.522	0.310	0.543	0.568	0.460	0.538

この集計とは別に，各項目について正答率（%），回答者数，小問数，入力方式，1 題当たり所要時間，ヒントの有無，一覧・順次の別，解説文の長さ（行数）を求めておき，表２の値に対する影響の大きさを調べた.

6. 考察

以下は測定の信頼性という視点から見たときの調査結果の要約である.

- 小項目（小問）数が 10 問以下であるとき小項目数が増えれば測定の信頼性はわずかに良くなる.
- 回答者数が増えても信頼性が増すとは限らないが，回答者数が増えると信頼性を表す平均相関係数が収束してくる傾向が見られる.
- 正答率が 50%から 90%の間では正答率による信頼性の差異は認められない.
- 入力方式による違いとして空欄書き込み方式は必ずしも選択方式よりも優れているとは言えない．むしろ選択肢の個数が 10～20 個あるような多対多選択が優れていることが多い.
- １小項目当たりの所要時間の長短は信頼性にほとんど影響していないが所要時間が増えると平均相関係数が収束してくる傾向がある.
- 問題の表示方法による差異として，概して順次表示の信頼性が少し高くなる.
- 解説文の有無，ヒントの有無は信頼性にほとんど影響しない.

7. まとめと今後の課題

現役の中学生から見れば学年や季節によって行うことのできる単元が変わるので，年間データで集計すればより正確な分析が得られると考えられる.

選択問題にすべきか空欄書き込みにすべきかとなど出題形式で迷うことが多かったが，調査の結果によりある程度の目安が得られた.

調査の結果として，信頼性が大きく損なわれる条件が「なかった」ことが成果であると考える．すなわち，通常想定されている範囲内で測定が行われる限り，測定条件として不適当といえるものは見当たらない.

参考文献

植野真臣・永岡慶三 （2009）「ｅテスティング」．培風館

豊田秀樹（2006）「項目反応理論［入門編］」.朝倉書店