

選択肢形式テスト問題の使用状況の可視化

～ 単純な確率モデルに基づくシナリオの利用 ～

○劉 東岳（学研ホールディングス／学研教育総合研究所）

〈はじめに〉

客観テストで利用される選択肢形式テスト問題 (MCQ) には、常に「ランダム(いい加減)に回答しても正解する確率」の議論が付きまとう。良質の不正解選択肢を作成する労力を考えると、選択肢の数を増やすことで“その確率”を下げればよいと、単純に言い切れないケースも多い。また、1 つのテスト問題に注目するのではなくテスト全体を考えると、出題数と選択肢数は「一部のテスト問題をランダムに回答しても合格する確率」と、どのような関係があるのか。この議論を正確あるいは緻密に進めようとする、テスト専門家以外のステークホルダーには、分かりにくい内容になりがちである。本発表では、MCQ の使用状況の可視化を試みる際のコミュニケーションツールとして、極めて単純で直観的な確率モデルに基づくシナリオの利用について検討する。

〈基本シナリオ〉

互いに独立した MCQ のみで構成されているテストに対し、受験者は「実力分」を確実に正解できるが、「実力を超える分」は全く歯が立たずランダムに回答する、というシナリオを考える。本発表では、議論の本質を損ねることなく、一緒に出題される全ての試験問題は選択肢数が同じと仮定し、かつ 1 問 1 点で正解選択肢は 1 つとする。また、個人の能力は「得点率〇%」と表現する。

・例 1

2 択 MCQ が 50 問出題される 50 点満点のテストを、仮に実力得点率 40% の受験者が受けた場合を

考える。まず「実力分」の 20 問には本人の実力で解答し、全てに正解できる。そして「実力を超える分」の 30 問には“運”に任せて全てにランダムに回答することになり、ここで測定誤差が発生する。ランダムに回答する 30 問それぞれに対し、本人の実力とは関係なく正解する確率は 2 分の 1 であり、測定誤差の平均は $30 \times 0.5 = 15$ 問と計算される。従って、実力得点率 40% の受験者の平均得点率は 70% (実力点 20 点 + 平均測定誤差 15 点で合計 35 点) になる。

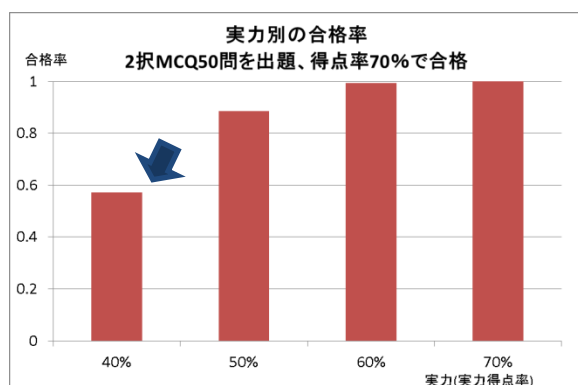
・例 2

5 択 MCQ が 200 問出題される 200 点満点のテストを、仮に実力得点率 70% の受験者が受けた場合を考える。本人の実力で 140 問には正解でき、本人の実力とは関係ない測定誤差の平均は $60 \times 0.2 = 12$ 問と計算される。従って、実力得点率 70% の受験者の平均得点率は 76% (152 点) になる。

・測定誤差と合格率

このシナリオでの得点は実力点と測定誤差の合計で表される。いわゆる古典的テスト理論の True-Score モデルとは異なり、ここでの測定誤差は負の値を取ることがない。測定誤差により発生するのは、プラスの方向に振れる“嵩上げ効果”のみである。従って、実力が合格基準以上の受験者は、必ずテストに合格する。一方で、実力が合格基準に足りない受験者も、状況によっては“かなりの確率”で合格することが容易に想像できる。仮に例 1 で合格基準を得点率 70% に設定すると、実力得点率 40% の受験者の半分以上 (57.2%) が合格する (図 1)。

図 1:



〈確率モデル〉

この合格率は上述のシナリオに従って、次のような確率モデルに基づいて計算される。まず、特定の受験者に対し、実力点は定数である。むろん、個人の実力得点率の本当の値は知ることはできないが、本人は試験で必ず「実力分」の得点を確保する。一方で、「実力を超える分」の問題をランダムに回答することで発生する測定誤差には、実力とは関係ない“運”のよし悪しのみ反映されるが、その値(正解する問題の数)は二項分布に従う確率変数である。

・二項分布

受験者は、自分の「実力を超える分」の問題一つ一つに、ランダムに回答していく。その結果、問題一つ一つに対して正解あるいは不正解の、いずれか片方の結果が得られる。正解が得られる確率は選択肢数によって決まる。二種類の結果(この場合は正解と不正解)が決まった確率で発生する事象が独立して繰り返されていくとき、片方の結果(この場合は正解)が発生する回数は、二項分布 $\text{Binomial}(n, p)$ に従う確率変数となる。 n は事象が繰り返される試行回数で、 p は各試行での成功確率である。

・試行回数と成功確率

このシナリオにおいて、試行回数 n は「実力を超える分」の問題数、つまり満点と実力点の差である。従って n の値は受験者の実力に依存し、実力が低

い受験者ほど大きくなるが、そのような受験者ほど“運に頼る機会が多くなる”と解釈できる。また、 n は出題数の割合で定まるため、出題数にも依存する。出題数が倍になれば n も倍になる。一方で、各試行の成功確率 p はランダムに回答して正解する確率であるが、これは全受験者に対して同じ値である。2 択 MCQ なら 0.5、4 択 MCQ なら 0.25 と、 p は選択肢数の逆数となる。また、 p は出題数にも依存せず、あくまでも個々の問題に対する確率である。なお、「実力を超える分」の中で“運”に任せて正解できる問題数(測定誤差)の平均は $n \times p$ で求められる。

出題数(満点): N_I
 選択肢数: N_A
 合格基準(得点率): PS
 実力(得点率): R
 実力点: $T_R = R \times N_I$
 測定誤差: $E_R \sim \text{Binomial}(N_I - T_R, 1/N_A)$
 得点: $S_R = T_R + E_R$

・合格率

合格とは得点(S_R)が合格基準点($PS \times N_I$)以上の状態である。実力不足でも合格するのは、実力不足分の得点を測定誤差が補いきれるとき、つまり測定誤差(E_R)が合格基準点と実力点の差($PS \times N_I - T_R$)以上のときである。 E_R は二項分布 $\text{Binomial}(N_I - T_R, 1/N_A)$ に従う確率変数である。その値は最小 0 と最大 $N_I - T_R$ の間に収まる整数となる。 E_R が x という具体的な値になる確率は

$$P(E_R = x) = \frac{(N_I - T_R)!}{x! (N_I - T_R - x)!} \left(\frac{1}{N_A}\right)^x \left(1 - \frac{1}{N_A}\right)^{N_I - T_R - x}$$

で表され、測定誤差が合格基準点と実力点の差以上となる確率は

$$P(E_R \geq PS \times N_I - T_R) = \sum_{x=PS \times N_I - T_R}^{N_I - T_R} P(E_R = x)$$

で求められる。実際の計算は、広く市販されている表計算プログラムの多くで、容易に対応できる。プロ

グラムによっては、二項分布の確率計算をする関数を提供していることもある。

この計算式を整理すると、最終的に合格率はテストの出題数(N_I)、選択肢数(N_A)、合格基準(PS)、そして受験者の実力得点率(R)と、互いに独立した4つの値に依存する。このシナリオでは、実力が合格基準以上のときの合格率は1で計算の必要が無いが、「実力不足でも合格する確率」を計算するには、これら4つの値を決める必要がある。これら4つの値のうち、テストを実施する側がコントロールできるのは、テストの出題数、選択肢数、そして合格基準の3つである。この3つはテスト構成を決定するパラメータとして、テスト開発の段階で互いに独立して定めることが可能である。

〈合格率曲線〉

特定のテスト構成の“状態”を表す視覚的な情報として、合格率曲線というものを実力得点率の関数で表す。

$$f_{pc}(R|N_I, N_A, PS) = \begin{cases} 1 & R \geq PS \\ P(E_R \geq PS \times N_I - T_R) & R < PS \end{cases}$$

図 2:

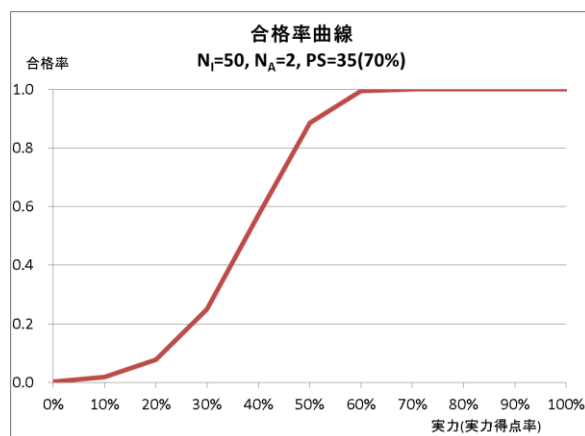


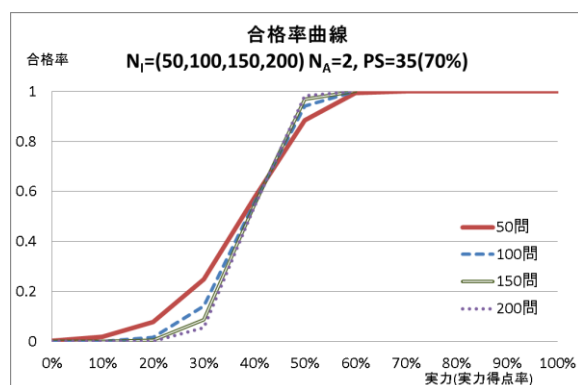
図2は図1と同じく、例1で合格基準を得点率70%に設定した場合の合格率曲線である。2択MCQを50問出題するテスト構成は、実力不足の受験者にとって非常に“緩い”仕組みであることが、この曲線

の形から見て取れる。テスト構成を変えると、合格率曲線も形を変え、実力不足の受験者の合格状態の変化が、視覚的に捉えられる。

・出題数 N_I

選択肢数と合格基準を2択と70%に保って、出題数を50問、100問、150問、200問と変動させたときの合格率曲線を図3に示す。出題数が増えると実力が“大幅”に不足している受験者の合格率は下がるが、実力が“多少”不足している程度の受験者の合格率は逆に上がっている。なぜか。得点比率で考えれば、問題数が増えるても測定誤差の平均は変わらないが、標準偏差は減少し、測定誤差の値は“より”平均値近くに集中するからだ。従って、この場合の“大幅”と“多少”の違いは、平均点(実力点+測定誤差の平均)が合格基準点より上か下かということになる。もともと平均点が合格基準点を超えている“多少”実力不足の受験者にとっては、問題数が増えることで、測定誤差が自分の実力不足を補うという効果が“より確実に”期待できることになる。

図 3:

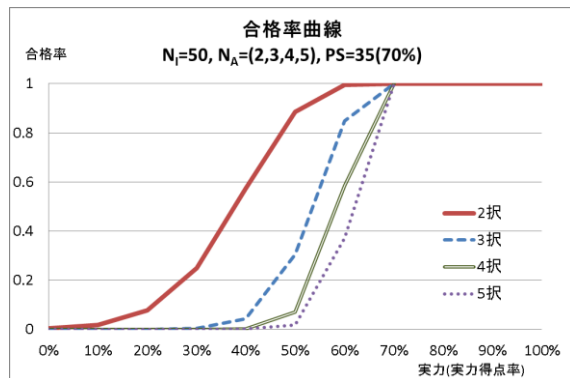


・選択肢数 N_A

出題数と合格基準を50問と70%保って、選択肢数を2択、3択、4択、5択と変動させたときの合格率曲線を図4に示す。選択肢を1つ増やすごとに、実力不足の全範囲で合格率が下がる。問題数を増やす場合と違って、平均点が合格基準を超えている

“多少”実力不足の受験者が合格してしまう状況の改善にも、選択肢数を増やすのは有効である。

図 4:



・合格基準点 SP

出題数と選択肢数を 50 問と 2 択に保って、合格基準を 60%、70%、80%、90%と変動させたときの合格率曲線を図 5 に示す。

図 5

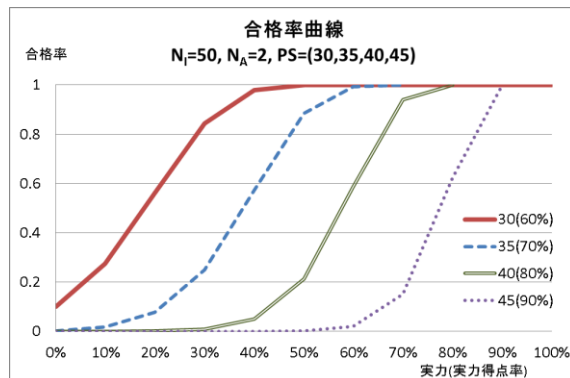


図 3～図 5 は図 2 からの派生で、あくまでも例であり、これらの図によって、この種の操作の一般的な特徴を代表させようとするものではない。

〈上限合格率曲線〉

MCQ のみが出題されるテストでは、どんなテスト構成であっても、必ず実力不足の合格者が発生する。従って「実力不足の合格者が発生する事態」への対策は、その発生確率を 0 にすることではなく、

“許容可能な確率水準”内に抑えることである。全ての実力得点率 R の値において、合格率はこの確率水準を下回っている必要がある。その意味で、この“許容可能な確率水準”を上限合格率と呼ぶ。

上限合格率曲線は、合格率曲線と同じく受験者の実力得点率 R の関数で表される。ただし、テスト構成には依存しない(べき)という点では、合格率曲線と異なっている。どの程度の実力不足の受験者であれば、“運”で合格してしまう確率をどの程度まで許容できるのかは、テスト制度の目的や役割、あるいは実施団体のポリシーや社会的通念によって、テスト構成の外部に定められるものである。

MCQ の利用により実力不足の合格者が発生する事態への対策の一つは、本発表で紹介したシナリオベースの合格率曲線が、この上限合格率曲線を常に下回するようなテスト構成を採用することである。ただ、現実のテスト開発の現場では、必ずしも上限合格率曲線が先に定まらない場合も多く見られる。

〈今後の課題〉

- ・このシナリオが、どこまで現実的なのか、データを踏まえた検証が十分にされていない。
- ・このシナリオが、どこまで説明の役に立っているのか、客観的な検証がされていない。
- ・このシナリオでは、測定誤差の期待値は 0 ではない。得点から受験者の実力を推定するには？
- ・コミュニケーションツールとして分かり易いことが重要であるが、いくつかの発展形は可能と思われる。単純に指摘できる点として 2 点挙げる。
 - (1) 受験者は分からない問題でも、完全にランダムな回答はしないのではないか。
 - (2) 受験者が実力分を確実に正解できる前提は、測定の品質を語る上でナイーブ過ぎないか。
- ・この先、さらに探求する価値があるか。用語など発表者個人が正しい加減に使用している面もある。