

選択肢形式テスト問題の使用状況の可視化 その2

—単純な確率モデルに基づいて合格基準を見直すことについての予備的考察—

劉東岳 (学研ホールディングス/学研教育総合研究所)

概要: 単純な確率モデルを用いて、選択肢形式テスト問題のみで構成されるテストにおける受験者の実力点、得点、そして合否判定結果の関係の可視化を試みた。また、合格基準の見直しを検討するための準備として、同モデルにおける受験者群の実力点分布と得点分布について予備的考察を行った。

キーワード: MCQ, 合格基準, Borderline Candidate (境界線上の受験者), 二項分布, 正規分布, 歪度

<はじめに>

選択肢形式テスト問題(MCQ)には、常に「ランダムに回答しても正答する確率」の議論が付きまとう。この議論を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シナリオに基づく確率モデルを利用し、合格基準を見直すことについて予備的考察を行う。

<基本シナリオ>

劉 [2015] は独立したMCQのみで構成されるテストにおいて「受験者は実力分の問題を確実に正答できるが、実力を超える分の問題はランダムに回答する」というシナリオを使って、「実力が合格基準に満たない受験者が合格する確率」の可視化を試みた。

・ 確率モデル

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

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

このシナリオにおける測定誤差は二項分布に従い負の値を取らないため、嵩上げ効果のみが期待できる。本発表では、議論の本質を損ねることなく、テストの出題数は100問で100問全ての選択肢数が同じとし、かつ1問1点で正解選択肢は1つと仮定する。

・ 合格率曲線

このシナリオにおいて、ある受験者がテストに合格する確率は、出題数(N_I)、選択肢数(N_A)、合格基準(PS)、そして本人の実力(R)の、互いに独立した4つの値で決まる。テストを設計する側からコントロールできるのは N_I 、 N_A 、 PS であるが、特定のテスト構成(N_I 、 N_A 、 PS)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(R)の関数として定義する。

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

図1

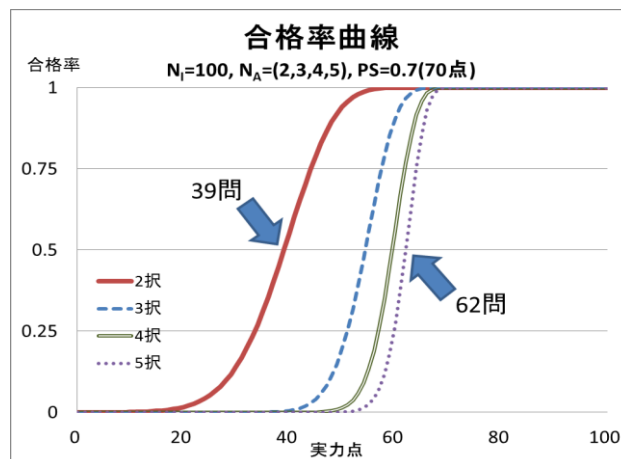


図1は、合格点を70点、選択肢数を2択、3択、4択、そして5択と設定した場合の合格率曲線である。たとえば2択のテストの場合は、実力点が39点の受験者でも合格率50%を期待できるような、実力不足の受験者にとって寛容な仕組みであることが分かる。

・実力点からの議論

以上は受験者からの視点である。受験者が確実に正答できる問題のみを真面目に解答し(そして正答し)、残りの問題は放棄してランダムに回答しても期待できる合格率の議論の起点となっているのは、本人の実力点である。一方、テストの結果を見る側にとって、受験者の実力点は知りえない情報であり、実際に入手可能なのはテスト結果としての得点だけである。このシナリオの下では、受験者の得点から本人の実力点については、どのような“逆算”が可能なのか。

<得点からの議論>

このシナリオにおいて測定誤差は嵩上げ効果としてしか現れず、受験者の実力点が本人の得点を上回ることには無い。たとえば、ある受験者の得点が65点だった場合、「実力点0点+誤差65点」から「実力点65点+誤差0点」までの66通りの可能性しかない。ここで必要なのは、これらの可能性の確からしさを比較し可視化する方法である。たとえば、「実力点60点+誤差5点」の結果としての65点である可能性と、「実力点40点+誤差25点」の結果としての65点である可能性は、どのように比較できるのか。

・標準化

特定の受験者の実力に関しては、何も事前情報がないことを前提とする。すなわち、受験者の得点が判明するまでは、本人の実力点が0点か、1点か、2点か、...、あるいは100点かの可能性は全て同じであると仮定する。すると、受験者の得点が $S_R=s$ という値だと判明した時に、本人の実力点が $T_R=t$ という値である可能性については、次のような標準化の計算が可能になる。測定誤差である E_R は二項分布に従う確率変数であり、右辺の分母は異なる試行回数と共通の成功確率によって算出される確率の和である。

$$f_{st}(T_R = t | S_R = s, N_I, N_A)$$

$$= \begin{cases} \frac{P(E_R = s - t | T_R = t)}{\sum_{x=0}^s P(E_R = s - x | T_R = x)} & s \geq t \\ 0 & s < t \end{cases}$$

図2aと図2bは、この計算式に従って、ある受験者の得点が確定した時の本人の実力点について標準化された“逆算”結果を相対的な確からしさを表している。たとえば図2aからは、2択のテストでは、得点が65点の受験者の実力点が50点以上であることは、ほとんど期待できないと分かる。

図2a

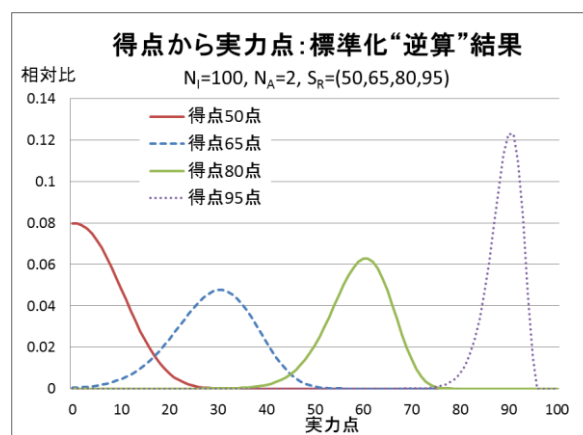
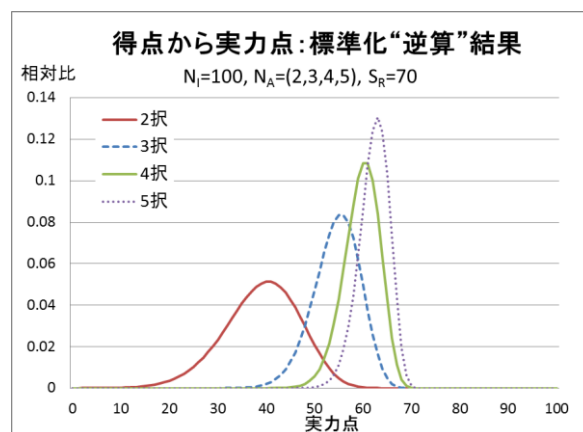


図2b



MCQのみで構成されるテストにおいて、受験者の得点が本人の実力点と等しいことをほとんど期待できない状況は、合格基準を決める過程で **Borderline Candidate**(境界線上に位置する受験者)について言及する場面では、十分に考慮されるべきであろう。

仮に、合格点が70点のテストにおいて、得点70点を取る受験者の大部分が70点を大幅に下回る実力しか持っていないと示されたら、テスト実施者が取る対策として単純に考えられるのは、合格基準点を引き上げる検討であろう。ここで“適切な引き上げ幅”を判断するのに求められるのは、ある得点(たとえば75点)を取った受験者の実力点が、ある基準(たとえば70点)に満たない可能性を評価する方法である。

・ Slip-Past Candidate 曲線(SPC 曲線)

Slip Past とは「すり抜ける」といった意味だが、ここでは実力以上の得点を取るというニュアンスを込めて、SPC 曲線を次のように定義する。これは、先ほど求めた「得点から実力点へ、標準化された“逆算”結果」の累積関数として、得点 s を取った受験者の実力点が t を下回っている可能性を表すものである。

$$f_{\text{SPC}}(S_R = s | T_R = t, N_I, N_A) = \sum_{x=0}^{t-1} f_{\text{st}}(T_R = x | S_R = s, N_I, N_A)$$

図 3

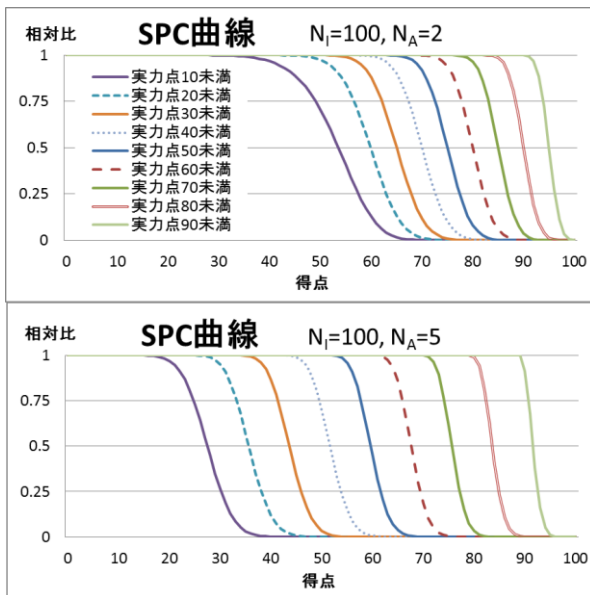


図 3 は 2 択と 5 択のテストについて、SPC 曲線を得点 10 点ごとに示したものである。たとえば 2 択のテストの場合、得点が 70 点の受験者の実力が 50 点を上回っていることは、ほとんど期待できないことが示されている。図 1, 図 2a, 図 2b とは異なり、図 3 の横軸は実力点ではなく得点である。

<合格基準見直しの予備的考察>

今までの考察・議論を踏まえて合格基準の見直しをするに当たっては、次のような事態が発生する確率が低くなるように、合格基準を引き上げる検討をすることが考えられる。

- ① 個人の実力点から得点へ: 実力は合格基準を“許容できないほど”下回る受験者が、結果として合格基準を満たす得点を取ってしまう事態
- ② 個人の得点から実力点へ: 得点は合格基準を満たしている受験者の実力が、合格基準を“許容できないほど”下回っている事態

・ 受験者群

実際に合格基準見直しを検討する際には、受験者一人の実力点と得点についての考察・議論だけでは不十分で、受験者群の実力分布と得点分布についても、考察・議論の範疇に含める必要がある。ここでは、その理由を 3 つ挙げる。

- 1) 受験者一人のパフォーマンスの議論だけからでは、このシナリオが現実的なのか(妥当なのか)を検証するのは難しい。我々が普段目にするような得点分布が、このシナリオの下で実際に得られるのかについて、確認する必要がある。
- 2) 制度的には「合格基準を満たさない合格者の割合」を低くしたいという需要があり、受験者群の能力分布・得点分布を考える必要がある。
- 3) 測定誤差による嵩上げ効果を念頭に置いて合格基準を引き上げることを検討するのであれば、「実力は本来の合格基準を超えているのに不合格となる受験者」が出てくることを考慮する必要がある。Livingston と Zieky[1982]は「合格するべきでない受験者を合格させる判断」と「合格するべき受験者を合格させない判断」の 2 種類の合否判断ミスについて言及している。

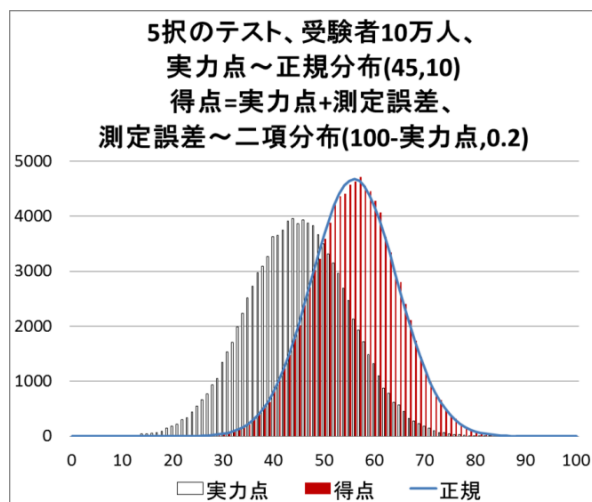
・ 受験者群の実力分布と得点分布

先ほどの標準化計算において、特定の受験者の実力点については 0 点から 100 点のいずれの可能性も同じと仮定した。しかし、受験者群の実力点分布を考える時には、その値が一様分布に従う(いずれの実

力点の受験者も同じぐらい存在する)と仮定するのは、現実的ではない。本発表では、受験者群の実力点が正規分布(得点分布の一般的なモデル)に従うものと仮定する。たとえば、実力点が正規分布(45, 10)に従う受験者群が 5 択のテストを受けると仮定すると、どのような形状の得点分布が得られるのだろうか。

正規分布の線形変換した結果は正規分布であり、測定誤差が従う二項分布の試行回数は正規分布(55, 10)に従う確率変数になる。代数的なアプローチは別の機会に譲るとし、ここではシンプルなシミュレーションの結果を紹介する。実力点が平均 45、標準偏差(SD)10 の正規分布に従う 10 万人の受験者群を想定し、本シナリオの下で 5 択のテストの得点をシミュレートした結果を図 4 に示す。得られた得点分布と同じ値の平均とSDを持つ正規分布グラフを重ねたところ、この得点分布は正規分布で近似できそうな形(我々が普段目にする、あるいは前提とするような形)をしており、少なくとも得点分布の形状からは、このシナリオの妥当性が否定されることはなかった。

図 4



無論これだけでは、「このシナリオの下で、実力点が正規分布に従うと仮定したら、得点分布は正規分布で近似できる」という結論は出てこない。前述のシミュレーション(ケース A)を 10 回、ならびに平均は同じ 45 で SD を 1 にした場合のシミュレーション(ケース B)を 10 回繰り返して得られた得点分布の基本統計量を表 1 にまとめる。ケース A では得点分布の歪度が 10

回とも負の値であるのに対し、ケース B では歪度が 10 回とも正の値だった。本シナリオにおける実力点分布と得点分布の関係については、代数的なアプローチも含めて更なる調査が必要である。

表 1 得点分布の基本統計量(負の値に網掛け)

【A】実力点～正規分布(45,10)				【B】実力点～正規分布(45,1)			
平均	SD	尖度	歪度	平均	SD	尖度	歪度
56.00	8.54	-0.01	-0.04	56.00	3.08	0.01	0.17
55.99	8.52	-0.02	-0.06	56.00	3.07	-0.01	0.17
55.98	8.54	0.00	-0.07	56.02	3.08	0.00	0.17
55.95	8.52	0.02	-0.05	55.99	3.09	-0.02	0.16
56.05	8.54	0.01	-0.07	56.01	3.09	0.00	0.18
55.99	8.52	-0.02	-0.04	56.01	3.07	0.01	0.17
56.01	8.55	-0.02	-0.06	56.00	3.08	0.01	0.17
55.96	8.50	0.01	-0.05	56.00	3.09	0.01	0.16
55.99	8.54	-0.01	-0.05	56.00	3.08	-0.01	0.17
56.02	8.85	-0.01	-0.05	56.00	3.09	0.01	0.16

<まとめと今後の課題>

テストの専門家以外のステークホルダーに MCQ の使用状況を伝える可視化ツールとして、単純なシナリオに基づく確率モデルの利用を試みた。また、合格基準の見直しを検討する上で、受験者群の得点分布に関する予備的考察を進めた。今後の課題は、本シナリオにおける実力点分布と得点分布の関係の代数的および数値的なアプローチによる解析、実際のデータに基づく本シナリオの検証、そして、適切な合格基準の引き上げ幅の算出・選定方法の確立である。

参考文献

- ◆ 劉東岳 (2015), 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～, 日本テスト学会第 13 回大会発表論文抄録集, pp100-103, 日本テスト学会
- ◆ Livingston, Samuel A. & Zieky, Michael J., (1982) Passing Scores – A Manual for Setting Standards of Performance and Occupational Tests, Educational Testing Services