

# イブサティブ回答形式を用いた性格特性尺度分析に関する考察

酒匂志野<sup>1</sup> 今城志保<sup>1</sup> 北原瑞穂<sup>2</sup> 荘島宏二郎<sup>3</sup>

<sup>1</sup>株式会社リクルートマネジメントソリューションズ <sup>2</sup>東京大学大学院教育学研究科 <sup>3</sup>大学入試センター

## 【背景】

性格適性検査は、性格によって職務への適性を推測する検査である。入社試験、昇進昇格試験などの場面で活用されることが多いため、フェイクング、すなわち、自分をよく見せようとする傾向が現れやすいことが知られている。

フェイクングを統制するための方法として、一対比較法やランキング法などのイブサティブ形式でのデータ収集が知られている。ここでイブサティブ形式とは、回答値の合計がどの回答者においても同一となるデータ形式のことである(Thurstone, 1927)。伝統的なイブサティブデータの分析法では個人間の比較に焦点があてられることはなかったのだが、近年、多次元 IRT の技術を応用することにより、イブサティブデータから個人間比較を可能にする方法が提案されている(Maydeu-Olivares & Brown, 2010)。また、川端(2007)においても個人間の比較が可能となるような SEM のモデルが提案されている。川端(2007)では、「企業の採用試験」という場面の想起を求める条件とそうではない条件の2条件で性格検査への回答を求めており、分析の結果、リッカート形式での得点は前者において高まるがイブサティブ形式での得点は両者においてさほど大きな違いが見られないことを明らかにしている。

上記の議論の中では、「リッカート形式とイブサティブ形式では同一の概念が測定されている」と考えられている。たとえば、Joubert, et al. (2015)では、両者の相関係数が.50~.84であったと報告されているが、一般に、フェイクングの操作を行わない場合に、イブサティブ形式とリッカート形式での測定内容間に強い相関関係はみられるのだろうか。そこで本研究では、9尺度で構成されるイブサティブ形式の性格特性質問紙を構成し、得られたデータを Maydeu-Olivares & Brown(2010)のモデルを用いて多次元 IRT により分析した。同じ尺度についてリッカート形式でも同一の回答者に回答を求め、両者の回答の特徴をみることにした。

## 【方法】

日本のビジネスパーソンを対象に調査を実施した。調査はインターネット上で実施し、調査日は2015年3月17日~23日であった。調査対象者は日本のホ

ワイトカラーのビジネスパーソン 800名であった(性別:男性のみ/年齢:30~49歳/勤務先企業の従業員規模:500名以上)。そのうち、英語の理解度に問題がないと思われる621名のデータを用いて以後の分析を行った。なお、本調査で実施した尺度は、以下の通りである。

・性格を問う4件法(Strongly agree から Strongly disagree まで)のリッカート形式の尺度(「外向」「統率」「変革」「強靱」「大胆」「行動」「維持」「調整」「思索」の9尺度から成る。大胆に関する項目数は8、維持に関する項目数は9、調整に関する項目数は10、それ以外の概念に関する項目数は5)。以下「リッカート形式の9尺度」とする。

・性格を問うイブサティブ形式の9尺度(上記と同じ9尺度から成り、各尺度に属する4つの項目を1セットとして提示し、回答者自身にあてはまる順に順位をつけさせる課題が27セット)。以下「イブサティブ形式の9尺度」とする。

・虚偽尺度(リッカート形式、7件法(Strongly agree から Strongly disagree まで)、5項目)。Balanced Inventory of Desirable Responding (BIDR-J)を参考に作成したもので、項目は「I am fully in control of my own fate.」などである。回答者ごとに5項目の平均値を算出し、それを以下の分析に用いた。

## 【結果】

### イブサティブ形式の9尺度

IRT モデルによる推定には Mplus(ver 7.2)を、それ以外の分析には主に SPSS(ver22.0)を用いた。

イブサティブ形式では、項目は4つセットで提示され、回答者はその中から最も当てはまる項目(most like me)と最も当てはまらない項目(least like me)を1つずつ選択することが求められた。分析時には、それらのデータを2項目ずつ比較した場合への反応(2値データ)に変換した(例:外向が most、統率が least の場合、統率よりも外向のほうが自分に当てはまっているので、これらの比較では外向のほうが選択されると考えられる。1セットにつき6ペア作られる)。なお、各セットにつき1ペアずつ値が定まらないペアが発生する(例:強靱と大胆ではどちらがより自分に当てはまるか不明)。よって、IRT モデルにより推定を行う際には、多重代入法により 60

個のデータセットを生成し、それぞれにおいて項目パラメータと個人特性値の推定を行って、それらの平均をとった。

なお、あてはめたモデルは以下の通りである。すなわち、特性 $a$ (例 外向性)を測定する項目 $i$ と特性 $b$ (例 統率性)を測定する項目 $k$ とを比べる比較対 $l$ に対する反応を $y_l$ とする。このとき、比較対 $l$ への反応は、2項目の効用の差 $y_l^*$ によって、

$$y_l = \begin{cases} 1, & y_l^* \geq 0 \text{ のとき} \\ 0, & y_l^* < 0 \text{ のとき} \end{cases}$$

と決まると考える。さらに、効用の差 $y_l^*$ に対して以下のような因子分析モデルを考える。

$$y_l^* = -\gamma_l + (\lambda_{ia}\eta_a - \lambda_{kb}\eta_b) + (\epsilon_i - \epsilon_k) \quad (1)$$

ここで、 $\eta_a, \eta_b$ は潜在特性 $a, b$ の値を表しており、 $\lambda_{ia}, \lambda_{kb}$ はそれぞれの項目への各特性の因子負荷量を、 $\epsilon_i, \epsilon_k$ は項目 $i$ と $k$ の独自性を、 $\gamma_l$ は比較対に固有の切片を、それぞれ表している。

### リッカート形式の9尺度

クロンバックの係数は、外向で.852、統率で.757、変革で.781、強靱で.723、大胆で.805、行動で.733、維持で.743、調整で.856、思索で.791であった。また、尺度間の相関係数は.048～.677であった。なお、IRTモデルの母数の推定には段階反応モデルを用い、尺度ごとに推定を行い、推定にはR(for windows 3.0)を利用した。

### 分析1

#### 相関係数

イプサティブ形式の9尺度(IRT推定値)と、リッカート形式の9尺度(IRT推定値)でそれぞれ対応する特性間の相関を求めたところ、平均値は0.610、標準偏差は0.077であり、最小値は0.444、最大値は0.734であった(Table1参照)。ちなみに、Joubert et al. (2015)では32尺度の相関が求められており、それらの平均値は0.713、標準偏差は0.073、最小値は0.500、最大値は0.840であった。

今回は特に「入社試験」のような場面の想起を求めず、フラットな条件で回答を求めている。フェイキング条件を設定したわけではないのに相関がさほど高くないため、イプサティブ形式はリッカート形式とは異なるものを測定している可能性がある。以下、その点の検証を試みた。

### 分析2 イプサティブ形式とリッカート形式の回答のずれのパターン

回答者ごとにイプサティブ形式の9尺度と、リッカート形式の9尺度の差分を算出し、それをクラスタ分析した。以下の4つのクラスタが表れた。

全体的にイプサティブ形式の9尺度(IRT推定値)のほうが得点が低いクラスタ(「フェイキングクラスタ」、68名)

イプサティブ形式・リッカート形式ともに得点が平均値付近にあるクラスタ(「中庸クラスタ」、149名)

全体的にイプサティブ形式の9尺度(IRT推定値)のほうが得点が高いクラスタ(「謙遜クラスタ」、215名)

維持のみイプサティブ形式の9尺度(IRT推定値)のほうが得点が高いクラスタ(「維持が特徴的なクラスタ」、189名)

各クラスタにおいて、イプサティブ形式およびリッカート形式のIRT推定値の平均値はFigure1のとおりである。

Figure 1 各クラスタのイプサティブ・リッカートのIRT推定値

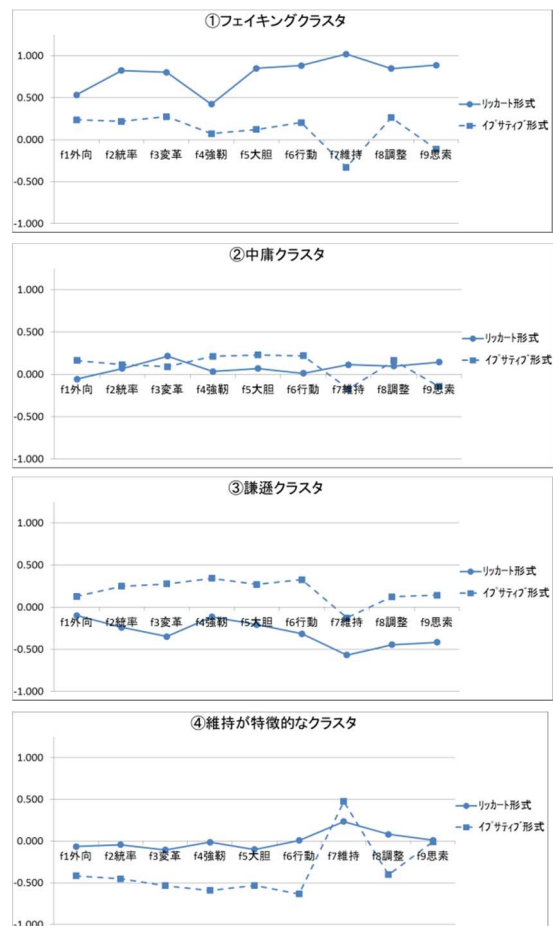


Table1 リッカート形式IRT推定値とイプサティブ形式IRT推定値の相関

外向	統率	変革	強靱	大胆	行動	維持	調整	思索
0.714	0.673	0.650	0.665	0.642	0.567	0.444	0.577	0.560

Table2 各クラスターの虚偽尺度得点の平均値・標準偏差

	虚偽尺度得点	
	平均値	標準偏差
フェイキングクラスター	4.512	0.965
中庸クラスター	4.243	0.779
謙遜クラスター	3.872	0.840
維持が特徴的なクラスター	4.047	0.748

虚偽尺度得点( =0.649)の平均値と標準偏差はそれぞれ Table2 のとおりであった。虚偽尺度得点を分散分析し多重比較(Tukey法)を行ったところ、「フェイキングクラスター」のほうが「謙遜クラスター」、「維持が特徴的なクラスター」よりも有意に虚偽尺度得点が高かった(いずれも  $p < .001$ )。また、「中庸クラスター」、「維持が特徴的なクラスター」のほうが「謙遜クラスター」よりも虚偽尺度得点が高い(いずれも  $p < .001$ )。

### 各クラスターにおける回答の詳細

以上のように、リッカート形式とイブサティブ形式の得点の違いにはいくつかのパターンがみられた。そこで、各クラスターにおいて両者の回答の関係性の違いを見ていくことにした。具体的には、イブサティブの各項目セットにおいて、横軸にリッカートのIRT推定値、縦軸にイブサティブ項目における「most like me」の選択率をとり、各クラスターで両者に比例関係が見られるか(右肩上がりの図になっているか)を項目単位で出力した。特に、ここでは、クラスター間でのずれが見られた維持と、それ以外の尺度(代表として、統率)の特性値への負荷の高かったイブサティブ項目においてそれらを見ることにした。なお、本分析にも R(for windows 3.0)を使用した。

### 維持

リッカートのIRT推定値を、-1、-0.5、0.5、1で区切り5つの群に分け、それぞれの群においてイブサティブ形式の項目セットにおいて維持を選択している回答者の割合を算出した(Figure2参照)。なお、各クラスターのリッカートIRT推定値群の人数はTable3のとおりであった。

Table3 クラスター別リッカートIRT推定値の各群の人数(維持)

	IRT推定値の各群				
	-1未満	-1以上 -0.5未満	-0.5以上 0.5未満	0.5以上 1未満	1以上
フェイキングクラスター	1	1	16	19	31
中庸クラスター	10	16	84	23	16
謙遜クラスター	61	46	98	7	3
維持が特徴的なクラスター	6	22	114	24	23

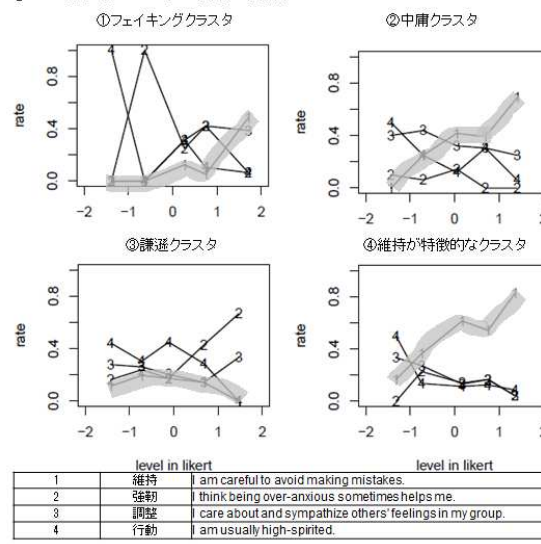
維持を含む12個の項目セットのうち、反転項目を除く10個の項目セットを本分析の対象とした。なお、検討対象となった項目の因子への負荷の最大値は1.212、最小値は0.049であった。

項目セット17において、「フェイキングクラスター」はリッカートのIRT推定値(横軸)の値の割にイブサティブにおける選択率(縦軸)が低く、リッカートIRT

の推定値が「0.5以上1未満」の群においてイブサティブにおける選択率は0.1程度であった。一方、「中庸クラスター」は、リッカートのIRT推定値が「0.5以上1未満」の群においてイブサティブにおける選択率が0.4程度であった。「謙遜クラスター」「維持が特徴的なクラスター」では、リッカートのIRT推定値が高くなるほどイブサティブにおける選択率も高くなり、リッカートのIRT推定値が「0.5以上1未満」の群においてイブサティブにおける選択率は0.5程度であった(Figure2参照)。

つまり、クラスターにおいてはリッカートの推定値が高い人であってもイブサティブにおける選択率がそれほど高くないが、クラスターにおいてはリッカートの推定値が高い人であればイブサティブにおける選択率も高くなっていた。

Figure2 項目セット17における「維持」の選択率



### 統率

維持と同様にリッカートのIRT推定値を、-1、-0.5、0.5、1で区切り5つの群に分け、それぞれの群においてイブサティブ形式の項目セットにおいて維持を選択している回答者の割合を算出した(Figure3参照)。各クラスターのリッカートIRT推定値群の人数はTable4のとおりであった。

Table4 クラスター別リッカートIRT推定値の各群の人数(統率)

	IRT推定値の各群				
	-1未満	-1以上 -0.5未満	-0.5以上 0.5未満	0.5以上 1未満	1以上
フェイキングクラスター	1	2	19	14	32
中庸クラスター	13	21	72	25	18
謙遜クラスター	41	30	104	26	14
維持が特徴的なクラスター	31	26	82	36	14

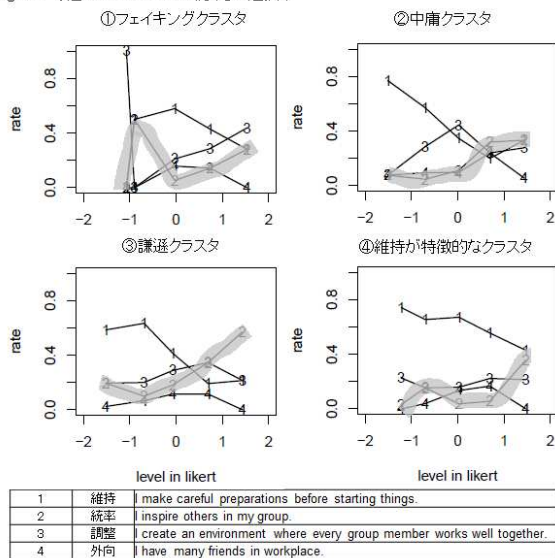
統率を含む12個の項目セットのうち、反転項目を除く10個の項目セットを本分析の対象とした。なお、検討対象となった項目の因子への負荷の最大値は1.533、最小値は0.134であった。

項目セット2において、「フェイキングクラスター」

「維持が特徴的なクラスタ」はリッカートのIRT推定値(横軸)の値の割にイプサティブにおける選択率(縦軸)が低く、リッカートのIRT推定値が「0.5以上1未満」の群において、イプサティブにおける選択率が0.1程度であった。一方、「中庸クラスタ」「謙遜クラスタ」では、リッカートのIRT推定値が「0.5以上1未満」の群においてイプサティブにおける選択率が0.3程度であった(Figure 3参照)。

つまり、クラスタではリッカートの推定値が高い人であってもイプサティブにおける選択率がさほど高くないが、ではリッカートの推定値が高い人はイプサティブにおける選択率も高くなっていった。

Figure3 項目セット2における「統率」の選択率



### 【考察】

イプサティブ形式は一般にフェイキングを抑える方法として知られている。しかし、今回の調査においては、特にフェイキングの条件を設定しなくとも、リッカート形式とイプサティブ形式において対応する概念間の相関係数は0.4~0.7程度しかなかった。フェイキング条件を設定したわけではないのに相関がさほど高くないため、イプサティブ形式はリッカート形式とは異なるものを測定している可能性があると考えられる。

そこで、リッカート形式におけるIRTの推定値とイプサティブ形式におけるIRTの推定値の差分を出し、差分をクラスタ分類した。そしてクラスタ別に個別項目の回答傾向を確認した。リッカートにおける推定値の割にイプサティブ形式における項目選択率が高いクラスタもあれば(謙遜クラスタ)、その逆のクラスタもあった(フェイキングクラスタ)。「維持が

特徴的なクラスタ」の人たちは、維持においてはリッカートにおける推定値が高い人であればイプサティブにおける選択率が非常に高かったが、維持以外の尺度においてはリッカートの推定値が高い人であってもイプサティブにおける選択率がさほど高くなかった。その理由として以下のようなことが考えられる。  
・リッカート項目は単独で提示されるため、基準を外側の他者または自分の理想において回答する可能性がある。維持すなわち着実さについて身近に優れた人がいる場合、自分はまだまだだという思いが強くなりリッカート得点が低くなる。一方イプサティブでは個人内での比較が求められるため、2つの行動のうちどちらをとることが多いのかという基準で回答する可能性がある。例えば、業務で着実さが求められる人の場合、他の選択枝よりは維持を選択する可能性がある。

本研究で明らかになったことは、リッカート形式の得点水準を固定したときに、イプサティブ形式の項目を選択しやすい人と選択しにくい人がいるということである。さらに、それがすべての尺度において等しく表れるわけではなく、特定の尺度でのみリッカートの得点水準の割にイプサティブ形式の項目を選択しやすくなるといった特徴があることである。

なぜリッカートの得点の水準がイプサティブ項目の選択率に反映される度合いが人によって異なるのかについては、本研究では明らかにすることができなかった。今後は、本人の理想とする姿や業務上求められている役割など、測定結果に影響を与える他要素との関連性についても明らかにしていきたい。

### 【文献】

- Joubert, T. et al. (2015) A comparison of the psychometric properties of the forced choice and likert scale versions of a personality instrument. *International Journal of Selection and Assessment*, 23(1) 92-97.
- 川端一光(2008) 構造方程式モデリングによる一対比較型尺度構成法の開発. 早稲田大学博士論文.
- Maydeu-Olivares, A., & Brown, A. (2010). Item Response Modeling of Paired Comparison and Ranking Data. *Multivariate Behavioral Research*, 45 935-974.
- L. L. Thurstone. (1927) A law of comparative judgment. *Psychological Review*, Vol 34(4) 273-286.