性格検査のフェイキング耐性に関する研究
－多尺度一対比較形式と評定尺度形式の比較－

○坂本　亜沙美[1]、○宇佐美　慧[2]、内藤　淳[1]
[1]株式会社リクルートキャリア、[2]日本学術振興会・南カリフォルニア大学

目的

採用選考場面における性格検査の実施が一般的になるとともに、社会的望ましさに代表されるフェイキング耐性への頑健性は性格検査の質を評価する上で重要な観点となってきている。フェイキング耐性の高い検査形式として強制選択方式（Forced-choice format）が採用される場合が数多くあるが、これにおいてはイプサティブなデータ構造となる回答形式が用いられることが多い。その結果、データからは各性格特性の評価点についての相対的な順位に関する情報しか得られないため、評価結果を受検者間で直接比較できないことや、他にもイプサティブデータにより各性格特性の評価点の間の内部相関の推定にバイアスが生じることが指摘されている（Guilford,1954; Hicks, L. E,1970）。

そこで、受検者間比較が可能で、尚且つフェイキングが抑えられる新たな方法として、Stark et al.(2005)は、MUPP（the multi-unidimensional pairwise preference）によるアプローチを提案した。この方法は、下記の 6 ステップの手続きを経て実行される。

1. 評定尺度法に基づく、各性格特性を測定するための尺度項目を複数作成する。
2. ある受検者集団に対し、各項目がどれくらい自分に当てはまるかを 5 件法で実験的に回答させる（回答データ A）。また、（別の受検者集団に対し）各項目がどれくらい社会的に望ましいものかを回答させる(回答データ B)。
3. 回答データ A を利用して、各尺度の回答データに（一次元）項目反応モデルを当てはめて各項目の（困難度）母数の推定を行う。もしくは単に項目ごとに評定平均値を計算する。
4. 回答データ B を利用して、望ましさの評価平均点が近く尚且つ互いに異なる性格特性を評価する項目を対にして一対比較型の項目を作る。また、3.の結果を利用して、望ましさの評価平均点が近く尚且つ困難度母数の推定値の差異が比較的大きい､同じ性格特性を評価する項目を対にした一対比較型の項目を一部作る。
5. 4.で作成した一対比較型の項目からなる検査を実施し､どちらの項目が自分により当てはまるかを（二値型の形式で）回答させる。
6. 5.で得られた一対比較データに対して、一対比較モデル(Bradley-Terry Model)と展開型の項目反応モデル(Generalized Graded Unfolding Model, GGUM; Roberts et al, 2000; Usami, 2011)を統合した項目反応モデルを当てはめ、各尺度に関する潜在特性値の推定値を得る。

Chernyshenko et al.(2009)では、フェイキングをせず正直に回答するという実験的条件の下で、評定尺度形式、同尺度から構成される一対比較形式、多尺度から構成される一対比較形式の計三形式から得られた性格検査得点の互換性を検討し、これらの間で各尺度間の内部相関構造が概ね等質であることを報告している。ただし、実際の選抜場面で活用するためには、フェイキング条件下での同様の有効性の確認、およびフェイキングにより生じうるバイアスが一対比較形式により実際にどの程度改善しうるのかに関する定量的な検証が必要である。そこで本研究では、Stark et al.(2005)の手続きに則って実験的にデータを収集して、フェイキング条件と正直に回答する条件の両者の結果を比較する。具体的には、評定尺度形式と比べて、一対比較形式の方法が尺度間の内部相関の等質性を保ちながらも、推定値のバイアスを抑制可能であるかどうか検証する。

方法

1．予備調査Ａ（ステップ 1-ステップ 3）

1-1．材料　株式会社リクルートキャリア保有の項目プールより、情動性、支配欲求、抑うつ性、シャイネス、粘り強さ、奉仕性、自己信頼の 7 尺度、計 481 項目を用いて研究用検査を作成した(上記のステップ 1)。

1-2．手続き　項目プールの中から、検査項目に対する回答データ 15000 件をランダムサンプリングし抽出した（回答データ A：上記のステップ 2)。この回答データＡは、20 代の大学生を対象に、インターネット上で、各項目がどれくらい自分に当てはまるかを 4 件法で回答させたものを 2 値に変換した

ものである。そして、この回答データＡを元に各項目について評定平均を計算し、項目困難度を推定した（上記のステップ 3)。

2. 予備調査Ｂ（ステップ 2)

2-1. 調査対象者　大学 4 年生から社会人 3 年目までの 736 名。

2-2. 材料　1-1 で作成した研究用検査を使用した。

2-3. 手続き　インターネット上で、各項目に対し、どのくらい社会的に望ましいと感じるかを「望ましい」から「望ましくない」までの 5 件法で回答させた（回答データ B：上記のステップ 2)。

3. 本調査(ステップ 4-ステップ 5)

3-1. 調査対象者　大学 4 年生から社会人 3 年目までの 130 名。

3-2. 材料　回答データ B をもとに、望ましさの評価点の平均値が同程度で、かつ異なる性格特性を反映した項目を対にした一対比較型項目を 67 項目作成した。また、回答データ A も利用して、同一の性格特性を反映し尚且つ評定平均の差異が比較的大きいと考えられる項目を対にした一対比較型項目を 21 項目作成した。これらの項目をランダムに配列して作成された検査を検査Ⅰとした（上記のステップ 4)。また、比較対象として、計 62 項目からなる評定尺度型の検査を作り、これを検査Ⅱとした。

3-3. 手続き　調査対象者を会場に集め、正直な回答およびフェイキングを促す指示を通して検査Ⅰ・Ⅱを受検させた。指示の順序と検査Ⅰ・Ⅱの実施順序についてはランダム化した（上記ステップ 5)。

4. 潜在特性値の推定(ステップ 6)

上述の手続きに拠り、二種類の測定方法（検査Ⅰの一対比較形式、検査 II の評定尺度形式）と二種類の回答態度（フェイキング条件（以降 faking))、正直条件（以降 neutral))からなる 4 種の回答データ(一対比較・faking，一対比較・neutral，評定尺度・faking，評定尺度・neutral)が得られた。

二種類の一対比較データに対する各尺度の母数の推定は、一対比較モデル(Bradley-Terry model)および（GGUM ではなく）2 パラメタロジスティックモデルを併用したモデルである、

$$P_x(x_{jili'l'}=1|\theta_{jl},\theta_{jl'},\eta_{il},\eta_{i'l'})=\frac{P_y(y_{jil}=1|\theta_{jl},\eta_{il})\times P_y(y_{ji'l'}=0|\theta_{jl'},\eta_{i'l'})}{P_y(y_{jil}=1|\theta_{jl},\eta_{il})\times P_y(y_{ji'l'}=0|\theta_{jl'},\eta_{i'l'})+P_y(y_{jil}=0|\theta_{jl},\eta_{il})\times P_y(y_{ji'l'}=1|\theta_{jl'},\eta_{i'l'})}$$

$$P_x(x_{jili'l'}=0|\theta_{jl},\theta_{jl'},\eta_{il},\eta_{i'l'})=\frac{P_y(y_{jil}=0|\theta_{jl},\eta_{il})\times P_y(y_{ji'l'}=1|\theta_{jl'},\eta_{i'l'})}{P_y(y_{jil}=1|\theta_{jl},\eta_{il})\times P_y(y_{ji'l'}=0|\theta_{jl'},\eta_{i'l'})+P_y(y_{jil}=0|\theta_{jl},\eta_{il})\times P_y(y_{ji'l'}=1|\theta_{jl'},\eta_{i'l'})} \quad (1)$$

を利用して行った（上記ステップ 6)。ここで、$P_x(x_{jili'l'}=1|\theta_{jl},\theta_{jl'},\eta_{il},\eta_{i'l'})$は、調査対象者 j が，l 番目の尺度の i 番目の項目と l' 番目の尺度の i' 番目の項目を比較したときに、前者の項目を選択する（この選択を表すダミー変数 $x_{jili'l'}$ において $x_{jili'l'}=1$ となる）確率である。また、$P_y(y_{jil}=1|\theta_{jl},\eta_{il})$は、$l$ 番目の尺度の i 番目の項目に関して仮に評定尺度法により「当てはまる」・「当てはまらない」の 2 値で回答した場合に、2 パラメタロジスティックモデルの観点から「当てはまる」を選択する（この選択を表すダミー変数 y_{jil} において $y_{jil}=1$ となる）確率である。θ_{jl},η_{il} は l 番目の尺度における調査対象者 j の潜在特性値、および同様の尺度における i 番目の項目についての（識別力母数と困難度母数をまとめた）項目母数ベクトルを表す。このように、Stark et al.(2005)の基本的な考え方は、2 値である一対比較型の反応確率を、評定尺度法により測定した場合に一方を支持する（しない）確率と他方を支持しない（する）確率の積和を利用して表現し，母数の推定を複数の尺度に亘って行うというものである。また、母数の推定は MCMC 法に基づくベイズ推定を利用した。

最後に、2 種類の評定尺度データに関しては、GPCM(Generalized Partial Credit Model; Muraki, 1992)を当てはめて、ベイズ法により各項目の項目母数を推定した。なお，各尺度における尺度得点と対応する潜在特性値の相関はいずれも 0.95 以上と非常に高く、本研究のデータにおいて，累積型のモデルである GPCM を当てはめて問題ないと考えられる。これらより、本研究で Stark et al.(2005)の手続きと特に異なるのは、ステップ 6 において展開型ではなく累積型の項目反応モデルを用いて、さらに母数の推定にベイズ推定を利用した点であると言える。

結果と考察

(1) 尺度別にみた 4 種類の回答データ間の相関関係

得られた潜在特性値の推定値を利用し、4 つの回答データ間の相関係数を算出した（表 1)。各尺度の推定値について、一対比較・faking と一対比較・neutral の間の相関係数の平均は 0.55 程度であり、また評定尺度・faking と評定尺度・netural の間の同様の相関係数の平均は 0.30 程度であった。また、

各尺度の推定値について、一対比較・faking と評定尺度・neutral、および評定尺度・faking と一対比較・neutral の間の相関係数をそれぞれ見てみると、前者の相関は平均 0.40 程度で後者は平均 0.30 程度であった。これらより、一対比較・neutral、評定尺度・neutral のいずれの結果を、フェイキングの影響を受けない“正しい”推定値とみなして比較したとしても、一対比較・faking の方が評定尺度・faking よりもそれらとの相関が平均的に高い。

表 1　4 種類の回答データに基づく潜在特性値間の尺度別の相関係数

			一対比較		評定尺度	
			faking	neutral	faking	neutral
情動性	一対	faking				
		neutral	0.67			
	評定	faking	0.59	0.50		
		neutral	0.49	0.76	0.52	
抑うつ性	一対	faking				
		neutral	0.61			
	評定	faking	0.53	0.33		
		neutral	0.41	0.77	0.28	
粘り強さ	一対	faking				
		neutral	0.55			
	評定	faking	0.55	0.36		
		neutral	0.41	0.73	0.44	
自己信頼	一対	faking				
		neutral	0.58			
	評定	faking	0.44	0.07		
		neutral	0.44	0.67	0.24	

			一対比較		評定尺度	
			faking	neutral	faking	neutral
支配欲求	一対	faking				
		neutral	0.47			
	評定	faking	0.74	0.32		
		neutral	0.37	0.81	0.30	
シャイネス	一対	faking				
		neutral	0.31			
	評定	faking	0.69	0.06		
		neutral	0.21	0.72	0.18	
奉仕性	一対	faking				
		neutral	0.58			
	評定	faking	0.40	0.25		
		neutral	0.37	0.71	0.36	

(2)　4 種類の各回答データ内の尺度間相関の等質性

各尺度の潜在特性値から計算される 4 種類の尺度間分散共分散行列について、Ⅰ）尺度間の分散共分散行列が互いに異なることを仮定したモデル、Ⅱ）同一測定方法内の分散共分散行列が等しいことを仮定したモデル、Ⅲ）同一回答態度内の分散共分散行列が等しいことを仮定したモデル、Ⅳ）尺度間の分散共分散行列が全て等しいことを仮定したモデルをそれぞれデータに当てはめ、各モデルの当てはまりの良さを各種適合度指標および情報量規準を用いて比較した（表 2）。等質な共分散構造をもつ回答データを積極的に見出す分析目的と一連の統計的指標の結果を踏まえると、モデルⅡを選択することが適当と考えられる。つまり、同一測定方法内の相関構造は類似しているが、測定方法が変われば等質性が満たされないことが示唆される。この点は Chernyshenko et al. (2009)とは異なるものである。

表 2　モデル I-モデル IV を回答データに当てはめた際の各種適合度指標および情報量規準

ﾓﾃﾞﾙ	母数の数	χ二乗値	自由度	RMR	GFI	CFI	RMSEA	AIC	BIC
ﾓﾃﾞﾙI	112	0	0	0	1	1	0.158	224.000	545.164
ﾓﾃﾞﾙII	56	139.657	56	0.099	0.934	0.922	0.054	251.657	412.239
ﾓﾃﾞﾙIII	56	238.621	56	0.154	0.897	0.830	0.079	350.621	511.203
ﾓﾃﾞﾙIV	28	337.141	84	0.175	0.857	0.765	0.076	393.141	473.432

*モデル I は飽和モデルに相当する。

表 3　モデル II から推定された尺度間相関行列
（左下対角成分が一対比較法，右上対角成分が評定尺度法による推定値）

	情動性	支配欲求	抑うつ性	シャイネス	粘り強さ	奉仕性	自己信頼
情動性		-0.236	0.552	0.402	-0.509	-0.336	-0.305
支配欲求	0.213		-0.527	-0.593	0.270	0.507	0.754
抑うつ性	-0.028	-0.207		0.698	-0.199	-0.345	-0.714
シャイネス	-0.294	-0.475	0.259		-0.319	-0.665	-0.570
粘り強さ	-0.237	-0.065	0.232	0.064		0.430	0.228
奉仕性	0.183	0.139	-0.023	-0.310	-0.010		0.314
自己信頼	0.217	0.419	-0.232	-0.248	-0.106	-0.115	

測定手法の違いにより測定している構成概念がどの程度変わりうるかが問題であるが、推定された相関行列からすると、一対比較では評定尺度に比べて全体的に相関が低くなっている（表 3）。これには、今回の一対比較データにおいて項目数が少ないことによる推定精度の問題やそれに伴う相関係数の希薄化の影響、さらには(1)式の統計モデル自体の複雑さや当てはまりの程度が影響した可能性も考慮すべきであるが、異なる測定手法間で評価している構成概念には質的な差異があると示唆される。

(3) フェイキングが潜在特性値の推定に与えるバイアスの大きさの評価

フェイキングが潜在特性値の推定に与えるバイアスを評価するため、ここでは項目反応モデルを 4 種類のデータにそれぞれ当てはめる際に識別力母数がデータ間で同じであることを仮定して推定し、各尺度内の潜在特性値の推定値の平均と分散を比較した。(2)の結果からもこの識別力母数の等値制約はやや強い仮定である事は考慮すべきであるが、測定方法の違いにより生じうるバイアスの量を大まかに推定する目的からは有効であろう。ここでは評定尺度・neutral の各尺度の潜在特性値の平均を 0、分散を 1 にして比較した（表 4）。ボールド体で示されている数字は、平均が 0 もしくは分散が 1 であるという帰無仮説が棄却された結果を意味する。

表 4 より、faking の影響がある場合、一対比較と評定尺度のいずれの手法においてもバイアスはみられるが、評定尺度・faking と比べると一対比較・faking においては、統計的に有意なバイアスが生じている尺度数は少なく、そのバイアスの大きさも相対的に小さい。

表 4　各回答データにおける、潜在特性値の平均と分散の推定値

		情動性	支配欲求	抑うつ性	シャイネス	粘り強さ	奉仕性	自己信頼
一対比較・faking	平均	-0.104	0.280	-0.012	**-0.598**	**0.550**	0.108	0.314
	分散	**2.300**	1.421	**2.914**	1.572	**2.456**	1.555	**3.003**
一対比較・neutral	平均	0.084	-0.299	0.116	0.299	0.171	-0.201	-0.101
	分散	1.482	**1.992**	**1.681**	1.559	**1.839**	1.627	**2.534**
評定尺度・faking	平均	**-0.405**	**1.455**	0.020	**-0.576**	**1.742**	**1.553**	**1.606**
	分散	**0.682**	**0.487**	1.083	**0.622**	0.926	1.100	1.041

まとめ

本研究の結果から、フェイキング場面において、一対比較形式のほうが評定尺度形式に比べて、正直に回答した場合の結果とより高い相関関係を有し、またフェイキングによる潜在特性値の推定のバイアスに対してもより頑健であった。この点は今回の一対比較形式に基づく測定手法の有効性を示す一方で、一対比較形式と評定尺度形式では評価している構成概念には質的な差異がある可能性も示された。一対比較形式における尺度間の内部相関が相対的に低くなるという点は Stark et al. (2011)とも一致するが、そのことが真の傾向であるのか、或いは今回の回答データや統計モデル特有の影響を受けた結果であるのかを判断することは現時点では難しいであろう。そのため、今後に向けては、項目数やサンプルサイズを増やし、より大規模なデータを利用して追加検証を行うことや、一対比較データを表現する他の統計モデルの可能性を検討することが第一に求められる。

引用文献

Chernyshenko,O.S.,Stark,S.,Prewett, M. S., Gray, A. A., Stilson,F.R.,&Tuttle, M.D.(2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Emprical comparisons with other formats. *Human Performance*, 22, 1-23.

Guilford,J.P.(1954). Psychometric methods(2nd ed.). New York: McGraw-Hill.

Hicks,L.E.(1970).Some properties of ipsative, normative and forced choice normative measures. *Psychological Bulletin,* 74, 167-184.

Johnson, C. E., Wood, R. & Blinkhorn, S. F.(1988).Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology*, 61,153-162.

Muraki, E. (1992). A generalized partial credit model : Application of an EM algorithm. *Applied Psychological Measurement , 17*, 351-363.

Roberts, J.S., Donoghue, J.R., & Laughlin, J.E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

Stark,S.,Chernyshenko,O.S., & Drasgow,F.(2005).An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: An application to the problem of faking in personality assessment. *Applied Psychological Measurement,* 29, 184-201.

Stark, S., Chernyshenko, O.S., & Drasgow, F. (2011). Constructing fake-resistant personality tests using item response theory: High stakes personality testing with multidimensional pairwise preferences. In Matthias Ziegler, Carolyn MacCann, & Richard D. Roberts (Eds.). *New Perspectives on Faking in Personality Assessments*. NY: Oxford University Press.

Usami,S. (2011). Generalized graded unfolding model with structural equation for subject parameters. *Japanese Psychological Research*, *53*, 221-232.