

項目反応理論によるパフォーマンス測定ツールの 項目分析の試み

Item analysis of performance measurement tool
by item response theory

西 田 豊 昭*

Toyoaki NISHIDA

1. はじめに

人事アセスメントツールは、採用・昇進昇格といった選抜を目的として、また教育・育成を目的として広く利用されている。実際、企業の人事・教育担当者は、その企業内での役割（人材の選抜・配置・育成など）を果たすために各種の人事アセスメントのツールを導入・活用するだけではなく経営者やそのツールを受験しフィードバックを受ける従業員に対して、ツールの内容について説明をする責任が生じる。しかし、パフォーマンスを測定するアセスメントツールは、その開発の経緯、開発過程、保証すべき信頼性・妥当性について十分に知ること非常に困難であるといえる。

一般的に、人事アセスメントツールにおいて、同じ批評家者に対してどの評価者も同一の評価を行うことは困難であり、これまで、これらは評価者の評定誤差として捉えられてきた。

一方、従業員のパフォーマンス評価、入社試験、教育効果評価などで得られたデータの解析には、これまで一般的に素点の「平均値 (mean)」「標準偏差 (standard deviation : SD)」「分散 (variance)」「通過率 (proportion correct)」などの統計量が使用されてきた。テスト理論において、これらはいずれも「古典的テスト理論 (Classical Test Theory : CTT)」に依拠して用いられるもので、これまで多くの研究者達によって議論されてきた概念である。このように、データ解析の手段として一般的に使用されてきたこうした方法に、再検討の余地があるということが、最近、主張されるようになってきた (Embretson & Reise, 2000)。それは、このような古典的テスト理論に依拠した方法にはいくつかの問題点があるためである。

例えば、人事評価において10項目中6項目で「～できる」「十分能力がある」「実行

*経営総合学科－組織心理学

Department of Management Synthesis – Organizational Psychology

している」など肯定的な評価を受けた場合、60点という点数が個人の能力や成績として示されるとする。このようにチェックされた項目数をそのまま点数として示される得点は「素点 (raw score)」と呼ばれている。しかし、このように「肯定された評価項目の数に基づいた得点 (number-right score)」には、1) 同一測定単位の問題、2) 項目弁別力指数が不明確という問題、3) 平行テスト作成の問題、4) 被評価者全体に対する測定の精度の問題、5) 被評価者の現状の能力に合わない評価項目の問題など多くの問題¹があると指摘されている (Embretson & Reise, 2000)。

本研究では、ある企業の従業員に対して実施した調査データを元に、インタビュー調査に基づいて作成されたパフォーマンス測定尺度項目に対して新たな理論に基づく項目分析を実施するとともに、パフォーマンス測定尺度の尺度の構成を行い、古典的テスト理論に依拠しない方法により測定尺度の開発を試みる。

なお、本研究では項目分析において項目反応理論 (Item Response Theor, 以下IRT) を用いた分析を行う。IRTを用いることにより、これまでの古典的テスト理論では行うことができなかった分析が可能になるからである。本研究の構成としては、多値型IRTモデルを用いることで、項目の特徴を精査することができる。また、テスト情報量による尺度の精度をきめ細かく記述することが可能となる。以上の点から、今回は項目分析にIRTを用いることとした。

2. 調査方法

(1) 尺度

本研究では、従業員の総合的なパフォーマンスを診断する測定尺度として、複数の企業において、仕事上成果をおさめているハイパフォーマーと一般の社員を対象に、インタビューを実施し、その結果をもとに、実際の職務における行動、スキル、態度、思考、性格を測定する新たな測定尺度を作成した。インタビューから得られた回答をもとに同一の意味内容をもつものを1項目としてまとめる作業を行い、全体として回答が行動50項目、スキル50項目、態度50項目、思考50項目、性格50項目の合計250項目に整理された。

(2) 調査手続きおよび回答者の構成

本研究のデータは、関西地区と九州地区の総合人材サービス企業、技術系派遣会社に所属する従業員305名 (男性206名 女性99名) から得られたものである。調査方法は、協力企業の社員がWeb上の評価システムを使用して評価が行われ、集計された。

¹ ①同一測定単位の問題、②項目弁別力指数が不明確という問題、③平行テスト作成の問題、④被評価者全体に対する測定の精度の問題、⑤被評価者の現状の能力に合わない評価項目の問題などが指摘されている。

3. パフォーマンス測定尺度のIRT分析

(1) 次元性の確認

今回使用したパフォーマンス測定尺度のIRT分析を行う前に、その前提条件となる尺度の1次元性を確認しておく必要がある。

まず、作成した行動項目、スキル項目、態度項目、思考項目、性格項目の内的一貫性としての信頼性の推定を行い、続いて、因子分析による次元性の検討を行った。

内的一貫性としての信頼性には、Cronbach (1951) の α 係数を採用した。その結果、尺度の信頼性係数は、行動 (.97)、スキル (.97)、態度 (.96)、思考 (.97)、性格 (.92) であり、いずれも尺度の信頼性が十分高いことが明らかとなった。

次に、有効回答として得られた305名分のデータを用いて、主因子解を用いた探索的因子分析を行った。分析の結果、各尺度の第1因子の累積寄与率は、行動尺度では39.174%、スキル尺度では39.558%、態度尺度では36.614%、思考尺度では41.106%、性格尺度では24.290%であった。これらの値は、Reckase (1979) によって示された20%という基準よりも高い値である。また、各尺度の累積寄与率をスクリープロットした結果、第1因子の以後の因子によって説明される分散の割合が急激に低下していることがわかった。そこで、行動尺度、スキル尺度、態度尺度、思考尺度、性格尺度は、IRT分析の前提条件である1次元性が確認された。

以上の結果から、本研究で使用するどのデータに対してもIRT分析の前提条件の1つである尺度の1次元性の仮定が満たされたことになる。また、その他のIRT分析の前提条件である局所独立の仮定に関しても、尺度を構成するどの項目も他の項目への回答の影響を受けない構造となっていることから、局所独立の仮定も満たされているといえよう。

(2) IRT分析

2パラメタ・ロジスティック・モデルに基づいて、周辺最尤法を用いてBILOG-MG3で項目パラメタ値の推定を行った。分析に際しては項目反応に対して肯定的な2件（やや行っている、十分行っている）に2を与え、否定的な2件（やや行っていない、十分行っていない）に1を与えるといった2値化を行った。

表1から表5は、行動尺度、スキル尺度、態度尺度、思考尺度、性格尺度それぞれの尺度ごとに、2パラメタ・ロジスティック・モデルに基づいてパラメタ値の推定を行った結果である。表1から表5には識別力パラメタ値 a 、困難度パラメタ値 b 、肯定反応の比率 p が示されている。識別力および困難度の水準については様々な議論があるが、ここではRoznowski (1989) の研究を参考に ($a > 0.50$)、($-4.00 < b < +4.00$) を基準として用いることにした。その結果、行動50項目、スキル50項目、態度50項目、思考50項目、性格50項目の全項目においてRoznowski (1989) の識

別力の基準である $a > 0.50$ を満たしていることが明らかとなった。また、困難度パラメタに関しても、すべての項目が $-4.00 < b < +4.00$ という基準を満たしていた。

この結果、パフォーマンス測定尺度の下位尺度として行動（50項目）、（スキル50項目）、（態度50項目）、（思考50項目）、性格50項目が完成した。

表1 行動測定尺度：
項目パラメタの推定値

項目	識別力パラメタ(a)	困難度パラメタ(b)	通過率(p)
1	1.668	-3.641	86.4
2	2.164	-0.839	69.8
3	2.185	-1.423	65.1
4	2.209	-1.704	82.2
5	2.011	-0.479	61.5
6	1.397	-1.181	81.7
7	1.975	-0.515	65.1
8	1.111	-1.610	81.1
9	1.209	0.167	46.2
10	1.322	-0.929	71.6
11	1.435	-2.025	97
12	1.548	-3.121	122.4
13	1.661	-4.217	147.8
14	1.774	-5.313	173.2
15	1.887	-6.409	198.6
16	2.000	-7.505	224
17	2.113	-8.601	249.4
18	2.226	-9.697	274.8
19	2.339	-10.793	300.2
20	2.452	-11.889	325.6
21	2.565	-12.985	351
22	2.678	-14.081	376.4
23	2.791	-15.177	401.8
24	2.904	-16.273	427.2
25	3.017	-17.369	452.6
26	3.130	-18.465	478
27	3.243	-19.561	503.4
28	3.356	-20.657	528.8
29	3.469	-21.753	554.2
30	3.582	-22.849	579.6
31	3.695	-23.945	605
32	3.808	-25.041	630.4
33	3.921	-26.137	655.8
34	4.034	-27.233	681.2
35	4.147	-28.329	706.6
36	4.260	-29.425	732
37	4.373	-30.521	757.4
38	4.486	-31.617	782.8
39	4.599	-32.713	808.2
40	4.712	-33.809	833.6
41	4.825	-34.905	859
42	4.938	-36.001	884.4
43	5.051	-37.097	909.8
44	5.164	-38.193	935.2
45	5.277	-39.289	960.6
46	5.390	-40.385	986
47	5.503	-41.481	1011.4
48	5.616	-42.577	1036.8
49	5.729	-43.673	1062.2
50	5.842	-44.769	1087.6

表2 スキル測定尺度：
項目パラメタの推定値

項目	識別力パラメタ(a)	困難度パラメタ(b)	通過率(p)
1	0.535	-3.641	86.4
2	1.335	-0.839	69.8
3	0.459	-1.423	65.1
4	1.101	-1.704	82.2
5	1.300	-0.479	61.5
6	2.102	-1.181	81.7
7	1.975	-0.515	65.1
8	1.111	-1.610	81.1
9	1.209	0.167	46.2
10	1.322	-0.929	71.6
11	1.435	-2.025	97
12	1.548	-3.121	122.4
13	1.661	-4.217	147.8
14	1.774	-5.313	173.2
15	1.887	-6.409	198.6
16	2.000	-7.505	224
17	2.113	-8.601	249.4
18	2.226	-9.697	274.8
19	2.339	-10.793	300.2
20	2.452	-11.889	325.6
21	2.565	-12.985	351
22	2.678	-14.081	376.4
23	2.791	-15.177	401.8
24	2.904	-16.273	427.2
25	3.017	-17.369	452.6
26	3.130	-18.465	478
27	3.243	-19.561	503.4
28	3.356	-20.657	528.8
29	3.469	-21.753	554.2
30	3.582	-22.849	579.6
31	3.695	-23.945	605
32	3.808	-25.041	630.4
33	3.921	-26.137	655.8
34	4.034	-27.233	681.2
35	4.147	-28.329	706.6
36	4.260	-29.425	732
37	4.373	-30.521	757.4
38	4.486	-31.617	782.8
39	4.599	-32.713	808.2
40	4.712	-33.809	833.6
41	4.825	-34.905	859
42	4.938	-36.001	884.4
43	5.051	-37.097	909.8
44	5.164	-38.193	935.2
45	5.277	-39.289	960.6
46	5.390	-40.385	986
47	5.503	-41.481	1011.4
48	5.616	-42.577	1036.8
49	5.729	-43.673	1062.2
50	5.842	-44.769	1087.6

表3 態度測定尺度：
項目パラメタの推定値

項目	識別力パラメタ(a)	困難度パラメタ(b)	通過率(p)
1	0.535	-3.641	86.4
2	1.335	-0.839	69.8
3	0.459	-1.423	65.1
4	1.101	-1.704	82.2
5	1.300	-0.479	61.5
6	2.102	-1.181	81.7
7	1.975	-0.515	65.1
8	1.111	-1.610	81.1
9	1.209	0.167	46.2
10	1.322	-0.929	71.6
11	1.435	-2.025	97
12	1.548	-3.121	122.4
13	1.661	-4.217	147.8
14	1.774	-5.313	173.2
15	1.887	-6.409	198.6
16	2.000	-7.505	224
17	2.113	-8.601	249.4
18	2.226	-9.697	274.8
19	2.339	-10.793	300.2
20	2.452	-11.889	325.6
21	2.565	-12.985	351
22	2.678	-14.081	376.4
23	2.791	-15.177	401.8
24	2.904	-16.273	427.2
25	3.017	-17.369	452.6
26	3.130	-18.465	478
27	3.243	-19.561	503.4
28	3.356	-20.657	528.8
29	3.469	-21.753	554.2
30	3.582	-22.849	579.6
31	3.695	-23.945	605
32	3.808	-25.041	630.4
33	3.921	-26.137	655.8
34	4.034	-27.233	681.2
35	4.147	-28.329	706.6
36	4.260	-29.425	732
37	4.373	-30.521	757.4
38	4.486	-31.617	782.8
39	4.599	-32.713	808.2
40	4.712	-33.809	833.6
41	4.825	-34.905	859
42	4.938	-36.001	884.4
43	5.051	-37.097	909.8
44	5.164	-38.193	935.2
45	5.277	-39.289	960.6
46	5.390	-40.385	986
47	5.503	-41.481	1011.4
48	5.616	-42.577	1036.8
49	5.729	-43.673	1062.2
50	5.842	-44.769	1087.6

表4 思考測定尺度：
項目パラメタの推定値

項目	識別力パラメタ(a)	困難度パラメタ(b)	通過率(p)
1	0.535	-3.641	86.4
2	1.335	-0.839	69.8
3	0.459	-1.423	65.1
4	1.101	-1.704	82.2
5	1.300	-0.479	61.5
6	2.102	-1.181	81.7
7	1.975	-0.515	65.1
8	1.111	-1.610	81.1
9	1.209	0.167	46.2
10	1.322	-0.929	71.6
11	1.435	-2.025	97
12	1.548	-3.121	122.4
13	1.661	-4.217	147.8
14	1.774	-5.313	173.2
15	1.887	-6.409	198.6
16	2.000	-7.505	224
17	2.113	-8.601	249.4
18	2.226	-9.697	274.8
19	2.339	-10.793	300.2
20	2.452	-11.889	325.6
21	2.565	-12.985	351
22	2.678	-14.081	376.4
23	2.791	-15.177	401.8
24	2.904	-16.273	427.2
25	3.017	-17.369	452.6
26	3.130	-18.465	478
27	3.243	-19.561	503.4
28	3.356	-20.657	528.8
29	3.469	-21.753	554.2
30	3.582	-22.849	579.6
31	3.695	-23.945	605
32	3.808	-25.041	630.4
33	3.921	-26.137	655.8
34	4.034	-27.233	681.2
35	4.147	-28.329	706.6
36	4.260	-29.425	732
37	4.373	-30.521	757.4
38	4.486	-31.617	782.8
39	4.599	-32.713	808.2
40	4.712	-33.809	833.6
41	4.825	-34.905	859
42	4.938	-36.001	884.4
43	5.051	-37.097	909.8
44	5.164	-38.193	935.2
45	5.277	-39.289	960.6
46	5.390	-40.385	986
47	5.503	-41.481	1011.4
48	5.616	-42.577	1036.8
49	5.729	-43.673	1062.2
50	5.842	-44.769	1087.6

表5 性格測定尺度：項目パラメタの推定値

項目	識別力パラメタ(a)	困難度パラメタ(b)	通過率(p)
1	0.535	-3.641	86.4
2	1.335	-0.839	69.8
3	0.459	-1.423	65.1
4	1.101	-1.704	82.2
5	1.300	-0.479	61.5
6	2.102	-1.181	81.7
7	1.975	-0.515	65.1
8	1.111	-1.610	81.1
9	1.209	0.167	46.2
10	1.322	-0.929	71.6
11	1.435	-2.025	97
12	1.548	-3.121	122.4
13	1.661	-4.217	147.8
14	1.774	-5.313	173.2
15	1.887	-6.409	198.6
16	2.000	-7.505	224
17	2.113	-8.601	249.4
18	2.226	-9.697	274.8
19	2.339	-10.793	300.2
20	2.452	-11.889	325.6
21	2.565	-12.985	351
22	2.678	-14.081	376.4
23	2.791	-15.177	401.8
24	2.904	-16.273	427.2
25	3.017	-17.369	452.6
26	3.130	-18.465	478
27	3.243	-19.561	503.4
28	3.356	-20.657	528.8
29	3.469	-21.753	554.2
30	3.582	-22.849	579.6
31	3.695	-23.945	605
32	3.808	-25.041	630.4
33	3.921	-26.137	655.8
34	4.034	-27.233	681.2
35	4.147	-28.329	706.6
36	4.260	-29.425	732
37	4.373	-30.521	757.4
38	4.486	-31.617	782.8
39	4.599	-32.713	808.2
40	4.712	-33.809	833.6
41	4.825	-34.905	859
42	4.938	-36.001	884.4
43	5.051	-37.097	909.8
44	5.164	-38.193	935.2
45	5.277	-39.289	960.6
46	5.390	-40.385	986
47	5.503	-41.481	1011.4
48	5.616	-42.577	1036.8
49	5.729	-43.673	1062.2
50	5.842	-44.769	1087.6

4. 考察

今回作成したパフォーマンス測定尺度の下位尺度における項目パラメタ値の推定結果から言える事は、各項目の識別力パラメタaの値は高く、それぞれの項目が個人のパフォーマンスを識別する上で感度のよいことが明らかとなった。また、識別力パラメタaの値が、Roznowski（1989）の識別力の基準である $a > 0.50$ を下回ったものや、困難度パラメタbについても、極端に困難度が高かったり低かったりといった項目もなく、標準的な従業員のパフォーマンスの測定に十分適していると判断できる。

5. 結論

インタビュー調査に基づいて作成されたパフォーマンス測定尺度項目に対して新たな理論に基づく項目分析を実施し、新たなパフォーマンス測定尺度の開発を試みた。

その結果として、識別力パラメタ (a) や、困難度パラメタ (b) を確認することで、効率的に項目分析を行うことが可能であり、また、完成したパフォーマンス測定尺度が、標準的な従業員のパフォーマンスの測定に適していることも明らかとなった。

今後の課題としては、作成したおパフォーマンス測定尺度が、個人の業績を測定する上で、安定した外的基準になり得るかどうかについて調べる必要があるということである。つまり、人事考課を用いた、基準連関妥当性の確認を実施したり、より広範な妥当性の検証を行うために、横断的なデータを用いた予測的妥当性 (predictive validity) などとも検討していく必要があるだろう。

6. 参考文献

- Angoff, W. H. (1993). "Perspectives on differential item functioning methodology." In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*, 3-23. Hillsdale, NJ: Lawrence Erlbaum.
- Ashford, S. J. (1989). Self-assessment in organizations: A literature review and integrative model. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior*. Greenwich, CT: JAI Press, 11, 113-174.
- Bass, B. M. (1990). *Bass and Stogdill's handbook of leadership*. New York: Free Press.
- Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias*. Baltimore: Johns Hopkins University Press.
- Borman, W. C. (1979). Format and training effects on rating accuracy and rating errors. *Journal of Applied Psychology*, 64, 410-421.
- Cascio, W. F. (1991). *Applied psychology in personnel management (4th ed.)*. Englewood Cliffs, NJ: Prentice Hall.
- Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology*, 74, 130-135.
- Collins, W. C., Raju N. S., & Edwards, I. E. (1997, April). Assessing DIF in a polytomously-scored satisfaction scale. Symposium paper presented at the 1997 SIOP annual convention in St. Louis, MO.
- Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218-244.
- Drasgow, F. & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, 70, 662-680.
- Drasgow, F. (1982). Biased test items and differential validity. *Psychological Bulletin*, 92, 526-531.
- Drasgow, F. (1987). Study of the measurement of bias of two standardized psychological tests. *Journal of Applied Psychology*, 72, (1), 19-29.
- Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory For Psychologists*, New Jersey: LEA.
- Engelhard, C., & Wilson, M. (1996). (Eds.). *Objective measurement III: Theory into practice*. Norwood, NJ: Ablex.
- Flanagan, W. J. (1997). *Measurement equivalence between high and average impression management groups: An IRT analysis of personality factors*. Unpublished doctoral

- dissertation, Georgia Institute of Technology, Atlanta, GA.
- 池田央 (1994). 現代テスト理論, 朝倉書店.
 - 楠田丘 (1981). 人事考課の手引き. 日本経済新聞社.
 - Ledvinka, J. & Scarpello, V. G. (1991). *Federal regulation of personnel and human resource management*. Belmont CA: Wadsworth inc.
 - Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate appraisal measurement equivalence. *Journal of Applied Psychology*, 83, (5), 693-702.
 - Mulqueen, C. M., Stahl, J., & Gershos R. C. (1996). *A Multifacet Rasch Analysis of 360-degree Performance Appraisal Results*. Paper presented at SIOP annual conference, St. Louis, MO.
 - Reckase, M. D. (1979). Unifactor latent trate models applied to multi-factor tests: Results and implications. *Joyrnal of Educational Statistics*. 4, 207-230.
 - Reise, S. P., & Wailer, N. C. (1990). Fitting the two-parameter model to personality data: The parameterization of the Multidimensional Personality Questionnaire. *Applied Psychological Measurement*, 14, 45-58.
 - Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, pp. 413-428.
 - Samejima, F. (1972). A general model for free-response data. *Psychometric Monograph*, 18.
 - Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, 38, 203-219.
 - Samejima, F. (1996). The graded response model. In W. J. van der Linden & Hamblton, R. K.(Eds.), *Handbook of modern item response theory*. New York: Springer.
 - Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*. 17.
 - 関本昌秀 (1983). 適正多面評価の安定性と妥当性に関する一考察 慶応経営論集, 4(2), 1-17.
 - 芝祐順 (1991). 項目反応理論 —基礎と応用—, 東京大学出版.
 - Smith, P. C. , & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 47, 149-155.
 - Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of Satisfaction in Work and Retirement*. Chicago IL: Randy McNally.
 - Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In Shrout, P. E., & Fiske, S. T. (Eds.). *Personality Research Methods, and Theory: A Festschrift Honoring: Donatd W. Fiske*, pp. 161-181. Hillsdale NJ: Lawrence Erlbaum Associates.
 - Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, 49, 501-519.
 - Thissen, D., Steinberg, L., & Wainer. H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity*. 147-169. Hillsdale, NJ: Lawrence Erlbaum Association
 - 渡辺直登・野口裕之(1999). 組織心理測定論 —項目反応理論のフロンティア—, 白桃書房.
 - Zimowski, M. F., Muraki, E., Robert J. Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items*. Scientific Software.