

項目反応理論に基づく翻訳等価性分析のアプローチ

Translation equivalence analysis based on item response theory

西 田 豊 昭*

Toyoaki NISHIDA

1. はじめに

経済活動のグローバル化に伴い、言語的・文化的背景の異なる従業員を採用したり、配置したり、教育・訓練を行ったり、さらには考課を行う機会が増加している。こうしたプラクティスの一翼を担うのが人的資源アセスメントである。国際的なEEO論議や、有能人材の獲得競争が高まる中、グローバルなコンテキストで人的資源のアセスメントをいかに行うかは、企業にとって悩みの深い課題となってきた。これは、海外に直接投資を行っている企業だけの課題ではない。国内にあって母語や文化・習慣の異なる従業員の人的資源管理を行っている企業についてもいえることである。

グローバル化時代の人的資源アセスメントでは、言語的・文化的背景の異なる人々の持つさまざまな属性を、精度の高いツールで測定することが求められる (Hulin, Drasgow, & Komocar, 1982; Hulin & Mayer, 1986; Ellis & Kimmel, 1992)。しかし、わが国で用いられている「言語による」人的資源アセスメント・ツールの多くは、グローバルな、より控えめに言えばクロス・カルチャルな (交差文化的な) 場面での使用さえほとんど想定せずに開発されている。内田クレベリン検査に代表されるような、ごくわずかな「非言語的」アセスメント・ツールが、唯一グローバル時代に耐えうるものとしてあげられるのみである。

もちろん、外国、特に米国で開発されたアセスメント・ツールを日本語に翻訳し、日本人の被験者を用いて標準化し、実用に供しているケースは多く見られる。しかし、そうした日本語版アセスメント・ツールが、英語と日本語を話す従業員にとって真に「等価 (equivalent)」であるかどうかは疑わしい。これまで日本語版開発にあたって、手間隙のかかる原版と翻訳版間の精緻な等価性分析はほとんど行われてこなかったし、またグローバル化がこれほどまでに進展していない時代にあっては、その必要性はほとんどなかった。「日本語版」はまさに、日本の言語的・文化的環境にふさわ

* 経営総合学科－組織心理学

Department of Management Synthesis – Organizational Psychology

しく「日本化」して開発・使用されてきたのである。

本研究では、項目反応理論に基づく翻訳等価性分析のアプローチを応用し、このような国際間の測定的等価性の問題を解決するための方法について論じる。

2. 16PF（第5版）について

本研究の対象としたのは、50余年前にR. Cattellによって開発され、世界30カ国以上の言語に翻訳され使用されている16PFパーソナリティ質問紙（Sixteen Personality Factor Questionnaire）である（以降16PFと略記）。本研究では、その第5版（185項目）の元となったForm S 249項目（Cattell, Cattell, & Cattell, 1993）と、B因子（推理）に関するカルチャー・フリー項目群16項目を加えた、合計265項目を用いた。

16PFについては、日本でも1982年に第4版が開発され日本文化科学社から発売されている。第5版は1993年に行われた大規模な改訂の結果誕生したもので、①質問項目の表現を現代的な米語に大きく変えたこと、②中間反応をすべて「？」で統一したこと、③5つのグローバル因子（基礎の16因子の2次因子分析したもの）が同定されており、Big5概念と対応していること、④産業場面での使用を念頭に開発されていること、⑤標準化のためのサンプルとして、国勢調査結果を参考に厳密なサンプリングが行われていること、⑥第4版まで項目の中に埋め込まれていた嘘尺度を独立させ、新たに印象管理（Impression Management：IM）尺度を設けたこと、などが特徴としてあげられる。

Cattellはパーソナリティを表す言葉に対する自己評価や他者評価データを因子分析することにより、質問項目のデータを、16種類の基本的尺度に分類している（表1）（Cattell, 1957）。

本研究では16種類の基本因子尺度に、印象管理（Impression Management：IM）を測定する尺度を加えた、計17の尺度を研究対象とした。

16PFでは、16尺度中15尺度と印象管理（IM）を測定する尺度については、二者択一の質問項目に対して、aかcのどちらかの回答を選ぶようになっており、どうしても答えられない場合はbの“？”を選択するようになっている。16尺度中の1尺度（B因子：推理）だけは、知的能力を測定する質問からなっており、3つの選択肢a, b, c, の中から1つの正解を選ぶ形式となっている。

表1 Sixteen Personality Factor (16PF) Questionnaire

尺度	
F1	Warmth (A) 親近
F2	Reasoning (B) - Problem solving questions 推理
F3	Emotional Stability (C) 適応
F4	Dominance (E) 支配
F5	Liveliness (F) 躍動
F6	Rule Consciousness (G) 規則
F7	Social Boldness (H) 大胆
F8	Sensitivity (I) 感度
F9	Vigilance (L) 警戒
F10	Abstractness (M) 抽象
F11	Privateness (N) 隔絶
F12	Apprehension (O) 懸念
F13	Openness to Change (Q1) 変革
F14	Self-Reliance (Q2) 自立
F15	Perfectionism (Q3) 完璧
F16	Tension (Q4) 緊張
IM	Impression Management (IM)

3. 被調査者

本研究では、16PF日本語版作成にあたり、米国でオリジナル版の16PFを販売するIPATから、アメリカ人のデータ3,755名分データの提供を受け、それを参照グループのデータとして使用した。また、日本人のデータは大手企業（総合電気メーカー、自動車メーカー、製薬メーカー）と、大学生、大学院生などから収集したデータ5,720名分を使用した。日米合わせたサンプル数の合計は9,475名であった。

4. 英語版の翻訳

まず始めに、すべての項目について英語から日本語への翻訳作業を行った。翻訳にあたっては準再翻訳法（quasi-back translation）を用いた。手順は以下のとおりである。

- ① 16PFについて何もその背景を知らないプロの翻訳家に、日本語への翻訳を依頼。
- ② 知的能力を測定するB因子に含まれる言語的に特殊な項目は、日本語で書かれた問題で代替。

- ③ プロの翻訳家が訳した結果を、日本人心理学者に提示し、適当でない表現の修正を依頼。
- ④ プロの翻訳家が訳した結果を、別の日本人のプロの翻訳家に提示し、表現の修正を依頼。
- ⑤ 日本人バイリンガルに上記③④の結果を提示し、さらに修正を依頼。
- ⑥ 本人心理学者2人と、日本人心理計量学者1人が最終的な修正・調整を行った。

翻訳にあたっては、2名以上の人がテストの翻訳に携わることが推奨されていることから (Hambleton & Kanjee, 1995; Van de Vijver & Hambleton, 1996), 本研究では6人が翻訳に携わった。

5. 分析方法と結果

【第1段階：内的整合性による信頼性の検討】

- (1) 1回目日本トライアルデータ (N=193) と米国データ (N=3,755) の比較

日本語に翻訳された最初のバージョンの尺度の特性を調べるために、パイロット・スタディーが行われた。日本人193名のデータは、大学の卒業生と学部生から得られたものである。表2はそれぞれの尺度の信頼性係数 (α 係数) を示している。(尺度は英語のオリジナルのバージョンの信頼性係数との比較で記載されている)。表2に示される結果から、①それぞれの尺度の内的一貫性はかなり高い。②しかし、17尺度のうち、3尺度がIPATが国際的に要求する基準 ($\alpha = .60$) を満たしていない (すなわち、B, I, およびIMの3つの尺度は改善の余地がある)。3尺度について信頼性係数が低くなったのは、サンプルサイズが小さいことも原因であると考え、このまま項目を変更せずに、より大きいサンプルを使用して2回目の調査を実施した。

表2 信頼性分析 (1 回目日本人
トライアルデータ)

	US	JAPAN
A	0.722	0.714
B	0.725	0.470
C	0.823	0.687
E	0.749	0.775
F	0.773	0.789
G	0.806	0.785
H	0.878	0.865
I	0.798	0.436
L	0.786	0.792
M	0.767	0.718
N	0.825	0.799
O	0.803	0.852
Q1	0.694	0.664
Q2	0.795	0.798
Q3	0.796	0.804
Q4	0.782	0.755
IM	0.600	0.382

US: N=3755 JAPAN: N=193

(2) 2回目日本大規模データ (N=4,591) と米国データの比較 (N=3,755)

2回目の調査データは、大手電機メーカーとその子会社および関連企業の従業員4,591名から得られたものである。表3はそれぞれの尺度の信頼性係数 (α 係数) を示している。この結果を見ると、若干の係数の改善は見られるものの、1回目のトライアルデータと同様に、依然としてB, I, およびIMの3つの尺度において要求されている基準である $\alpha = 0.60$ を下回っていることが明らかとなった。しかし、BとIの2つの尺度に関しては、その中のいくつかの項目を取り除くことにより、要求されている基準を満たすことも明らかとなった。

表3 信頼性分析 (2回目日本人大規模データ)

The Second Exploratory-Study			The First Pilot-Study		
	US	JAPAN		US	JAPAN
A	0.722	0.746	A	0.722	0.714
<u>B</u>	0.725	<u>0.529</u>	<u>B</u>	0.725	<u>0.470</u>
C	0.823	0.777	C	0.823	0.687
E	0.749	0.801	E	0.749	0.775
F	0.773	0.758	F	0.773	0.789
G	0.806	0.686	G	0.806	0.785
H	0.878	0.898	H	0.878	0.865
<u>I</u>	0.798	<u>0.553</u>	← <u>I</u>	0.798	<u>0.436</u>
L	0.786	0.724	L	0.786	0.792
M	0.767	0.722	M	0.767	0.718
N	0.825	0.766	N	0.825	0.799
O	0.803	0.805	O	0.803	0.852
Q1	0.694	0.675	Q1	0.694	0.664
Q2	0.795	0.809	Q2	0.795	0.798
Q3	0.796	0.786	Q3	0.796	0.804
Q4	0.782	0.764	Q4	0.782	0.755
<u>IM</u>	0.600	<u>0.429</u>	<u>IM</u>	0.600	<u>0.382</u>
US:N=3755 JAPAN:N=4591			US:N=3755 JAPAN:N=193		

(3) 再翻訳の実施

大量のデータを収集したにもかかわらず、日本語版B, I, IMの3つの信頼性が十分に高いとは言えないことが明らかになったため、日本語版の英語への再翻訳を行った。手順は以下のとおりであった。

- ① IPATが米国に住むバイリンガルの翻訳家に、日本語版の英語への再翻訳を依頼。
- ② 米国IPATの研究者2名が、再翻訳されたものとオリジナル版の意味的な齟齬

を指摘し、われわれ（日本側）に通知。

- ③ 日本側2人の研究者が再度日本語版を見直し、項目を修正。

- (4) 項目改訂後、3回目日本大規模データ (N=936) と米国データの比較 (N=3,755)

改訂版を用いて、3回目の調査を実施した。表4は、2回目の調査と、項目改定後の3回目の調査における各尺度の信頼性係数 (α 係数) を示している。この結果を見ると、3回目の分析結果では、2回目の分析結果から、約半数の尺度において信頼性係数に改善が見られた。特にI尺度の問題は解決された。また、BとIM尺度における問題もほとんど改善された。

表4 信頼性分析 (3回目日本人大規模データ)

The Third Confirmatory-Study			The Second Exploratory-Study		
	US	JAPAN		US	JAPAN
A	0.722	0.802	A	0.722	0.746
B	0.725	0.553	B	0.725	0.529
C	0.823	0.769	C	0.823	0.777
E	0.749	0.761	E	0.749	0.801
F	0.773	0.758	F	0.773	0.758
G	0.806	0.686	G	0.806	0.686
H	0.878	0.888	H	0.878	0.898
I	0.798	0.601	I	0.798	0.553
L	0.786	0.741	L	0.786	0.724
M	0.767	0.697	M	0.767	0.722
N	0.825	0.767	N	0.825	0.766
O	0.803	0.801	O	0.803	0.805
Q1	0.694	0.630	Q1	0.694	0.675
Q2	0.795	0.801	Q2	0.795	0.809
Q3	0.796	0.795	Q3	0.796	0.786
Q4	0.782	0.748	Q4	0.782	0.764
IM	0.600	0.499	IM	0.600	0.429
US:N=3755 JAPAN:N=939			US:N=3755 JAPAN:N=4591		

【第2段階：再検査法による信頼性の検討】

日本人の特定の被験者を対象に、再検査法による信頼性係数 (test-retest reliability) の推定を実施した。4週間のインターバルにおいて、同一の質問紙を同一の被験者に対して2度実施し、得られた得点間の相関係数から信頼性係数を推定した。再検査法による信頼性係数は、表5に示されている。この結果を見ると、B尺度を除く、他の16尺度の信頼性係数は0.72から0.92の間に分布しており、いずれの下

位尺度とも信頼性の高いことが明らかになった。

表5 日本語版16PF質問紙（第5版）の再検査信頼性係数

再検査信頼性係数	
A	0.862
B	0.430
C	0.752
E	0.723
F	0.923
G	0.886
H	0.894
I	0.797
L	0.843
M	0.770
N	0.879
O	0.838
Q1	0.850
Q2	0.834
Q3	0.907
Q4	0.896

JAPAN: N=48

【第3段階：項目反応理論によるパラメタ値の推定と最終採用項目の選択】

次に、各項目に関して項目反応理論（IRT）に基づく項目分析を行った。方法として、周辺最尤推定法による2パラメタ・ロジスティック・モデルを用い、BILOG-MG プログラム（Zimowski, Muraki, Mislevy, & Bock, 1996）で項目パラメタの推定を行った。項目分析では、aパラメタが低く（傾きが緩やか）、bパラメタ（困難度）が極端に高いか低い項目が削除の対象となった。そのうえで信頼性係数への影響を考慮し、最終的に採用される項目が決定された。さらに、最終的に採用された項目に対しても、IRTに基づく項目分析を行った。表6は最終項目を選択する前と後の分析結果を示している。

この結果を見ると、aパラメタ（識別力パラメタ）に関しては、最終的に選択されたバージョンの方が、選択前のバージョンよりも全ての尺度で改善されていた（大きな値を示した）。

bパラメタ（困難度パラメタ）に関しては、17尺度中9尺度で改善されていた。

以上の結果は、最終バージョンの方がそれぞれの尺度の識別力が増し、なおかつ、より一般的な困難度を測定できる尺度になっていることを示している。

表6 最終項目選択前と後のIRTに基づく項目分析の結果

項目選択後			項目選択前		
	a(slope)	b(threshold)		a(slope)	b(threshold)
A	<u>1.619</u>	0.130	A	1.424	0.115
B	<u>1.229</u>	<u>-2.982</u>	B	0.997	-3.205
C	<u>1.347</u>	-0.756	C	1.143	-0.517
E	<u>1.225</u>	0.054	E	1.177	0.020
F	<u>1.288</u>	<u>-0.551</u>	F	1.196	-0.858
G	<u>1.031</u>	<u>-0.069</u>	G	0.924	0.103
H	<u>2.149</u>	0.041	H	1.976	0.001
I	<u>0.845</u>	<u>0.430</u>	I	0.797	0.602
L	<u>1.412</u>	0.753	L	1.295	0.599
M	<u>1.138</u>	1.054	M	1.069	0.768
N	<u>1.513</u>	<u>0.591</u>	N	1.346	0.924
O	<u>1.435</u>	<u>0.015</u>	O	1.343	-0.018
Q1	<u>1.053</u>	<u>-0.785</u>	Q1	0.963	-0.886
Q2	<u>1.507</u>	<u>1.058</u>	Q2	1.447	1.157
Q3	<u>1.470</u>	-0.324	Q3	1.324	-0.237
Q4	<u>1.292</u>	0.973	Q4	1.212	0.647
IM	<u>0.880</u>	<u>1.385</u>	IM	0.832	1.933
JAPAN:N=939			JAPAN:N=939		

5. 考察

以上、本研究では項目反応理論に基づく翻訳等価性分析のアプローチを応用し、言語的・文化的に等価で、信頼性の高いツール開発の試みを提示した。以上、本研究の結果を要約すると以下ようになる。

- 1回目日本トライアルデータ（N=193）と米国データ（N=3,755）の比較では、それぞれの尺度の内的一貫性は十分高いが、17尺度のうち、3尺度だけがIPATが国際的に要求する基準（ $\alpha = .60$ ）を満たしていなかった。
- 2回目日本大規模データ（N=4,591）と米国データの比較（N=3,755）の比較では、若干の係数の改善は見られるものの、1回目のトライアルデータと同様に、依然としてB, I, およびIMの3つの尺度において要求されている基準である $\alpha = 0.60$ を下回っていることが明らかとなった。
- 項目改訂後、3回目日本大規模データ（N=936）と米国データの比較（N=3,755）では、2回目の分析結果と比較して、約半数の尺度において信頼性係数に改善が見られた。特にI尺度の問題は解決され、BとIM尺度における問題もほぼ改善された。

- d) 再検査法による信頼性係数 (test-retest reliability) の推定を実施した結果、B尺度を除く他の16尺度の信頼性係数は0.72から0.92の間に分布しており、いずれの下位尺度とも信頼性の高いことが明らかになった。
- e) 各項目に関して項目反応理論 (IRT) に基づきパラメタ値を推定し、最終採用項目を選択した結果、aパラメタ (識別力パラメタ) に関しては、最終的に選択されたバージョンの方が、選択前のバージョンよりも全ての尺度で改善され (大きな値を示した)、bパラメタ (困難度パラメタ) に関しても、17尺度中9尺度で改善されていた。

以上の結果は、本研究で使用した項目反応理論に基づくアプローチが国際比較研究を行う上で、翻訳された尺度や項目の測定的等価性を診断する方法として、有効であることを示しており、これまで古典的テスト理論に従い素点に基づく集計や結果の比較を行ってきた全ての国際比較研究、翻訳、項目分析の結果に対して再検討の必要性を示唆するものである。つまり、これまで実施されてきた、国際比較研究で使用されてきた尺度や翻訳された項目は、その大半が素点による集計・計算を基に、得点の高低が比較・検討が行われてきたが、その中で使用されてきた項目や測定尺度の中には測定的等価性が保たれていない項目が含まれている可能性があるということである。今後は、IRTに基づくDIF検出のアプローチを使用し、現在使用されている様々な翻訳された尺度や国際比較研究で使用されている尺度を全て再検討し、測定的等価性を確認する必要がある。

6. 参考文献

- Cattell, R. B. (1957). *Personality and motivation: Structure and measurement*. New York: World Book.
- Cattell, R. B., Cattell, A. K., & Cattell, H. E. P. (1993). *Sixteen Personality Factor Questionnaire Fifth Edition*. Champaign, IL: Institute for Personality and Ability Testing.
- Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology*, 77, 177-184.
- Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of scale translations. *Journal of Applied Psychology*, 67, 818-825.
- Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of the Job Descriptive Index into Hebrew. *Journal of Applied Psychology*, 71, 83-94.