Development of a Culturally and Linguistically Equivalent Personality Test: A case of the Japanese version of 16PF 5th edition

Naotaka Watanabe, Tomoko Ito, and Toyoaki Nishida

Abstract

In recent trend toward globalization in the field of business, workforce diversity has become a major issue in Human Resource Management. In order to deal with this issue, various types of psychological tests which can be used in multiple languages and multiple cultures have been administered in many HRM situations, such as the situation of employee recruitment, selection, orientation, training, appraisal, and so on. For psychologists who are engaged in test development, translating and adapting psychological tests across cultures and languages already became a common practice in the era of diversity workforce management. In this paper, the process of translating and adapting the US English version of Sixteen Personality Factor Questionnaire (16PF5) into Japanese version is presented. Results of reliability and validity studies which were conducted for examining cultural and linguistic equivalence of the test are also described.

Introduction

In line with the recent trend toward the globalization of business activities, a lot of cross-cultural practices have been done in the field of Human Resource Management (HRM) (Watanabe,1999). Combined with this trend is an increasing acceptance for using psychological tests for use in global HRM settings. These two factors are combining to create a demand for psychological tests that can be used in multiple languages and multiple cultures. In order to achieve high quality measurement by psychological test, we should resolve the serious problem of how we translate the original language into a target language to eventually develop culturally and linguistically equivalent test (Tracey, Watanabe, N., and Schneider,1997;Watanabe, Bedwell, and Williams, 2006). Such tests will contribute to make global HRM practices effective as well as to fulfill the legal requirement in global context, such as equal employment opportunity (EEO)(Long, Watanabe, and Tracey,2006;Watanabe,1988).

In this paper, we will pick up the Sixteen Personality Factor Questionnaire,16PF, which is one of the most widely used personality test in the world, and provide comprehensive information about what kind of process was taken for translating the US English version into the Japanese version in aiming at attaining cultural and linguistic equivalence.

Short History of the 16PF

The Sixteen Personality Factor Questionnaire,16PF, is a comprehensive measure of personality. It has been used in a variety of settings, particularly in industrial and organizational setting, where an in-depth assessment of the whole personnel is needed. The 16PF questionnaire was originally published in 1949 by Raymond Cattell following an extensive series of factor analytical

explorations into the adult personality structure. Cattell's initial exploration reduced a list of thousands of descriptive adjectives to sixteen elemental personality traits (Cattell, 1945). Such research is distinguished by (1) attempting to cover the domain of human personality through human language, and (2) a commitment to factor analytic methods for the discovery of the elemental units of personality.

Since its first publication in 1949, four major revisions were made so far. The latest revision was for releasing the 16PF fifth edition (Cattell et al., 1993). In 1988 a six-year project was begun in the US to improve the questionnaire (which had evolved into five different adult forms and numerous forms intended for children and adolescents). This Fifth Edition of the 16PF questionnaire,16PF5, involved an initial pool of over 750 questions (items) and the participation of 6,220 pilot testing participants in four iterative studies. The main goals of the revision were to develop updated, refined item content and collect a large, new norm sample. The 16PF5 has been translated and adapted into many languages, such as UK English, French, German, Chinese, Spanish, and so on. Through this endeavor, the 16PF5 has become one of the most widely used personality tests across the world. This article has been prepared for users of the Japanese 16PF5 questionnaire to get comprehensive information about the test development process.

Background Information

When Raymond Cattell and his colleagues set out to measure the broad range of normal personality approximately 70 years ago, they reasoned that adjectives relating to personality had to correspond to adjectives commonly used to describe people. They therefore began researching, on the basis of the Allport and Odbert (1936) trait lexicon, a set of some 18,000 adjectives.

Initially, Cattell and his colleagues asked observers to rate subjects well known to them on the basis of a subset of adjectives reduced to eliminate similar terms in the Allport and Odbert set. The researchers then subjected the observers' ratings to factor analysis. Cattell performed factor analysis with the intent of identifying the 'primary' personality traits, or those that could explain the entire personality domain.

Factor analyses of the observers' rating data, termed 'Life-data' or 'L-data', identified twelve traits that could encompass the range of descriptors in the trait lexicon. These traits, called 'factors', were named using letters of the alphabet, such as Factors 'A', 'B' and 'C'. (Within the alphabetical listing of factor names, some letters are missed out. Factors corresponding to these missing letters were found in parallel studies of child and adolescent personality, but were not found in descriptions of adults.)

The adjectives rated for the factors were translated into multiple-choice questionnaire items and were termed 'Questionnaire-data' or 'Q-data'. In a series of studies, responses to the questionnaire items were factor-analyzed, and the resultant data was used in constructing the sixteen primary scales of the 16PF instrument. Twelve of the 16PF scales measure the factors labelled alphabetically that were originally identified through analyses of the L-data. The remaining four scales measure factors labelled Q1, Q2, Q3 and Q4 that originated from analyses of the Q-data.

For the first edition of the 16PF questionnaire, as well as for the subsequent editions, Cattell factor-analyzed the sixteen primary scales to derive global factors on which related primary scales cluster together. (The global factors were called 'second-order factors' in previous 16PF literature.) The most-replicated 16PF global factors are Extraversion, Anxiety, Tough-Mindedness, Independence, and Self-Control. These global factors show how the sixteen factor scales are interrelated, and also allow personality to be viewed at a simpler, broader level than do the individual factor scales.

Historically, the basic scales of the 16PF questionnaire have been labelled with letters (for example, Factor A, Factor B, and so on, through to Factor Q4). As shown in Table 1, the 16PF scales are bipolar in nature; that is, both high and low scores have meaning. The right-side pole, or high-score range, of a factor is described as the plus (+) pole. The left-side pole, or low-score range, is the minus pole (–). For example, high scorers on Factor A are described as Warm (A+); low scorers are described as Reserved (A–).

Table 1. Primary Factor scale descriptors

Factor		**Left meaning/Low scores**	**Right meaning/High scores**
A	Warmth	More emotionally distant from people	Attentive and warm to others
B	Reasoning	Fewer reasoning items correct	More reasoning items correct
C	Emotional Stability	Reactive, emotionally changeable	Emotionally stable, adaptive
E	Dominance	Deferential, cooperative, avoids conflict	Dominant, forceful
F	Liveliness	Serious, cautious, careful	Lively, animated, spontaneous
G	Rule-Consciousness	Expedient, non-conforming	Rule-conscious, dutiful
H	Social Boldness	Shy, threat-sensitive, timid	Socially bold, venturesome, thick-skinned
I	Sensitivity	Objective, unsentimental	Subjective, sentimental
L	Vigilance	Trusting, unsuspecting, accepting	Vigilant, suspicious, sceptical, wary
M	Abstractedness	Grounded, practical, solution-oriented	Abstracted, theoretical, idea-oriented
N	Privateness	Forthright, straightforward	Private, discreet, non-disclosing
O	Apprehension	Self-assured, unworried	Apprehensive, self-doubting, worried
Q1	Openness to Change	Traditional, values the familiar	Open to change, experimenting
Q2	Self-Reliance	Group-oriented, affiliative	Self-reliant, individualistic
Q3	Perfectionism	Tolerates disorder, unexacting, flexible	Perfectionistic, organised, self-disciplined
Q4	Tension	Relaxed, placid, patient	Tense, high energy, impatient, driven

In addition to the primary scales, the 16PF tool contains a set of five scales that combine related primary scales into global factors of personality (See Table 2).

Table 2. Global factor scale descriptors

Factor		Left meaning/Low scores	Right meaning/High scores
EX	Extraversion	Introverted, socially inhibited	Extraverted, socially participating
AX	Anxiety	Low anxiety, unperturbed	High anxiety, perturbable
TM	Tough-Mindedness	Receptive, open-minded	Tough-minded, resolute
IN	Independence	Accommodating, agreeable, selfless	Independent, persuasive, wilful
SC	Self-Control	Unrestrained, follows urges	Self-controlled, inhibits urges

Development of the US English 16PF5

Since the first edition of the 16PF instrument was published in the United States in 1949, four revisions have followed, with scale refinements distinguishing each one (1956, 1962, 1967–69, 1993). The 1993 revision, which resulted in the fifth edition, reflects improved psychometric characteristics and gives attention to cultural changes and advances within the profession.

A 248-item questionnaire was trialed, which consisted of seventeen scales, each scale having approximately fourteen items. 15 per cent of the questionnaire's items were new, while 85 percent of the items were drawn from the existing forms A, B, C, D and E of the fourth edition of the 16PF tool. Many items had been rewritten to reduce their ambiguity and grammatical complexity and hence increase their readability. This had a further effect of making the language of the items more 'international English' and less 'American'.

All items except the Reasoning (Factor B) items were revised so that '?' was the middle response. This allowed respondents to choose the middle response when they thought that both 'a' and 'b' responses were equally applicable or when they thought neither 'a' nor 'b' applied to them. All the Reasoning (B) items were placed together at the end of the questionnaire, enabling them to have separate administration instructions pointing out their differences from the other items (Conn and Rieke, 1994).

Development of the Japanese 16PF5

The project to translate and adapt the 16PF5 into Japanese language started in October 2001, when the Test Development Agreement between IPAT(The Institute for Personality and Ability Testing) and Naotaka Watanabe, one of the authors of this paper, was signed up at Champaign, Illinois. After IPAT had been merged to OPP (Oxford Psychologist Press), an amendment was made between the author, Naotaka Watanabe, and OPP in July 2004 (Watanabe and Nishida, 2003, 2004).

All the test development procedures described in this paper were based on the recommendations of the International Test Commission's "Guidelines for Adapting Psychological Tests" (Hambleton, 2001) and IPAT's "Standards for Test Translations" in principle (See Table 3). Some

parts of the procedures, however, did not necessarily meet the standard, due to the Japanese corporates' policies which had provided the research fields to the authors. Since the Japanese 16PF5 questionnaire is also a broad measure of normal personality, it can be used in a variety of settings (clinical/counselling, occupational and research) to measure a wide range of life behaviors.

Table 3. IPAT's Standards for Test Translations

The following information details IPAT's best practices policy for translating tests. These standards were created to help you and IPAT develop new versions of tests that are psychometrically sound and well-translated. However, we realize that not everyone will be able to meet all of these standards. Please review the standards listed below and determine which standards you will be able to complete and which ones you will not be able to successfully complete. We will then discuss the standards you will not be able to complete and determine how they will be addressed.

- ***Licensee will have the IPAT test translated by at least two translators.*** *The translation should be done by people whose native language is the one into which the IPAT test items are being translated and who are also fluent in English. All translators should be educated about issues surrounding test development and the testing process, and at least one of the translators should be a psychologist familiar with the IPAT test, preferably with experience in the factor meanings. The other can be a professional translator or linguist from that country or ethnic region.*
- *Each translator will translate the IPAT test items independently.*
- *They will discuss and try to resolve disagreements. When disagreements cannot be resolved, both translations of the item(s) in question should be included for preliminary item evaluation studies. Item analyses will determine which version of the translated item(s) is selected.*
- ***Licensee will write additional questions to address any potential cultural differences between the U.S. and the country involved.*** *Special attention should be given to writing extra items to extend the range of scales that have shown elevated levels in a particular culture (i.e., where the American edition might not cover the whole range found in the culture).*
- ***Licensee will have the final draft translation back-translated.*** *The back translator should be a different person(s) from the individuals that did the initial translation. The back translator should be a native English-speaking psychologist who is fluent in the language into which the IPAT test items are being translated. The person conducting the back translation should take the translated items and then translate them back into English without previously seeing the English version of the items.*
- ***Back translator and initial translators will resolve any differences.*** *If resolution is not possible, both translations of the item(s) in question should be included for preliminary item evaluation studies. Item analyses will determine which version of the translated item(s) is selected.*
- ***At this point, the translation will be sent to IPAT to check for appropriate copyrights and trademarks.*** *Note: the correct wording for the standard copyright notice is included in your license agreement.*
- ***Licensee will administer the test to at least 500 people in order to do initial item***

analyses.

- ***Licensee will conduct item analyses and IPAT will review the results:***
- *Reliability analyses - most scale reliabilities should be above .70; some reliabilities can be lower, but all must be above .60;*
- *DIF analyses - most items should not be working differently for males and females, or for different races (if that is an issue), or for other groups;*
- *Factor analyses – the analysis should result in the same factor structure as the U.S. version of the IPAT test. Licensee should use oblique rotation in the factor analysis program when examining the factor structure of the 16PF® Questionnaire;*
- *Construct validity analyses - to compare the IPAT tests to other measures of the construct, when possible (e.g., compare the 16PF scales to other measures of personality);*
- *Cross-validation analyses - to check the reliability results, factor analysis results, etc.;*
- *Confirmatory factor analysis - ideal, but could be done later; this step requires additional data collection, resulting in a second sample of people that took the test.*
- ***Licensee will have the poor items rewritten or new items written to supplement the item bank, if necessary:***
- *The target percentage of items on a scale that overlap with the U.S. version should be 75%.*
- ***Licensee will collect more data to either redo item analyses above, if necessary, or to start norm data collection:***
- *Collect sample size of at least 1,000 to develop norms. The norm sample should consist of various types of participants – various ages, sex, educational levels, occupations, etc. (i.e., NOT all university students).*
- ***Licensee will adapt parts of the IPAT test manual that make sense to adapt, and include other information that would be useful to the test user:***
- *Administration and Scoring*
- *Interpretation of IPAT test*
- *Reliability*
- *Results of factor analyses*
- *Results of DIF analyses*
- *Other information that Licensee collected*

Initial Item Translations: The development of the first Japanese version

In the development of the Japanese 16PF5, a Form S Research Version was used with an extended number of trial items per scale. This extended item set was supplemented with some additional factor B general reasoning that were taken from the “Culture Fair Intelligence Test” published by IPAT. In total, 256 items were translated into Japanese by a professional bilingual translator.

Certain items were identified as being difficult to translate directly into an equivalent item in Japanese. The author applied his knowledge of the 16PF5 scales in order to adapt such items appropriately. A Japanese psychologist checked the translation and highlighted those expressions that he considered to be inadequate. A different professional Japanese translator compared the original questionnaire with the translated one. Her recommendations were passed to the author. Finally, a reconciliation meeting was held by the author with two Japanese psychologists and a psychometrician to discuss the item translations and suggestions for amendments. Together they

agreed on the most appropriate Japanese translation to use for each item, thus providing the item content of the first version of the Japanese 16PF5 was produced (Watanabe and Nishida,2003).

Trial sample sizes

The number of respondents who participated in our projects and the type of version administered are shown in Table 4.

Table 4. Sample size and questionnaire version used for each stage of development

Stage of Questionnaire Development	Sample Size	Version***
The first (pilot) stage	241*	I
The second (exploratory) stage	4,591	I
The third (confirmatory) stage	939	II
Validation studies	1,209**	II
Standardisation stage of norm collection	1,142	III

* Including 48 test-retest takers.

** Including the respondents to the second (exploratory) study.

*** We modified the Japanese version two times by referring to the results obtained in each stage.

First (Pilot) study

The first version of the Japanese 16PF5 was administered to Master's students at Keio University and undergraduates at Ube University (N=193) between June-July 2002.

As shown in Table 5, analysis of these pilot study results revealed that the internal consistency of most scales was already fairly high. The following scales had unacceptable values for Cronbach's Alpha that were below 0.60: scale B (0.47); scale I (0.44); and for the Impression Management scale (0.38).

Table 5. Reliability coefficient of each scale

	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

Test-retest reliability was assessed by administering the first Japanese 16PF version to 48 Master's students at Keio University with four weeks interval. As shown in Table 6, the test-retest reliability was sufficiently high for all scales, except for B (Pearson's product moment correlation coefficient, r= 0.43).

Table 6. Test-retest reliability coefficient of each scale

	JAPAN
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
IM	---

N=48

From the results shown in Table 5 and Table 6, we can observe the following facts:
(1) Internal consistency of each scale was fairly high. But, three out of seventeen scales did not fulfill the standard. Namely, B, I, and IM had some problems which should be resolved.
(2) Test-retest reliability of each scale was very high, except for B scale. The Pearson's product moment correlation coefficients of all the scales except for Factor B were larger than 0.70 (Watanabe,2012).

Second (Exploratory) Study: Extension of the pilot study

Prior to modifying any items as a result of the pilot study, a second (exploratory) study was conducted. The main reason for this was that the low reliability results for the B scale might have been due to the pilot study's small sample size.

In September 2002, the first Japanese 16PF version was administered to the employees of a large Japanese electronics company, including those of the company's subsidiaries, via an Intranet

common to the company and its subsidiaries (N=4,591). This second (exploratory) study resulted in comparable reliability results to the first pilot study. Once again, the internal reliability of the B, I and IM scales – as measured by Cronbach's Alpha – was less than 0.60. However, each coefficient had improved: B = 0.53; I = 0.55; and IM = 0.43 (See Table 7).

Table 7. Reliability coefficient: Second (Exploratory) Study

	US	JAPAN
A	0.722	0.746
B	0.725	**0.529**
C	0.823	0.777
E	0.749	0.801
F	0.773	0.758
G	0.806	0.686
H	0.878	0.898
I	0.798	**0.553**
L	0.786	0.724
M	0.767	0.722
N	0.825	0.766
O	0.803	0.805
Q1	0.694	0.675
Q2	0.795	0.809
Q3	0.796	0.786
Q4	0.782	0.764
IM	0.600	**0.429**

US:N=3755, JAPAN:N=4591

At this stage, a decision was taken to create a second version of the Japanese 16PF. Specific factor B and I items were deleted in order to improve internal consistency, although this was not a viable option for the IM scale. Factor B items on the second questionnaire version were too easy for the working Japanese respondents. About 2/3rds of the Factor B items included in the second version had been replaced in the third questionnaire version with more difficult items, including new items and modified Form S items.

In addition, we created the Japanese second version by referring to the following comments of IPAT staff on the results of back translation (The back translation was done by a bilingual person whom IPAT had designated). Two Japanese psychologists were engaged in this process (Watanabe and Nishida,2004).

(1)Statements are close but not quite the same

Item Number	Comments
2	Changes “upset” to “Unintimidated”
5	Changes “doesn’t have too many rules” to “Relaxed atmosphere of freedom”
7	Changes “willing to help” to “Want to help”
10	Wording is more transparent but on target
14	Changes “interruptions” to “asking other people” for input
16	Changes “trained myself” to “have confidence”
32	Adds “the opinion of other people” in the stem; option c misses the willingness to change plans
42	Adds “actually make a difference in option a; missing daydreams in option c
46	Changes “games” to “sports”
53	Changes “moral standards” to “morals”
58	Changes from object of thoughts to evaluation of thoughts
63	Past tense changed to present tense
69	Changes “things that aren’t proper” to “tricks”
75	Changes “good friend” to “friend”
76	Changes “done something wrong” to “mistake”
88	Changes a preference to specific behaviors
90	The use of the term “other” changes the meaning a little
91	Changes “providing information” to “talking”
97	In option c, changes manager to receptionist
101	Changes “guilt” to “regret” in stem
117	Changes “more important” to “important” in the item stem
123	Option c is not quite the same as responding “not true”
126	Doesn’t explicitly state “alone” in option a
137	Changes “act friendly” to “nice”
142	Missing the notion of “around me”
145	Changes “personal feelings” to “problems”; also missing the as opposed to other things concept
146	Changes “get down” with “overwhelmed”
149	Uses institutions rather than authority figures
157	Changes “more concerned” to “concerned”
158	Option a is missing the concept of asking others for suggestions
159	Adds “at tasks and”
162	Changes “needs” to “feelings”
177	Changes “affectionate” to “nicest”
21 (side 2)	Changes “people interrupt me” to “something comes up”
25	Missing “enjoy in the item stem; also Changes “social events or parties” to “what’s going on”
27	Changes “feel shy and unsure” to “freeze up and be withdrawn”
38	Loses the “in my personal life” context

(2)Statements are not the same

Item Number	Comments
11	Adds “and ideas”
12	Introduces a time component and missing the too sensitive component
13	Changes “conventional” to “common sense”, also specific to a business context
27	Changes “I am careful in choosing” to “I judge another person carefully”
49	Changes “formal” to “organized” in option c
52	Uses “Going to” rather than “in the middle of”
55	Changes “volunteer” to “not part of my job”
57	Changes “frank” to “not serious”
67	More specific; incorporates tough after asking nicely **and** being told no
80	Changes meaning from being patient to waiting patiently
89	Changes “seems friendly” to “looks nice”
92	Changes “bad news” to “negative gossip”
95	Changes “keep in tip-top shape” to “organize”
105	Changes “toughen up” with be stronger and protect themselves
112	Changes the context from home to concerts where loud obnoxious behavior could be expected
151	Adds “who is getting punishment which I think they deserve”
155	Changes “be reserved” to “not blend in”
169	Changes “form opinions too quickly” to “ talk about me when they don’t know me”
179	There seems to be something missing from the item stem
18 (side 2)	Changes “conventional” to “sensible”
29	Changes focus to being efficient rather than enjoying routine tasks

(3)Statements are more specific or extreme

Item Number	Comments
3	Causes a problem is more extreme than bothers me
20	Explicitly specifies by myself in option a
30	Replaces “have an office to myself” with “work alone in my own office”
82	Specifies feelings as opposed to presenting examples; but on target
127	Seems focused specifically on at work behaviors

(4)Frequency issues

Item Number	Comments
39	Replaces frequently with sometimes in the stem
87	Drops the term sometimes
122	The response options are missing frequency of occurrence

(5)Factor B items

Item Number	Comments
46	Technically this works, but would it be the same difficulty level

48	This seems too simple
50	Changes “quickest” to “fastest” when “fast” is already a option response
51	Changes “terminal” to “directly”; changes “cyclical” to “periodical” which has a duel meaning of regular intervals and intermittent
53	Is missing the second opposite term in the stem, also the target word does not have the double meaning reflected in the response options
54	Not the same thing
56	Is missing the second opposite term in the stem, also the target word does not have the double meaning reflected in the response options
63	Changes “distance” to “separate”
69	Changes “pale” to “light” which has multiple meanings; options b and c are not accurate

(5) Reverse Keyed items
66,104

Third (Confirmatory) Study
Following the pilot study extension, a second version of the Japanese 16PF was administered to 939 subjects, who were working for Japanese large automobile and electronics companies. Item-level analyses and reliability analyses were conducted using both Classical Test Theory (CTT) and Item Response Theory (IRT).

The goal of the item selection (deletion) process was to attain acceptable internal reliability using an appropriate number of items per scale. In some instances this process required a compromise to be made in terms of certain aspects of a scale’s psychometric properties. In summary, these were the criteria upon which item deletion was based:

- Based upon Classical Test Theory, those items with particularly low item-total correlations compared to the other items on that scale;
- Based upon Item Response Theory, those items with particularly low *a*-parameters is low and/or extremely high or low *b*-parameters;
- Each scale, except for Factor B, should consist of 12 items; and
- Factor B should have 24 items.

(1) Classical Test Theory approach
The application of these criteria enhanced the internal reliability of each 16PF scale. In particular, the Cronbach’s Alpha for Factor B became 0.63, for factor I became 0.60 and for IM became 0.57. From this third version a single Impression Management item was removed and the most appropriate factor B item set was selected, giving an acceptable internal consistency (0.63). Table 8 shows the results of item selection. This work was done in collaboration with OPP/IPAT. This version became the 215-item published Japanese 16PF5 Questionnaire.

Table 8. Reliability coefficient after item selection: Third (Confirmatory) Study

	US	JAPAN
A	0.722	0.812
B	0.725	0.630
C	0.823	0.772
E	0.749	0.752
F	0.773	0.761
G	0.806	0.699
H	0.878	0.885
I	0.798	0.601
L	0.786	0.756
M	0.767	0.698
N	0.825	0.796
O	0.803	0.794
Q1	0.694	0.641
Q2	0.795	0.793
Q3	0.796	0.791
Q4	0.782	0.753
IM	0.600	**0.573**

US:N=3755 JAPAN:N=939

(2) Item Response Theory approach

Item Response Theory is known as a powerful psychometric tool for language translation, especially for detecting differential item functioning (Watanabe,1992,1994,1996).

Table 9 shows the results of the data analyses for the selected items version (216 items version) based on Item Response Theory. Below described Two-parameter Logistic Model (2PL model) was adopted with marginal maximum likelihood estimation by BILOG-MG program.

$$p_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

Where

θ = latent trait parameter

a_i = the slope of parameter of item i, characterizing its sensitivity to proficiency (attitude), where $a_i>0$;

b_i = the threshold parameter of item i, characterizing its difficulty;

D = an arbitrary scaling constant typically set to 1.7 to approximate results from the normal ogive model.

Concerning *a*-parameter (slope), all of the scales included in the selected item version have larger values than those of the original revised version. It means the scale has become more discriminating power than before the item selection. Concerning *b*-parameter (threshold), 10 out of 17 scales has smaller absolute value. It means that we succeeded to obtain more general difficulty power through the item selection.

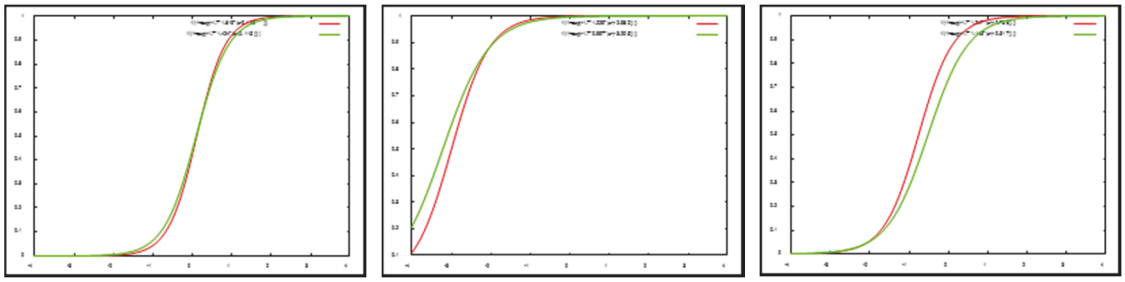
Table 9. Changes of the mean of item parameters before and after item selection.

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

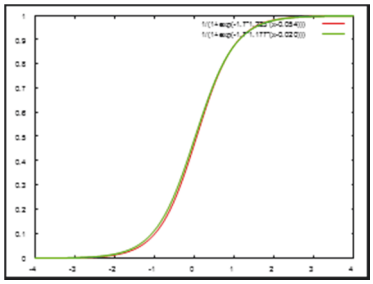
Note: Bold and underlined value indicates more appropriate item parameter.

Figure 1 shows the change of the shape of item characteristic curve (ICC) of each scale.

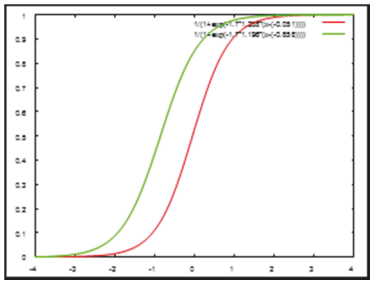
A Warmth B Reasoning C Emotional Stability



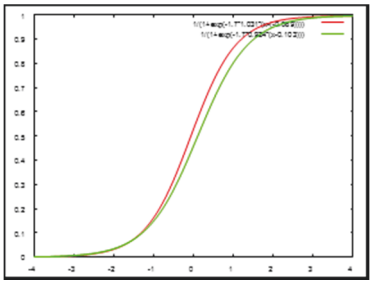
E Dominance



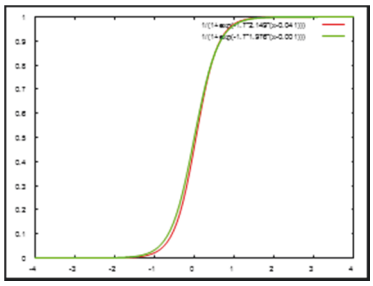
F Liveliness



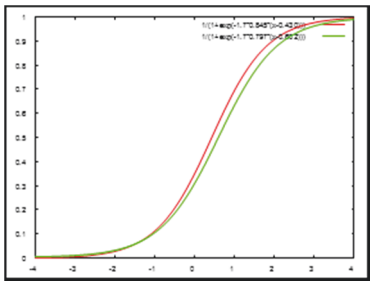
G Rule-Consciousness



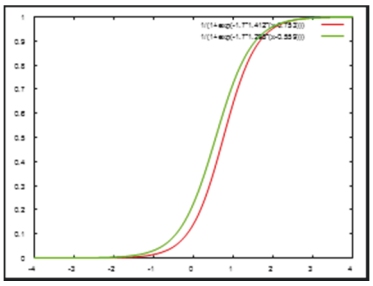
H Social Boldness



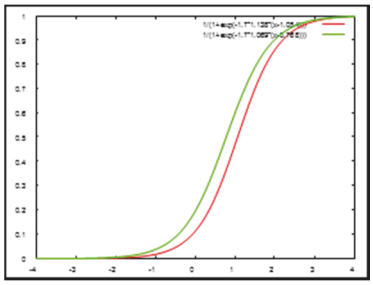
I Sensitivity



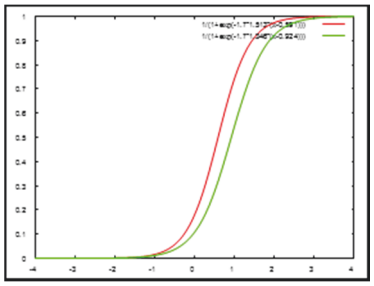
L Vigilance



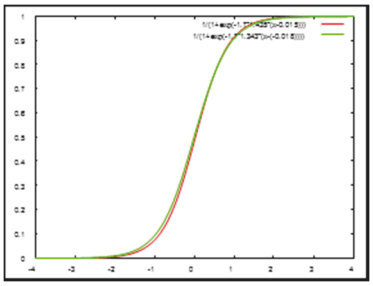
M Abstractedness



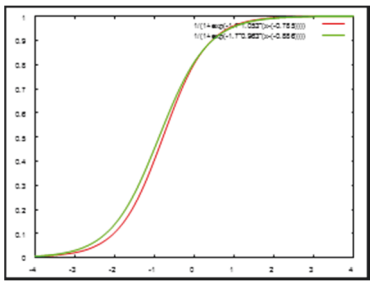
N Privateness



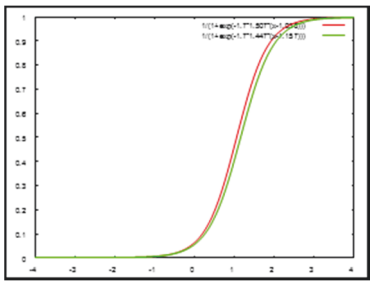
O Apprehension



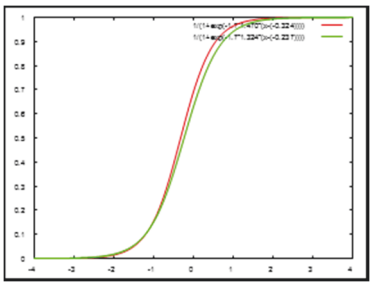
Q1 Openness to Change



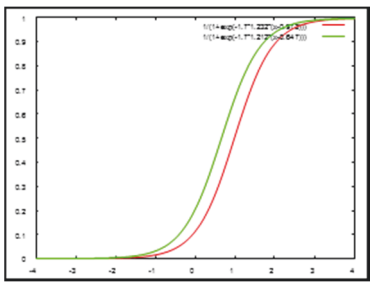
Q2 Self-Reliance



Q3 Perfectionism



Q4 Tension



IM Impression Management

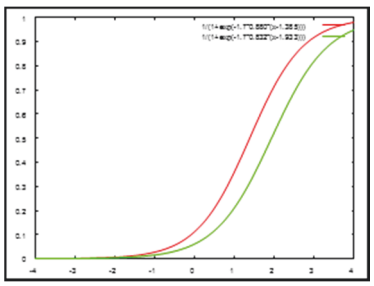


Figure 1. Item characteristic curves before and after item selection

Note: Thin line and thick line mean “before” and “after” item selection respectively.

Norm Creation

The third version of the Japanese 16PF was administered to a wide variety of samples (N=1,142; including 881 males, 253 females and 8 of unspecified gender). There were 940 employees from various companies and 202 college or graduate students of three universities. Age range = 19-58 years. Mean age = 27.3 years.

Statistical properties of the 16PF5 scales

Statistical information (means, standard deviation) for the published version of the Japanese 16PF can be found in Table 10.

Table 10. Means and standard deviations for each 16PF5 factor

Primary factor		N = 1142	
		Mean	SD
A	Warmth	10.46	5.51
B	Reasoning	17.70	3.30
C	Emotional Stability	16.43	5.94
E	Dominance	12.99	6.27
F	Liveliness	14.72	5.54
G	Rule-consciousness	14.03	5.06
H	Social Boldness	12.50	7.82
I	Sensitivity	11.47	4.95
L	Vigilance	9.14	5.37
M	Abstractness	8.74	5.44
N	Privateness	10.28	6.03
O	Apprehension	13.09	6.40
Q1	Openness to change	15.84	4.77
Q2	Self-reliance	8.30	6.17
Q3	Perfectionism	14.02	6.34
Q4	Tension	8.34	6.13

The norm tables for the combined standardization sample and for the male and female sub samples are shown in Tables 11, 12 and 13 respectively.

Table 11. Norm group in stens (N=1142)

Scale	Stens									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	0-1	2	3-5	6-7	8-10	11-14	15-17	18-20	21-22	23-24
B	0-10	11-12	13-14	15-16	17-18	19	20	21	22	23-24
C	0-3	4-5	6-10	11-13	14-17	18-19	20-21	22-23	24	-
E	0-1	2-3	4-6	7-9	10-13	14-16	17-19	20-21	22-23	24
F	0-3	4-6	7-9	10-12	13-15	16-18	19-20	21-22	23-24	-

G	0-3	4-5	6-8	9-10	11-13	14-16	17-19	20-21	22	23-24
H	-	0	1-3	4-7	8-13	14-18	19-21	22-23	24	-
I	0-2	3-4	5-6	7-8	9-11	12-14	15-16	17-18	19-20	21-24
L	0	1-2	3-4	5-6	7-9	10-11	12-14	15-18	19-21	22-24
M	-	0-1	2-3	4-5	6-8	9-10	11-13	14-17	18-19	20-24
N	-	0-1	2-3	4-5	6-9	10-13	14-16	17-19	20-21	22-24
O	0	1-2	3-6	7-9	10-13	14-16	17-19	20-21	22-23	24
Q1	0-5	6-8	9-11	12-13	14-16	17-18	19-20	21-22	23	24
Q2	-	0	1-2	3-4	5-7	8-10	11-14	15-17	18-21	22-24
Q3	0-1	2-3	4-7	8-10	11-14	15-18	19-21	22	23-24	-
Q4	-	0	1-2	3-4	5-7	8-10	11-14	15-18	19-21	22-24

Table 12. Norm group in stens for males (N=881)

Scale	Stens									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	0-1	2	3-4	5-7	8-9	10-13	14-16	17-20	21-22	23-24
B	0-11	12-13	14-15	16-17	18	19-20	21	22	23	24
C	0-4	5-7	8-11	12-14	15-18	19-20	21-22	23	24	-
E	0-1	2-4	5-7	8-10	11-13	14-17	18-19	20-21	22-23	24
F	0-2	3-5	6-9	10-12	13-15	16-18	19-20	21-22	23-24	-
G	0-3	4-5	6-8	9-11	12-14	15-17	18-19	20-21	22-23	24
H	-	0	1-3	4-7	8-13	14-18	19-21	22-23	24	-
I	0-1	2-3	4-5	6-7	8-10	11-12	13-14	15-17	18-20	21-24
L	0-1	2	3-4	5-6	7-8	9-11	12-14	15-18	19-21	22-24
M	-	0-1	2-3	4-5	6-9	10-12	13-15	16-18	19-20	21-24
N	-	0-1	2-3	4-5	6-9	10-13	14-17	18-19	20-22	23-24
O	0	1-2	3-5	6-9	10-13	14-16	17-19	20-21	22-23	24
Q1	0-5	6-8	9-11	12-13	14-16	17-18	19-20	21-22	23	24
Q2	-	0	1-2	3-4	5-7	8-10	11-14	15-18	19-21	22-24
Q3	0-1	2-4	5-7	8-10	12-15	16-18	19-21	22	23-24	-
Q4	-	0	1-2	3-4	5-7	8-10	11-14	15-17	18-20	21-24

Table 13. Norm group in stens for females (N=253)

Scale	Stens									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
A	0-1	2-4	5-6	7-9	10-13	14-17	18-19	20-21	22	23-24
B	0-7	8-9	10-11	12-13	14-16	17-18	19-20	21	22	23-24
C	0-1	2-3	4-7	8-11	12-15	16-19	20-21	22-23	24	-
E	0-1	2	3-4	5-6	7-10	11-14	15-18	19-20	21-22	23-24
F	0-4	5-7	8-10	11-14	15-17	18-19	20-21	22	23-24	-
G	0-4	5-6	7-8	9-11	12-13	14-16	17-18	19-20	21-22	23-24

H	-	0	1-2	3-6	7-12	13-17	18-20	21-23	24	-
I	0-4	5-8	9-10	11-13	14-15	16-17	18-19	20-21	22	23-24
L	0	1-2	3-4	5-6	7-9	10-12	13-15	16-18	19-21	22-24
M	-	0-1	2-3	4	5-7	8-10	11-13	14-16	17-19	20-24
N	0	1	2-3	4-5	6-9	10-12	13-16	17-19	20-22	23-24
O	0	1-3	4-7	8-10	11-14	15-17	18-19	20-21	22-23	24
Q1	0-5	6-7	8-9	10-11	12-15	16-17	18-19	20-21	22	23-24
Q2	-	0	1-2	3-4	5-7	8-10	11-13	14-16	17-20	21-24
Q3	0-1	2-4	5-7	8-9	10-13	14-16	17-19	20-21	22-23	24
Q4	-	0	1-2	3-5	6-8	9-13	14-16	17-19	20-22	23-24

Reliability

Reliability gauges the consistency of test results. As a generic term, it relates to a number of different aspects of consistency. Essentially, a reliable test yields the same approximate results when administered repeatedly under similar conditions.

Internal consistency reliability

The aspect of reliability addressed in this chapter is that of internal consistency, or homogeneity, of the test items, as measured by Cronbach's coefficient alpha. Internal consistency of the sixteen factors measured by the 16PF5 questionnaire reflects the degree to which that set of scale items are sampling the same personality domain. In statistical terms, internal consistency reliability is how large the inter-correlation is between the items that make up each of the sixteen personality scales. Internal consistency can be viewed as reliability estimated from a single test administration. Measurement of the internal reliability of a test provides a source of evidence that all items on a given scale assess the same personality construct. As the inter-correlations among items within a scale increase, reliability of the scale itself increases. Internal consistency is lowered to the degree that items on the same scale measure different traits or to the extent that scale items are not inter-correlated.

As a measure of scale internal consistency, Cronbach's coefficient alpha essentially calculates the average value of all possible split-half reliabilities. Cronbach alpha coefficients for the 16PF questionnaire were calculated. Table 14 compares the values for the Japanese 16PF5 scales (and item numbers per scale) with those of the U.S.'s fifth edition of the 16PF (Watanabe,2012).

Table 14. Cronbach alpha coefficients and item numbers for the Japanese 16PF5 (by factor) compared to the US

	JAPAN			**US**	
	Cronbach alpha coefficients	Cronbach alpha coefficients (cross validation study)	Number of items	Cronbach alpha coefficients (N=2500)*	Number of items
A	.81	.76	12	0.69	11
B	.63	.58	24	0.77	15
C	.77	.82	12	0.78	10

E	.75	.81	12	0.66	10
F	.76	.76	12	0.72	10
G	.70	.67	12	0.75	11
H	.89	.90	12	0.85	10
I	.60	.66	12	0.77	11
L	.76	.75	12	0.74	10
M	.70	.75	12	0.74	11
N	.80	.80	12	0.75	10
O	.79	.81	12	0.78	10
Q1	.64	.67	12	0.64	14
Q2	.79	.82	12	0.78	10
Q3	.79	.81	12	0.71	10
Q4	.75	.81	12	0.76	10
IM	.57	.58	11	0.60	12

** See Conn and Rieke (1998).*

Overall the results show that the Japanese 16PF5 has fairly high reliability. In addition, the third column shows the reliability coefficients calculated from cross-validation analyses. Tentative items selected in the process of validity studies were used for examining the reliability of each scale. The reliability coefficients based upon the cross-validation data are always higher, except for Factor B. As described earlier, the Factor B item set was changed quite drastically for the third research version.

Standard Errors of Measurement

Test users often use the standard error of measurement, termed SE_M, to establish confidence intervals around an obtained score (for a particular scale). Adding plus or minus (+/–) 1 standard error to the obtained score provides a 68% confidence interval for an individual's true score. As a generally conservative estimate, the sten score standard error of measurement for most 16PF scales is near 1. Thus, one can be 68% confident that for people who obtain a sten score of 6, their true score will fall within a sten score band of 5–7 (i.e. 6 +/– 1). Similarly, adding +/– 2 standard error units to an obtained score provides a 95% confidence interval. Table 15 shows the standard error of measurement (SEM) for both raw scores and sten scores (Watanabe and Nishida,2004).

The SEM is calculated from the following equation: where *SD* is the standard deviation of scores for that specific scale and *r* is the reliability coefficient, which in this particular case is the Cronbach's alpha coefficient:

$$SE_M = SD\sqrt{1-r}$$

Table 15. Raw score and sten scores standard errors of measurement

Primary Factor SEM		
	Raw scores	Sten scores
A Warmth	2.40	0.87
B Reasoning	1.86	1.21

C Emotional Stability	2.85	0.96
E Dominance	3.14	1.00
F Liveliness	2.71	0.98
G Rule-Consciousness	2.77	1.10
H Social Boldness	2.59	0.66
I Sensitivity	3.13	1.26
L Vigilance	2.63	0.98
M Abstractness	2.72	1.10
N Privateness	2.70	0.89
O Apprehension	2.93	0.92
Q1 Openness to Change	2.86	1.20
Q2 Self-Reliance	2.83	0.92
Q3 Perfectionism	2.91	0.92
Q4 Tension	3.01	1.00
IM Impression Management	1.92	1.51

Validity

Generally, validity refers to a test's ability to measure the concept it was designed to measure. There are a number of different types of validity concerned with the usefulness of the inferences that can be made from test scores.

Construct Validity Study using the SPI

Construct validity – when applied to a personality tool – refers to the extent to which a test captures the personality domain that it claims to. Construct validity of the 16PF5 questionnaire demonstrates that the test measures sixteen distinct personality traits.

A large Japanese electronics corporate provided SPI (Synthetic Personality Inventory) data, which had been administered in the past and these subjects went on to complete the 16PF for the purposes of this validation study (N=336). The SPI has 18 subscales: 13 basic scales (e.g. "introspection", "endurance", uniqueness", "achievement motivation"); 4 primary scales measuring type, and one ability scale (called the General Ability Test). The results are shown in Table 16 in the form of correlation coefficients between the two psychometric tools. Based upon the SPI study it can be generally concluded that the 16PF has high construct validity (Watanabe,2012).

Table 16. Summary of Correlations between the 16PF and the SPI (N=336, p<.05)

Negative correlation (-)	**Correlation Coefficient**	**Primary Factor**	**Positive correlation (+)**	**Correlation Coefficient**
Introversion		**A Warmth**	Extroversion	.440
Sensing	-.277		Intuitive	
Thinking	-.255		Feeling	
Low Ability(GAT)		**B Reasoning**	High Ability(GAT)	.281
		C Emotional Stability		

Introversion		**E Dominance**	Extroversion	.342
Introversion Sensing	 -.261	**F Liveliness**	Extroversion Intuitive	.605
Perceptive		**G Rule-Conscious**	Judging	.407
Introversion		**H Social Boldness**	Extroversion	.598
Sensing	-.280	**I Sensitivity**	Intuitive	
		L Vigilance		
Sensing Judging	-.323	**M Abstractness**	Intuitive Perceptive	.308
Extroversion	-.450	**N Privateness**	Introversion	
		O Apprehensive		
Sensing	-.367	**Q1 Openness to Change**	Intuitive	
Extroversion	-.444	**Q2 Self-Reliance**	Introversion	
Perceptive		**Q3 Perfectionism**	Judging	.538
		Q4 Tension		

Construct Validity Study using the OPQ

436 employees at a large Japanese electronics company completed both the Japanese 16PF5 and a version of the Japanese OPQ (Occupational Personality Questionnaire)– a version of the OPQ containing 30 subscales (i.e. not the OPQ32). Pearson's product moment correlation coefficients were calculated between the 16PF scales and the OPQ subscales (see Table 17).

Table 17. Significant correlations between 16PF and OPQ (N=436, p<.05)

Negative correlation (-)	Correlation Coefficient	Primary Factor	Positive correlation (+)	Correlation Coefficient
Practical	-0.40	**A Warmth**	Persuasive	0.33
Data rational	-0.40		Controlling	0.34
Conceptual	-0.40		Outgoing	0.36
Conscientious	-0.40		Socially confident	0.36
Worrying	-0.30		Caring	0.35
			Active	0.38
Traditional	-0.30	**C Emotional Stability**		
Worrying	-0.30			
Modest	-0.30	**E Dominance**	Persuasive	0.36
Traditional	-0.40		Controlling	0.37
Worrying	-0.30		Socially confident	0.31
Emotional control	-0.40			
Modest	-0.30	**F Liveliness**	Outgoing	0.50
Traditional	-0.30		Affiliative	0.40
Conceptual	-0.30		Socially confident	0.37
Conscientious	-0.30		Active	0.32
Worrying	-0.30			
Modest	-0.40	**H Social Boldness**	Persuasive	0.49
Practical	-0.30		Controlling	0.40

Data rational	-0.30		Outgoing	0.47
Traditional	-0.40		Socially confident	0.64
Worrying	-0.30			
Persuasive	-0.30	**N Privateness**	Modest	0.30
Outgoing	-0.30		Worrying	0.32
Affiliative	-0.30			
Socially confident	-0.40			
Active	-0.30			
Tough-minded	-0.30	**O Apprehensive**	Worrying	0.41
Optimistic	-0.40			
Traditional	-0.40	**Q1 Openness to Change**		
Worrying	-0.30			
Outgoing	-0.40	**Q2 Self-Reliance**	Independent	0.30
Affiliative	-0.40		Conscientious	0.30
Socially confident	-0.30			
Active	-0.30			
Relaxed	-0.30	**Q3 Perfectionism**	Forward planning	0.44
Optimistic	-0.30		Conscientious	0.42

The summary of the validation results in Table 17 shows that the 16PF is moderately correlated with the OPQ. The key findings are as follows (Watanabe,2012):

(1) Factor B is not correlated with any subscale of OPQ – as would be expected.
(2) Scales A, F, H, N, and Q2 exhibited high correlations with the OPQ's subscales.
(3) Scales B, G, L, M, Q4 exhibited low correlations with the OPQ's subscales.

Construct Validity Study through Examining Factor Structure

Factor-analytic results also provide evidence about the construct validity of the 16PF questionnaire, and how distinct the sixteen personality traits are. As explained earlier, Raymond Cattell's original development of the 16PF questionnaire used factor analysis to identify sixteen primary factors. Factor analysis was also used to identify a set of global factors that explain the sixteen primary factor scales at a broad level.

One particular aspect of Cattell's factor-analytic method merits explanation because it represents a departure from that used in the development of some other personality inventories. Cattell anticipated that distinct personality traits might, nonetheless, be related to one another. Therefore, rather than extracting factors forced to be independent of one another and consequently uncorrelated (orthogonal factors), Cattell chose to use oblique factors, which are allowed to inter-correlate. Cattell's assumption is reflected at the global factor level, where related primary factors cluster along the five global scales. As shown in Table 18, almost all the scales of the Japanese 16PF5 are also significantly correlated each other (Watanabe and Nishida,2003,2004).

Table 18. Correlation matrix of the factors of Japanese 16PF5

	F1:A	F2:B	F3:C	F4:E	F5:F	F6:G	F7:H	F8:I	F9:L	F10:M	F11:N	F12:O	F13:Q1	F14:Q2	F15:Q3	F16:Q4
F1:A	1															
F2:B	−.014	1														
F3:C	.261**	.026	1													
F4:E	.207**	.051	.327**	1												
F5:F	.493**	.030	.337**	.327**	1											
F6:G	−.049	−.050	.078*	.013	−.143**	1										
F7:H	.444**	.046	.437**	.479**	.576**	−.020	1									
F8:I	.406**	.009	−.023	−.005	.174**	−.118**	.169**	1								
F9:L	−.129**	−.023	−.415**	−.070*	−.204**	−.167**	−.191**	.024	1							
F10:M	−.105**	.002	−.402**	−.101**	−.152**	−.187**	−.195**	.121**	.401**	1						
F11:N	−.398**	−.041	−.327**	−.372**	−.532**	.025	−.541**	−.125**	.328**	.198**	1					
F12:O	−.109**	−.041	−.600**	−.248**	−.259**	.151**	−.332**	.006	.335**	.366**	.246**	1				
F13:Q1	.193**	.041	.306**	.408**	.394**	−.154**	.330**	.149**	−.085**	.088**	−.306**	−.218**	1			
F14:Q2	−.399**	−.004	−.316**	−.185**	−.555**	−.051	−.374**	−.074*	.344**	.339**	.487**	.175**	−.198**	1		
F15:Q3	−.056	.017	.087**	.195**	−.062	.401**	.080*	−.075*	−.026	−.059	.040	.169**	−.008	−.002	1	
F16:Q4	.020	−.045	−.558**	−.013	−.138**	−.178**	−.195**	.100**	.402**	.344**	.149**	.391**	−.157**	.226**	−.089**	1

** p＜.01

* p＜.05

Results of Factor Analyses

The factor structure of the final set of items was examined for the norm sample according to the specifications of Conn and Rieke (1994). Items within each factor were put into ‘parcels’ based upon the strength of their correlations with items contained within the same factor. Hence the term ‘parcels’ refers to small groupings of items within a scale.

The twelve items per scale (except Factor B) were grouped into six parcels by referring to the correlation coefficients. Factor B’s twenty-four items were also grouped into six parcels. The total number of parcels was 96. 19 factors were extracted by PAF with the Kaiser-Guttman’s criteria and rotated by Equamax oblique rotation. (See Table 19).

Table 19. Results of primary factor analysis

	A	B	C	E	F	G	H	I	L	M	N	O	Q1	Q2	Q3	Q4
A1	0.67															
A2	0.71															
A3																
A4	0.65															
A5	0.56															
A6					0.35											
B1		0.44														
B2		0.36														
B3		0.45														
B4		0.52														
B5		0.45														
B6																
C1			0.32													
C2			0.41													
C3			0.53													
C4			0.31													
C5			0.45													
C6																
E1				0.43			0.31									

E2		0.57										
E3		0.61										
E4		0.48										
E5		0.46										
E6		0.39										
F1												
F2			0.39									
F3												
F4			0.41									
F5			0.43									
F6			0.50									
G1				0.42								
G2				0.56								
G3												
G4				0.49								
G5				0.32								
G6				0.46								
H1					0.60							
H2					0.50							
H3					0.55							
H4					0.58							
H5					0.65							
H6					0.50							
I1	0.35					0.35						
I2	0.41					0.37						
I3						0.48						
I4						0.41						
I5												
I6	0.46					0.36						
L1							0.53					
L2							0.46					
L3							0.52					
L4							0.37					
L5							0.48					
L6							0.57					
M1								0.46				
M2								0.64				
M3						0.35		0.49				
M4												
M5												
M6								0.40				
N1									0.65			
N2		-0.40			-0.30				0.36			
N3									0.49			
N4									0.46		0.31	

N5										0.61										
N6										0.39										
O1											0.64									
O2											0.61									
O3											0.44									
O4											0.45									
O5											0.32									
O6											0.47									
Q1_1												0.47								
Q1_2												0.39								
Q1_3																				
Q1_4												0.34								
Q1_5												0.52								
Q1_6			0.36									0.35								
Q2_1				-0.40									0.51							
Q2_2													0.43							
Q2_3													0.46							
Q2_4													0.58							
Q2_5													0.53							
Q2_6													0.55							
Q3_1																				
Q3_2																				
Q3_3																0.57				
Q3_4																0.46				
Q3_5																0.58				
Q3_6																0.56				
Q4_1																			0.46	
Q4_2																			0.40	
Q4_3																			0.48	
Q4_4																			0.56	
Q4_5																			0.55	
Q4_6																			0.57	

Factor loadings >.30, N=2081

The expected factor pattern was observed with the parcels for factors E, H, L, N, O and Q2 showing very clear loadings on their own factors. The parcels for factors A, B, C, G, Q1, and Q4 parcels also had quite clear loadings on their own factors, together with a little overlap with other factor(s). These factor analysis results provide some construct validity evidence for the 16-factor structure associated with the US fifth edition of the 16PF. Although the parcels for factors F, I, M and Q3 parcels did overlap with other factors (Watanabe,2012).

Construct validity of the global factor scores

A factor analysis of the primary factors was conducted to examine the second order structure of the global factors. The previously described results from the factor analysis at the item level (see Table 19) were used as the starting point for the global factor analysis. An Exploratory Factor

Analysis (EFA) with Principal Axis Factoring and a Promax Rotation with Kaiser Normalization (delta=3) was run (N=2081).

The results from requesting a five factor solution are shown in Table 20. The pattern of factor loadings are quite close, but not same as that of the US 16PF5 Questionnaire. There are several plausible explanations for the discrepancies indicated in Table 20.

Table 20. Global factor-analytic results

	Anxiety	**Independence**	**Extraversion**	**Self-Control**	**Tough-Mindedness**
	1	**2**	**3**	**4**	**5**
Warmth			-.222		.622
Emotional Stability	-.801	.177			
Dominance		**.834**		.193	
Liveliness		**.317**	**-.573**	**(-.154)**	
Rule-Consciousness	-.187	-.124		**.698**	
Social Boldness	-.145	**.573**	**-.215**		
Sensitivity		-.119	.114		**.656**
Vigilance	**.556**	**(.123)**	.198		
Abstractedness	**.582**		.100	**(-.176)**	**(<.10)**
Privateness		**-.327**	**.479**		
Apprehension	**.716**	-.175	-.168	**.317**	
Openness to Change		**.570**			**(<.10)**
Self-Reliance	.144		**.794**		
Perfectionism		.232		**.656**	
Tension	**.645**		.107		

N=2081,

Note: (1)Squared are those factor loadings in the Japanese version which match those in the US version of the 16PF Questionnaire; (2) Underlined are factor loadings that are higher in the Japanese data, compared to the US; (3) Parenthesized are factor loadings that are lower in the Japanese data, compared to the US.

The first and most obvious explanation is that the second order factor structure of the 16PF5 Questionnaire is different in Japanese culture than in the US culture. However, examining the sample upon which the current analyses were conducted causes some concern as to whether the discrepancies may be due to an overly homogenous sample. For example, the sample is comprised largely of males working in professional positions and having graduated from college. In comparison, the US sample is more evenly balanced among men and women, and individuals in professional occupations versus service, labor, and administrative/clerical occupations. Given these concerns about the influence of a homogenous sample on the resulting second order factor structure, the decision was taken to adopt the US 16PF factor structure and global factor scale equation weights (see Table 21) (Watanabe and Nishida,2004).

Table 21. Gobal Factor scale equations

EXTRAVERSION = 0.3A + 0.3F + 0.2H – 0.3N – 0.3Q2 +4.4
ANXIETY = -.4C + 0.3L + 0.4O + 0.4Q4 +1.6
TOUGH-MINDEDNESS = -0.2A -0.5I -0.3M – 0.5Q1 + 13.8
INDEPENDENCE = 0.6E + 0.3H + 0.2L + 0.3Q1 -2.2
SELS-CONTROL = -0.2F + 0.4G – 0.3M + 0.4Q4 + 3.80

Criterion-Related Validity Study

Criterion-related validity refers to how effectively the 16PF5 questionnaire measures the external criteria, such as effective behavior in the workplace. This section presents the findings of a criterion validity study at a large Japanese company that investigated whether the Japanese 16PF could discriminate high-level performers at the company from lower-level performers.

The HR department at the company was asked to select two groups of employees: those who had shown high-level performance based upon their appraisal ratings and those who had average performance. Performance data was available for four roles, as follows: “Sales”, “R&D”, “Production”, and “Administration” and the numbers of subjects within each are shown in Table 22.

Table 22. The number of people categorized into each condition

	High-level performer	**Average performer**
Administration	45	38
Sales	41	42
R&D	143	40
Production	51	37

The mean sten scores for each 16PF scale are shown in Table 22 for the 8 performance groups. These results are shown in bold when t-tests revealed a significant difference in means between the Above Average (High-level performer) and the Average groups.

Table 22 t-test results for each performance group across the 4 departments

STEN MEANS	**Administration**		**Sales**		**R&D**		**Production**	
	High	Average	High	Average	High	Average	High	Average
Warmth (A)	5.98	5.71	6.76	6.52	4.90	4.78	**4.29**	**5.00**
Reasoning (B)	5.29	5.71	4.83	4.52	5.29	5.38	5.45	5.27
Emotional Stability (C)	6.40	6.71	7.15	6.69	6.37	5.63	6.47	5.73
Dominance (E)	5.96	5.61	6.41	5.95	**5.90**	**4.78**	5.71	5.14
Liveliness (F)	5.82	5.08	6.22	6.10	**5.28**	**4.68**	4.94	5.49
Rule-Consciousness (G)	**6.13.**	**5.11**	5.12	5.14	5.44	5.70	5.76	5.27
Social Boldness (H)	**6.04**	**5.13**	**7.20**	**5.74**	5.64	5.20	5.25	5.65

Sensitivity (I)	6.27	6.13	6.00	6.05	5.27	5.80	5.06	5.43
Vigilance (L)	5.76	5.26	5.34	5.33	5.41	5.53	5.31	5.59
Abstractedness (M)	**6.22**	**4.95**	5.10	5.29	5.25	5.48	5.00	5.43
Privateness (N)	5.27	5.50	4.49	4.64	**5.45**	**6.48**	5.53	5.43
Apprehension (O)	5.64	5.21	5.29	5.21	5.28	5.70	5.39	5.38
Openness to Change (Q1)	**6.67**	**5.74**	6.24	6.10	**5.95**	**5.33**	5.61	5.38
Self-Reliance (Q2)	5.47	5.74	5.07	5.14	**5.43**	**6.25**	6.04	5.27
Perfectionism (Q3)	4.69	4.74	5.39	4.60	5.10	5.60	5.16	5.00
Tension (Q4)	5.80	5.08	5.49	5.64	5.16	5.73	5.43	5.43
Extraversion	5.93	5.29	6.86	6.40	**5.32**	**4.46**	4.75	5.46
Anxiety	5.34	4.61	4.66	4.87	4.85	5.58	4.94	5.31
Self-Control	**4.27**	**5.24**	4.80	4.84	5.64	5.64	6.11	5.76
Independence	**6.34**	**5.48**	6.75	5.99	**5.90**	**4.93**	5.55	5.31
Tough-Mindedness	**4.69**	**5.65**	5.23	4.89	5.39	5.74	5.68	5.18

Significant results are shown in highlighted bold (p<.05)

The 16PF5 was more sensitive at discriminating improved performance for those:
- In Administration (higher Rule-Consciousness, higher Social Boldness, higher Abstractedness, lower Self-Control, higher Independence and higher Tough-Mindedness); and
- In R&D (higher Dominance, higher Liveliness, lower Privateness, higher Openness to Change, lower Self-Reliance, higher Extraversion and higher Independence) (Watanabe and Nishida,2004).

Differential Item Functioning

Differential item functioning (DIF) occurs under the condition that groups (such as defined by gender, race, age, education, etc.) have different probabilities of endorsing a given item on a multi-item scale after controlling for overall scale scores. Gender is a more pertinent issue to test for differential item functioning on a Japanese personality questionnaire. The reason is that gender is widely recognized as playing a significant role in Japanese businesses. DIF should be studied since we are often interested in comparing groups.

DIF Summary and Results

DIF analyses were conducted by using Item Response Theory. The reference group (group 1) and the focal group (group 2) were male and female respectively. A two-parameter logistic model (2PL model) was adopted with marginal maximum likelihood estimation by the BILOG-MG program. Item parameters were estimated separately on DIF and non-DIF models. The log likelihood function of the fit of the DIF and non-DIF models was compared for each scale. The differences were examined using a Chi-square test with 12 degrees of freedom (24 for Factor B). Table 23 shows a summary of the results.

Table 23 DIF Summary for the male and female subgroups

	Difference of -2 log likelihood (Non-DIF model)-(DIF model)	DIF or Non-DIF	Difference of *b* (threshold) Means	Equated *a* (Slope)	Ratio of DIF items
A	12191.39-12081.35=**110.04**	DIF	**-0.529**	1.876	3/12
B	10293.10-10747.82=**-454.72**	Non-DIF	0.123	1.018	**11/24**
C	11517.75-11436.46=**81.29**	DIF	-0.122	1.374	4/12
E	12864.64-12811.03=**53.61**	DIF	0.174	1.344	2/12
F	12424.02-12330.67=**93.35**	DIF	**-0.499**	1.329	**5/12**
G	13534.51-13489.16=**45.35**	DIF	-0.085	1.066	3/12
H	11814.67-11752.26=**62.41**	DIF	-0.174	2.440	1/12
I	13039.28-12879.32=**159.96**	DIF	**-1.241**	0.776	**8/12**
L	11375.29-11334.60=**40.69**	DIF	0.155	1.546	1/12
M	10913.12-10847.57=**65.55**	DIF	0.157	1.210	4/12
N	11706.05-11637.74=**68.31**	DIF	0.224	1.701	1/12
O	13068.80-13017.90=**50.90**	DIF	-0.037	1.512	2/12
Q1	11942.51-11904.73=**37.78**	DIF	0.043	1.051	3/12
Q2	10645.85-10592.57=**53.28**	DIF	0.082	1.558	2/12
Q3	12326.56-12301.92=**24.64**	Non-DIF	-0.046	1.589	0/12
Q4	11420.23-11376.28=**43.95**	DIF	-0.032	1.303	2/12
IM	11935.18-11897.81=**37.37**	DIF	-0.171	0.863	**5/12**

Note: DIF or non-DIF judgment was made by 1% level of significance.

DIF was shown in 15 of the 17 scales. However Thissen, Steinberg, and Wainer (1988) criterion was also applied (the difference of b-parameter mean should be less than 0.40) and this only reveals 3 problematic scales (A,F and I). Also, the equated a-parameter (slope) is > 0.75 for each scale. This means that every scale has acceptable discrimination.

In summary, the scales that showed differential item functioning were B, F, I, and Impression Management. The reason why a small amount of DIF was found could be due to:
(1) Subgroup imbalances: male sub group (N=769) compared to the female sub (N+166); and/or
(2) The procedure taken in this DIF analysis might be too strict. Research is ongoing and once more female data is collected other DIF analysis procedures, such as SIBTEST, D-FIT and/or Mantel-Haenszel will be run.

Concluding Remarks

The recent trends in globalization of business activities require the development of culturally and linguistically equivalent personnel assessment tools in order to fulfil the increasing demand of sustaining fairness and/or justice in global HRM settings. As an attempt to respond to this social and business demand, we tried to develop the Japanese version of 16PF Questionnaire through the collaboration with the US and UK publishers. By repeating translation, data gathering, item and scale analyse until having attained quantitative and qualitative equivalence of the test, we could consequently succeed to develop the Japanese version 16PF5, which is likely to be culturally and

linguistically equivalent to its US English version (Watanabe, Bedwell, and Williams,2006).

Through these endeavour, we learned some lessons about a test translation study.
(1)A literal translation of test items from the source language to a target language is often not possible.
For the Reasoning scale (Factor B), there were quite a few items which could not be translated into Japanese due to the unique characteristics of Japanese characters (script), which includes some meaning in the character itself. As a result some of the items had to be replaced or re-written to match the Japanese language context.

(2)Due to the complications involved in translating items, it is advisable to begin the adaptation process using more items than the test-developer anticipates needing for the final version.
The final Japanese draft of the questionnaire was sent to IPAT to be back-translated. The back-translated draft was then compared with the original research version by IPAT. Out of total of 265 items, 64 items were judged to have a different meaning in the back-translated version than the original. A further 9 items were judged to be either more extreme or were problematic with regard to the response options. In spite of a lengthy and involved translation procedure, approximately 30% of the items were identified as having problems.

(3) A reciprocal process of item translation and item analysis is important to successfully adapt a test for use in a target culture and language (Tsutsumi, Iwata, Watanabe, de Jonge,J. et al.,2009). Reciprocal process of item translation and item analysis resulted in the final version with no aberrant item characteristics for both CCT and IRT methods, except for the Reasoning items. As a result, quite a few of the Reasoning items were replaced. Then reliability coefficient alpha of all the factors finally exceeded .60 after item selection.

References

Allport,G.W. and Odbert, H.S. (1936). Trait-names: A psychotexical study. *Psychological Monographs*, **47**, 171.

Cattell, R.B. (1945). The description of personality: Principles and findings in a factor analysis. *American Journal of Psychology,* 58, 69-90.

Cattell,R.B., Cattell,A.K., and Cattell,H.E.P. (1993). *16PF Fifth Edition Questionnaire.* Champaign, Illinois: Institute for Personality and Ability Testing.

Conn,S.R. and Rieke,M.L. (1994). *The 16PF Fifth Edition Technical Manual.* Champaign, IL: Institute for Personality and Ability Testing.

Conn, S.R. and Rieke, M.L. (1998). *The 16PF Fifth Edition Technical Manual (2nd ed.).* Champaign, IL: Institute for Personality and Ability Testing.

Hambleton,R.K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessme*nt. 17(3), 164-172. DOI: 10.1027//1015-5759.17.3.164.

Long,L., Watanabe,N., and Tracey,T.J.G. (2006). Structure of interests in Japan: Application to the Personal Globe Inventory occupational scales. *Measurement and Education in Counseling and Development*, 38(4), 222-235.

Thissen,D.,Steinberg,L., and Wainer,H. (1988). Use of item response theory in the study of group diffences in trace lines. In H.Wainer & H.I.Braun (Eds.) *Test Validity.*

Hillsdale,NJ:Erlbaum.

Tracey, T. J. G., Watanabe, N., & Schneider, P. L. (1997). Structural invariance of vocational interests across Japanese and American Cultures. *Journal of Counseling Psychology*, 44, 346-354.

Tsutsumi,A., Iwata,N., Watanabe,N., de Jonge,J. et al. (2009). Application of Item Response Theory to achieve cross-cultural comparability of occupational stress Measurement. *International Journal of Methods in Psychiatric Research*, **18(1)**,58-67.

Watanabe,N. (1988). Cross-cultural job training in Japanese automobile companies in the United States. Working Paper Series No.8806, *Nanzan University Center for Management Studies.*

Watanabe,N. (1992). Item response theory and language translation: An aid to comparative studies of management in the United States and Japan . *Nanzan Review of American Studies, 14, 20-32.*

Watanabe,N. (1994). Application of item response theory to questionnaire translations. *Best Papers Proceedings of Association of Japanese Business Studies*, 687-707.

Watanabe,N.(1996). Examination of job interest structure through a development of a job interest index(JII) by item response theory (IRT). *Keio Business Forum*, 13(3), 179-198.

Watanabe,N. (1999). Changing human resource management practice in Japanese companies and its effects to employee well-being. *The Journal of Tokyo Medical University*, 58(3), 392-397.

Watanabe,N. (2012). Reliability and validity of the Japanese version of 16PF 5th edition: For personnel decision making in the era of globalization. *Keio Business Forum,* 29-1, 63-73.

Watanabe,N., Bedwell,S., and Williams,R. (2006). Development of culturally and linguistically equivalent tests: Some lessons learned from a test translation studies. *Paper presented at the 21st Annual Conference of Society for Industrial and Organizational Psychology(SIOP)*, Dallas:Texas,US.

Watanabe,N. and Nishida,T. (2003). Development of the Japanese version of 16PF Questionnaire (5th edition): Interim report. *Unpublished technical report.*

Watanabe,N. and Nishida,T. (2004). Development of the Japanese version of 16PF Questionnaire (5th edition): Final report. *Unpublished technical report.*