JOURNAL OF APPLIED MEASUREMENT, *13*(3),

Using Rasch Measurement to Validate the Big Five Factor Marker Questionnaire for a Japanese University Population

Matthew T. Apple
Nara National College of Technology

Peter Neff
Doshisha University

In recent years, psychological studies have increasingly come to support the so-called "Big Five" or "Five-factor Model" (FFM) of human personality. However, the vast majority of research in this field has been undertaken in Western contexts, thus raising the question of how applicable the Big Five is to Asian populations. Moreover, nearly all research into the Big Five has relied on traditional techniques of statistical analysis (e.g., factor analysis, correlation) to validate their results, despite the limitations of such methods. This study examined instrument validation of a widely-used Big Five instrument (the Factor Markers questionnaire) given to a Japanese population (n = 283) by using the Rasch rating scale model (Andrich, 1978). Rasch principal components analysis of the item residuals indicated the possible existence of additional factors within the Intellect/Imagination and Agreeableness factors, as well as additional item fit problems within each hypothesized construct.

Requests for reprints should be sent to Matthew T. Apple, Department of Liberal Studies, Nara National College of Technology, 22 Yata-cho, Yamato-koriyama, Nara, Japan 630-0212, e-mail: mapple@libe.nara-k.ac.jp.

In recent years, psychological studies have increasingly come to support the so-called "Big Five" or "Five-factor Model" (FFM) of human personality traits. Although the nomenclature may vary slightly according to the researcher, the generally agreed upon names of the five psychological constructs theorized to comprise this model are Extraversion-Introversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect/Imagination. In order to indirectly measure these traits, personality trait researchers typically rely on questionnaires developed from either of two competing taxonomies for the five factor model: The Five-factor Model supported by the 240-item Revised Neuroticism-Extraversion-Openness Personality Inventory (NEO-PI-R) and its related short-form 60-item version, Neuroticism-Extraversion-Openness Five Factor Inventory (NEO-FFI) (Costa and McCrae, 1992), and the Big Five model supported by Goldberg's "Factor Markers" 50-item and 100-item questionnaires (Goldberg, 1992, 1993, 1999).

While proponents of the two taxonomies have argued that the Big Five model and the Five-factor Model are slightly different (McCrae and Costa, 1996; Saucier and Goldberg, 1996), the Big Five moniker is now widely regarded as a "template" for personality trait researchers, and thus both terms have come to be used interchangeably (De Raad and Perugini, 2002; Fruyt, McCrae, Szirmak, and Nagy, 2004). In practice, personality trait researchers worldwide generally translate the Big Five Factor Markers questionnaire, the NEO-PI-R, or the NEO-FFI, or create their own instruments and validate their findings through correlation with existing questionnaires.

In order to encourage the exchange of personality trait instrument creation and data correlation, Goldberg and colleagues established an Internet site called the International Personality Inventory Pool, or IPIP for short (Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, and Gough, 2006). Although the site contains information regarding a Japanese version of the Factor Markers questionnaires, and despite the fact that the use of Five Factor Model questionnaires for research has become increasingly internationalized in recent decades, there has been surprisingly little evidence of its use inside Japan. The present paper is designed to provide an overview of the present state of Big Five personality trait studies in Japan, as well as the problem of Big Five personality trait instrument validation for a Japanese population.

Big Five Studies in Japan

Although the first possible indication of the existence of five factors for a Japanese population was a study conducted by Bond, Nakazato, and Shiraishi (1975), it was not until 1990s when further research into the factors gained momentum. The first sign of the Big Five in Japan was Isaka (1990), who conducted three successive studies based on the premise that adjectives embodied underlying human personality traits (the "psycholexical tradition"), and ultimately concluded that there were ten factors that roughly corresponded to the same five factors as the Bond et al study. After the advent of the Big Five moniker, several Big Five studies appeared in Japanese-language psychology journals during a six-year span from 1993 to 1999 (Kashiwagi, 1999; Kashiwagi and Wada, 1996; Kashiwagi, Wada, and Aoki, 1993; Kashiwagi and Yamada, 1995; Wada, 1996). Further evidence of increasing scholarly attention came when the NEO-PI-R was subsequently translated into Japanese (Shimonaka, Nakazato, Gondo, and Takayama, 1999). Kashiwagi (2002) additionally described the results of a Big Five study with 218 Japanese university participants and 200 Japanese adjectives, relying on a dichotomous rating scale. Additional Big Five studies using Japanese populations have also appeared as part of larger studies (Yik, Russell, Ahn, Dols, and Suzuki, 2002) or as part of small-scale, computer-assisted learning studies (Nakayama, Yamamoto, and Santiago, 2006; Santiago and Nakayama, 2006).

Other personality trait studies conducted in Japan have used the Eysenck Personality Questionnaire (Iwawaki, Eysenck, and Eysenck, 1980), the Yawate-Guilford Personality Inventory (Brown, Robson, and Rosenkjar, 2001), and the Myers-Briggs Type Indicator (Busch, 1982). Results from many of the above-mentioned studies, while somewhat informative, could also

be argued to have limited applicability to the larger Japanese population due to a combination of insufficient participant sizes (often less than 150-200) and at-times excessive item variable numbers (more than one item per case).

Confirming Construct Validity

From a statistical standpoint, one potential concern in Big Five personality trait studies conducted in Japan has been the use of personality trait measurement instruments without validating for construct unidimensionality. Big Five personality trait research in Japan has hitherto typically relied on two classical statistical methods to examine construct validity: exploratory factor analysis (EFA) and Cronbach's alpha. Characteristically, EFA is first conducted on the data, after which the resulting factors are then correlated with results from other measurement instruments using Cronbach alpha reliability estimates to validate the results.

However, the use of EFA and Cronbach's alpha has been criticized as being insufficient for determining construct validity in psychological survey instruments. The heart of these criticisms has often focused on the inherent limitations of each method. While commonly employed as a method of instrument validation, it has been argued that the use of EFA alone does not compare data sets to any other criteria and that, through multiple rotations, data can be easily manipulated (Tabachnick and Fidell, 2007). Waugh and Chapman have further commented that, "…just because scores on items are correlated doesn't mean that one has a conceptual scale of items even if there is a strong loading on a single factor" (Waugh and Chapman, 2005, p. 81). Moreover, Kline noted that a major drawback in reliance on EFA is the tendency of personality trait researchers to write items that are essentially paraphrases of each other in order to form a factor based on correlational analysis; thus, "[i]t is hardly surprisingly…that if such items are factored, they load onto a factor" (Kline, 2000, p. 340). Such factors are termed "tautologous factors" that fail the test of construct validity and cannot be generalized across samples.

As for the use of Cronbach's alpha, personality trait questionnaires are often listed as having Cronbach reliability estimates that are assumed to be consistent regardless of the sample population. However, "it is known that the size of an internal consistency index (e.g., coefficient alpha) is irrelevant to dimensionality" (Embretson and Reise, 2000, p. 231). Furthermore, high Cronbach's alpha reliability estimates have no relation to construct validity (Cortina, 1993; Green, Lissitz, and Mulait, 1977; Schmitt, 1996), and Cronbach's alpha is "persistently and incorrectly taken to be a measurement of the internal structure" of constructs that researchers intend to measure (Sijtmas, 2009, p. 107), despite the fact that alpha is highly dependent on the number of items in the construct (Nunnally, 1978).

Because previous Big Five studies in Japan have relied exclusively on EFA and Cronbach's alpha, the lack of instrument validation may have led to data and results whereby constructs were not measured as intended. In other words, the issue of instrument validation has yet to be addressed for Big Five personality trait research in a Japanese context. In this paper, we propose to confirm construct dimensionality for a Big Five questionnaire instrument through the use of Rasch model measurement analysis.

A Possible Solution: Rasch Model Measurement Analysis

The Rasch model (Rasch, 1960) is a unidimensional measurement model that calculates the relationship between item "difficulty" and person "ability" as the ratio of positive or negative endorsement of an item and expresses the difference in log-odds, or logits (Embretson and Reise, 2000). The Rasch model states that the probability a person will endorse an item is logistically related to the difference between the level of the latent construct present in the person and difficulty level of endorsing the item. In other words, the latent constructs being measured are determined by the probability of questionnaire respondents answering agree or disagree to items of varying degrees of endorsability ranged along the construct. Data is fit to the Rasch model by mathematically transforming raw scores on items

into logarithms and then by placing both item responses and persons on the same log-odds scale, which theoretically extends from negative to positive infinity but in practice typically extends from −5 to +5 logits. By doing so, the unrelated percentages of weakly and strongly endorsed items become transformed into a linear scale that can estimate probabilitistically how the questionnaire respondents are likely to answer to similar items on a future implementation of the questionnaire.

The original dichotomous Rasch model has the form:

$$\ln [P_{ni} / 1 - P_{ni}] = B_n - \delta_i,$$

where P_n is person n, i is the item being answered, B_n is the ability level of person n, and δ_i is the difficulty level of the item (Wright, 1999). Rasch model measurement analysis of Likert-type categorical data is typically conducted using the rating scale model (RSM), which is a special case of the Rasch polytomous model (Andrich, 1978; Ostini and Nering, 2006) and is thus frequently referred to as the Rasch rating scale model. The polytomous form of Rasch is:

$$\ln [P_{nij} / P_{ni(j-1)}] = B_n - D_i - F_j,$$

where P_n is person n, i is the item being answered, j is the response category of the Likert scale, B_n is the difficulty level of the construct on person n, D_i is the endorsability difficulty level of the item, and F_j is the Rasch-Andrich thresholds across categories j (Linacre, 2003a).

Analysis of personality questionnaires using the Rasch rating scale model can be valuable for five reasons. First, while such data obtained from questionnaires are often analyzed as if they were representative of an interval scale, in fact such data are ordinal. Using raw scores from Likert-type categorical data obtained from questionnaires in correlational analyses can thus potentially lead to erroneous conclusions (Bond and Fox, 2007; Wolfe and Smith, 2007a).

Second, traditional analyses such as exploratory factor analysis and correlation tend to treat each item on a measurement instrument as though it contributes equally to the measurement of the construct, regardless of whether some items are easier to endorse than others. RSM analysis can demonstrate the relative endorsability of items through the use of item-person maps, which display both items and persons on the same logit scale (Wilson, 2005, p. 96) according to item difficulty estimates.

Third, Rasch principal components analysis (PCA) of the item residuals is useful for determining the unidimensionality of Likert-type categorical data such as that obtained from a questionnaire. In PCA of item residuals, data obtained from a multidimensional instrument are examined by conducting an unrotated PCA on the item residuals that remain after extracting a linear measure (Bond and Fox, 2007; Wright, 1996a). The item residuals whose variance is not accounted for by the Rasch model may be correlated strongly enough to form spurious factors, thus reducing the validity of the factor analysis (Wright, 1996a, p. 10). Item residuals that correlate highly (e.g., above a factor loading of .40) have variance unexplained by the Rasch model and may need further examination to determine unidimensionality of construct (Bond and Fox, 2007; Linacre, 1998; Smith, 2002).

Fourth, Rasch model measurement analysis provides reliability figures both for items in the measurement instrument and for persons (i.e., study participant responses). Rasch model measurement analysis uses the statistical concept of separation to measure not only the conventional person reliability as an indication of person response consistency, but also item reliability, which indicates how well the items measured the sample population (Fisher, 1992). Separation is the ratio of error-free variance and observed variance and refers to the number of difference groups among the sample population distinguishable by the measurement instrument (Wilson, 2005; Wright, 1996b). Thus, by accounting for measurement error, Rasch model measurement analysis of person and item separation reliability statistics report a more accurate calculation of measurement instrument items for the sample population.

Fifth, Rasch model measurement analysis uses the concept of item fit rather than the reliance on Cronbach's alpha to demonstrate the quality

of items measured by the hypothesized constructs (Smith, 2001). Item and person responses that misfit the model may be the result of carelessness, response set answering, or item bias (Wolfe and Smith, 2007b, p. 211). Bond and Fox (2007) also recommend the use of Rasch model measurement analysis for Likert-type categorical data due to its ability to determine the level of endorsability of questionnaire items and the degree to which participants are being accurately measured. Through the combination of PCA of item residuals and item fit analysis, Rasch model measurement analysis can thus be used to support claims of generalizability of questionnaire results across samples (Wolfe and Smith, 2007a).

The lack of instrument validation for items measuring the Big Five personality traits severely hampers claims of cross-sample validation for Big Five-based personality trait measurement instruments. Thus, the aim of the present study was to address the lack of instrument validation of a Big Five questionnaire by examining the Factor Marker questionnaire with the Rasch rating scale model for Likert-type categorical data (Andrich, 1978; Bond and Fox, 2007; de Ayala, 2009). An additional aim was to examine the appropriateness of an American-developed five-factor model personality trait instrument for measuring a Japanese population.

Methodology

Participants

The participants in this study were 283 first and second year students studying English at a private, four-year undergraduate university in Kyoto, Japan. Students' majors included engineering, psychology, commerce, and philosophy. All participants were 18 or 19 years of age. Although the sample size was smaller than that recommended for traditional factor analysis ($n < 300$; Tabachnick and Fidell, 2007), the Rasch rating scale model functions with a minimum of 10 observations per category (Linacre, 2002). As a Likert-type scale of four categories was used for the present study, the minimum sample size for Rasch model measurement analysis for the measurement instrument in this study was $n = 40$. Linacre (1994) also noted a sample size of between $n = 108$ and $n = 243$, dependent upon the number of items, as sufficient for Rasch model measurement analysis, while de Ayala (2009, p. 199) suggested a sample size of no fewer than $n = 250$, provided that responses are distributed across categories in "reasonable numbers" (p. 199). These recommendations suggest the appropriateness of the sample size of $n = 283$ in the present study for Rasch measurement model analysis of the Big Five Factor Markers questionnaire instrument.

Instrumentation

Whereas the NEO-PI and NEO-FFI are proprietary questionnaires that must be purchased through a psychological assessment company, the 50-item version of Goldberg's Big Five Factor Markers questionnaire (Goldberg, 1993, 1999) is freely available through the aforementioned IPIP internet site (Goldberg, et al., 2006) and was utilized for this study (Appendix). The five human personality factors intended to be measured by the Factor Marker questionnaire are: Extraversion-Introversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect/Imagination. Of the 50 items comprising the Big Five Factor Markers questionnaire, 24 were negatively-worded.

The instrument used was a Japanese translation of the original 50-item English version. Although the original English version had a five-point Likert-type category scale, the version in the present study used four response categories (1, "Strongly disagree," to 4, "Strongly agree"). The Winsteps program used to analyze data obtained from the instrument recoded response categories from 0 to 3, as required for the Rasch model. The Japanese translation of the questionnaire instrument was provided by a native speaker of Japanese and was back-translated by a different native speaker to ensure accuracy of the statements. Participants completed the questionnaire in class at the midpoint of the academic year.

Analysis Procedures

In order to examine the validity of the Factor Marker questionnaire as a personality trait

measurement instrument for the sample population, Rasch model measurement analysis was performed using the Rasch rating scale model, or RSM (Andrich, 1978; de Ayala, 2009) in Winsteps 3.63 computer software (Linacre, 2006). The RSM was selected over other forms of categorical data analysis such as the partial credit model (PCM; Masters, 1982). The Factor Markers questionnaire was created with the intention of using the same response categories applied equally to all items; while RSM assumes that response categories remain consistent across all measurement instrument items (Ostini and Nering, 2006, p. 34). PCM allows for the thresholds between response categories to vary according to each item (Bond and Fox, 2007, p. 123). Therefore, RSM was felt to be a more appropriate method of analysis due to the assumptions of this measurement instrument.

Prior to data analysis, 24 negatively-worded items were recoded into positive items. An initial Rasch residual analysis was conducted using all 50 items using PCA of Rasch item residuals, and an item-person map of all items was produced to examine overall targeting of the sample. However, as the Factor Marker questionnaire was developed with the intention of measuring five independent psychological variables, further PCA of Rasch item residuals and Rasch item analysis were used to examine each of the five hypothesized constructs separately by inputting data from the ten items intended to measure each individual construct.

A PCA of the Rasch item residuals was conducted for each separate construct to determine individual construct unidimensionality. Objective measurement researchers have argued that random data in a factor analysis can have eigenvalues as high as 1.5 (Smith and Miao, 1994) to 2.0 (Wright, 1996a), and furthermore seldom produce factor loadings that can be reproduced by different sample populations. Thus, if the residual errors resulting from factor extraction were truly random, principal contrasts to the factor should have low eigenvalues (<1.5) and low percentages (<5%) of the total variance compared to that of the expected factor in order to demonstrate construct unidimensionality (Linacre, 1998). Higher percentages of variance may indicate that errors are correlated, masking the existence of a potential factor within the residuals left over from factor extraction.

A further analysis was conducted by correlating disattenuated person measures from principal contrast positive loading items and negative loading items as an additional verification of construct validity (Smith, E., 2002). *Disattentuated* refers to removing the item residual errors on person measures from item loadings on the factor contrast, after which the person measures from positively loading items are then correlated with measures from negatively loading items to determine whether the item responses indicate the same latent construct. High correlation of the person measures (>.7) suggest construct validity.

Rasch Likert scale category functioning analysis was also conducted for each individual construct to examine the effectiveness of the 4-point response scale employed. The following criteria proposed by Linacre (1999, 2002) were considered:

1. At least 10 observations should be present for each step of the scale.
2. Average person measures for each step should be higher than the average person measures of the previous step.
3. Outfit means squares of each step should be less than 2.0.
4. Gaps in step difficulties should be no fewer than .59 and no greater than 5 logits.

Finally, Rasch item fit analysis was conducted on each construct to examine item fit. An item mean squared fit statistic of 1.0 denotes an item that perfectly fits the expected model. High scores indicate items that overfit the model (i.e., a possible result of random answering or participants not seriously engaging in the task of taking the questionnaire); low scores indicate items that underfit the model (i.e., a possible indication of the "halo effect" of participants answering to please the researcher). Of particular interest are those items that may be underfitting the model, as the inclusion of these may be impacting measurements of reliability and unidimensionality.

For this study, item fit was considered good if the fit statistics fell between .6 to 1.4 for mean squares and ±2.0 for standardized *z*-scores (Bond and Fox, 2007; Smith, R., 2000). We should note that mean squares fit statistics have been criticized as sample-dependent, i.e., in large sample sizes items with little misfit can be identified as having larger misfit, thus increasing the likelihood of Type I error (Linacre, 2003b; Smith, Schumacker, and Bush, 1998). However, because the present study sample size was 283, neither very small nor very large, we will give both traditional mean squares in addition to standardized *z*-scores, with our interpretation of misfitting items focusing on outfit *z*-scores.

Results

An initial PCA of the item residuals was conducted on all 50 questionnaire items combined (Table 1). According to the results, the Rasch model explained 23.4% of the variance (eigenvalue = 15.3). The first contrast explained 8.3% of the variance (eigenvalue = 5.4).

An item-person map utilizing all 50 items from the Factor Marker questionnaire was also produced during the initial analysis (Figure 1). The item-person map indicated overall satisfactory targeting of the sample by the questionnaire items. The most difficult item to endorse was Emo 3 ("Worry about things," Rasch item difficulty measure = 1.00) and the easiest item to endorse was Int 4 ("Am not interested in abstract ideas," Rasch item difficulty measure = −.82). Item and person separation and reliability estimates were then calculated for the questionnaire as a whole. The resulting estimates indicated overall good item separation and reliability (Rasch item separation = 6.09, Rasch item reliability = .97), and poor to unacceptable person separation and reliability (Rasch person separation = 1.99, Rasch person reliability = .80). The low person reliability can be at least partially attributed to the fact that 24 of the 50 items on the questionnaire were negatively-worded.

Following initial analysis for all items combined, a Rasch PCA of the item residuals for conducted for each hypothesized construct to examine construct unidimensionality. Results (Table 2) indicate that none of the constructs are explained to a strong degree by their items, with explained variances ranging from 33% to 56.3%.

To confirm the multidimensional nature of each construct, an additional bivariate correlation analysis was performed using the disattenuated person measures from the positively and negatively loading items on the construct. Correlations of disattenuated person measures on each construct ranged from $r = .32$ to $r = .62$ (Table 2). Three constructs (Extraversion-Introversion, Conscientiousness, and Emotional Stability) had medium-strength correlations, and two constructs (Agreeableness and Intellect/Imagination) had weak correlations. As correlations differed significantly from the expected $r = 1.00$, the disattenuated person measures correlation analysis demonstrated the lack of unidimensionality for each construct.

Following the Rasch PCA of item residuals, item and person separation and reliability estimates were calculated for each hypothesized construct (Table 3). Person separation and reliability were minimally acceptable, with only Extraversion-Introversion reaching a score above 2.0 for the separation measure. According to Linacre (2006), person separation of 1.5 or above indicates an acceptable level of two different levels of separation, while a person separation of 2.0 indicates the construct to have two to three levels of separation, a more desirable result for psychological measurements. Person reliabilities, which indicate the consistency of the participant responses, ranged between .64 and .81.

On the other hand, item separation and reliability results were mixed. Although each factor consisted of ten items, only three of the five constructs (Conscientiousness, Emotional Stability, Intellect) had item separation above 5.0, an indication of the ability for the items to represent different levels of the construct within the sample population. Two of the constructs (Extraversion-Introversion and Agreeableness) had items that did not separate from each other well, with scores of 3.57 and 3.20, respectively. The item separation scores for these factors may indicate redundant

Table 1

Rasch Principal Components Analysis for the Big Five Factor Markers Questionnaire (50 Items)

Item	Item Description	Wording	Loading	Measure	Outfit MNSQ
EI 9	Don't mind being the center of attention	P	.58	.41	1.08
EI 8	Don't like to draw attention to myself	N	.56	−.03	.93
EI 1	Am the life of the party	P	.56	.49	.78
EI 2	Don't talk a lot	N	.52	−.22	.96
EI 7	Talk to a lot of different people at parties	P	.51	.23	.91
EI 5	Start conversations	P	.48	.03	.79
EI 6	Have little to say	N	.41	.12	1.23
Int 3	Have a vivid imagination	P	.41	−.39	1.00
Int 6	Do not have a good imagination	N	.37	−.33	.99
Int 5	Have excellent ideas	P	.36	.55	.72
EI 4	Keep in the background	N	.35	.12	.88
EI 10	Am quiet around strangers	N	.35	.54	1.19
Agr 2	Am interested in people	P	.30	−.54	.85
Int 10	Am full of ideas	P	.30	.03	.95
EI 3	Feel comfortable around people	P	.26	.04	.78
Agr 1	Feel little concern for others	N	.26	−.81	.92
Agr 5	Am not interested in other people's problems	N	.23	−.52	.87
Agr 10	Make people feel at ease	P	.19	−.07	.62
Agr 4	Sympathize with others' feelings	P	.18	−.52	1.05
Agr 9	Feel others' emotions	P	.04	−.25	.70
Agr 7	Am not really interested in others	N	.04	−.65	.87
Agr 8	Take time out for others	P	.01	−.32	.75
Con 4	Make a mess of things	N	−.57	−.04	1.30
Con 2	Leave my belongings around	N	−.55	−.06	1.29
Con 1	Am always prepared	P	−.51	.28	.90
Con 6	Often forget to put things back in their proper place	N	−.50	.50	1.54
Con 8	Shirk my duties	N	−.43	−.64	1.21
Emo 7	Change my mood a lot	N	−.43	.47	1.19
Emo 8	Have frequent mood swings	N	−.38	.13	1.19
Con 9	Follow a schedule	P	−.37	−.57	1.09
Con 5	Get chores done right away	P	−.36	.29	1.10
Con 7	Like order	P	−.32	.05	1.35
Emo 6	Get upset easily	N	−.30	.56	1.05
Agr 3	Insult people	N	−.30	−.70	.98
Emo 9	Get irritated easily	N	−.24	.08	.97
Int 9	Spend time reflecting on things	P	−.21	−.61	1.14
Int 7	Am quick to understand things	P	−.15	−.27	.73
Emo 10	Often feel blue	N	−.13	.42	1.14
Emo 5	Am easily disturbed	N	−.12	.72	1.11
Emo 4	Seldom feel blue	P	−.12	.84	1.10
Emo 1	Get stressed out easily	N	−.11	.02	1.11
Con 3	Pay attention to details	P	−.10	−.28	.90
Int 4	Am not interested in abstract ideas	N	−.07	−.82	.95
Con 10	Am exacting in my work	P	−.06	−.46	.91
Emo 2	Am relaxed most of the time	P	−.05	.01	.72
Int 1	Have a rich vocabulary	P	−.04	.87	.88
Int 2	Have difficulty understanding abstract ideas	N	−.03	.01	.85
Int 8	Use difficult words	P	−.03	.88	1.24
Agr 6	Have a soft heart	P	−.02	−.65	.86
Emo 3	Worry about things	N	.00	1.00	1.38

Notes: P = positive wording; N = negative wording; MNSQ = mean squared; EI = Extraversion-Introversion; Agr = Agreeableness; Con = Conscientiousness; Emo = Emotional Stability; Int = Intellect/Imagination; *n* = 283.

items that may not contribute meaningfully to overall measurement of the intended constructs, or may indicate that negatively- and positively-worded items were measuring separate constructs.

Next, Rasch Likert scale category functioning analysis was conducted for each individual construct (Table 4). Results indicated that there were no disordered thresholds. Outfit mean squares were within the fit criterion of below 2.0. All steps for all constructs were well within the criterion of .59 to 5 logits. However, the first category ("Strongly disagree") for the Agreeableness construct attracted only 4% of the possible person responses, while the third category ("Agree") attracted 52%, a possible indication that the

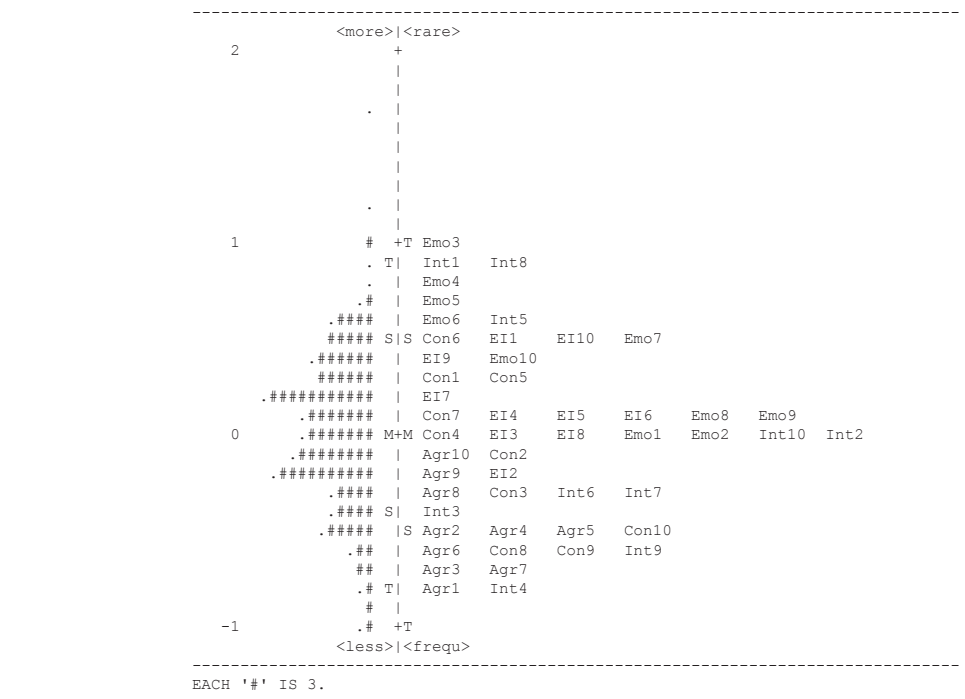

Figure 1. Item-person map for all 50 items of the Factor Markers questionnaire. Each # represents three persons. Each . represents one person. M stands for mean. S stands for one standard deviation from the mean. T stands for two standard deviations from the mean. Emo = Emotional Stability; EI = Extraversion-Introversion; Con = Conscientiousness; Int = Intellect/Imagination; Agr = Agreeableness; $n = 283$.

Table 2

Explained and Unexplained Variance for Each Big Five Factor Including Variance Explained by the First Contrast and Disattenuated Person Measures Correlations

Construct	Variance Explained	Unexplained Variance	First Contrast Explained Variance	First Contrast Eigenvalue	Disattenuated Person Measure Correlations
Extraversion	54.1 %	45.9 %	8.7 %	1.9	$r = .62$
Agreeableness	31.4 %	68.6 %	15.7 %	2.3	$r = .32$
Conscientiousness	46.8 %	53.2 %	9.3 %	1.7	$r = .50$
Emotional Stability	52.2 %	47.8 %	10.6 %	2.2	$r = .52$
Intellect/ Imagination	53.3 %	46.7 %	13.2 %	2.8	$r = .36$

*$n = 283$.

construct items were too easy for participants to endorse.

The final iteration of Rasch model measurement analysis carried out for each of the five constructs from the questionnaire involved looking at item measures, specifically those of fit (Table 5). First, infit and outfit means squared and standardized *z*-scores were obtained for the ten items in

Table 3

Person and Item Separation and Reliability Scores for Individual and Combined Constructs for a Japanese Population

Construct	Person Separation	Person Reliability	Item Separation	Item Reliability
Extraversion	2.08	.81	3.57	.93
Agreeableness	1.34	.64	3.20	.91
Conscientiousness	1.68	.74	5.20	.96
Emotional Stability	1.91	.78	5.00	.96
Intellect/Imagination	1.66	.73	8.61	.99
All Constructs	1.99	.80	6.09	.97

Table 4

Rasch Likert Scale Category Functioning Analysis for the Five Constructs Comprising the Big Five Factor Markers

Category	Observed Score Count	Percentage	Outfit MNSQ	Structure Calibration	Category Measure
			Extraversion-Introversion		
1 SD	396	14	1.01	—	(−3.25)
2 D	1148	41	.99	−2.06	−1.05
3 A	963	34	.91	.05	1.07
4 SA	312	11	1.13	2.02	(3.21)
			Agreeableness		
1 SD	109	4	1.24	—	(−3.26)
2 D	725	26	.93	−2.06	−1.15
3 A	1461	52	.85	−.17	1.07
4 SA	532	19	1.05	2.23	(3.40)
			Conscientiousness		
1 SD	387	14	.99	—	(−2.67)
2 D	867	31	1.03	−1.40	−.86
3 A	1091	39	.91	−.17	.80
4 SA	485	17	1.08	1.56	(2.78)
			Emotional Stability		
1 SD	656	24	1.04	—	(−2.86)
2 D	900	35	.92	−1.59	−1.00
3 A	916	33	.95	−.30	.87
4 SA	228	8	1.10	1.89	(3.07)
			Intellect/Imagination		
1 SD	334	12	1.11	—	(−3.23)
2 D	1067	38	1.00	−2.06	−1.01
3 A	1003	35	.87	.15	1.07
4 SA	423	15	1.05	1.91	(3.12)

Notes: SD = Strongly Disagree; D = Disagree; A = Agree; SA = Strongly Agree; *n* = 283.

Table 5

Rasch Item Fit Statistics for the Five Constructs Comprising the Big Five Factor Markers

Item	Wording	Measure	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD	PMC
			Extraversion-Introversion				
EI 6	N	−.09	**1.37**	**4.2**	**1.44**	**4.8**	.50
EI 10	N	.56	1.24	**2.8**	1.21	**2.4**	.66
EI 3	P	−.21	1.12	1.5	1.20	**2.3**	.48
EI 9	P	.36	1.08	1.0	1.05	.7	.65
EI 4	N	−.08	.97	−.4	.97	−.4	.66
EI 8	N	−.31	.91	−1.1	.92	−1.0	.64
EI 2	N	−.60	.87	−1.7	.86	−1.7	.66
EI 7	P	.07	.87	−1.7	.84	−2.0	.71
EI 1	P	.47	.77	**−3.1**	.81	**−2.4**	.67
EI 5	P	−.18	.77	**−3.1**	.76	**−3.1**	.68
			Agreeableness				
Agr 3	N	−.29	**1.34**	**3.8**	**1.37**	**4.1**	.36
Agr 4	P	−.03	1.25	**2.8**	1.26	**3.0**	.41
Agr 6	P	−.22	1.11	1.3	1.11	1.3	.45
Agr 9	P	.38	1.02	.3	1.05	.7	.42
Agr 7	N	−.22	1.02	.3	1.03	.4	.52
Agr 8	P	.27	.91	−1.1	.91	−1.2	.51
Agr 1	N	−.46	.88	−1.5	.88	−1.7	.67
Agr 10	P	.65	.87	−1.6	.87	−1.6	.51
Agr 2	P	−.06	.82	**−2.3**	.81	**−2.5**	.64
Agr 5	N	−.02	.79	**−2.7**	.79	**−2.7**	.65
			Conscientiousness				
Con 4	N	.06	1.14	1.7	1.14	1.8	.48
Con 7	P	.17	1.10	1.4	1.12	1.6	.59
Con 6	N	.71	1.11	1.5	1.10	1.2	.69
Con 10	P	−.44	1.08	1.0	1.11	1.2	.41
Con 8	N	−.65	1.08	1.0	1.11	1.4	.56
Con 3	P	−.22	1.02	.3	1.08	1.3	.41
Con 2	N	.03	1.02	.2	1.00	1.0	.59
Con 5	P	.46	.91	−1.2	.93	.1	.61
Con 9	P	−.57	.91	−1.2	.91	−.8	.55
Con 1	P	.45	**.57**	**−6.7**	**.58**	**−6.5**	.72
			Emotional Stability				
Emo 7	N	.06	**1.30**	**3.6**	**1.31**	**3.6**	.50
Emo 8	N	−.39	1.28	**3.3**	**1.30**	**3.4**	.55
Emo 3	N	.77	1.16	1.9	1.18	1.9	.58
Emo 6	N	.19	1.08	1.1	1.10	1.3	.58
Emo 2	P	−.56	.84	−2.1	.93	−.9	.58
Emo 5	N	.40	.93	−.9	.93	−.9	.64
Emo 1	N	−.55	.91	−1.2	.89	−1.5	.68
Emo 9	N	−.47	.88	−1.6	.88	−1.5	.65
Emo 4	P	.56	.81	**−2.6**	.77	**−2.8**	.69
Emo 10	N	.00	.75	**−3.5**	.76	**−3.3**	.75
			Intellect/Imagination				
Int 9	P	−.85	**1.46**	**5.2**	**1.50**	**5.4**	.33
Int 8	P	1.29	**1.37**	**4.1**	**1.36**	**4.0**	.45
Int 7	P	−.39	1.14	1.7	1.25	**2.9**	.39
Int 1	P	1.27	.93	−.8	.98	−.2	.56
Int 4	N	−1.13	.97	−.4	.96	−.5	.51
Int 2	N	.01	.90	−1.2	.90	−1.3	.59
Int 10	P	.04	.89	−1.4	.90	−1.2	.68
Int 5	P	.80	.77	**−3.1**	.76	**−3.3**	.65
Int 3	P	−.56	.76	**−3.3**	.75	**−3.3**	.72
Int 6	N	−.47	.76	**−3.4**	.75	**−3.4**	.75

Notes. P = positive wording; N = negative wording; Bolded numerals indicate misfit. EI = Extraversion-Introversion; Agr = Agreeableness; Con = Conscientiousness; Emo = Emotional Stability; Int = Intellect/Imagination; MNSQ = mean squares; ZSTD = standardized *z*-scores; PMC = Part-measure correlation; *n* = 283.

each construct separately. Based on the criteria of ±2.0 for Outfit standardized *z*-scores (Smith, 2000), Intellect/Imagination had six misfitting items, Extraversion-Introversion had five misfitting items, Agreeableness and Emotional Stability had four misfitting items, and Conscientiousness had only one misfitting item.

In order to further examine the match of items to participants, item-person maps based on item difficulty measures were generated for each construct (Figures 2, 3, 4, 5, and 6). Visual analysis indicates that some constructs were more successful than others. Items concerning Extraversion-Introversion grouped tightly around the mean, whereas the participants were revealed to be spread out along the construct to a greater degree than was targeted by the items. The Conscientiousness, Emotional Stability, and Intellect/Imagination constructs were more successful and resulted in a greater spread of both items and participants; however, there were still many participants not targeted by the items. Moreover, there was a gap of almost one standard deviation not covered by Intellect/Imagination items, potentially demonstrating the multidimensional nature of the construct (as its very name suggests).

The item-person maps also indicated both ceiling and floor effects. Emotional Stability items had a moderate floor effect with approximately half the participants falling below the range of the items, an indication that some items for this construct may have been too difficult for participants to endorse. On the other hand, the Agreeableness item-person map suggests a strong ceiling effect with a majority of participants above the range of the items, thus indicating that the items for this particular construct were too easily endorsable by the participants.

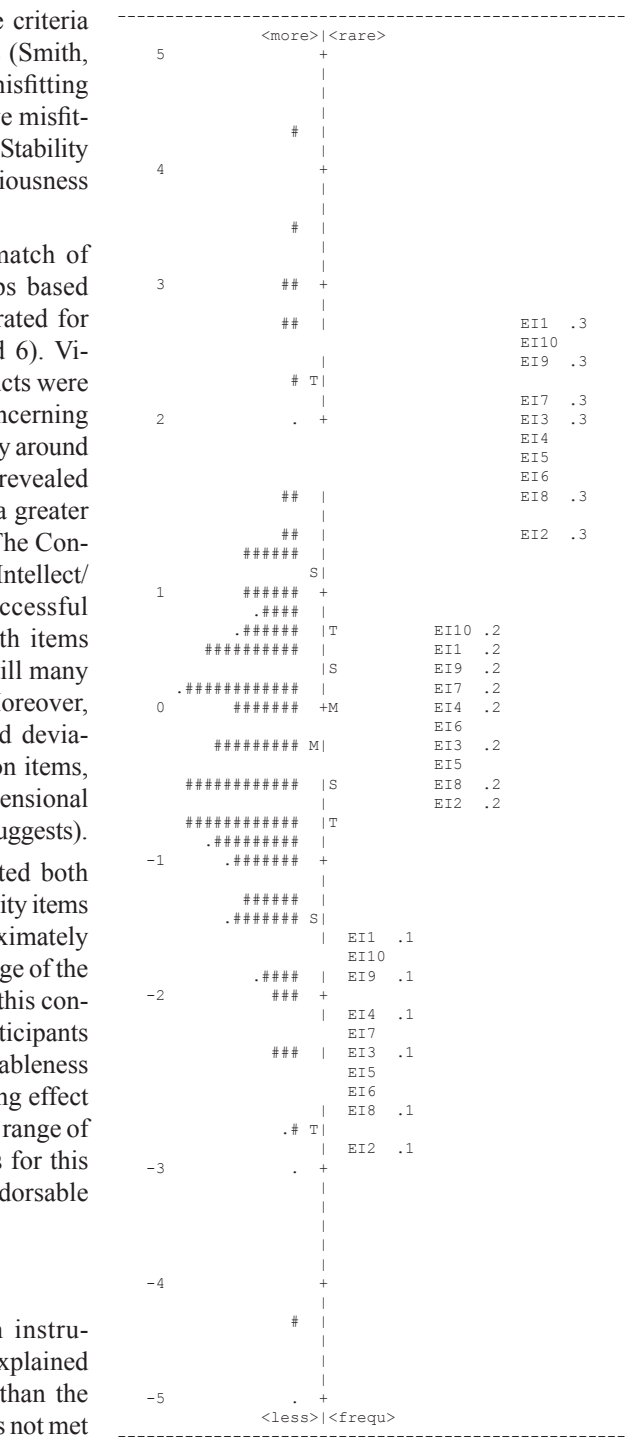

Figure 2. Item-person map for the Extraversion-Introversion construct. Each # represents two persons. Each . represents one person.

Discussion

According to Linacre (2006), an instrument can be considered reliable if the explained variance is at least four times greater than the unexplained variance. This condition was not met by any of the constructs in the Big Five Factor Markers instrument. Using the strict criteria by Smith (2002), who considered any first contrast

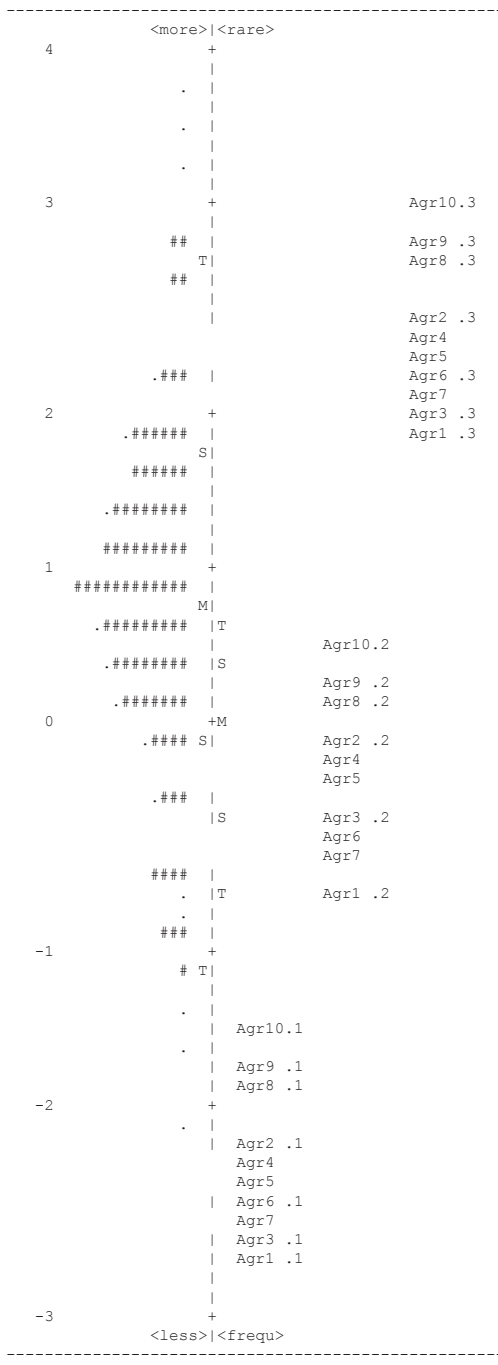

Figure 3. Item-person map for the Agreeableness construct. Each # represents three persons. Each . represents one person.

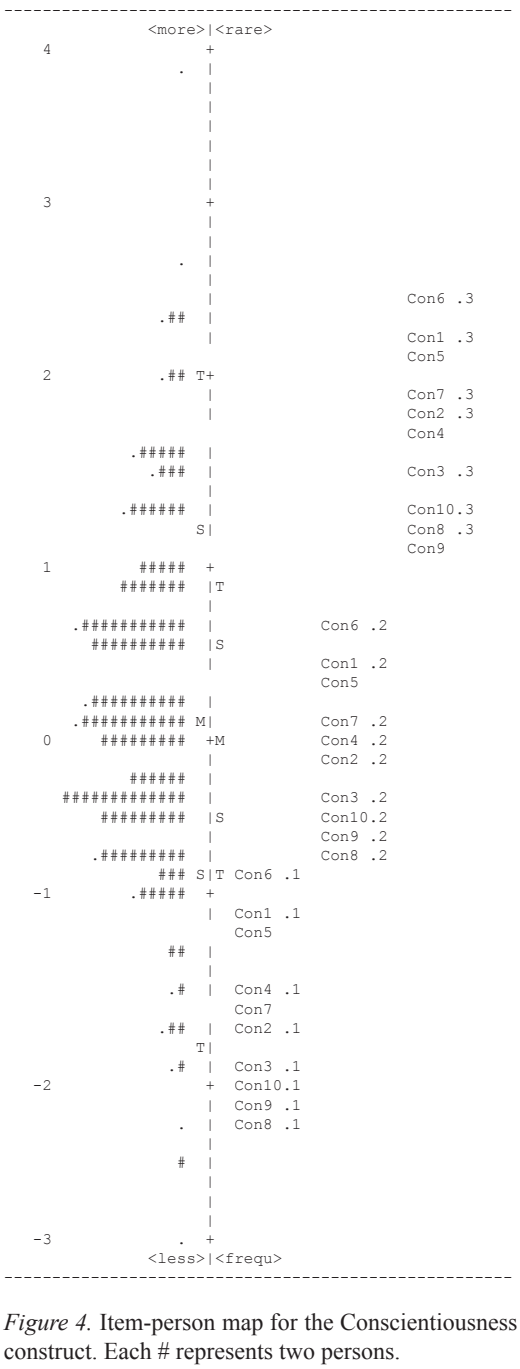

Figure 4. Item-person map for the Conscientiousness construct. Each # represents two persons.

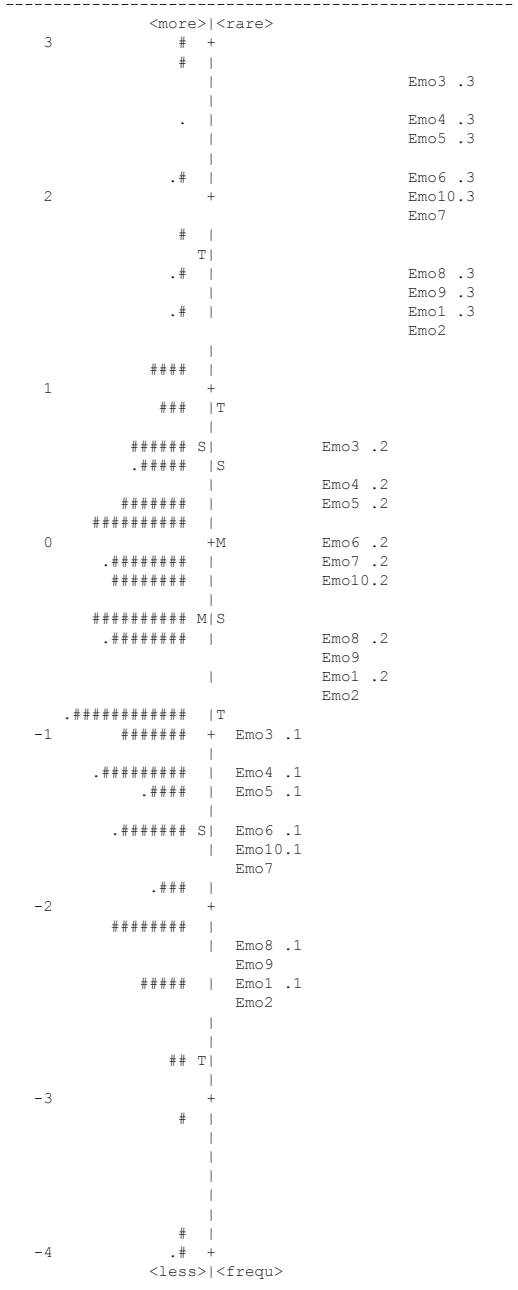

Figure 5. Item-person map for the Emotional Stability construct. Each # represents two persons.

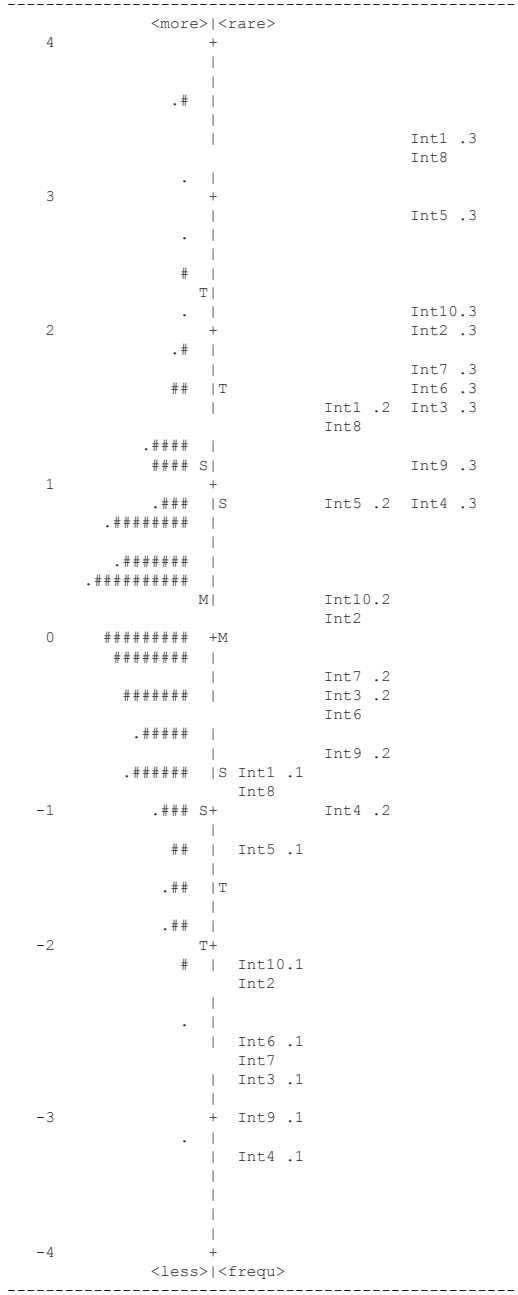

Figure 6. Item-person map for the Intellect/Imagination construct. Each # represents three persons. Each . represents one person.

eigenvalue over 1.5 to be an indication of a second construct, all of the measured factors would likewise fail the test of unidimensionality. On the other hand, Linacre (2006) recommended a less stringent level of 3.0 as one potential condition for bidimensionality, in which case none of the constructs would fail the test of dimensionality. However, because ideal first contrast values should have variances of less than 5% (Linacre, 1998), the large variances of first contrasts, ranging from 8.7% to 15.7%, hinted at the existence of correlated item errors and thus the presence of additional constructs.

Rasch model measurement analysis of the Big Five Factor Marker questionnaire demonstrated problems, to varying degrees, with the application of the Big Five to the Japanese sample in this study. The inferences can be drawn that none of the constructs from the Big Five Factor Markers demonstrated unidimensionality, and that at least two of the constructs contained at least one additional construct due to correlated residual errors that formed the first contrast in PCA of Rasch item residuals. Rasch item fit analysis further demonstrated low item separation scores across four of the five factors, with those for Extraversion-Introversion and Agreeableness in particular indicating that items failed to distinguish between endorsability levels of participants. Moreover, items intended to measure the Agreeableness construct did not adequately determine levels of the construct for the Japanese sample in this study. The item-person map and Likert category analysis suggested that the Agreeableness construct items were too easily endorsable by study participants.

Considering that the questionnaire items were designed for use with traditional factor analysis methods, the overall results were not too surprising. In other words, because the questionnaire items were constructed based on traditional statistical techniques, redundant positively- and negatively-worded items with approximately the same meanings were designed to produce high correlations and thus high Cronbach's alpha reliability estimates.

However, objective measurement studies have repeatedly shown that negatively keyed items do not necessarily measure the same construct as positively keyed items (Chang and Wright, 2001; Smith 1996), and that the inclusion of multiple items with virtually identical meanings in order to increase the Cronbach's alpha does not create a valid measurement instrument with unidimensional constructs. The PCA of Rasch item residuals compellingly demonstrated the weakness of reliance on correlation and Cronbach's alpha for construct validity in this Big Five Factor Marker questionnaire. All five constructs were shown to include at least one extra construct in the examination of item residuals, leading to the conclusion that the Big Five constructs for the sample in this study did not demonstrate construct validity or unidimensionality.

Conclusion

In order to adapt the Big Five Factor Markers questionnaire to target a Japanese population with greater precision, the following recommendations can be made based on Rasch model measurement analysis of the categorical data in this study. First, there is a need for a broader scope of items on all constructs, between the range of easy to endorse and difficult to endorse, in order to more accurately distinguish between participant levels. Second, the use of positively- and negatively-worded items on the same construct should be avoided in order to prevent item redundancy and to target a wider range of participant ability levels. Third, several individual items need to be either rewritten or eliminated. Items from the Agreeableness construct are most in need of item revision. For example, Item 12 (Agr 3, "Insult people") is an obvious candidate for elimination or revision.

Finally, as a Confucian-influenced society, Japanese culture is traditionally considered to be quite group-oriented compared to the Western population for whom the Big Five Factor Marker questionnaire was initially intended (Church and Lonner, 1998; Hendry, 2003). Personality psychologists often refer to the sense of "interdependent self" among members of Japanese society (Cross and Markus, 1994; Heine, Kitayama, and Lehman, 2001; Heine, Kitayama, Lehman,

Takata, Ide, Leung, and Matsumoto, 2001; Kitayama, 2000; Kitayama and Markus, 1999; Kitayama, Markus, and Lieberman, 1995; Markus and Kitayama, 1998). It may therefore be second nature for many Japanese to consider others' feelings in daily life to a greater degree than can be measured using the current instrument. This sense of attention to interpersonal relations is reflected in the floor effect shown in the item-person map of Agreeableness items, in which the endorsability difficulty level of the items was well below that of the questionnaire respondents.

The overall lack of construct unidimensionality and the inability of the questionnaire items to distinguish participant levels of the constructs in this study seem to support the conclusions in other Confucian-influenced countries (China and Korea) that for Asian populations the existing Big Five five-factor model structure may not be sufficient (Cheung, Leung, Zhang, Sun, Gan, Song, and Xie, 2001; Yoon, Schmidt, and Ilies, 2002). Although the participants in this study represent a fairly limited segment of Japanese society, we hope that the results from this study can be used to make revisions on future iterations of the instrument for use in measurements of a broader cross-section of Japanese society. The Factor Markers questionnaire needs to be adapted if it is to being meaningfully used with Japanese participants. More items should be added, or existing items should be revised or reworded, to reflect greater subtlety and a broader range of endorsability which may lead to more telling results and improve future psychological research involving Japanese participants.

Acknowledgements

The initial translation into Japanese of the Factor Markers questionnaire was supplied by Dr. Minoru Nakayama. The authors are grateful for his assistance. The authors also acknowledge Dr. Tim McNamara and Dr. David Beglar for their comments on an earlier draft of this paper.

References

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665-680.

Bond, M. H., Nakazato, H., and Shiraishi, D. (1975). Universality and distinctiveness in dimensions of Japanese person perception. *Journal of Cross-Cultural Psychology, 6*, 346-357.

Bond, T. G., and Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Brown, J. D., Robson, G., and Rosenkjar, P. R. (2001). Personality, motivation, anxiety, strategies, and language proficiency of Japanese students. In Z. Dörnyei and R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 361-391). Honolulu, HI: University of Hawaii.

Busch, D. (1982). Introversion-extraversion and the EFL proficiency of Japanese students. *Language Learning, 32*, 109-132.

Chang, C.-H., and Wright, B. D. (2001). Detecting unexpected variables in the MMPI-2 Social Introversion Scale. *Journal of Applied Measurement, 2*, 227-240.

Cheung, F. M., Leung, K., Zhang, J.-X., Sun, H.-F., Gan, Y.-Q., Song, W.-Z., and Xie, D. (2001). Indigenous Chinese personality constructs: Is the Five-Factor Model complete? *Journal of Cross-Cultural Psychology, 32*, 407-433.

Church, S. E., and Lonner, W. J. (1998). The cross-cultural perspective in the study of personality: Rationale and current research. *Journal of Cross-Cultural Psychology, 29*, 32-62.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Costa, P. T., and McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Cross, S. E., and Markus, H. R. (1999). The cultural constitution of personality. In L. A.

Pervin and O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed.) (pp. 378-396). New York: Guilford.

De Raad, B., and Perugini, M. (2002). Big Five factor assessment: An introduction. In B. De Raad and M. Perugini (Eds.), *Big Five Assessment* (pp. 1-26). Göttingen, Germany: Hogrefe and Huber Publishers.

Fisher, W. P., Jr. (1992). Reliability statistics. *Rasch Measurement Transactions, 6*, 238.

Fruyt, F. D., McCrae, R. R., Szirmak, Z., and Nagy, J. (2004). The Five-Factor Personality Inventory as a measure of the Five-Factor Model: Belgian, American, and Hungarian comparisons with the NEO-PI-R. *Assessment, 11*, 207-215.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26-42.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26-34.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. D. Mervielde, I. DeFruyt, and F. Ostendorf (Eds.), *Personality psychology in Europe, Vol. 7* (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. C. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Green, S. B., and Hershberger, S. L. (2000). Correlated errors in true score models and their effect on coefficient alpha. *Structural Equation Modeling, 7*, 251-270.

Green, S. B., Lissitz, R. W., and Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test undimensionality. *Educational and Psychological Measurement, 37*, 827-838.

Heine, S. J., Kitayama, S., and Lehman, D. R. (2001). Cultural differences in self-evaluation: Japanese readily accept negative self-relevant information. *Journal of Cross-Cultural Psychology, 32*, 434-443.

Heine, S. J., Kitayama, S., Lehman, D. R., Takata, T., Ide, E., Leung, C., and Matsumoto, H. (2001). Divergent consequences of success and failure in Japan and North America: An investigation of self-improving motivations and malleable selves. *Journal of Personality and Social Psychology, 81*, 599-615.

Hendry, J. (2003). *Understanding Japanese society* (3rd ed.). New York: Routledge.

Isaka, H. (1990). Factor analysis of trait terms in everyday Japanese language. *Personality and Individual Differences, 11*, 115-124.

Iwawaki, S., Eysenck, S. B. G., and Eysenck, H. J. (1980). Japanese and English personality structure. *Psychologia, 23*, 195-205.

Kashiwagi, S. (1999). The trait theoretic evaluation of the TEG from the view of the Five-Factor Model. *The Japanese Journal of Psychology, 69*, 468-477.

Kashiwagi, S. (2002). Japanese adjective list for the Big Five. In B. De Raad and M. Perugini (Eds.), *Big five assessment* (pp. 305-326). Seattle, WA: Hogrefe and Huber.

Kashiwagi, S., and Wada, S. (1996). A study on the concurrent validity of personality inventory from the view of the Five-Factor Model concerning personality traits. *The Japanese Journal of Personality, 67*, 300-307.

Kashiwagi, S., Wada, S., and Aoki, T. (1993). The Big Five and the oblique primary pattern for the items of the ACL Japan version. *The Japanese Journal of Personality, 64*, 153-159.

Kashiwagi, S., and Yamada, K. (1995). Evaluation of the Uchida-Kraepelin test based on the Five Factor Model of personality traits. *The Japanese Journal of Personality, 66*, 24-32.

Kitayama, S. (2000). Collective construction of the self and social relationships: A rejoiner and some extensions. *Child Development, 71*, 1143-1146.

Kitayama, S., and Markus, H. R. (1999). Yin and Yang of the Japanese self: The cultural

psychology of personality coherence. In D. Cervone and Y. Shoda (Eds.), *The coherence of personality: Social-cognitive bases of consistency, variability, and organization* (pp. 242-302). New York: Guilford.

Kitayama, S., Markus, H. R., and Lieberman, C. (1995). The collective construction of self esteem: Implications for culture, self, and emotion. In J. A. Russell and J.-M. Fernandez-Dols (Eds), *Everyday conceptions of emotion: An introduction to the psychology, anthropology, and linguistics of emotion* (pp. 523-550). New York: Kluwer Academic/Plenum.

Kline, P. (2000). The future of personality measurement. In J. Mohan (Ed.), *Personality across cultures: Recent developments and debates* (pp. 336-351). New Delhi, India: Oxford University Press.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-types work best? *Journal of Outcome Measurement, 2*, 266-283.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*, 193-212.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85-106.

Linacre, J. M. (2003a). Estimating 50% cumulative probability thresholds. *Rasch Measurement Transactions,16,* 901.

Linacre, J. M. (2003b). Rasch power analysis: Size vs. significance: Standardized chi-square fit statistic. *Rasch Measurement Transactions, 17,* 918.

Linacre, J. M. (2006). WINSTEPS Rasch measurement [Computer program]. Chicago.

Markus, H. R., and Kitayama, S. (1998). The cultural psychology of personality. *Journal of Cross-Cultural Psychology, 29*(1), 63-87.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

McCrae, R. R. (2002). NEO-PI-R data from 36 cultures. In R. R. McCrae and J. Allik (Eds.), *The Five-factor model of personality across cultures* (pp. 105-125). New York: Kluwer Academic.

McCrae, R. R., and Costa, P. T. (1996). Toward a new generation of personality theories: Theoretical contexts for the Five-Factor Model. In J. S. Wiggins (Ed.), *The Five-Factor Model of personality: Theoretical perspectives* (pp. 51-87). New York: Guilford.

Nakayama, M., Yamamoto, H., and Santiago, R. (2006). Investigating the impact of learner characteristics on blended learning among Japanese students. In D. Remenyi (Ed.), *International conference on e-Learning* (pp. 361-370). London: Academic Conferences Press.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Ostini, R., and Nering, M. L. (2006). *Polytomous item response theory models.* Thousand Oaks, CA: Sage.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Santiago, R., and Nakayama, M. (2006). Understanding e-learners' characteristics and performance in online courses. In R. Mizoguchi, P. Dillenbourg, and Z. Zhu (Eds.), *Frontiers in artificial intelligence and applications, Vol. 151: Learning by effective utilization of technologies: Facilitating intercultural understanding* (pp. 521-524). Washington, DC: IOS Press.

Saucier, G., and Goldberg, L. R. (1996). The language of personality: Lexical perspectives on the Five-Factor Model. In J. S. Wiggins (Ed.), *The Five-Factor Model of personality: Theoretical perspectives* (pp. 21-50). New York: Guilford.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*, 350-353.

Shimonaka, Y., Nakazato, K., Gondo, Y., and Takayama, M. (1999). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-factor Inventory (NEO-FFI) manual for the Japanese version*. Tokyo: Tokyo Shinri. Cited in McCrae, R. R. (2002). NEO-PI-R data from 36 cultures. In R. R. McCrae and J. Allik (Eds.), *The Five-factor Model of personality across cultures* (pp. 105-125). New York: Kluwer Academic.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measurement interpretation: A Rasch measurement perspective. *Journal of Applied Measurement*, *2*, 281-311.

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205-231.

Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling, 3*, 25-40.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1*, 199-218.

Smith, R. M., and Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice, Vol. 2* (pp. 316-327). Greenwich, CT: Ablex.

Smith, R. M., Schumacker, R. E., and Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.

Tabachnick, B., and Fidell, L. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson Education.

Wada, S. (1996). Construction of the Big Five Scales of personality trait items and concurrent validity with NPI. *The Japanese Journal of Personality, 67*, 61-67.

Waugh, R. F., and Chapman, E. S. (2005). An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: What is the difference? Which method is better? *Journal of Applied Measurement, 6*, 80-99.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Wolfe, E. W., and Smith, E. V., Jr. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I —Instrument development tools. *Journal of Applied Measurement, 8*, 97-123.

Wolfe, E. W., and Smith, E. V., Jr. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II —Validation activities. *Journal of Applied Measurement, 8*, 204-234.

Wright, B. D. (1996a). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3*, 3-24.

Wright, B. D. (1996b). Reliability and separation. *Rasch Measurement Transactions, 9*, 472.

Yik, M. S. M., Russell, J. A., Ahn, C.-K., Dols, J. M. F., and Suzuki, N. (2002). Relating the five-factor model of personality to a circumplex model of affect. In R. R. McCrae and J. Allik (Eds.), *The Five-Factor Model of personality across cultures* (pp. 74-104). New York: Kluwer Academic.

Yoon, K., Schmidt, F., and Ilies, R. (2002). Cross-cultural construct validity of the Five-Factor Model of personality among Korean Employees. *Journal of Cross-Cultural Psychology, 33*, 217-235.

Appendix: The 50-item Factor Marker instrument (Goldberg, 1999)

Construct Item Label	Item Number	Item Description
EI 1	1	Am the life of the party.
Agr 1	2	Feel little concern for others.*
Con 1	3	Am always prepared.
Emo 1	4	Get stressed out easily.*
Int 1	5	Have a rich vocabulary.
EI 2	6	Don't talk a lot.*
Agr 2	7	Am interested in people.
Con 2	8	Leave my belongings around.*
Emo 2	9	Am relaxed most of the time.
Int 2	10	Have difficulty understanding abstract ideas.*
EI 3	11	Feel comfortable around people.
Agr 3	12	Insult people.*
Con 3	13	Pay attention to details.
Emo 3	14	Worry about things.*
Int 3	15	Have a vivid imagination.
EI 4	16	Keep in the background.*
Agr 4	17	Sympathize with others' feelings.
Con 4	18	Make a mess of things.*
Emo 4	19	Seldom feel blue.
Int 4	20	Am not interested in abstract ideas.*
EI 5	21	Start conversations.
Agr 5	22	Am not interested in other people's problems.*
Con 5	23	Get chores done right away.
Emo 5	24	Am easily disturbed.*
Int 5	25	Have excellent ideas.
EI 6	26	Have little to say.*
Agr 6	27	Have a soft heart.
Con 6	28	Often forget to put things back in their proper place.*
Emo 6	29	Get upset easily.*
Int 6	30	Do not have a good imagination.*
EI 7	31	Talk to a lot of different people at parties.
Agr 7	32	Am not really interested in others.*
Con 7	33	Like order.
Emo 7	34	Change my mood a lot.*
Int 7	35	Am quick to understand things.
EI 8	36	Don't like to draw attention to myself.*
Agr 8	37	Take time out for others.
Con 8	38	Shirk my duties.*
Emo 8	39	Have frequent mood swings.*
Int 8	40	Use difficult words.
EI 9	41	Don't mind being the center of attention.
Agr 9	42	Feel others' emotions.

(Appendix continues on the next page.)

(Appendix continues from the previous page.)

Appendix: The 50-item Factor Marker instrument (Goldberg, 1999)

Construct Item Label	Item Number	Item Description
Con 9	43	Follow a schedule.
Emo 9	44	Get irritated easily.*
Int 9	45	Spend time reflecting on things.
EI 10	46	Am quiet around strangers.*
Agr 10	47	Make people feel at ease.
Con 10	48	Am exacting in my work.
Emo 10	49	Often feel blue.*
Int 10	50	Am full of ideas.

Notes. Questionnaire items are listed in their original order (Greenberg, 1993). Construct item labels have been added for ease of the reader. Items marked with an asterisk (*) were negatively-worded items recoded prior to analysis. EI = Extraversion-introversion; Agr = Agreeableness; Con = Conscientiousness; Emo = Emotional Stability; Int = Intellect / Imagination.