

University of Groningen

The construction of the five-factor personality inventory (FFPI)

Hendriks, Anja Alide Jolijn

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

1997

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Hendriks, A. A. J. (1997). *The construction of the five-factor personality inventory (FFPI)*. [s.n.].

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

8. Summary and Discussion

This study deals with the construction and a first psychometric evaluation of the Five-Factor Personality Inventory (FFPI; Hendriks, Hofstee, & De Raad, in preparation; Hendriks, Hofstee, De Raad, & Angleitner, 1995), which instrument efficiently assesses five broad dimensions (factors) of individual differences in behavior: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Autonomy. The FFPI consists of 100 brief and concrete statements in the third person singular, and can be administered in 10-15 minutes. In addition to the five factor scores, the FFPI may be used to assess 40 bipolar facet scores that arise as blends of the five factors.

The FFPI was developed within the psycholexical paradigm in which it is assumed that natural language provides for, or else will incorporate, all elements that people find important in their communications about individual differences in behavior. In line with this assumption, which is also known as the "sedimentation hypothesis" (Brokken, 1978) or "lexical hypothesis" (Goldberg, 1990; McCrae, 1990), adherents of the paradigm take the (unabridged) dictionary as their main source for delineating the "personality trait domain". Presumably the English scientist and writer Sir Francis Galton (1884) was the first person to scan a dictionary for personality trait descriptors (John, Angleitner, & Ostendorf, 1988). For the lexical approach to personality, however, Allport and Odbert's (1936) dictionary scan has been of more significance. These authors were the first to systematically list and categorize all terms in the English language that can "... distinguish the behavior of one human being from that of another" (p. 24). Following Allport and Odbert, dictionary scans have been undertaken by researchers in many different countries.

These "(multi-)trait psychologists" especially selected and subsequently empirically examined the interrelationships of the "stable traits", which were taken to be represented primarily by trait adjectives (*talkative*, *docile*, and so on), with the ultimate goal of revealing the *structure* of personality. Nowadays, more and more consensus is emerging that four or five replicable factors adequately summarize people's variance on most of these traits. Moreover, while researchers have used partly different labels in the past, it is now becoming common practice to refer to the first four personality factors as Extraversion, Agreeableness, Conscientiousness, and Emotional Stability. Concerning the fifth factor, debate is still going on, not only with respect to its label, but also on whether this factor proves to be replicable at all. While it was originally referred to as Culture (Tupes & Christal, 1961), and subsequently as Intellect (Goldberg, 1990; Ostendorf,

1990), recently suggested labels for this fifth factor are Creativity (Johnson, 1994; Saucier, 1992), Originality (Saucier, 1992), or Imagination (Saucier, 1992, 1994). Another widely used label is Openness to Experience (Costa & McCrae, 1985, 1992); this label, however, does not follow from studies in the psycholexical tradition.

From the moment it became available, Principal Components Analysis (PCA) followed by varimax rotation has been *the* method to reveal the structure of personality traits. The classical "Big-Five" simple-structure model, in which trait terms are assigned to the one factor in the five-dimensional trait space on which they have their largest projection (loading), has been refined to the Abridged Big-Five Dimensional Circumplex (AB5C) Model (Hofstee & De Raad, 1991; Hofstee, De Raad, & Goldberg, 1992) some years ago. Most trait terms appeared to be blends of *two* of the Big-Five factors: apart from having a high loading on one factor, they have a quite substantial loading on a second, which means that these trait terms are close to a vector in the five-space which is positioned somewhere in between the pertaining two Big-Five factors. The AB5C model takes this finding into account. The model distinguishes between factor-pure variables, which have no secondary loading of any significance, and two blends: one in which a specific Big-Five factor plays a primary role and one in which that same factor plays a secondary role. Examples are the two blends I+II+ ("cheerfulness") and II+I+ ("kindness"), as distinguished from factor-pure I+ (Extraversion) and factor-pure II+ (Agreeableness) variables. The result is a partitioning of the five-dimensional trait space into 90 "facets" containing clusters of trait terms that are much more homogeneous than the five large clusters in the Big-Five simple-structure model. The AB5C model thus offers a much better anchor point for factor interpretation, by virtue of its representation of nuances in trait meaning.

We took the AB5C model as a point of departure for the construction of an inventory that covers the five-dimensional trait space. Based on the content of all well-filled (65) facets, brief concrete statements (sentence items) were written (see section 2.1). We considered a facet to be well-filled if it contained at least three trait adjectives having a projection of .40 or more on the pertaining facet vector. The instructions to team members were to define the meaning of each AB5C facet in a recursive way, that is, by taking the shared meaning of its composing cluster of trait terms, while contrasting it to the meaning of its opposite cluster and centering it between its two adjacent clusters. Each team member independently wrote as many items as he or she could think of --typically some seven per facet-- given the content of the AB5C facets that were taken as a point of departure for a particular team session. Per team session, which were held once or twice a week, two to three facets were thus addressed. All independently produced items were

examined one by one for their fit to the facet for which they were written, and on whether they fulfilled explicit guidelines for item production (Hofstee, 1991). These guidelines served the purpose of creating an instrument that can be used for a broad range of educational levels, avoids discrimination of certain people or groups of people, and elicits ratings as objective as possible. This first stage of the project resulted in a preliminary Dutch item pool of 909 sentence items. Examples are: *Has a good word for everyone* and *Makes friends easily*.

Additional sentence items were produced on the basis of personality descriptive verbs (see section 2.2). This word class is fit par excellence for use in brief behaviorally concrete statements because, in contrast with trait adjectives and nouns which may be used in simple rating lists, verbs need further specification in order to make clear what is meant. For instance, *Listens to others* is something totally different from *Makes others listen to him/her*. Furthermore, items based on personality descriptive verbs may add meaning that is not contained in trait adjectives (De Raad, Mulder, Kloosterman, & Hofstee, 1988). In total 136 such sentence items were added to the Dutch item pool. Examples are: *Insults people* and *Knows how to convince others*.

At this stage of the project, the items were translated into American-English, in cooperation with Lewis R. Goldberg from the Oregon Research Institute, and into German, by Alois Angleiter and his team from the University of Bielefeld. Translation difficulties are a notorious problem when tests which are developed in one language are translated into other languages. For this reason, we considered the translatability of the items to be a prerequisite for their inclusion in the final item pool. This strategy has the advantage that any difficulties in this respect are encountered before, and not after, the final item selection has taken place. Because of the time-consuming nature of such a procedure, we had to limit ourselves to American-English and German. But, if items proved to be translatable into these two languages, their chance of being translatable into still other languages was expected to be greatly enhanced.

Against this background, the initial Dutch item pool was extended from a third source (see section 2.3). As the Dutch Factor V is best interpreted as Rebelliousness or Spirit (Hofstee et al., 1992), whereas the American and German Factors V are much more Intellectual, additional sentence items referring to Intellect were written in order to link up the initial Dutch item pool to the American and German Factors V. In total 266 Intellect sentence items were added to the pool. Examples are: *Wants to understand things* and *Wants to form his/her own opinions*.

The initial Dutch item pool eventually contained a total of 1,311 sentence items. The translation of this item pool into American-English and German appeared to be an intricate

process, and is best referred to as an "internationally interactive" way of constructing the final item pool (see Chapter 4). First of all, 397 items were discarded in the process because of one of the following reasons: (1) no good translation could be found in one or either of the other languages, or (2) the item was judged to be inferential, that is, requiring an inference by the rater, or (3) the item was judged to be a social effect, or (4) its translation in one or both of the other languages revolved around a trait adjective, or (5) the item was judged to be ambiguous, or (6) the item was judged to be too specific. Several other items needed adaptations, which could mean replacing them by their closest back-translation. After a final round of cleansing with regard to the guidelines for item production, a total of 914 sentence items remained to constitute the trilingual item pool.

In The Netherlands, self- and others' ratings were collected on the 914 sentence items, together with ratings of this sample on a 225-item trait-adjective rating list that covers the five-space. The aim of the study was to determine the relationships of the sentence items with the adjective-based Big-Five dimensions, and to evaluate the items' psychometric quality (see Chapter 5). Target subjects were 167 first-year students and staff members of psychology. With the exception of eight of them, each target provided a self-rating, and was independently rated by two to four others who knew the target person well, giving a total of 790 raters.

A principal components analysis (PCA) was performed on the pooled ($N = 790$) sample of self- and others' ratings on the 225 trait adjectives; each subject's raw scores were corrected beforehand for acquiescence response set (Hofstee, 1994b). The appropriateness of pooling the sample was checked in a pre-analysis (see section 5.1.4). The scree test clearly indicated five factors, which were subsequently varimax rotated. Next, Pearson correlations were calculated between the 914 sentence items and the five varimax-rotated factors, which were identified as Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Autonomy. The AB5C facet projections of the sentence items appeared to range from $-.66$ to $.71$; their median absolute value was found to be $.47$. In other words, we found a clear demonstration of the relationship between the sentence items on the one hand and the adjective-based Big-Five dimensions on the other hand. In addition, Pearson correlations were calculated between self- and averaged others' ratings, indicating the self-peer validity of the sentence items. These values were found to be very promising (median value: $.35$).

We further investigated the effects of social desirability and observability on interjudge agreement (self-[averaged]others correlations), following John and Robins (1993). The sentence items had been rated on these two variables (see section 3.2), as well as on comprehensibility

(see section 3.1), by independent samples at an earlier stage of the project. These ratings were collected in order to be used as secondary criteria for item selection. We only partly replicated John and Robins' findings: a moderately strong positive relationship (.30) between observability and interjudge agreement, but no relationship of any significance between social desirability and interjudge agreement was found.

Our intention was to use measurement equivalence in the three languages as one of the primary criteria for item selection (see section 6.1), next to the items' factor loadings, self-peer validity, and comprehensibility. Data ($N = 766$, self-ratings) on the English items were made available by Lewis R. Goldberg, and data ($N = 118$, peer ratings) on the German items were made available by Alois Angleitner. According to a pilot study (Houtman, 1994), a Mokken scale analysis (Molenaar, Debets, Sijtsma, & Hemker, 1994) followed by an analysis on Differential Item Functioning and Differential Test Functioning (Stout & Roussos, 1992, 1994), both in their versions for polytomous items, seemed a promising combination of methods for our goal. We had to conclude, however, that the undertaking met a dead end. The main reason to abandon measurement equivalence as a selection criterion was a substantial loss of items, and item variety, due to large differences in the scalability of the items between the Dutch and German data on the one hand and the American data on the other hand, for which differences we had no good explanation. A second thought in this respect was that one is not necessarily interested in item equivalence on a one-to-one basis, but in equivalence of scale scores.

Instead, we used the results in the Dutch data set to make a preselection of 284 sentence items, on which self- and others' ratings were collected in a fresh Dutch sample. This strategy enabled us to enlarge the sample of subjects on which the final item selection would take place, while having reduced the number of items to a more manageable (for subjects) proportion which only contained the most promising ones, and to collect data on convergent measures in order to be able to perform a preliminary validity study. Target subjects were 125 first-year students of psychology. With the exception of eight of them, each target gave a self-rating, and each target was rated by two to four others who knew the target person well, giving 606 raters in total. Self-ratings were collected on the 284 sentence items, the 225-item trait-adjective rating list, and the NEO-PI-R (Costa & McCrae, 1985, 1992; Dutch translation: Hoekstra, Ormel, & De Fruyt, 1996). Others' ratings were collected on the 284 sentence items only. This sample was pooled with the available ($N = 790$) data set, giving a total of 1,311 raters ($N = 292$ targets), after deletion of subjects with suspect response profiles. Because they had had different instructions for their ratings, a check on the appropriateness of pooling the two samples was performed. An

analysis of variance in intra-individual spreads across the 284 sentence items revealed significant main effects for Instruction and Rater (self- or other-rating). Consequently, in all analyses using this pooled sample, each subjects' raw scores were corrected beforehand not only for acquiescence, but also for differences in intra-individual spreads.

In order to determine their AB5C-facet positions, a PCA was performed on the 284 sentence items, followed by a varimax rotation of the first five factors (in accordance with the scree test). The final positions of the axes, however, were subsequently based on the joint results in the Dutch and American data sets; not on the German data set, by reason of its much smaller number of raters. We decided to take the American-English structure into account for the final item selection because, in spite of the disappointing results in the analysis of measurement equivalence, the two structures appeared to be remarkably similar: the congruence coefficients for the varimax-rotated principal components were all found to be (well) above .87. In order to establish the final positions of the axes, a Procrustes rotation to optimal agreement was performed on the Dutch and American matrices of loadings, whereupon the consensus matrix was varimax rotated once more (Evans, 1971, p. 43; Hakstian, 1973, p. 226; see also Kiers, 1995, for a comparison of methods). Next, the items' AB5C facet positions were determined by taking their two highest loadings.

Several item sampling plans were considered (see section 6.4.1). We finally decided to select 20 good sentence items per factor having their primary loading on that factor, to be spread across the different facets of the factor so as to avoid redundancy. Consequently, the unweighted sum scores of the items for the five scales are mostly positively correlated; an orthogonalization procedure (cf. Costa & McCrae, 1992) is needed to obtain orthogonal factor scores.

The 100 items that constitute the FFPI were primarily selected on the basis of having a large projection, a substantial self-(averaged)others correlation (self-peer validity), and a low Difficulty score (high comprehensibility). Observability and non-extreme social desirability served as marginal criteria. The items meet the standards for item quality to a considerable extent, as Table 10 (Chapter 6) shows.

A first psychometric evaluation (see Chapter 7) revealed that the FFPI scale and factor scores show high internal consistencies, substantial (six months) stabilities, and very promising self-peer validities. These results can only be taken to be tentative, however, as this first psychometric evaluation was conducted within the available data set in which also the item selection took place. Also, these values may be expected to shrink somewhat in other-language versions.

Additional indications of construct validity were obtained from correlations between FFPI factor scores and scores on two other five-factor personality inventories. We found a clear-cut convergent validity between the FFPI and the 225-item trait-adjective rating list. Although far from being trivial, these results serve mainly to confirm an expectation, since both instruments stem largely from the same source, namely the AB5C model of personality traits. With respect to the NEO-PI-R, clear-cut convergent validity was found for Extraversion, Agreeableness, Conscientiousness, and Emotional Stability (reversed). FFPI-Autonomy and NEO-Openness to Experience appeared to have almost nothing ($r = .10$) in common. Although this correlation of .10 will almost certainly prove to be an underestimate due to a sample peculiarity, we do expect a lack of common variance between these two variables to be a robust finding (see, for instance: McCrae, 1990). This is not necessarily a problem in itself; the question is more generally what to think of this lack of common variance.

Reflections on Factor V

McCrae and Costa (McCrae, 1990) believe that the fifth factor identified in psycholexical studies is best interpreted as a variant of Openness to Experience, which they consider to be a psychologically fundamental dimension; other interpretations are taken to be confoundings of intelligence, education and sophistication with this more basic factor (p. 122). According to McCrae and Costa, the reason why the lexical tradition has failed to portray adequately Factor V as Openness to Experience is the customary rigid ("single-word") application of the lexical hypothesis, which, to their opinion, results in an underrepresentation of certain aspects of the personality domain. As McCrae (1990) states: "When examined by facet, it appears that English has many words that express Openness to Ideas, but fewer that capture other facets of Openness" (p. 124). There is no simple answer to this remark as yet, at least with respect to the FFPI. If one adheres to the lexical tradition in its strongest (single-word) formulation, the obvious reply would be that people most probably find these other facets of Openness (to Fantasy, Aesthetics, Feelings, Actions, and Values) of less importance. But if one, like the Groningen team, adheres to the essence of the lexical hypothesis, which merely states that individual differences in behavior that people find important to communicate about will eventually become encoded into the natural language, the situation is less clear-cut. Did we fail to portray adequately Factor V as Openness to Experience? When it does take sentences to convey individual differences in (facets of) Openness to Experience as it is suggested by McCrae and Costa, indeed there may be an

omission on the part of the FFPI: its basis has been (clusters of) single trait terms, Intellect terms, and personality descriptive verbs. De Raad (1994), however, points at the inconsistent findings on which the postulation of Openness to Experience as one of the basic personality dimensions was based. So, for the time being, Openness to Experience is but one of the candidates for a universal Factor V, no more so than Intellect, or Creativity, or Imagination, or, for instance, Autonomy. That is to say, if a fifth factor proves to be replicable at all.

Although we did our best to come up with Intellect as a fifth dimension (see Chapter 2), we ended up with Autonomy instead, as factor-pure items like *Links facts together*, *Wants to form his/her own opinions*, and *Analyses problems* loading on the positive pole, versus *Follows the crowd*, *Copies others*, and *Does what others do* loading on the negative pole, reveal. The large majority of the extra Intellect sentence items that were included in the final item pool appeared to be blends of Factors V and III (Conscientiousness). Also trait adjectives like *intelligent*, *wise*, *quick*, *clever* as opposed to *unintelligent*, *unwise*, *gullible*, *stupid* were found to be associated with the V+III+/V-III- blend in the final structure. Interestingly enough, it can be noted that of the three Germanic languages only the German trait structure (Ostendorf, 1990, Table 50) shows a clear Intellect factor when one applies an AB5C-modelling to the data. The core of the American-English Factor V (see Saucier & Goldberg, 1996) appears to be represented by traits that rather refer to Autonomy (*philosophical*, *inquisitive*, *insightful*), whereas traits referring to Intellect (*intelligent*, *intellectual*, *smart*) have their largest projections on the V+III+ blend, like in the present findings. An exciting thought in this respect is that it also is more easy to conceive of Creativity, Imaginativeness, and Openness to Experience as facets of Autonomy than it is vice versa (Autonomy as a facet of one of the others).

It would be interesting to clarify the relationship between FFPI-Autonomy and, for instance, measures of *self-efficacy* (Bandura, 1977) and *locus of control* (Levenson, 1974; Rotter, 1966). An abundance of publications on the topic indicates that these variables are viewed to be important concepts in developmental, educational, clinical, and health psychology (e.g., Bekker, 1993; Clark, Steer, Beck, & Ross, 1995; Cronbach, 1977; Mills, 1994; Ryff, 1989). Preliminary findings with respect to the Italian translation of the FFPI (Marino, Perugini, & Ercolani, 1996) suggest that the three concepts are moderately to strongly related.

Also in the realm of organizational psychology would one be interested in the relationship between the personality variables autonomy and locus of control, as well as their relationship with job characteristics and job performance. Based on a meta-analytic review of the literature with respect to the relationship between the Big-Five dimensions and job performance, Barrick

and Mount (1991) report Conscientiousness to be about the only personality variable with some predictive validity. In a subsequent study, Barrick and Mount (1993) investigated the role of *job autonomy* as a moderating variable influencing the validity of personality predictors. Indeed, they found degree of autonomy on the job to slightly moderate the validity of at least three of the Big-Five dimensions (Conscientiousness, Extraversion, and Agreeableness). Barrick and Mount conclude that future research should concentrate on the generalizability of their results in order to further clarify the conditions under which personality constructs are likely to be related to performance in management jobs. However, personality variables and situational (job) variables may have been confounded, thereby attenuating the relationships due to restriction of range. Spector and O'Connell (1994), for instance, found a correlation of $-.31$ between job autonomy and locus of control: "internals" are found in jobs with higher autonomy. Furthermore, results of a small study ($N = 50$) of our own with respect to the ideal applicant's Big-Five profile for different types of jobs (Dottinga, Hoekendijk, & Raaijmakers, 1995) suggest that a person's standing on Autonomy may be more important for high-level jobs with managerial tasks, while his/her standing on Conscientiousness may be more important for lower-level jobs with no managerial tasks. Note, however, that this latter study has no empirical basis other than *judged* importance of concrete Big-Five related behaviors with respect to job performance. Nevertheless, these findings seem to suggest that studies with respect to person-environment fit, concentrating on Big-Five profiles of applicants and jobs, may offer an alternative route to enhance the predictive validity of personality variables with respect to job performance. Autonomy may certainly be an interesting variable in this respect.

Conclusion

In conclusion, the FFPI efficiently assesses five broad dimensions (factors) of individual differences in behavior: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Autonomy. The first four factors are well-established and robust, as the literature extensively illustrates. The fifth factor, Autonomy, is less well-established, but appears to be an interesting concept in different fields of psychology. According to a first psychometric evaluation, the FFPI scale and factor scores show high internal consistencies, substantial stabilities, and good construct validity. For applied purposes, AB5C facet scores can readily be obtained from the five factor scores, with reliabilities that are in the same order as those of the factor scores. The special item format, consisting of brief concrete behavioral statements in the third person singular which were

explicitly selected for their comprehensibility, makes the FFPI applicable for self- and other's ratings, and for a broad range of educational levels. Its shortness (administration time 10-15 minutes) makes the FFPI an almost ideal instrument, not only for applied purposes, but also to be included as a standard device in studies investigating the relationship between personality and other variables of interest. Naturally, further validity studies are needed including ones with other-language versions of the instrument, in order to make clear whether the FFPI really merits its place among the available instruments in the domain of personality. Based on the preliminary results so far, however, we have every confidence it does.