2021-02

The long and winding road: A comprehensive analysis of 50years of Eysenck instruments for the assessment of personality

Ruch, W

10.1016/j.paid.2020.110070
Personality and Individual Differences
Elsevier BV

The long and winding road: A comprehensive analysis of 50 years of Eysenck instruments for the assessment of personality

Willibald Ruch, Sonja Heintz, Fabian Gander, Jennifer Hofmann, Tracey Platt, & René T. Proyer

Author note

Willibald Ruch, Sonja Heintz, Fabian Gander, and Jennifer Hofmann are at the Department of Psychology at the University of Zurich, Switzerland. Tracey Platt is at the School of Psychology at the University of Sunderland, United Kingdom. René T. Proyer is at the Department of Psychology at the Martin-Luther-University Halle-Wittenberg, Germany.

Address for correspondence: Willibald Ruch, University of Zurich, Department of Psychology, Personality and Assessment, Binzmuehlestrasse 14, Box 7, CH-8050 Zürich, Switzerland, w.ruch@psychologie.uzh.ch

This manuscript has been published as

Ruch, W., Heintz, S., Gander, F., Hofmann, J., Platt, T. & Proyer, R. T. (2020). The long and winding road: A comprehensive analysis of 50 years of Eysenck instruments for the assessment of personality. *Personality and Individual Differences, Special Issue "PAID 40 years"*. Advance online publication. https://doi.org/10.1016/j.paid.2020.110070

Abstract

Instruments for the assessment of the Eysenckian superfactors of personality, Psychoticism (P), Extraversion (E), and Neuroticism (N), were developed over the course of almost 50 years. Typically the convergence with the precursor was examined when a new scale was published. In the present study the continuity and change of the substance of P, E, and N is tested by administering all instruments to a sample simultaneously, together with measures of the Five-Factor Model. A factor analysis of the 19 markers of the PEN model clearly yielded three factors, with higher loadings for E and N compared to P. The superfactors typically were measured purely after the historically second (or third, for P) instrument. Analysing the item difficulty confirmed that the P items were softened during the revisions but this created a confounding of item difficulty and content: The earlier "tough" items (mostly low Agreeableness) were gradually complemented by "softer" items representing the presumed obverse of P, superego strength (mostly low Conscientiousness). Finally, a part of the observed heterogeneity of P was due to these differences in item difficulty. Overall, the EPQ-R seems to be the most valid single measure of the PEN model.

Keywords: PEN model, Extraversion, Psychoticism, Neuroticism, questionnaires

1. Introduction

Personality is a construct that helps highlighting the overlap of patterns in behaviour, feelings, thoughts, and motivations, which can be measured and used to look into the causes and consequences of these individual differences(a). Many personality models have been put forward in the 20th century, with the one by H. J. Eysenck being particularly influential over a very long time span (e.g., Eysenck, 1947, 1994, 1997; Eysenck & Eysenck, 1985). Eysenck laid the path for a scientific approach to the study of personality involving genetic and psychometric studies, followed by experiments testing hypotheses from the causal models (typically biological theories) put forward. His approach also received cross-cultural support (see Bowden, Saklofske, Van de Vijver, Sudarshan, & Eysenck, 2016). The assessment tools and the substance of the core traits co-developed, involving a process that covered a span of almost half a century.

The aim of the present study is to examine assessments of Psychoticism (P), Extraversion (E), and Neuroticism (N) in terms of change in substance and psychometric properties over time. These include the *Maudsley Medical Questionnaire* (MMQ; Eysenck, 1947), *Maudsley Personality Inventory* (MPI; Eysenck 1959a), *Eysenck Personality Inventory* (EPI; Eysenck, 1970, 1974), *Eysenck Personality Questionnaire* (EPQ; Eysenck & Eysenck, 1975), *EPQ – Revised* (EPQ-R; Eysenck & Eysenck, 1991, 1992; Eysenck, Eysenck, & Barrett, 1985), and the *Eysenck Personality Profiler* (EPP; Eysenck & Wilson, 1991). These measures were supplemented by various precursors of the P scale and sets of adjectives for self-ratings composed of markers for P, E, and N. For a better mapping of the changes in the different scales for measuring P, the Five-Factor-Model (FFM) of personality is used to investigate the relative contributions of Agreeableness (A) and Conscientiousness (C) to P. This also contributes to a controversy started by Goldberg and Rosolack (1994) and continued by Costa and McCrae (1995) and more recently by Heaven, Ciarrochi, Leeson, and Barkus (2013). While the development of the Eysenckian instruments has been described before

(Furnham, Eysenck, & Saklofske, 2008), the present manuscript will extend this review by actually studying and comparing all the available material at once in a sample.

1.1 The PEN System

The PEN system of personality is a factor-analytically based descriptive taxonomy of personality containing the three superfactors Psychoticism, Extraversion, and Neuroticism (Eysenck & Eysenck, 1985). The PEN system assumes a hierarchical arrangement of personality characteristics with Psychoticism (versus Impulse Control), Extraversion (versus Introversion), and Neuroticism (versus Emotional Stability) located at the highest level. They are referred to as *types* (or second-order factors in factor-analytic terms) as opposed to *traits* (or first-order factors) defining them. The type concept of *Psychoticism* is made up of traits like being aggressive, cold, egocentric, impersonal, impulsive, antisocial, unemphatic, creative, and tough-minded. The traits whose intercorrelations give rise to the type concept of *Extraversion* are sociable, lively, active, assertive, sensation-seeking, carefree, dominant, surgent, and venturesome. Finally, *Neuroticism* is made up of traits like anxious, depressed, tense, irrational, shy, moody, emotional, and proneness to guilt feelings and low self-esteem (Eysenck & Eysenck, 1985).

Eysenck started studying basic personality types by using ratings and objective tests applied to individuals of chosen clinical groups (such as neurotics, psychotics) and later designed questionnaires for their measurement. The first instrument, the MMQ (Eysenck, 1947) measured N (with 40 items), the MPI (Eysenck, 1959a) measured E and N with 24 items each, and two forms of the EPI (Eysenck & Eysenck, 1964) measured E and N (with 24 E and 24 N items in each form). The first studies of P used unpublished instruments containing items of all factors. P items needed to be identified that fulfil three criteria; specifically, the items had to a) intercorrelate together to define a common factor; b) discriminate between non-clinical groups and psychotic and criminal groups; and c) not correlate to any noteworthy extent with E and N. In the two studies using the PI (Eysenck, &

Eysenck, 1968) and PEN (Eysenck & Eysenck, 1972), 20 items each fulfilled the criteria. Further work then led to the publication of the EPQ (Eysenck & Eysenck, 1975) measuring P, E and N (with 25 P, 21 E, and 23 N items) and the psychometrically improved EPQ-R (Eysenck et al., 1985) measuring P, E, and N (with 32, 23, and 24 items, respectively). Finally, the EPP (Eysenck, Barrett, Wilson, & Jackson, 1992; Eysenck & Wilson, 1991) was published that contained facets (7 per superfactor with 20 items each).

1.2 The Sequential Development of the Structural Model: Some Inherent Consequences

The sequential development of concepts (compared to a simultaneous one) bears some predictable hurdles to master. A first one is that the substance of a factor most likely needs adjustments once a further factor is added. The initial definition of N in the MMQ was strongly influenced by dysthymia, and once E was added the more introverted N items had to be eliminated to have E and N clearly separated. While with the EPI (but not the MPI) E and N were almost orthogonal, the addition of P to the model posed problems with impulsivity, which, together with sociability, formed E. Studies revealed that impulsivity could be broken down into four positively correlated components with one of them—non-planning impulsiveness—being mostly aligned with P, while other elements like venturesomeness remained with E (Eysenck & Eysenck, 1978; for a recent discussion, see Zuckerman & Glicksohn, 2016). Consequently, the EPQ shifted its focus to assessing mostly sociability, rather than a mixture of both impulsivity and sociability as in the EPI (Rocklin & Revelle, 1981).

A second hurdle was to keep the meaning of the superfactors constant when including facets for each superfactor, which started with the EPP (Eysenck et al., 1992; Eysenck & Wilson, 1991). Despite the fact that since 1985 nine traits were always listed as defining facets, the EPP uses seven facets. It is of course difficult to balance out the secondary loadings of the facets so that the total scores only measure P, E, and N. This appears to be

more problematic with the short version of the EPP (the EPP-S; Eysenck, Wilson, & Jackson, 1999), which only contains three facets for each factor. It should be noted that the first version of this scale employed partly different labels for the subscales and also assumed a partly different assignment of the primaries to the three factors (Eysenck & Wilson, 1976). Furthermore, it had 630 items compared to the 440 items.

Studies correlating the EPQ with the EPP-factors (Costa & McCrae, 1995) or the EPP-S total scales (Knyazev, Belopolsky, Bodunov, & Wilson, 2004) clearly showed that the correlations a) were highest for homologous scales (indicating the best correspondence for the markers of P, E, and N in the two instruments); b) were consistently higher for N and E compared to P; and c) showed some patterns in the off-diagonals (e.g., EPP P correlating positively with E, and EPP N correlating negatively with EPQ E). This is not surprising as, for example, the three scales defining P (Risk-taking, Impulsiveness, and Irresponsibility) seem to capture E as well, while the three scales defining N (Anxiety, Inferiority, and Unhappiness) seem to stem from the introverted side. This can be explored further by looking into the studies examining the factor structure of the standard and short versions, which yielded solutions with three (EPP, EPP-S) and five (EPP only) factors. The three-factor solution for the 21 facets of the EPP typically yielded factors of P, E, and N. Several scales with double loadings (and for aggression even triple loadings) emerged, and one or two scales (i.e., practical) that were outside the PEN model (Costa & McCrae, 1995; Eysenck et al. 1992; Jackson & Francis, 2004; Jackson, Furnham, Forde, & Cotter, 2000). The latter result gave rise to the idea to investigate whether the EPP scales might be better represented by the FFM (Costa & McCrae, 1995); a view later refuted by Jackson et al. (2000). However, most importantly, when P, E, and N were obtained most purely through a target rotation (Costa & McCrae, 1995; Eysenck et al. 1992), it became apparent that the secondary loadings did not even out; for instance, facets of P also tended to load on E, and more facets of N were on the introverted side of E facets.

A third hurdle was to preserve the substance of the two clinically orientated factors N and P while softening the item contents to make the scales applicable to the general population. While item contents can be maintained in a softer version for N (e.g., “have you ever thought of suicide” may be weakened to “are you occasionally really fed up” and medical symptoms may be reduced in intensity and frequency), this is more difficult for psychotic symptoms (e.g., paranoid ideas) or not possible at all for others, like hearing voices. Thus, it is important to see how the softening of P was undertaken and what the final outcome was. An empirical comparison among different Eysenckian questionnaires (MMQ, MPI, EPI, and EPQ) has only been conducted for the N scale (Ferrando, 2001). The unidimensionality of the 47 N items in these four questionnaires received support, and the MMQ items were found to be more difficult (i.e., had lower means) than the items from the other three questionnaires. These items often referred to the occurrence of physical symptoms, which were “softened” over time to make the items more suitable for non-clinical rather than clinical populations.

1.3 The P-scale: Two new Perspectives on some of the Existing Criticisms and Earlier Controversies

We argue that this process of softening items is related to two criticisms or controversies related to P; namely, the acclaimed heterogeneity of P and the lack of correspondence in an alternative system of personality description, the FFM, in which it covers two factors (A and C). The P-scale of the EPQ (Eysenck & Eysenck, 1975) had a lower internal consistency (i.e., Cronbach's alpha) than the E and N scales, despite the higher number of items. The EPQ-R raised the number of items from 25 to 32 to obtain a satisfactory alpha. Several explanations were put forward; for instance, the P facets might have a lower reliability (Eysenck & Eysenck, 1991) or the P-scale might be factorially heterogeneous (Roger & Morris, 1991). Additionally, the P-scales of different questionnaires were found to

show a lower convergence than the E and N scales, respectively (e.g., EPQ-A and EPQ-RS, Alexopoulos, & Kalaitzidis, 2004; EPP and EPQ, Knyazev et al., 2004).

Eysenck's (1992a) conceptual account of the P dimension shows the diversity of traits and syndromes; he lists (from the low P pole to the middle) traits like altruistic, socialized, empathic, conventional, and conformist, and locates (from above average to extreme) phenomena like being criminal, impulsive, hostile, aggressive, psychopathic, schizoid, unipolar depressive, schizoaffective, schizophrenic, or suffering from an affective disorder. Clearly, these lists are prone to show some heterogeneity when packed into one scale.

Eysenck (1992b) listed a narrower segment of primaries of P that should also explain why P relates to both low A and low C, despite the latter two being uncorrelated. In his controversy with Costa and McCrae, Eysenck declared A and C to be (narrow) primaries of P assuming the two outermost positions in the segment of primaries covering (low) A, coldness, Machiavellianism, hostility, aggression, (low) empathy, and (low) C. The alternative interpretation that P is an arbitrary combination of low C and low A was first raised in a factor analysis of the 25 P items by Goldberg and Rosolack (1994). They found the two sets of positively and negatively keyed items to be scattered in arcs of about 125 degrees in a space defined by two orthogonal components (see Figure 1.1 in Goldberg & Rosolack, 1994). This implies negative correlations and hence a large heterogeneity among the items. Additionally, they found the two factors to be correlated with A- and C- of the FFM and concluded that P is heterogeneous and a blend of low C and A.

We would like to add two new perspectives on this matter. The first perspective is based on the fact that while the Eysencks saw the typical high P-scorer as "cold, impersonal, hostile, lacking in sympathy, unfriendly, untrustful, odd, unemotional, unhelpful, antisocial, lacking in human feelings, inhumane, generally bloody-minded, lacking in insight, strange, with paranoid ideas that people were against him" (Eysenck & Eysenck, 1976, p. 47), they

also followed Royce (1973) who saw the third factor (beyond E and N) in personality to be superego, as championed by Cattell. They conceded that superego “is clearly the obverse of the psychoticism factor we are here hypothesizing; all the traits characterizing the 'high superego' person are characteristically absent in the high P scorer, as we shall see” (Eysenck & Eysenck, 1976, pp. 43–44). High superego, of course, makes the low pole of P closer to impulse control (but also C in the FFM).

We might therefore expect to find that earlier approaches to P yield items that have a low endorsement frequency (and low variance in case of binary answers), tougher content, and more overlap with low A–in line with the description of the typical high P-scorer mentioned above. Later approaches might include softer items, with higher endorsement rates (and hence a larger variance , thereby affecting the scale variance more than the older items) and item contents that also reflect C or impulse control. Thus, these two sources of heterogeneity (differences in item endorsement and contents) might be confounded. Findings on the relationships between earlier and later questionnaires and the FFM are in line with this interpretation of a shift from A- to C- in the P items. For instance, using 53 P items from the EPQ and new items, McCrae and Costa (1985) found correlations of -.20 to -.45 with A and -.29 to -.31 with C. Later studies using the EPP-P scale (e.g., Costa, & McCrae, 1995; Muris, Schmidt, Merckelbach, & Rassin, 2000), by contrast, found the largest correlations with C- and smaller or even non-significant correlations with A-.

The second perspective is that keeping the tough items (that mark the content of P well) and supplementing them with “softened” P items will lead to a) very skewed distributions for the tough items (as the EPQ uses a yes/no answer format) and to b) a large range in the item means. In the English norm data, the item with the lowest mean was Item 11 (M = 0.03; “Would it upset you a lot to see a child or an animal suffer?”) and the one with the

highest mean was Item 74 ($M = 0.30$[1]; "When you catch a train do you often arrive at the last minute?"). Thus, the correlation (PHI-max) between these two items can maximally be .27. Consequently, if one allows for more factors–as, for example, Goldberg and Rosolack (1994) did–the severely lowered upper limit for the correlation makes it likely that these two items will load on different factors. Thus, differences in item difficulty likely contribute to the heterogeneity of the P-scale, and it opens the possibility that item difficulty and content might be confounded.

1.3 Aims of the Present Study

A multitude of self-rating forms (adjectives from different articles describing P covering the entire time span, the nine primary traits depicted in the model, as well as adjectives from the German trait taxonomy studies; Ostendorf, 1994) and questionnaires (from the MMQ to the EPP-S) will be administered to examine the following questions: a) How did the three concepts P, E, and N develop over the years in terms of basic statistics (*M, SD*, Cronbach's alpha) as well as their factor loadings?; b) How does the correlation of different forms of the P scale and rating markers of P change in relation to C and A?; and c) Is the alleged heterogeneity of the P-scale in part an artefact due to the wide range in item means?

Regarding a), we expect that the means (and *SD*) for P increase from the early to the latest versions of the scale to the midpoint of the scale. At the same time, the factor loadings in a three-factor model should increase and display a clearer factor structure, reflecting the increased reliability and purification of the scales, respectively. Regarding b), we expect that the earlier versions of the scale will be mainly negatively related to A while the later versions will have an equal contribution of C- to P. This pattern should also emerge for adjectives describing the P scales sampled over the course of the development of the P scales based on

[1] The information about the British norm data was kindly provided by Paul Barrett.

expert ratings by 10 FFM experts (prototypicality for C, A, N, E, and Openness to experience). This analysis will be performed on items that received high prototypicality evaluations by a PEN expert. Regarding c), we expect that the difference in the item means explains most but not all of the observed heterogeneity of P in the EPQ, pertaining to a confound between the means and contents of the scale (extending the analyses of Goldberg & Rosolack, 1994).

2 Material and Methods

2.1 Participants and Procedure

Sample 1 comprised 629 adults (63.3% women) from the general population aged 17 to 91 years (M = 41.3, SD = 13.9). All participants completed the two latest versions of Eysenck's questionnaires (i.e., the EPQ-R and EPP), a questionnaire for the assessment of the FFM of personality (NEO-PI-R; Ostendorf & Angleitner, 2004), as well as self-ratings based on the 21 measured traits in the EPP scales (EPP-SR) and self-ratings on the 27 (9×3) adjectives depicted in the PEN-model (PEN-SR). Participants in Sample 1 were recruited through radio and newspaper reports, mouth-to-mouth propaganda, and a website dedicated to the project. The participants completed the questionnaires in the lab or received them via mail. Upon request, participants received personal feedback on their scores or a general feedback on selected findings of the study.

Sample 2 was a subsample of Sample 1 (338 adults; 60.9% women) who additionally completed the older versions of Eysenck's questionnaires (i.e., MMQ, MPI, and both forms of the EPI), and various precursors of the P scales (i.e., PI 68, PEN 72, and EPQ 75 as well as items tested for the EPQ-R that were excluded from the final scale). The non-redundant items of these scales were compiled into one longer instrument.

Sample 3 was composed of an expert sample of one PEN-expert (Sybil B. G. Eysenck), and 10 FFM-experts (Alois Angleitner, Peter Borkenau, Filip deFruit, Lewis R. Goldberg, A. A. Jolijn Hendriks, Willem K. B. Hofstee, John A. Johnson, Robert R. McCrae,

Ivan Mervielde, and Gerard Saucier). These experts rated all P items of the *Multiple Prototypicality Ratings Form*, which contained the P items of all P scales except the one in the EPP, regarding their prototypicality for the PEN-model (Sybil B. G. Eysenck) or the FFM (the 10 FFM-experts).

We additionally conducted analyses and simulations based on the British norm data of the EPQ kindly provided by Paul Barrett.

2.2 Instruments

The *Maudsley Medical Questionnaire* (MMQ; Eysenck, 1947; used in the German version by Eysenck, 1953) assesses N with 38 items. Additionally, it contains a lie scale (18 items). All items use a dichotomous response format (0 = “no”, 1 = “yes”).

The *Maudsley Personality Inventory* (MPI; Eysenck, 1959a; used in the German version by Eysenck, 1959b) assesses E and N with 48 items (24 items per scale). All items are rated on a three point-scale (0 = “no”, 1 = “can’t decide”, 2 = “yes”).

The *Eysenck Personality Inventory* (EPI; Eysenck, 1970; used in the German version by Eggert, 1974) assesses E and N with 48 items (24 items each). Additionally, it contains a lie scale (9 items). All items use a dichotomous response format (0 = “no”, 1 = “yes”). There are two parallel forms of the instrument (EPI-A and EPI-B).

The *Eysenck Personality Questionnaire revised* (EPQ-R; Eysenck & Eysenck, 1991, 1992; Eysenck et al., 1985; used in the German version by Ruch, 1999) contains 102 items for the assessment of P (32 items), E (23 items), N (25 items), while 22 items form a lie scale. All items use a dichotomous response format (0 = “no”, 1 = “yes”).

Psychoticism items from different precursors of the P scale that did not make it into the EPQ-R were combined in one instrument. These were P items from the PI (Eysenck & Eysenck, 1968), PEN (Eysenck & Eysenck, 1972), EPQ (Eysenck & Eysenck, 1975), and an unpublished instrument that finally led to the EPQ-R (Eysenck, Eysenck, & Barrett, 1985). All items use a dichotomous response format (0 = “no”, 1 = “yes”). These items, together

with the standard P items, were used to compute the total scores representing the P scales of 1968, 1972, and 1975, but also scores for the new items were derived.

The *Eysenck Personality Profiler* (EPP; Eysenck & Wilson, 1991; used in the German translation by Bulheller & Häcker, 1998) contains 420 items for the assessment of P, E, and N (140 items each). Additionally, 20 items form a lie scale. All items are rated on a three-point scale (0 = "no", 1 = "can't decide", 2 = "yes"). It should be noted that the three existing German adaptations (EPP-D) did not use all 21 facets, namely the long (EPP-D BH) and short version (EPP-DS BH) by Bulheller und Häcker (1998) and one (EPP-D M) with yet a different items scoring key (Moosbrugger, Fischbach, & Schermelleh-Engel, 1998). These forms were derived from the pool of 420 items.

The *NEO Personality Inventory-revised* (NEO-PI-R; Costa & McCrae, 1992; German version by Ostendorf & Angleitner, 2004) assesses the FFM traits (i.e., N, E, Openness to experience, A, and C) with 240 items (48 items per dimension). All items use a 5-point Likert-style scale (1 = "strongly disagree" to 5 = "strongly agree").

The *Multiple Prototypicality Ratings Form-PEN* contained 67 non-redundant P items used in precursors of the P Scale until the EPQ-R. The instruction read "Please judge the degree of prototypicality of each of the following 67 questions. For each question ask yourself: how prototypical is a 'yes'-answer to an item for the dimensions of Psychoticism (P), Extraversion (E), and Neuroticism (N). For your answer use a seven-point scale ranging from -3 (highly prototypical for the negative pole) to +3 (highly prototypical for the positive pole), with '0' meaning that this item is orthogonal/unrelated to that dimension." For the sake of the present study, these ratings will only be used to identify the items highly prototypical of P but not yielding high scores for E or N at the same time. This eliminated five items from the pool.

The *Multiple Prototypicality Ratings Form-FFM* contained the remaining 62 P items, and the instruction given to the FFM experts was the same except the beginning: "Please

judge the degree of prototypicality of each of the following 62 questions. For each question ask yourself: how prototypical would a 'yes'-answer to these items be for the dimensions of Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C)". The answers were then averaged.

The *PEN-SB* contained 207 trait adjectives presumably measuring P and coming from different sources, namely descriptions of the high P scorer in research papers on P and manuals (e.g., Eysenck, 1992a; Eysenck & Eysenck, 1972, 1975, 1976, 1992), the model with nine defining subtraits (Eysenck & Eysenck, 1985), and the German personality taxonomy project. As part of the latter, Ostendorf (1994) collected prototypicality ratings for P, E, and N from H. J. Eysenck (who judged 430 adjectives) as well as from 10 students (who judged 823 adjectives). The students were very familiar with the PEN system and they were also given materials to study. The judgments were done on a 7-point rating scale (-3 = prototypical for negative pole of the trait, 0 = not prototypical, +3 = prototypical for the positive pole of the trait). Adjectives were selected for the study if they were more prototypical for P than for E and N combined; that is, for slightly prototypical P-adjectives (+1/-1) the scores for E and N needed to be "0" to be included in the study. These ratings were then added to a total score, but also separate scores were used in the analyses; for example, one score for each of the six levels of prototypicality (-3, -2, -1, +1, +2, +3) in the students' rating as well as Hans Eysenck's rating. Likewise, separate scores were computed for how P was described in the above-mentioned six publications.

2.3 Data Analysis

To test research question a), we first computed Cronbach's alpha, item difficulties (means), corrected item-total correlations, and average inter-item correlations of the P-scales from the PI, PEN, EPQ, EPQ-R, EPP, and EPP-S. These items were administered in Sample 2 and covered versions of the P-scale from 1968 to 1999. Additionally, the item difficulties (means) of adjective descriptions of P (from 1972 to 1992) were investigated as well (also

completed by Sample 2). Second, to determine the factor structure and loadings of P, E, and N, we subjected the scales from 1947 to 1991 as well as the means of the adjective descriptions of P, E, and N (see Eysenck & Eysenck, 1985) to a principal components analysis with varimax-rotation.

To test research question b), we first correlated the different scales and adjectives of P (from 1968 to 1999) with the A and C scales of the NEO-PI-R, based on self-reports in Samples 1 and 2. Next, the ratings for the 62 non-redundant P items were averaged across the 10 FFM experts. Only ratings of A and C were used, and each item was coded as belonging to "A" or "C" (depending on which yielded the higher mean). The items were grouped according to their first appearance in a P scale, and each item was used only once even if it reappeared in later versions of the P scale. The use of non-redundant item sets helped to examine whether the relative importance of A and C shifted throughout the different versions of the P items.

Finally, the sources of the heterogeneity of the P-scale (research question c) was investigated by computing principal component analyses of 25 P items in the normative sample of the English EPQ (Phi), a simulated data set (Phi-max) based on a perfect Guttman scale, and the corrected correlation matrix (Phi-corr = Phi/Phi-max). These three factor solutions were compared using rank-order correlations and by plotting the factor loadings.

3 Results

3.1 How did the Three Concepts P, E, and N Develop over the Years?

3.1.1 Changes in basic psychometric properties

We analysed the psychometric properties (i.e., Cronbach's alpha, item difficulties, corrected item-total correlations, and average inter-item correlations) of different versions of the P-scale that were completed by Sample 2 (Figure 1).

--Insert Figure 1 about here--

Figure 1 shows that–as expected–item difficulty decreased over time from on average rather difficult items in the precursors of the P-scale (PI, PEN), over slightly less difficult items in the EPQ-R, to considerably easier items in the EPP and the EPP-S. At the same time, Cronbach's alpha increased. This increase was not only due to the inclusion of more items, since also the average inter-item correlations and the average corrected item-total correlations showed a similar increase over time.

Interestingly, different patterns were found for the E and N scales (not shown in detail): While the E scale remained considerably constant over time with regard to the relevant psychometric properties (e.g., item means between .45 and .54), the N scale also showed some decrease in item difficulty in one step from the MMQ (item mean of .29) to the MPI (and subsequent scales; .48 for the N scale of the EPQ-R). However, the items were more difficult again in the latest additions; that is, the EPP (item mean of .27) and the EPP-S (.32). Thus, there was a softening of N, but it was over a short time and not implemented in the EPP and EPP-S.

An even clearer picture was obtained when analysing the means of the self-ratings of adjectives used to describe the P concept in several publications and manuals over time: The item difficulty strongly decreased over time (see Figure 2), thus paralleling the picture obtained by the questionnaires. Interestingly, the typically observed gender difference for P (men with higher scores than women) was not found in the adjectives used in the last two instruments.

--Insert Figure 2 about here--

3.1.2 Factor structure of all used Markers for P, E, and N

For the main question of the continuity or change in the prime concepts between the MMQ (Eysenck, 1947) and the Eysenck Personality Profiler (Eysenck & Wilson, 1991), we computed a principal component analysis of all scales completed by Sample 2. To avoid overlap among P items, only the newly added items in later versions were considered. The

first three components explained 71.7% of the variance and were rotated to the varimax-criterion. Table 1 gives the factor loadings on all three components labelled in accordance to the theoretical expectations.

--Insert Table 1 about here--

Table 1 shows that the three factors clearly may be identified as P, E and N, explaining 12.7%, 26.2%, and 32.8% of the variance, respectively. With one exception every marker had its highest loading on the expected factor and most of these were high and pure. The core of the P items that were used in 1968 (from the PI) was actually loading on N. The sets of items added in 1975 and 1985 (i.e., new in EPQ and EPQ-R, respectively) were good markers of P. The P scale of the EPP had a high loading, but also loaded on E. The total score of adjectives for P (but also the one of E and N) was a clear marker. Thus, the model and the item contents converged. E was mostly clearly measured, but the MPI E scale, as well as the EPI-B E, were more on the emotional stable side. Extraversion in the EPI-A E scale tended to load positively on P, which was likely due to the impulsivity items (which were removed from the EPQ). N was clearly marked by the scales, with the exception that there were introverted elements in the MMQ and EPP N scales. Thus, overall there was continuity across the scales, with some unexpected secondary loadings for certain scales.

3.2 Did the Relation between P and C and A Change Over Time?

3.2.1 Self-ratings

The correlations between self-ratings in the various indicators of P (scales and adjectives) and the A and C scales of the NEO-PI-R were computed (Samples 1 and 2). The results regarding the adjectives and scales were very clear-cut. The correlations were used in Figure 3 as coordinates to place the marker in this two-dimensional space with A and C serving as axes.

--Insert Figure 3 about here--

The results were quite different for different sets of markers. For H. J. Eysenck the prototypical markers were exclusively A- (denoted by black squares in Figure 3). This was also the case when additionally considering the level of prototypicality (not shown in detail). Next, the adjectives based on the description of P from articles and manuals (top left panel) were primarily A- with a slight correlation with C-. There was no development; in fact, also the Eysenck and Eysenck (1985) model is A-, and only the description in the manual of the EPQ-R (Eysenck & Eysenck, 1991, 1992) yielded a slight involvement of C- (-.20). The entire span of traits as presented in Eysenck (1992a) included conformist and conventional and had a noticeable correlation with low C. Thus, in terms of the description of the concept P did not get softened. However, there was a slight mismatch between Eysenck's view of the concept and the adjectives used to describe the concept. This was picked up by the students that mostly saw prototypical P (+1, +2, +3) as mostly A- and slightly lower on C (white squares).

This pattern was different for measured P in questionnaires. A very strong change can be seen for the EPP (bottom right panel in Figure 3). While Eysenck and Wilson's (1976) total score was measuring only A-, the total score in the EPP contributed equally to C and A and the short scale EPP-S was even more C- than A-; the same held true for all German adaptations of the scale (EPP-D; the regular and short version by Bulheller & Häcker and the version by Moosbrugger et al., 1998), which lacked substantial correlations with A- (these versions can be derived from the translated version of the EPP used; results are not shown in detail). The inspection of the P items at different times (in the EPQ-R and its precursors) should be most interesting as this signified the development of P. In 1968, P was marginally negatively related to both C and A. In 1972, the correlation to A was twice as high as the one to C. The EPQ P scale correlated more with C (-.40) than with A (-.30), and for the EPQ-R scale the two correlations were equally high. The same pattern can be observed when only analysing those items that were newly added to the scales at the respective time points

(bottom left panel in Figure 3): While the items added in 1972 (PEN) only had slight loadings on C-, subsequently added items had considerably higher loadings on C-, while simultaneously their loadings on A- decreased.

3.2.2 Expert ratings of P-items: Prototypicality for A and C

Similar trends were found for the expert ratings conducted for the set of new P items introduced for different versions of the P scale. To highlight the change, two packages were distinguished: The early P scales of 1968 and 1972 had 15 (7 positively and 8 negatively keyed) items identified by the experts to represent A, and only 2 items relating to C (2 yes, 0 no-items). Later P scales (new items for EPQ and EPQ-R) that led to the EPQ-R had 9 items (4/5) relating to A and 13 to C (10/3). This clearly demonstrates that there was a shift in the substance of P once superego was noticed in the mid-1970s to be the obverse of P. Early items defined P purely as A-, and in later scales P was a mixture of C- and A-. Items of the EPP or EPP-S were not used, but it is evident that these would be more prototypical for C- than for A-.

3.3 Is the Alleged Heterogeneity of the P-scale in Part an Artefact due to the Wide Range in Item Means and Heterogeneity?

Before answering research question c), we first conducted a simulation to estimate the effects of impaired maximal correlations due to different item difficulties in binary data and then computed a principal component analysis. Based on the 25 item means taken from the English norm sample of the EPQ, a P-scale was simulated by forming a perfect Guttman scale. The results were equivalent to factors derived from an intercorrelation of PHI-max coefficients. Figure 4 shows how the loadings of the first two factors depended on item difficulty, and Figure 5 gives the factor plot of the first two factors extracted from the intercorrelation of PHI-max coefficients (equivalent to a perfect Guttman scale).

--Insert Figure 4 about here--

The first principal component was affected by differences in item difficulty in an inverted U-shape (Figure 4); items with higher and with very low means yield maximal loadings of about .60. (It should be noted that if all items had the same item mean, their intercorrelation would be perfect, and the loading on the first and only factor should equal 1 for each item.) The second unrotated factor of an item factor analysis was predominantly a linear function of the item difficulty. It should be noted that for the EPQ-R, the maximal correlation between the toughest and softest item will even be lower and the factor loading would even be more distorted.

--Insert Figure 5 about here--

The positively and negatively keyed P-items (simulated data of Figure 5) covered wide segments in the two-dimensional space, suggesting almost the same amount of heterogeneity as found for the empirical data (Goldberg & Rosolack, 1994). While the arc was lower than 90 degrees and thus not reaching the arcs of about 125 degrees (as reported by Goldberg & Rosolack, 1994), one can easily see that a substantial part of the reported heterogeneity was an artefact. Thus, the P scale might be less heterogeneous than it seemed to be. However, it is also unlikely that the observed extent of heterogeneity was caused solely by the differences in item difficulty, and it should be remembered that item difficulty and item content might be confounded as described above.

Next, principal components analyses were performed for the intercorrelations among the 25 P-items in three data sets, namely a) the normative sample of the English EPQ (Phi), b) the simulated data set (Phi-max), and c) the corrected correlation matrix (Phi-corr = Phi/Phi-max). The rank-order correlation between the loading on PHI-max based PC1 and the first PC based on Phi was .39 (p = .057); this coefficient reduced to -.10 (n.s.) when the factor was derived from the corrected intercorrelation matrix. The second artifactual factor (PC2 based on Phi-max) was predictive as well; loadings on Factors 3 and 5 (r[25] = .46; p < .05 and -.53;

$p < .01$, respectively) reflected the quasi linear trend in item difficulty, but no such relationship was found for factors based on corrected coefficients (rs[25] = .23 and .06, n.s.).

Thus, this opens up the possibility that the previously extracted second unrotated factor was independent of the variations in item difficulty and meaningful contentwise. To explore this further the factor plot is inspected. While the negatively keyed items were located in the two-dimensional space (see Figure 6) according to their difficulty (r[11] = .64; $p < .05$), this relationship could not be found for the positively keyed items (r[14] = .11; n.s).

--Insert Figure 6 about here--

The two-dimensional configuration was highly similar to the one reported by Goldberg and Rosolack (1994) in several ways. First, the negatively scored items were more homogeneous than the positively keyed items, which again spread over more than 90 degrees. Second, the order of the items was highly comparable; that is, the same items were located at the beginning and end of the segments in the two studies. For the positively keyed cluster these items were negatively correlated with A and portrayed core elements of P (such as P65 “Are there several people who keep trying to avoid you?”, P76 “Do your friendships break up easily without it being your fault?”, and P87 “Do people tell you a lot of lies?”) on one end. Risk-taking items that correlated with C- in the Goldberg and Rosolack (1994) study (P74 “When you catch a train do you often arrive at the last minute?”, P67 “Do you think people spend too much time on savings and insurances?”) were on the other end. The converse was found for the negatively keyed items, where the C+ items (P57 “Do you like to arrive at appointments in plenty of time?”, P9 “Do you lock up your house carefully at night?”) formed the one end and the A+ empathy items (P90 “Would you feel very sorry for an animal caught in a trap?”, P11 “Would it upset you a lot to see a child or an animal suffer?”) formed the other. Thus, we found the expected confounding of item difficulty and item content, with the

low mean items stemming from the earlier P scales being more related to A-, and the more frequently endorsed ones added to the later P scales having the relation to C-.

4 Discussion

This article examined the instruments for the assessment of personality traits developed by H. J. Eysenck. Spanning over 50 years, the instruments covered the different scales on Eysenck's PEN model, but they also included various precursors of the P-scale and ad-hoc measures based on trait adjectives considered to be markers of P. Thus, the current article shows the development of this momentous and often used model. As all questionnaires were completed by the same individuals simultaneously, the development of the scales over time could be observed. Moreover, we examined the relationship of the Eysenckian instruments with the FFM of personality.

Overall, the results show that over time, the scales became clearer (e.g., the N scale lost E loading contents). As shown in the joint factor analysis, the first P scale from 1968 was loading higher on N than on P. With the inclusion of more constructs, the scales were refined and a more or less "pure" version of the P dimension was obtained. This was also supported in the expert ratings by Sybil Eysenck, which reconfirmed the suitability of most of the presented items for assessing P. While all three superfactors became independent in the EPQ-R, in the EPP, the P and N scale were again slightly confounded with E.

At the same time, the P items subsequently became easier: While in the very first version (PI; Eysenck & Eysenck, 1968) they had very low average agreement rates and seemed predominantly addressed at clinical populations, later editions of the P scale added less difficult items in order to also allow to monitor variations in P in the general population. This development can not only be traced in the items, but also in the adjectives used to describe the concept of P. The last versions of scales and ratings had P item means close to the scale midpoint (i.e., between .40 and .45). This addresses research question a).

However, and important for research question b), the P scales did not only change over time with regard to difficulty but there were also considerable changes in content that go beyond the above-mentioned initial overlap with N: While early studies assessed P more as the opposite of A, later studies assessed P more as the opposite of C. As Eysenck and Eysenck considered P to be the obverse of the superego factor, this leads to the question of what other evidence is there to the relation between P and superego? In a study of the 16 PF and the EPQ, McKenzie (1988) found separate factors for superego and P, but the P scale loaded -.40 on the superego factor, which lends some support to Eysenck's contention that P is the obverse of superego. In a further analysis, a superego factor was found that was loaded substantially by some but not all items of the P scale. The 14 P-items not loading significantly on superego were those involving the cruelty or sadism element of the concept of P. This goes along very well with the finding that the definition of P was mostly related to A-, but the P scale of the EPQ related to both A- and C-, and the new P items introduced in the EPQ were primarily C-. Overall, based on these findings one might question Eysenck and Eysenck's (1976) depiction of P as the opposite of the superego factor as reported by Cattell.

The change in the nature of P s also reflected in the EPP. For Eysenck and Wilson (1976), toughmindedness (i.e., P) was composed of aggressiveness–peacefulness, assertiveness –submissiveness, achievement-orientation–unambitiousness, manipulation–empathy, sensation seeking–unadventurousness, dogmatism– flexibility, and masculinity–femininity. The facets of P in the later EPP were risk-taking, impulsivity, irresponsibility, manipulativeness, sensation seeking, tough-mindedness, and practicality. Using a confirmatory factor analysis on the German EPP, Moosbrugger and Fischbach (2002) only found three facets fitting to the concept, namely impulsivity, irresponsibility, and sensation seeking. This means that the selection of facets suitable to measure P were actually the ones of superego/impulse control/C, but not the ones anymore that had typically been used in the

descriptions of P; that is, facets that relate to A- (manipulation, empathy, risk-taking, and tough-mindedness).

Overall, the present study showed that the line of development of P can be traced using different methods, including the instruments (e.g., PI, PEN, EPQ, EPQ-R, EPP) and the adjectives used to describe the P concept in publications and manuals. Interestingly, this is in contrast to Eysenck's own characterization of prototypical adjectives of the P concept, which also in 1994 rather followed the original description of the P-scale as only A- than the later operationalization (and description) as A- and C-. In sum, three different clusters need to be kept apart: First, Eysenck's stipulation that P is only A-; second, a slight development in the description of the high P scorer that stopped at 1975; and third, the items of instruments that gradually increased the involvement of C- until it even dominated over A-. Interestingly, the students that studied the provided material also rated primarily A-, with the exception of slightly low expression of P, which was strongly C-. This is consistent with the finding that the gender differences were more prevalent in the early versions (see Figure 2), as there typically are gender differences in A but not in C (see e.g., Weisberg, DeYoung, & Hirsh, 2011).

With respect to the scale characteristics and research question c), the reported low alpha of P was identified to be a function of item heterogeneity regarding item difficulty (which will be ameliorated if using Likert-type scales, such as a 6-point answer format). Part of the heterogeneity of the P scale stemmed from the differences in item means (this span increased for the revised P-scale of the EPQ-R) and was a by-product of applying data analysis methods to data not fulfilling the requirements. However, these effects of item difficulty were confounded with changes in item content: The easier items more strongly related to C-, while the more difficult items showed stronger relationships to A-. Thus, the present analysis limits, but does not rule out, the interpretation that P is factorially

heterogeneous, and we tentatively conclude that both of these aspects contribute to the reported heterogeneity of P. Future studies should tackle this problem by using modified items that disentangle item content and item difficulty.

4.1 Recommendations and implications

Several recommendations can be made based on the present study. The results help to better integrate the findings derived with different P scales. They provide new perspectives on potential causes for the observed heterogeneity of P and on the debate of the relations between P and A and C. The study also demonstrates that the gap between the concept of P and measures of P got wider. We suggest to use the EPQ-R for the testing of the PEN model, since it provided the clearest factor structure with the least amount of secondary loadings and since it came closest to Eysenck's conceptualization of the P-scale as primarily low A, while later instruments increased the scale's relationships to low C. We know which of the scales measured E and N most purely (i.e., those that had no second loadings), and for P we know the relative contribution of the different facets in a measure. For "tough" individuals and applications in a forensic context, it is important that elements of being cold, hostile, and aggressive are assessed, and hence the earlier versions of the P scale are recommended. If weaker expressions of P are to be measured, then the EPP scales are best for representing "soft" P. Due to the high item difficulty in several measures, it is recommended to compute factor analyses based on corrected coefficients (e.g., PHI-corr = PHI/PHI-max) or tetrachoric correlations for obtaining unbiased findings. Furthermore, when evaluating the internal consistency of the P scale, it is advisable to use split half-reliability (with items matched for difficulty; see Ruch, 1999) rather than Cronbach's alpha, for which the P scale does not meet the requirements (Feldt, Woodruff & Salih, 1987).

4.2 Limitations

Of course, the present findings have to be interpreted while taking some strengths and weaknesses into account. While the present study used a large sample of mainly no-student

adults and relied on different methods (e.g., using items and adjectives rated by experts), only self-report measures were employed. While we would expect the same patterns to emerge in peer-ratings, we did not collect data for settling this question. Further, as the participants completed consecutive measures, the questionnaire was redundant in many places. Although this might have decreased the participants' motivation in the study, we observed a generally very low dropout rate.

4.3 Conclusions

The present study is the first to simultaneously look at all scales measuring the Eysenckian concepts of P, E, and N over half a century. The different scales and adjectives of P, E, and N could be well separated in principal components analyses, supporting the general viability of the different measures to represent the PEN model. Still, it appeared to be most difficult to transfer P to a construct to be measured in a general adult population, which lead to a shift in content over time. The "softening" of the P items during the revisions confounded item difficulty and content, shifting the content from low A to a mixture of low A and low C, and contributed to the heterogeneity of the P scale. Depending on the intended population under study or application, either earlier versions of the questionnaires (PI, EPQ) or later versions (EPP, EPPS-S) can be recommended. Overall, the EPQ-R seems to be the most valid single measure of the PEN model.

Acknowledgments

wThis research was supported by a grant of the Deutsche Forschungsgemeinschaft – DFG (HE 3143/1-1 “Untersuchungen zum PEN-Modell der Persönlichkeit“). Thanks to Olga Altfreder, Matthias Bergande, Kristina Dürscheidt, Cornelia Kirchhof, Gwendolin Linnenbrink, Nina Rambech for collecting part of the data and app. 1000 participants filling in up to 8 hours worth of questionnaires over the time span of two years. The authors are grateful to Dr. Paul Barrett for providing access to the norm data on the EPQ for analysis of means. Furthermore, to Dr. S. B. G. Eysenck for providing the expert rating on P and Drs. Alois Angleitner, Peter Borkenau, Filip deFruit, Lewis R. Goldberg, A. A. Jolijn Hendriks, Willem K.B. Hofstee, John A. Johnson, Robert R. McCrae, Ivan Mervielde, and Gerard Saucier for providing the FFM ratings.

5 References

Alexopoulos, D.S., & Kalaitzidis, I. (2004). Psychometric properties of Eysenck Personality Questionnaire-Revised (EPQ-R) short scale in Greece. *Personality and Individual Differences, 37*(6), 1205–1220. https://doi.org/10.1016/j.paid.2003.12.005

Bowden, S.C., Saklofske, D.H., Van de Vijver, F.J.R., Sudarshan, N.J., & Eysenck, S.B.G. (2016). Cross-cultural measurement invariance of the Eysenck Personality Questionnaire across 33 countries. *Personality and Individual Differences*, *103*, 53–60. https://doi.org/10.1016/j.paid.2016.04.028

Bulheller, S., & Häcker, H. (1998). Deutsche Bearbeitung. (German revision). In H.J. Eysenck, C.D. Wilson, & C.J. Jackson, *Eysenck Personality Profiler EPP-D: Manual (EPP, German version, manual)*. Frankfurt am Main, Germany: Swets Test Services.

Costa, P.T. & McCrae, R.R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI). Professional Manual.* Odessa, FL: PAR.

Costa, P.T., & McCrae, R.R. (1995). Primary traits of Eysenck's P-E-N system: Three- and five-factor solutions. *Journal of Personality and Social Psychology*, *69*, 308–317. https://doi.org/10.1037/0022-3514.69.2.308

Furnham, A., Eysenck, S.B.G., & Saklofske, D.H. (2008). The Eysenck personality measures: Fifty years of scale development. In G.J Boyle, G. Matthews, & D.H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment* (vol. 2, pp. 199-218). London, UK: SAGE Publications.

Eggert, D. (1974). Eysenck-Persönlichkeits-Inventar: E-P-I: Handanweisung für die Durchführung und Auswertung [Eysenck Personality Inventory (EPI): Directions for its implementation and evaluation]. Göttingen, Germany: Hogrefe.

Eysenck, H.J. (1947). *Dimensions of personality*. London, UK: Routledge & Kegan Paul.

Eysenck, H.J. (1953). Maudsley Persönlichkeitsfragebogen [Maudsley Personality Inventory]. Göttingen, Germany: Hogrefe.

Eysenck, H.J. (1959a). *Manual of the Maudsley Personality Inventory*. London, UK: University of London Press.

Eysenck, H.J. (1959b). Das "Maudsley Personality Inventory" (MPI). Göttingen, Germany: Hogrefe.

Eysenck, H.J. (1970). *EPI Eysenck Personality Inventory*. London, UK: University of London Press.

Eysenck, H.J. (1974). Eysenck-Persönlichkeits-Inventar E-P-I [Eysenck Personality Inventory EPI]. Göttingen, Germany: Hogrefe.

Eysenck H.J. (1992a). The definition and measurement of psychoticism. *Personality and Individual Differences, 13*, 757–785. https://doi.org/10.1016/0191-8869(92)90050-Y

Eysenck, H.J. (1992b). Four ways five factors are not basic. *Personality and Individual Differences, 13*, 667–673. https://doi.org/10.1016/0191-8869(92)90237-J

Eysenck, H.J. (1994). Normality–abnormality and the three-factor model of personality. In S. Strack & M. Lorr (Eds.), *Differentiating normal and abnormal personality* (pp. 3–25). New York, NY: Springer.

Eysenck, H.J. (1997). Personality and experimental psychology: The unification of psychology and the possibility of a paradigm. *Journal of Personality and Social Psychology, 73*, 1224–1237. https://doi.org/10.1037/0022-3514.73.6.1224

Eysenck, H.J., Barrett, P., Wilson, G., & Jackson, C. (1992). Primary trait measurement of the 21 components of the P-E-N system. *European Journal of Psychological Assessment, 8*(2), 109–117.

Eysenck, H.J., & Eysenck, S.B.G. (1964). *Manual of the Eysenck Personality Inventory*. London, UK: University of London Press.

Eysenck, S.B.G., & Eysenck, H.J. (1968). The measurement of psychoticism: A study of factor stability and reliability. *British Journal of Social & Clinical Psychology, 7*(4), 286–294. https://doi.org/10.1111/j.2044-8260.1968.tb00571.x

Eysenck, S.B.G. & Eysenck, H.J. (1972). The questionnaire measurement of psychoticism. *Psychological Medicine, 2*, 50–55. https://doi.org/10.1017/S0033291700045608

Eysenck, H.J. & Eysenck, S.B.G. (1975). *Manual of the Eysenck Personality Questionnaire.* London, UK: Hodder & Stoughton.

Eysenck, H.J. & Eysenck, S.B.G. (1976). *Psychoticism as a dimension of personality*. London, UK: Hodder and Stoughton.

Eysenck, S.B.G., & Eysenck, H.J. (1978). Impulsiveness and venturesomeness: Their position in a dimensional system of personality description. *Psychological Reports, 43*(3), 1247–1255. https://doi.org/10.2466/pr0.1978.43.3f.1247

Eysenck, H.J. & Eysenck, M.W. (1985) *Personality and individual differences: A natural science approach.* New York, NY: Plenum Press.

Eysenck, H.J. & Eysenck, S.B.G. (1991). M*anual of the Eysenck Personality Scales (EPS Adults)*. London, UK: Hodder & Stoughton.

Eysenck, H.J., & Eysenck, S.B.G. (1992). *Manual of the EPQ-R and the impulsiveness, venturesomeness and empathy scales.* London, UK: Hodder & Stoughton.

Eysenck, H.J., Eysenck, S.B.G. & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences, 6*, 21–29. https://doi.org/10.1016/0191-8869(85)90026-1

Eysenck, H.J. & Wilson, G.D. (1976). *Know your own Personality*. New York, NY: Penguin Books.

Eysenck, H.J., & Wilson, G. (1991). *The Eysenck Personality Profiler*. London, UK: Corporate Assessment Network.

Eysenck, H.J., Wilson, G.D. & Jackson, C. (1999). *The Eysenck Personality Profiler (short) (2nd ed.)*. Guildford, UK: Psi-Press.

Feldt, L.S., Woodruff, D.J., & Salih, F.A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93–103. https://doi.org/10.1177/014662168701100107

Ferrando, P.J. (2001). The measurement of neuroticism using MMQ, MPI, EPI and EPQ items: A psychometric analysis based on item response theory. *Personality and Individual Differences*, *30*(4), 641–656. https://doi.org/10.1016/S0191-8869(00)00062-3

Goldberg, L.R. & Rosolack, T.K. (1994). The Big-Five factor structure as an integrative framework: An empirical comparison with Eysenck's P-E-N model. In C.F. Halverson, G.A. Kohnstamm, & R.P. Martin (Eds.), *The developing structure of temperament and personality from infancy to adulthood* (pp. 7–36). Hillsdale, NJ: Erlbaum.

Heaven, P.C., Ciarrochi, J., Leeson, P., & Barkus, E. (2013). Agreeableness, conscientiousness, and psychoticism: Distinctive influences of three personality dimensions in adolescence. *British Journal of Psychology*, *104*, 481–494. https://doi.org/10.1111/bjop.12002

Jackson, C.J., & Francis, L.J. (2004). Primary scale structure of the Eysenck Personality Profiler (EPP). *Current Psychology*, *22*(4), 295–305. https://doi.org/10.1007/s12144-004-1035-9

Jackson, C.J., Furnham, A., Forde, L., & Cotter, T. (2000). The structure of the Eysenck personality profiler. *British Journal of Psychology*, *91*(2), 223–239. https://doi.org/10.1348/000712600161808

Knyazev, G.G., Belopolsky, V.I., Bodunov, M.V., & Wilson, G.D. (2004). The factor structure of the Eysenck Personality Profiler in Russia. *Personality and Individual Differences*, *37*(8), 1681–1692. https://doi.org/10.1016/j.paid.2004.03.003

McCrae, R.R., & Costa Jr, P.T. (1985). Comparison of EPI and psychoticism scales with measures of the five-factor model of personality. *Personality and Individual Differences, 6*, 587-597. https://doi.org/10.1016/0191-8869(85)90008-X

McKenzie, J. (1988). Three superfactors in the 16PF and their relation to Eysenck's P, E and N. *Personality and Individual Differences, 9*(5), 843–850. https://doi.org/10.1016/0191-8869(88)90002-5

Moosbrugger, H., & Fischbach, A. (2002). Evaluating the dimensionality of the Eysenck Personality Profiler—German version (EPP-D): A contribution to the Super Three vs. Big Five discussion. *Personality and Individual Differences, 33*(2), 191–211. https://doi.org/10.1016/S0191-8869(02)00095-8

Moosbrugger, H., Fischbach, A., & Schermelleh-Engel, K. (1998). Zur Konstruktvalidität des EPP-D. [On the construct validity of EPP-D]. In H.J. Eysenck, C.D. Wilson, & C.J. Jackson (Eds.), *Eysenck Personality Profiler EPP-D*. Manual. Frankfurt, Germany: Swets Test Services.

Muris, P., Schmidt, H., Merckelbach, H., & Rassin, E. (2000). Reliability, factor structure and validity of the Dutch Eysenck Personality Profiler. *Personality and Individual Differences, 29*(5), 857-868. https://doi.org/10.1016/S0191-8869(99)00237-8

Ostendorf, F. (1994). Zur Taxonomie deutscher Dispositionsbegriffe. [On the taxonomy of German disposition terms]. In W. Hager, & M. Hasselhorn (Eds.), *Handbuch deutschsprachiger Wortnormen* (pp. 382–441). Göttingen, Germany: Hogrefe.

Ostendorf, F., & Angleitner, A. (2004). *NEO-PI-R - NEO-Persönlichkeitsinventar nach Costa und McCrae [NEO-PI-R – NEO-Personality Inventory by Costa and McCrae]*. Göttingen, Germany: Hogrefe.

Rocklin, T., & Revelle, W. (1981). The measurement of extroversion: A comparison of the Eysenck Personality Inventory and the Eysenck Personality Questionnaire. *British*

Journal of Social Psychology, *20*(4), 279–284. https://doi.org/10.1111/j.2044-8309.1981.tb00498.x

Roger, D. & Morris, J. (1991). The internal structure of the EPQ scales. *Personality and Individual Differences*, *12*, 759-764. https://doi.org/10.1016/0191-8869(91)90232-Z

Royce, J.R. (1973). *Multivariate analysis and psychological theory*. London, UK: Academic Press.

Ruch, W. (1999). Die revidierte Fassung des Eysenck Personality Questionnaire und die Konstruktion des deutschen EPQ-R bzw. EPQ-RK [The Eysenck Personality Questionnaire-Revised and the Construction of German Standard and Short Versions (EPQ-R and EPQ-RK)]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *20*(1), 1–24. https://doi.org/10.1024//0170-1789.20.1.1

Weisberg, Y.J., DeYoung, C.G., & Hirsh, J.B. (2011). Gender differences in personality across the ten aspects of the Big Five. *Frontiers in Psychology, 2*:178. https://doi.org/10.3389/fpsyg.2011.00178

Zuckerman, M., & Glicksohn, J. (2016). Hans Eysenck's personality model and the constructs of sensation seeking and impulsivity. *Personality and Individual Differences, 103*, 48–52. https://doi.org/10.1016/j.paid.2016.04.003

Table 1

Factor Loadings on the Three Varimax-Rotated Factors

Scales	Psychoticism	Extraversion	Neuroticism
PI P ('68)	***.09***	-.01	*.67*
PEN P (new'72)	**.65**	*-.22*	*.22*
EPQ P (new'75)	**.61**	.18	.05
EPQ-R P (new'85)	**.68**	.13	-.02
EPP P	**.75**	*.25*	-.12
Adjectives P	**.68**	-.04	.06
MPI E	.04	**.89**	*-.22*
EPI-A E	*.23*	**.89**	-.04
EPI-B E	.13	**.85**	*-.30*
EPQ-R E	.03	**.93**	-.14
EPP E	.10	**.81**	.02
Adjectives E	-.06	**.87**	-.22
MMQ N	-.09	*-.32*	**.84**
MPI N	.07	-.06	**.91**
EPI-A N	-.06	-.10	**.93**
EPI-B N	.04	-.09	**.93**
EPQ-R N	.02	-.08	**.91**
EPP N	.03	*-.27*	**.88**
Adjectives N	.12	-.17	**.80**
Eigenvalues	2.41	4.98	6.23

Note. *N* = 305. Expected loadings in boldface; anomalies italicized.

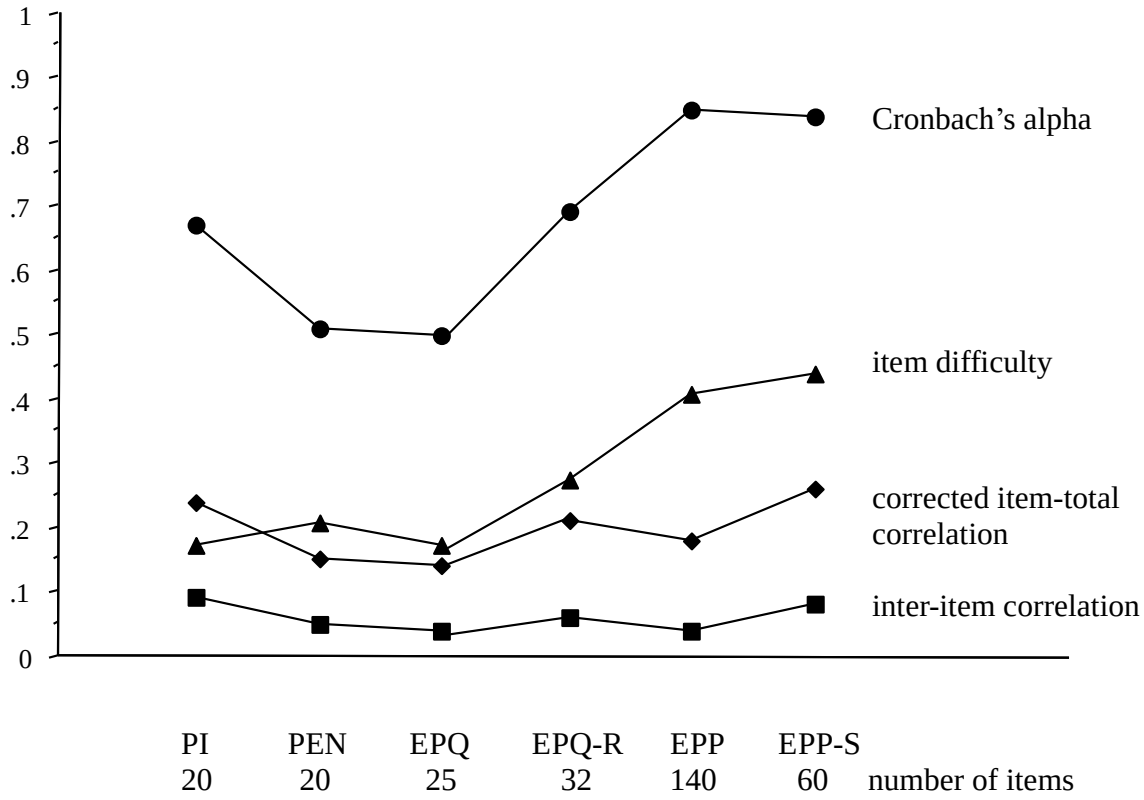

Figure 1. Psychometric properties of different versions of the P scale.

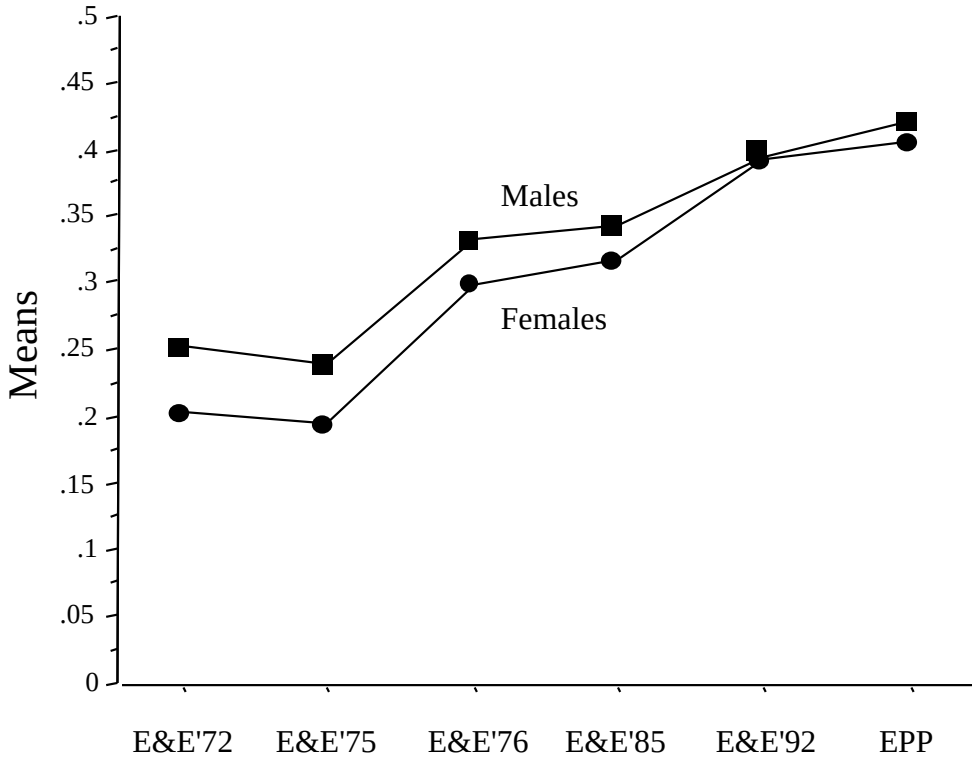

Figure 2. Item difficulties based on self-ratings of adjectives used to describe the P concept in publications and manuals. Abbreviations denote the publication year of the adjectives' source (e.g., E&E'72 = Eysenck & Eysenck, 1972).

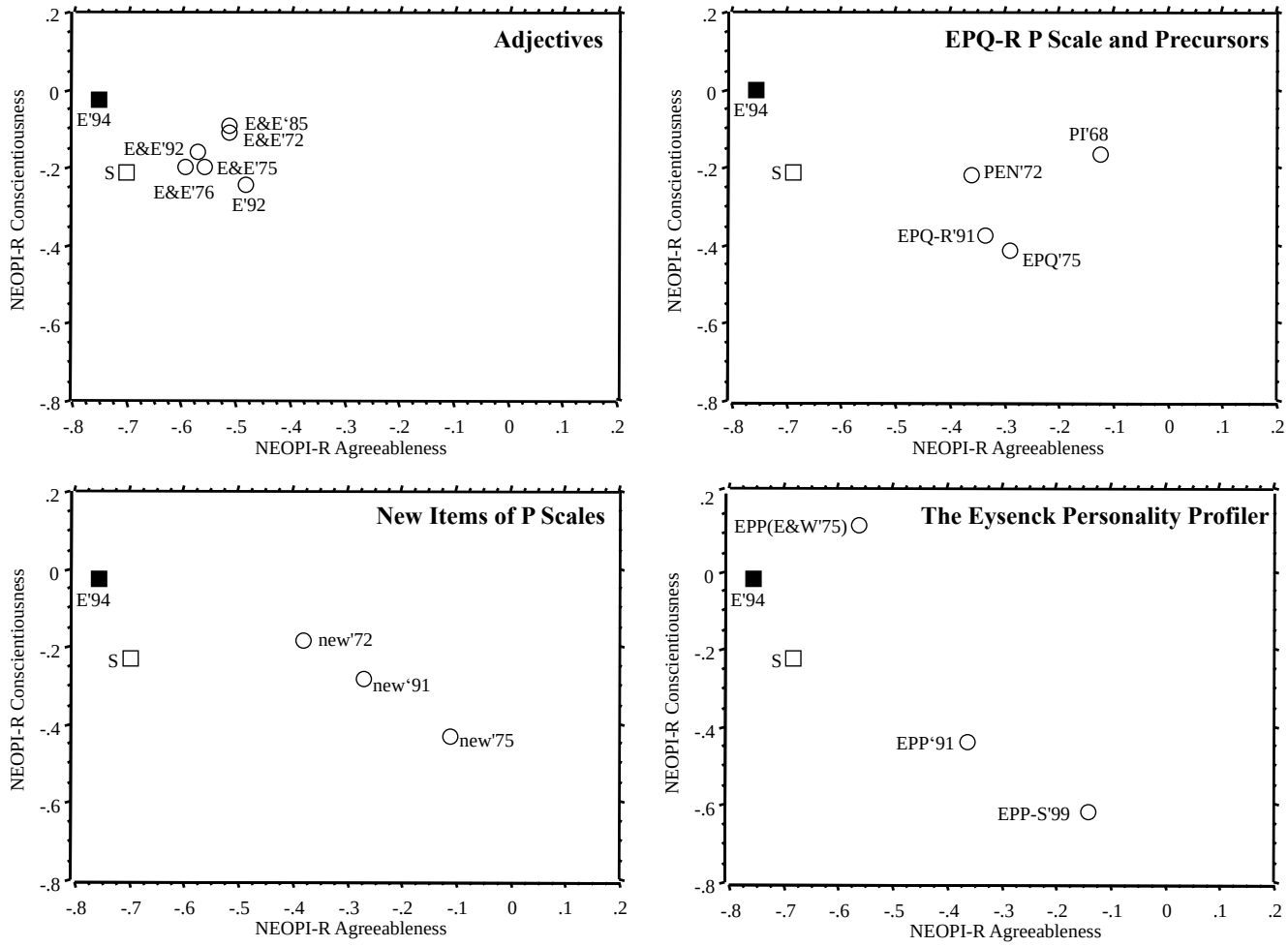

Figure 3. Relationships between self-ratings in several sets of P-adjectives (Top Left Panel), the EPQ-R P Scale and its Precursors (Top Right Panel), newly added items of P scales (Bottom Left Panel), and the Eysenck Personality Profiler P-Scale (Bottom Right Panel) with Agreeableness and Conscientiousness. Each panel also shows the relationships of prototypical adjectives for P as rated by H. J. Eysenck with Agreeableness and Conscientiousness (black squares; E'94) and prototypical adjectives as rated by students (white squares; S). Abbreviations denote the publication year of the adjectives' source (e.g., E&E'72: Eysenck & Eysenck, 1972).

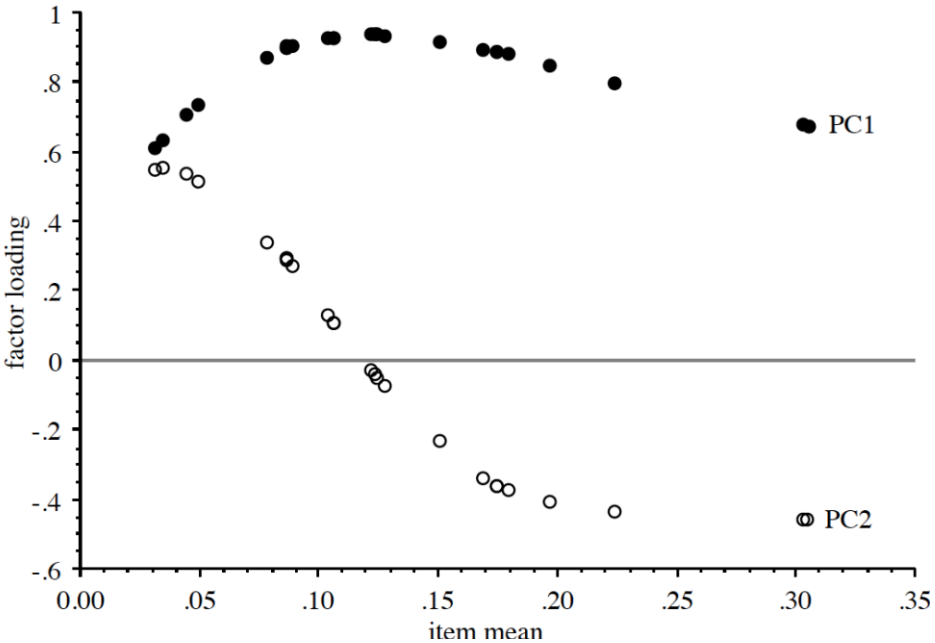

Figure 4. The loadings on the first two unrotated principal components as a function of variation in item difficulty.

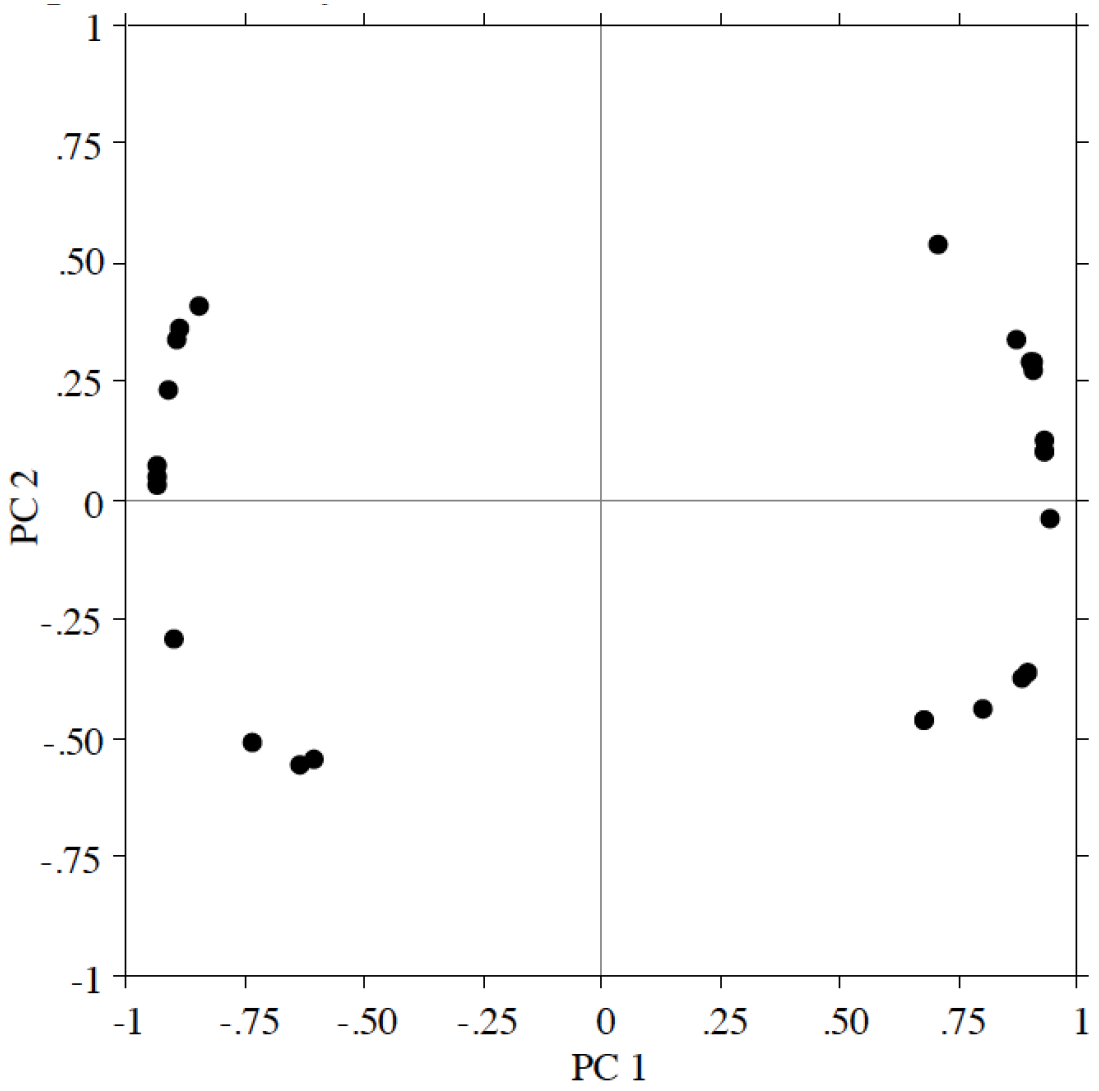

Figure 5. Loadings of the 25 P items on the first two unrotated factors (simulated data).

Note the "heterogeneity" of the item set, which theoretically is strictly unidimensional.

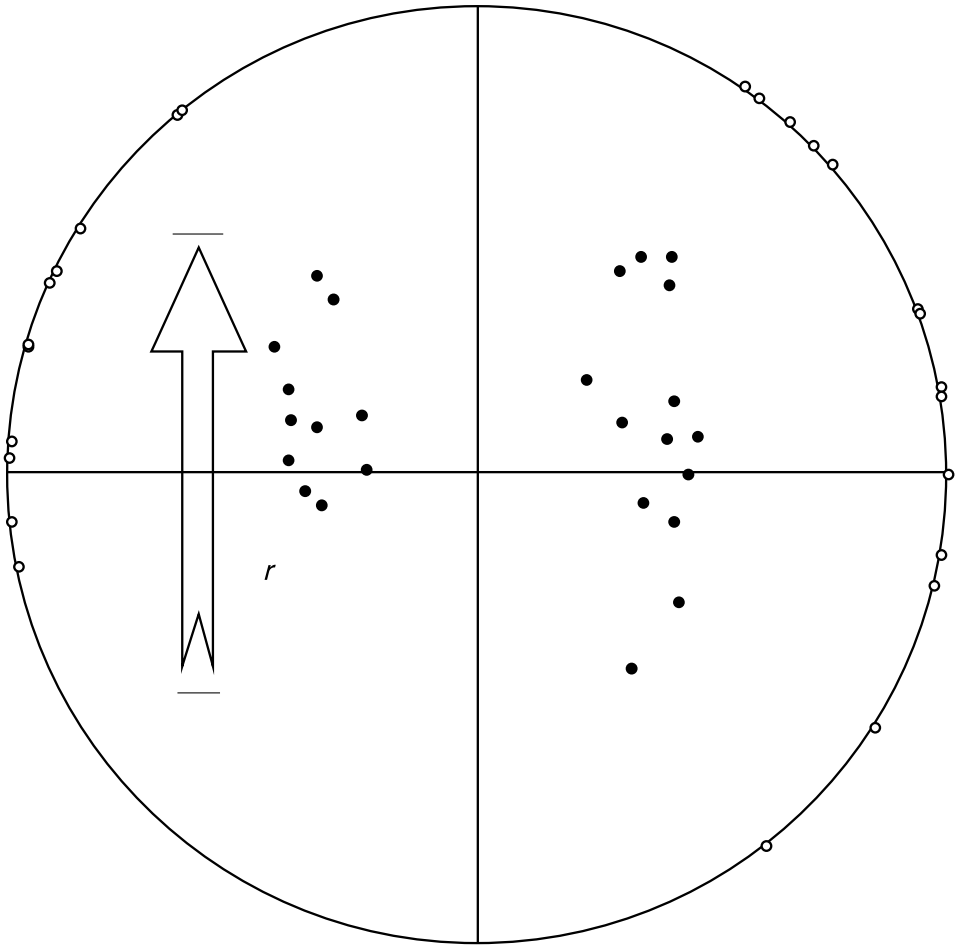

Figure 6. Factor space defined by the first and second (6.4 % of variance) principal component derived from the intercorrelation of the 25 P items in the English norm sample.