European Journal of Personality
Eur. J. Pers. **17**: S39–S66 (2003)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/per.484

The Construct Validity of Three Entry Level Personality Inventories Used in the UK: Cautionary Findings from a Multiple-Inventory Investigation†

NEIL ANDERSON*[1] and DENIZ S. ONES[2]*

[1]*Department of Work and Organizational Psychology, University of Amsterdam, Amsterdam, The Netherlands*
[2]*Department of Psychology, University of Minnesota, Minneapolis, USA*

Abstract

This paper reports psychometric analyses into the convergent and divergent validity of three popular entry-level measures of occupational personality in the UK and Continental Europe. A sample of 504 individuals completed all three measures: the British version of the Hogan Personality Inventory (HPI), the Occupational Personality Questionnaire (OPQ Version FS5.2), and the Business Personality Indicator (BPI). In addition, independent ratings of the conceptual loading of primary source scales onto the Five Factor Model (FFM) were obtained (n = 66). Data were used in a three-stage analytical procedure directed at examining psychometric and construct validity. Results are reported for descriptive statistics (means, standard deviations, Cohen's d), internal consistency reliability (Cronbach's alphas), and exploratory factor analyses. Findings into the construct validity of first-order scales (i.e. primary source scales) and second-order scales (i.e. FFM loadings) are presented in detail, including multitrait–multimethod (MTMM) analyses of convergent and divergent validity. For some scales, the observed variability in our sample suggested significant range restriction/enhancement. It was found that scale reliabilities were generally lower than those typically reported by the test publishers, and that published factor structures for these measures could not be replicated by the authors for this sample of individuals. Further independent construct validity research into occupational personality inventories is encouraged based upon our proposed model of single-, dual-, and multiple-inventory construct validation studies. Practically, our findings suggest that when IWO psychologists or personnel professionals aim to select/screen job applicants for a particular personality trait those who are selected may vary depending on (i) which personality inventory is used, (ii) the actual variability in the

*Correspondence to: Neil Anderson, Department of Work and Organizational Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands. E-mail: nanderson@fmg.uva.nl or
Deniz S. Ones, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, MN 55455-0344, USA. E-mail: Deniz.S.Ones-1@tc.umn.edu
†Both authors contributed equally to this paper and therefore order of authorship is arbitrary. Either author may be approached for reprints of, or further correspondence over, this paper.

Received 30 April 2002
Accepted 20 July 2002

applicant sample tested, and (iii) reliability of the scales under consideration. As such, this study sounds a note of caution. Future research is called for to replicate these findings.

INTRODUCTION

Recent surveys of employer usage of personality inventories in the UK, other European countries, and North America suggest their increasing popularity as a method of candidate assessment. Despite earlier more pessimistic qualitative reviews of the criterion-related validity of personality tests, more recent meta-analytical investigations have provided robust and generally unambiguous findings in support of their use in employee selection (for comprehensive recent reviews see Hough & Ones, 2001; Salgado, Viswesvaran, & Ones, 2001). The factors combined have stimulated renewed interest in different aspects of the contribution that personality inventories can make to assessment and selection procedures (Anderson & Herriot, 1997; Ones & Viswesvaran, 1996; Ones & Anderson, 2002). It is apparent that in several countries, but particularly in the UK, the use of inventories of work-relevant personality has become increasingly widespread not just for managerial and senior level appointments but for more junior, entry-level jobs as well (Bartram, Lindley, Marshall, & Foster, 1995; Shackleton & Newell, 1994; Herriot & Anderson, 1997; Hodgkinson & Payne, 1998; Ryan, McFarland, Baron, & Page, 1999).

Given this increasing popularity amongst employer organizations, it is important that commercially available personality inventories are subjected to independent scrutiny and validation. Moreover, this growing market for proprietary personality inventories also raises important questions over criterion-related validity (see e.g. Robertson & Kinder, 1993; see also Salgado, 1996), the generalizability of validities obtained in other cultural contexts (e.g. Salgado & Anderson, 2002), the psychometric properties and factor analytic structure of commercially available measures (e.g. Hough, 2001; Hogan & Roberts, 2001), potential for adverse impact and unfair discrimination on grounds of either gender or ethnic minority group membership (see e.g. Hough, Oswald, & Ployhart, 2001; Ones & Anderson, 2002), and appropriate professional standards for the construction and use of such measures (see e.g. Hough, 2001). Not surprisingly, therefore, this growth in the use of personality inventories has led to a concomitant increase in research interest in this method of candidate assessment by Industrial–Work–Organizational (IWO) psychologists in both North America and Europe.

Despite this growth in research by IWO psychologists, the fields of personality theory on one side, and occupational personality test usage on the other, have remained inexplicably isolated from one another. This is regrettable as developments in personality theory and research have been particularly rapid and fruitful over recent years, yet many of these advances have occurred within an almost completely separate literature and research tradition to the more pragmatic concerns of IWO psychologists (Hogan & Roberts, 2001). Hough (2001) succinctly states the problem:

> I/O psychologists have been lax in attending to the taxonomic structure of their variables, perhaps due partly to excessive empiricism, and perhaps partly the result of pragmatic attention to an immediate, applied goal (p. 21).

Only very recently have there been the first signs of a real convergence of agreement amongst IWO psychologists of the importance of construct validity research into

personality inventories (Hough, 2001; Hough & Ones, 2001) in spite of the previous lack of studies into these important questions. This area has been neglected by IWO psychologists for too long, we would argue, and more studies are called for that shed light upon construct validity issues in proprietary tests of occupational personality.

The Five Factor Model and occupational personality

Over the past decade a considerable body of evidence has been accumulated in personality research that provides support for the Five Factor Model (FFM), or so-called 'Big Five', latent structure of human personality (Digman, 1990). A descriptive model or taxonomy is vital if personality psychology is to advance as a science. What is most impressive about this body of evidence is that the FFM has emerged from a large number of studies conducted across different countries in Europe, North America, and the rest of the world (see e.g. McCrae & Costa, 1997), using a wide range of original source inventories (e.g. Conley, 1985; Costa & McCrae, 1988; Ferguson, Payne, & Anderson, 1994; McCrae & Costa, 1987; 1997; Lorr & Youniss, 1973), across different languages and cultures (e.g. Bond, Nakazato, & Shiraishi, 1975; Borkenau & Ostendorf, 1990; McCrae & Costa, 1997; Yang & Bond, 1990), and with different samples having completed the original scales (for a concise but comprehensive review see McCrae & Costa, 1997). A large body of literature has accumulated that provides compelling evidence for the robustness of the five-factor model of personality: across different theoretical frameworks (Goldberg, 1993) and using ratings obtained from different sources (e.g., Digman & Inouye, 1986; Digman & Takemoto-Chock, 1981; Fiske, 1949; McCrae & Costa, 1987; Norman, 1963; Norman & Goldberg, 1966; Watson, 1989). Based on the results of several studies using comprehensive sets of trait terms, using multiple replications, and different factor analytic techniques, the evidence for the Big Five factors is undeniably voluminous.

Additionally, and of direct relevance to pragmatic concerns within IWO psychology, the FFM has been used as the most parsimonious representation of the personality predictor space in major meta-analytic studies from Europe (see e.g. Salgado, 1997, 2002), the USA (Barrick & Mount, 1991; Barrick, Mount, & Judge, 2001), South Africa (Rothman, Meiring, & Barrick, 2002), and Korea (Yoo & Min, 2002). The dimensions of the FFM can be summarized as follows (see also Digman, 1990; McCrae & John, 1992; Goldberg, 1990):

(I) Emotional Stability (opposite pole of Neuroticism);
(II) Extraversion (including Surgency and Sociability);
(III) Openness to Experience (or Intellect, Imagination, or Culture);
(IV) Agreeableness (including Nurturance);
(V) Conscientiousness (including Will to Achieve and Dependability).

The Revised NEO Personality Inventory (NEO-PI-R) by Costa and McCrae remains the most cross-culturally researched FFM-based measure of normal adult personality to date. For example, McCrae and Costa (1997) present evidence from a seven-nation study where a large sample of subjects completed the NEO-PI-R (Costa & McCrae, 1992). German, Portuguese, Hebrew, Chinese, Korean, and Japanese samples were obtained and compared against the original US sample. Across these samples (total $n = 7134$) the FFM emerged as a universal, underlying structure with item loadings being similar across all countries.

McCrae (2001) has recently summarized and discussed many of the cross-cultural findings for the NEO-PI-R. Data came from 26 cultures measured from 1992 to 2000 (mean $N = 921$ respondents per country). Regarding its cross-cultural measurement equivalence, it

is worth stating that the NEO has been carefully translated for each particular culture with back-translation to confirm accuracy. For all versions, reliability is generally adequate as well as construct validity and factorial validity. The consistency of the NEO-PI-R's factor structure has been tested several times in multiple cultural contexts, all showing a consistent structure indicating universality, though at times the openness factor wavers somewhat and the specific subdimensions (facets) are not always precisely replicated in terms of loadings. All in all, despite marked cross-national differences in culture, language, and so forth, the FFM appears to be a parsimonious underlying structure of personality across the countries and cultures. Internationally, there is a remarkable convergence upon the FFM as a general model of personality trait structure by personality researchers.

Criterion-related validity and construct validity

Whilst considerable research has been conducted into the criterion-related validity of personality inventories in IWO psychology both at the level of individual validity studies and at the level of meta-analytic investigations, a paucity of studies critically restricts our understanding of important aspects of construct validity in IWO settings. Although there may be good reasons to explain why this disparity in the coverage of research has occurred, relative lack of construct validity evidence from applied settings is a definite shortcoming in our understanding of personality testing in workplace settings (see also Hough, 2001; Ones & Anderson, 2002). Moreover, questions of the predictive, or criterion-related, validity of particular measures tend to be driven by a highly pragmatic agenda to establish the veracity of particular instruments. This research orientation can fail to address more basic, fundamental, and theoretically driven concerns of the underlying structure of personality in IWO psychology settings. A case in point is integrity testing. In the early 1990s, although the criterion-related validity of integrity tests were meta-analytically demonstrated for two important criteria of both job performance and broadly assessed counterproductive behaviours (Ones, Viswesvaran, & Schmidt, 1993), few understood the personality underpinnings of integrity tests. Ones' doctoral dissertation (1993) was able to show, using both meta-analytic and primary data, that integrity tests assessed a combination of conscientiousness, agreeableness and emotional stability, in that order (i.e. Digman's (1997) higher order factor alpha (Ones & Viswesvaran, 2001)).

Construct validity issues regarding the most viable structure of personality may often be at the forefront of concerns amongst personality researchers but unfortunately rather in the background amongst some practitioners in IWO psychology. Why is construct validation important? At least three main reasons, all of which are independently compelling, can be cited. First, as noted earlier in this paper, most data supporting the use of personality scales in personnel selection have used the FFM as an organizing framework (Barrick et al., 2001; Salgado, 1997). This presumes that scales from different inventories can be organized into the FFM (Hough & Ones, 2001). Second, there is a need to develop a much more sophisticated appreciation of *why* and *how* different occupational personality inventories predict subsequent on-the-job behaviour (Hough, 2001). Amid the ever-increasing range of personality inventories on the market, IWO psychology appears particularly remiss in its parochial concentration upon criterion-related validity in the face of developments in personality psychology. Independent construct validation of personality inventories is the way forwards toward clearer understanding of the why and how of predictive validity. Third, and perhaps most importantly, independent research into personality inventory construct validity holds out the promise of beneficially

influencing standards of practice and professionalism in personality inventory test development. We should not ignore the possible negative effects that the enormous growth in market potential for personality inventories has created a lucrative market for measures both good and bad. This market is also one in which the customer, usually personnel practitioners in organizations, is relatively naive over standards of test construction and validation. Independent research into construct validity in applied settings is therefore vitally important. In the present study, we were fortunate to obtain the co-operation of three highly respected publishers, to whom we are grateful—ASE, SHL, and PCL.

Single-, dual-, and multiple-inventory studies in construct validation

Such has been the paucity of studies into the construct validity of commercially marketed measures of occupational personality that a detailed literature search by the present authors revealed only a handful of published studies. Nevertheless, a few isolated studies into the construct validity and factor structure of occupational measures of personality have been conducted over more recent years (see e.g. Payne, Ferguson, & Anderson, 1994). More studies have been carried out in applied social psychology (see e.g. Furnham, 1994, 1996), but here we focus specifically upon studies into occupational personality inventories. We propose the following typology as a framework to classify approaches to construct validation in this area.

(i) *Single-inventory studies* (factor analytic or psychometric properties studies into a single, target personality measure).
(ii) *Dual-inventory studies* (studies focusing on relations between two inventories).
(iii) *Multiple-inventory studies* (studies using three or more personality inventories as a basis for detailed construct validation).

Single-inventory studies typically present psychometric data and factor analysis results into the properties of a single target measure under scrutiny (see e.g. Ferguson et al., 1994; Thompson & Borrello, 1986). Ferguson et al. (1994), for instance, conducted a detailed exploratory and confirmatory factor analysis of the then newly published Occupational Personality Questionnaire (OPQ, FMXS-student version). Confirmatory factor analyses computed via LISREL procedures indicated that a modified, 'Big Five' model was the most parsimonious fit and that this solution displayed stronger goodness of fit statistics than the publishers' four-factor higher order model. Whilst single-inventory studies permit a detailed psychometric analysis of a particular measure, their design is inherently limited. Such studies can only address the personality space covered by the inventory itself, using primary scales from the inventory, and are thus dependent upon how well the measure covers the domain of interest. Single-inventory studies by themselves therefore are not sufficient in establishing a nomological net for measures.

Dual-inventory studies compare and contrast dimension inter-relations between two inventories. This design has been particularly common where a new test is being developed and its 'construct validity' checked against an existing, validated measure (see e.g. Hough, 2001). Again, the robustness of the research design is highly dependent upon the measures under investigation, how these were originally developed, and the dimensions measured by each inventory (Cellar, Miller, Doverspike, & Klawsky, 1996; Furnham, 1996). Even though dimension names may be similar or even identical, the construct being measured may be quite different (see our later discussion), and so the precise meaning of correlations between dimensions is actually less clear-cut than many

dual-inventory studies appear to suggest. A potential weakness in dual-inventory designs is the inability to fully examine construct validity (convergent and divergent validity) with any degree of depth or width.

Future research into construct validity, we argue, should adopt a *multiple-inventory* design based upon three or more measures of personality, where the inventories are chosen on clear criteria of appropriateness of item or scale contents, population for which the measures were intended, and underlying models of personality upon which the measures were originally developed (see for instance, Furnham, 1994, 1996). The present study therefore examines three population-specific measures of entry-level personality, all with claimed links to the FFM, but each comprising different dimensions at the first- and second-order factor levels.

In this study, we rely on a modified version of the traditional Multi-Trait, Multi-Method (MTMM) framework (Campbell & Fiske, 1959). A critical question that is yet to be fully addressed is the measurement equivalence of the different personality inventories used in personnel screening and selection. Before substantive construct level relationships can be addressed, measurement issues need to be satisfactorily answered. That is, in answering whether a given personality trait predicts job performance, or should be incorporated into theoretical models, we need to ascertain that different occupational personality instruments are equivalent in their measurement of the particular trait. The primary aims of the present study as an independent examination into three entry-level measures of personality used in the UK are as follows:

(i) to provide descriptive data on widely used personality scales using an appropriate sample in order to facilitate comparisons with the test publishers' original norm groups;
(ii) to evaluate the scale internal consistency reliabilities for primary scales comprising these inventories and to compare the findings of this study against those reported by the test publishers in the test manuals;
(iii) to examine the underlying factor structures of each measure;
(iv) to ascertain construct validity through modified application of MTMM procedures aimed at addressing both the convergent (i.e. multitrait–heteromethod or MTHM) and divergent (i.e. heterotrait–monomethod or HTMM) validity.

Psychometrically, the results are intended to assess the measurement adequacy of the three entry-level personality inventories widely used in the UK. Theoretically, the results are intended to clarify the meaning of the FFM and to provide convergent and discriminant validity evidence for scales assessing the Big Five personality factors. In addition, the results from this research are expected to have practical relevance to IWO psychologists and practitioners who use alternative measures of personality in individual and job applicant assessments.

METHOD

Sample

Five hundred and four British university students completed three occupational personality inventories (see below). Participants were recruited from several different university colleges in London. 315 (62.5%) were female and 189 (37.5%) were male, and mean age was 23.77 years (SD = 5.92 years).

Administration

Personality inventories were administered to groups of participants who took part in this study either in return for a feedback report on their personality profile or for course credits. Subjects completed all three inventories in a single sitting. The inventories were completed by subjects in randomized order across approximately 40 group testing sessions carried out under standardized conditions and in accordance with the test publishers' administration instructions. Participation was entirely voluntary and a de-briefing was given to subjects at the end of each administration session. The session was introduced as being part of a study into graduate selection and graduates' responses to personality tests in selection settings. Participants were informed that they were to complete three different measures, and any questions could be addressed to the administrator at the end of the testing session. Participants completed all measures in their own time and in accordance with the test publishers' administration instructions (for further details see Ones & Anderson, 2002).

Measures

The personality inventories used were the British versions of the Hogan Personality Inventory (HPI), the Occupational Personality Questionnaire (OPQ), and the Business Personality Indicator (BPI). It should be noted that although the versions of these measures used were the British (BPI, OPQ) or anglicized (HPI) versions of these questionnaires, many similarities were evident between these and the American versions of these measures.

Hogan Personality Inventory

The Hogan Personality Inventory (HPI) is a well recognized, highly regarded, widely used personality inventory that was initially developed in the United States for industrial–organizational and vocational applications (Hogan & Hogan, 1995). It is based on the Five Factor Model (FFM) of personality (cf. Costa & McCrae, 1992; Goldberg, 1990). The HPI makes use of homogenous item clusters (HICs) to construct seven primary scales that reflect aspects of the Big Five dimensions. The British version of the HPI was recently made available for use in personnel selection. Although there are only minor differences between the British and US versions, the test is supported by the US technical manual (Hogan & Hogan, 1995), with a number of items having been anglicized to British spellings and phraseology to be more acceptable to the British cultural context. It includes 206 dichotomous (true–false) items. The British HPI has 41 homogenous item clusters (HICs), which in turn load onto the seven primary scales (Adjustment, Ambition, Sociability, Likeability, Prudence, Intellectance, and School Success: Hogan & Hogan, 1995, p. 12). However, it should also be noted that four HICs cross-load onto more than one factor (these being Not Autonomous, Not Spontaneous, Impulse Control, and Avoids Trouble). Table 1 presents detailed descriptions of these primary scales as published in the HPI manual (Hogan & Hogan, 1995), including sample items for each scale.

Occupational Personality Questionnaire

The British standardized student version FS5.2 of the Occupational Personality Questionnaire (OPQ) was used. OPQ is a widely used family of measures of normal-adult personality designed specifically for work psychology applications. At least five non-parallel forms of the OPQ exist at present, assertedly measuring between six and 32 primary scales (OPQ Images, OPQ Customer Contact, OPQ Factor, OPQ Work Styles, and OPQ32: Anderson, 2000). In the present study we utilized version FS5.2 as it was

Table 1. HPI scales, scale descriptors, and sample items

HPI scale name	Scale description (from HPI manual, Hogan & Hogan, 1995, p. 13)	Sample items	FFM composite scale loading
Adjustment	'measures the degree to which a person appears calm and self accepting, or conversely, self-critical and tense'	• I am seldom tense or anxious • I keep calm in a crisis	I. Emotional Stability
Ambition	'measures the degree to which a person seems socially self-confident, leader-like, competitive, and energetic'	• I am an ambitious person • In a group, I like to take charge of things	II. Extraversion
Intellectance	'measures the degree to which a person is perceived as bright, creative, and interested in intellectual matters'	• I am interested in science • I am a quick-witted person	III. Openness to Experience
Likeability	'measures the degree to which a person is seen as perceptive, tactful, and socially sensitive'	• I work well with other people • I am sensitive to other people's moods	IV. Agreeableness
Prudence	'measures the degree to which a person seems conscientious, confirming, and dependable'	• I always practice what I preach • I strive for perfection in everything I do	V. Conscientiousness
School Success	'measures the degree to which a person seems to enjoy academic activities and to value educational achievement for its own sake'	• I have a large vocabulary • I can multiply large numbers quickly	III. Openness to Experience
Sociability	'measures the degree to which a person seems to need and/or enjoy interacting with others'	• I would go to a different party every night if I could • Being part of a large crowd is exciting	III. Extraversion

developed specifically for use on student populations and on applicants with little or no work experience. The FS5.2 version of the OPQ was developed specifically for entry-level applicants to the labour market and comprises 136 normative items using a five point response scale (strongly disagree–strongly agree.) One hundred and twenty-eight of the items loaded onto primary scales, and eight items made up the social desirability scale. Initially, the OPQ was designed around a four factor model of personality comprising Feeling, Relating, Thinking, and Energies domains (SHL, 1999), with acknowledged links to the FFM structure. There are 16 primary content scales on the OPQ, each containing eight items. Table 2 presents scale descriptors from the test manual (SHL, 1999), together with sample items.

Business Personality Indicator

The Business Personality Indicator (BPI), now updated and marketed by ASE under the title of Pin-Point, is also a measure of personality of entry-level applicants in the UK (Feltham & Woods, 1995). The BPI uses 128 trichotomous items (yes–?–no) to measure 11 primary content scales. These primary scales load onto five second order dimensions with acknowledged links to the FFM as follows: Dynamic, Extravert, Work Stamina, Worrying, and Controlled. Table 3 lays out descriptors for the 11 primary scales alongside sample items from the test and its user manual (Feltham & Woods, 1995).

Independent ratings of primary scales

In mapping the three inventories' scales onto the Big Five personality domain, independent ratings of primary personality scales were used. In addition to the sample of participants in this study who completed the three personality inventories, we also obtained ratings from a separate sample of independent raters ($n = 66$) of the primary source scales onto the FFM. For this procedure raters were presented with the scale descriptors appearing in the test publishers' manuals (see Tables 1–3) together with Goldberg's (1993) descriptors of the Five Factor Model scales. The latter was chosen because it presented relatively simplistic descriptions of the FFM scales that would be unambiguous to raters acting in this procedure. Raters were instructed to allocate each of the primary source scales across the three inventories to one of the FFM scales but allowing for a non-response category where raters could not clearly identify to which FFM scale a source scale should load. All raters were attending full- and part-time postgraduate masters courses in IWO Psychology in the UK, mean of years working experience was 7.64 (SD = 8.51), mean age was 28.9 years (SD = 8.17), and gender breakdown was 74.2% female and 25.6% male.

ANALYSES AND RESULTS

We adopted a three-stage analytical approach to this dataset. First, within-inventory psychometric properties were examined by computing scale descriptive statistics (means and standard deviations) and internal consistency reliabilities (Cronbach's alphas). These were compared to data from normative samples. Second, we examined the factor structure of each inventory through exploratory factor analyses (EFAs) at both the item and the primary scale levels of analysis (see also Anderson & West, 1998). Note that item level data was available for the OPQ and BPI whereas for the HPI only HIC level data was available, thus restricting our analysis of the latter. Third, construct validity was examined across all three measures combined at the level of primary source scales. Here

Table 2. OPQ scales, scale descriptors, and sample items

OPQ scale name	Scale description	Sample items	FFM composite scale loading
Achieving	'how important career ambition is in people's lives, how high they set their sights and how much precedence work takes over the commitments, such as family and social life'	• I strive to be the best in my field	V. Conscientiousness
Active	'how much people enjoy physical activities, particularly of a strenuous nature both in and out of work'	• I take part in energetic activities • I have less energy compared with other people	II. Extraversion
Competitive	'how much people need to win, hate to lose, and like to beat others at all costs	• I am always determined to come first • I have been called a poor loser	II. Extraversion
Conceptual	'how much people are interested in thinking about and discussing abstract theories, problems, and hypotheses'	• I prefer to deal with problems requiring abstract thinking	III. Openness to Experience
Data Rational	'how much people enjoy working with numbers and facts, looking for trends in data and conducting quantitative analyses'	• I find it easy to understand mathematical problems • I enjoy analysing statistical information	III. Openness to Experience
Decisive	'how quickly people arrive at conclusions, how rapidly they make up their minds and whether they are prepared to take risks'	• I am prepared to make fast decisions even at the risk of making an error	V. Conscientiousness
Emotionally controlled	'the extent to which people hide their feelings and exercise restraints in showing their emotions'	• I express my emotions freely • I am able to remain calm in emotionally charged situations	I. Emotional Stability
Empathic	'how prepared people are to listen to the personal problems of others as well as how concerned they are with feelings and welfare issues of others'	• I am not always very considerate to other people • I am interested in the welfare of others	IV. Agreeableness
Influential	'how much people like taking charge of others, managing, directing, and telling people what to do'	• I like directing the work of others • I often make decisions for the group	II. Extraversion
Innovative	'how much people feel that they generate new ideas and original solutions to problems'	• People come to me for creative ideas • I find I have an original approach to problems	III. Openness to Experience

Methodical	'how much people plan, how important tidiness, accuracy and precision is, and how conscientious they are'	• My paperwork is always neat and accurate • I like to see a job through once it has been started	V. Conscientiousness
Optimistic	'the extent to which people look on the bright side and remain happy and cheerful, even through difficulties'	• I find it difficult to stay cheerful • I keep my spirits up despite setbacks	I. Emotional Stability
Outspoken	'how strongly people hold their views and go their own way, and also how much they are prepared to voice those views, even if opposed'	• I tend not to have strong opinions on things • I tend to go with the consensus view	II. Extraversion
Relaxed	'how calm people are on an on-gong basis, how free they are of anxieties and how easily they can switch off'	• I generally remain calm about things • My feelings are rarely hurt	I. Emotional Stability
Sociable	'how comfortable people feel in the company of others and how talkative or sociable they tend to be'	• I am lively and talkative at social events • I am confident with strangers	II. Extraversion
Traditional	'a preference for more conventional, established methods and practices rather than new and radical ones'	• I like to preserve well-proven methods • I regret the passing of old traditions	III. Openness to Experience

Table 3. BPI scales, scale descriptors, and sample items

BPI scale name	Scale description	Sample items	FFM composite scale loading
Change Oriented	'active as opposed a passive approach to work and job tasks, interest in working in a variety of tasks, or in different ways'	• Be more interested in possibilities than practicalities, and open to change • Prefer the freedom to tackle their work in their own way	V. Conscientiousness
Competitive	'how important is it to individuals to be recognized as successful in their careers'	• Be ambitious • Value status more than popularity with colleagues	II. Extraversion
Limelight Seeking	'tendency to seek the limelight or be the centre of attention in group of social situations'	• Enjoy giving presentations • Feel comfortable having to think on their feet in group situations	II. Extraversion
Outgoing	'degree to which an individual seeks and enjoys the company of other people'	• Be talkative and outgoing • Have a wide circle of friends	II. Extraversion
Perfectionist	'extent to which individuals seek to carry out tasks to their own high standards, and not necessarily to other people's'	• Want to get all the detail right • Be seen as over-conscientious when carrying out tasks where speed is more important than total accuracy	V. Conscientiousness
Risk Taking	'the preference for taking risks, particularly those connected with finance and business'	• Be willing to take risks • Sometimes miss out on opportunities because of cautiousness	II. Extraversion
Stamina	'tendency to maintain mental energy levels when under pressure'	• Report they need less sleep • Be prone to pushing themselves too hard	III. Openness to Experience
Time Managed	'importance the individual places on planning and scheduling activities in advance'	• Attach a lot of importance to planning their day and scheduling time efficiently • Feel more comfortable when their schedule is predictable and under control	V. Conscientiousness
Warm	'degree of openness to other people'	• Relate to others by sharing their thoughts and feelings readily • Rated as easy to read by colleagues and business associates	II. Extraversion
Work Oriented	'degree of emphasis which the individual places upon work as opposed to leisure activities'	• Take work very seriously with a strong work ethic • Value more time spent at work than time outside of work	V. Conscientiousness
Worrying	'the individual's tendency to worry and feel anxious in different situations'	• Worry and blame themselves • Find it harder to keep things in perspective	I. Emotional Stability

Table 4. HPI descriptives and reliabilities

Primary scale	Study results			Publisher data[a]			Comparison	
	Mean	SD	Alpha[b]	Mean	SD	Alpha	*d* value	SD ratio
Adjustment	21.35	6.61	—	26.57	7.10	0.89	−0.74	0.93
Ambition	20.70	5.40	—	23.61	5.00	0.86	−0.58	1.08
Intellectance	15.51	4.78	—	14.68	4.90	0.78	0.17	0.98
Likability	18.16	2.90	—	19.62	2.36	0.71	−0.62	1.23
Prudence	15.61	4.46	—	20.28	4.62	0.78	−1.01	0.97
School Success	8.58	3.06	—	8.66	3.14	0.75	−0.03	0.97
Sociability	15.65	4.18	—	13.47	4.86	0.76	0.45	0.86
Mean alpha						0.79		

[a]From Hogan and Hogan, 1995.
[b]Could not be computed as items assessing each trait were not known.
d = standardized mean differences between the study sample and the publisher data; positive values indicate a higher mean for the sample of this study; negative values indicate a higher mean for the publisher data. *d* values were computed using the manual normative standard deviations. The SD ratios were computed by dividing the sample SD by the normative SD.

we utilized the independent ratings of the primary scales onto the FFM in order to examine convergent validity (i.e. monotrait–hetermethod correlations) and divergent validity (i.e. heterotrait–monomethod correlations) of the personality domain measured by these three FFM-based inventories. Results from each of these three stages of analysis are presented below.

Within-measure descriptive statistics and reliabilities

Tables 4–6 present descriptive statistics (means and SDs) and scale internal consistency reliability coefficients (Cronbach's alpha) for the HPI, OPQ and BPI, respectively. Each table compares the findings of this study against those reported for each measure in the test publishers' manual to allow direct comparisons between respective values. The last two columns in each table indicate (i) the standardized mean difference between means found in this study and manual norms (Cohen's *d*, standardized mean difference between means of the two groups compared—we used the manual national standard deviations in computing *d* values) and (ii) the ratio of sample to test manual national standard deviations. The former indexes how far apart our sample's means are from norms and the latter can be used to gauge the relative variability of the personality scales in our particular sample.

For the HPI, our sample appears to be lower on adjustment and prudence than the normative sample. Moderate mean differences can be observed on ambition, likability, and sociability (lower on the former two scales and higher on the latter). The SD ratios suggest that our sample was about 23% more variable in terms of likability, and 14% less variable in terms of sociability. Other variability differences were small. For the OPQ, our sample was moderately less achieving, active, decisive, empathic, and traditional than the normative sample. In terms of variability comparisons, our sample was about 27% more variable in terms of achieving, and 14% less variable in terms of innovative. Similar to the results for the HPI, other variability differences were negligible. For the BPI, moderate standardized mean differences were found for change oriented, outgoing, warm, worrying, stamina, and competitive. Our sample scored moderately higher in terms of the

Table 5. OPQ descriptives and reliabilities

Primary scale	Study results			Publisher data[a]			Comparison	
	Mean	SD	Alpha	Mean	SD	Alpha	*d* value	SD ratio
Achieving	21.23	6.41	0.53	23.59	5.03	0.71	−0.47	1.27
Active	24.72	5.92	0.68	27.75	6.42	0.85	−0.47	0.92
Competitive	18.14	6.04	0.58	19.18	5.73	0.81	−0.18	1.05
Conceptual	26.12	4.97	0.49	25.18	5.41	0.83	0.17	0.92
Data Rational	22.87	8.10	0.73	23.67	7.16	0.90	−0.11	1.13
Decisive	22.21	5.42	0.63	24.45	5.95	0.85	−0.38	0.91
Emotionally Controlled	25.28	6.26	0.64	25.02	6.46	0.84	0.04	0.97
Empathic	28.94	5.01	0.58	30.81	4.47	0.74	−0.42	1.12
Influential	26.84	5.88	0.69	27.96	6.01	0.87	−0.19	0.98
Innovative	24.84	4.97	0.69	26.44	5.78	0.86	−0.28	0.86
Methodological	28.02	5.41	0.65	28.31	5.53	0.82	−0.05	0.98
Optimistic	28.91	5.21	0.56	29.48	5.40	0.81	−0.11	0.96
Outspoken	28.01	5.12	0.56	27.57	5.10	0.77	0.09	1.00
Relaxed	22.41	5.44	0.43	21.87	5.20	0.76	0.10	1.05
Sociable	29.78	6.84	0.70	28.17	6.63	0.87	0.24	1.03
Traditional	20.04	5.12	0.45	22.29	4.71	0.73	−0.48	1.09
Social Desirability	19.42	3.96	0.57	20.64	4.30	0.71	−0.28	0.92
Mean alpha			0.59			0.81		

[a]From SHL (1999).

d = standardized mean differences between the study sample and the publisher data; positive values indicate a higher mean for the sample of this study; negative values indicate a higher mean for the publisher data. *d* values were computed using the manual normative standard deviations. The SD ratios were computed by dividing the sample SD by the normative SD.

Table 6. BPI descriptives and reliabilities

Primary scale	Study results			Publisher data[a]			Comparison	
	Mean	SD	Alpha	Mean	SD	Alpha	*d* value	SD ratio
Change Oriented	11.62	3.08	0.68	15.9	6.8	0.87	−0.63	0.45
Competitive	14.64	3.44	0.72	10.2	7.1	0.79	0.63	0.48
Limelight Seeking	11.81	3.51	0.77	12.0	6.7	0.85	−0.03	0.52
Outgoing	12.66	3.13	0.81	15.8	7.4	0.85	−0.42	0.42
Perfectionist	10.71	2.97	0.60	12.4	4.9	0.70	−0.34	0.61
Risk Taking	7.89	2.48	0.75	9.4	5.4	0.81	−0.28	0.46
Stamina	12.21	3.44	0.65	8.7	5.7	0.75	0.62	0.60
Time Managed	7.88	2.36	0.58	8.0	5.2	0.74	−0.02	0.45
Warm	11.98	3.09	0.56	14.8	5.0	0.74	−0.56	0.62
Work Oriented	9.15	2.61	0.70	4.6	4.0	0.78	1.14	0.65
Worrying	10.68	3.15	0.55	14.4	6.0	0.74	−0.62	0.53
Mean alpha			0.67			0.79		

[a]From Feltham and Woods (1995).

d = standardized mean differences between the study sample and the publisher data; positive values indicate a higher mean for the sample of this study; negative values indicate a higher mean for the publisher data. *d* values were computed using the manual normative standard deviations. The SD ratios were computed by dividing the sample SD by the normative SD.

latter two personality traits; higher mean scores were found for the former scales. For the work oriented scale, our sample's mean was drastically different than that of the normative sample. In fact, Table 6 indicates that the *d* value associated with this scale was 1.14. In other words, our sample was over a standard deviation more work oriented than the normative sample. Finally, and perhaps most significantly, SD ratios for all BPI scales suggested that our sample was about half as variable as the normative sample.

While significant mean differences between sample and population norms are to be expected, they have little consequence for rank ordering individuals. However, reductions in sample variability indicate range restriction; conversely, higher sample variabilities indicate range enhancement. Both have direct consequences for relationships that can be obtained with other measures (including criterion-related, convergent, and divergent validity). Range restriction downwardly depresses relationships with other variables. Reliability, convergent/divergent validity, and factor analyses can hence be directly influenced by the variability in the sample.

We turn our attention to differences in scale reliability coefficients reported in Tables 5 and 6 for the OPQ and BPI (note that for the HPI only scale level data was provided by the test publisher so scale reliability analyses could not be computed by the authors). For the OPQ, mean alpha was 0.59 in this study as compared with 0.81 reported in the test manual, with alphas ranging between 0.43 and 0.73 in this study as compared with 0.71 and 0.90 in the published manual. For the BPI, mean reliability was 0.67 in this study as against 0.79 in the test manual, and scale alphas ranged 0.55–0.81 in the study as compared with 0.70–0.87 in the published manual. On both inventories scale reliabilities were in all cases lower, in some cases substantially lower, than those reported in the publisher's manual. That scale reliabilities were found to be universally lower in this study also raised issues for subsequent data analysis, particularly the level of analysis of either the primary scale or alternatively at composite FFM level scales. We should note that part of the reason for lower reliabilities observed in our sample might be the reduced observed variability of some scales (this is especially true for BPI scales).

Reliability, homogeneity and error scores: some potential explanations

One possible explanation, and seemingly the most likely one in retrospect, concerns relationships between sample homogeneity, error scores, and scale reliability. In relatively homogenous samples, total variance by definition is small. However, error scores and true scores are uncorrelated and homogeneity does *not* affect the size of the error variance. Reliability *is* influenced by sample homogeneity as it is the ratio of the (1−error variance) to total variance. To estimate the expected reliability in a new (less variable) sample, one can make use of the standard error of measurement formula. Reliability expected in a more homogenous sample is given by the following formula (derived by equating SEMs for two samples):

$$\text{expected reliability}_{(\text{new sample})} = 1 - \{[\text{var}_{\text{old}^*}(1 - \text{reliability}_{\text{old}})]/\text{var}_{\text{new sample}}\}$$

To estimate whether or not the reduced reliabilities we found in our sample were simply due to reduced variability, we computed the expected reliabilities in our sample using the above formula. For the BPI, we found that the reliabilities we obtained in our sample were actually *higher* than those expected, given that the standard deviations are drastically smaller in our sample than in the normative sample. Under the influence of range

restriction alone, the expected mean reliability of BPI scales in a sample with the same variability as our sample was 0.21, compared to the mean reliability of 0.67 that we actually obtained. In other words, range restriction should have reduced reliability of BPI scales even further than it actually did. This suggests that a second effect may be working to produce greater than anticipated lower reliability levels in our sample. There are a number of potential explanations for this.

Knowles (1988) describes a phenomenon when subjects complete increasing number of items assessing a personality trait: repeated consideration of items on personality scales increases accessibility, clarity, and content of test-related schema and thus produces responses that are increasingly polarized, consistent, and reliable. This is a different effect than the so-called 'bloated specifics', that is asking essentially the same item content several times over in order to artificially increase scale reliability. Bloated specifics, of course, occur on primary scales within a single measure. By asking the virtually the same item many times over, relatively high levels of internal consistency reliabilities can be achieved due to increased inter-item correlations, with relatively few items, at the cost of construct coverage. The effect described by Knowles (1988) appears to occur even with a large number of items, offering relatively broad and full coverage of the construct domain being assessed.

In our study, by having participants complete three personality inventories in one sitting, we may have contributed to the finding of higher internal consistency reliability despite lower levels of reliability that should have resulted on the basis of relative homogeneity of our sample. Although subjects completed the three measures in a randomized order, due to randomization in two-thirds of the administration conditions, the BPI was completed either as the second or third personality inventory. Further, it could have been that items on the BPI were more prone to this response effect. For example, its use of double-barrelled items that perhaps could be argued to be more extreme in tone (see examples in Table 3) and trichotomous response options could have created higher internal consistency reliability despite lower levels of reliability that should have resulted on the basis of relative homogeneity of our sample on the BPI. Item content, scaling, and order need to be isolated to disentangle these potential effects, something that we unfortunately cannot do in this study. Future research should explicitly examine these effects in personnel selection settings.

Although our explanations regarding the effects of reduced variability in our sample and potential benefits of repeated responses to similar items are *post hoc* and at best speculative, the main point that we wish to make, nonetheless, holds: when the BPI is used with samples similar to ours, the means, SDs, and reliabilities can be quite different from the normative data reported by the test publisher. Nevertheless, all in all, given range restriction/enhancement on some of the scales and somewhat lower reliabilities than found for normative samples, we expected that both factor analytic findings and MTMM analyses would suffer.

Within-measure EFAs

Initial exploratory factor analyses were directed at the item level of analysis on the OPQ and BPI (for the HPI item level data were not usable as we did not know which items were part of which scales and this would hamper our interpretational capabilities). Pre-analysis checks as recommended by Comrey (1978) indicated that both datasets were suitable for EFAs to be run. For the OPQ, Bartlett's test of sphericity was significant ($p = 0.000$) and the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy was acceptable at 0.812. Likewise on the BPI, Bartlett's test was significant ($p = 0.000$) and KMO was 0.734. In both cases the

ratio of *n* cases to *n* items was acceptable, approaching 5:1 (Gorsuch, 1983; Kline, 1986) (OPQ = 504:136 and BPI = 504:128). We therefore computed two EFAs without rotation initially. Examination of the scree test (Cattell, 1966) and eigenvalues in excess of 1.0 suggested a five or possibly six factor solution in both cases. Nevertheless, we then attempted to replicate the factor solutions reported in the test publishers' manuals using maximum likelihood estimates, restricting the *n* of factors to 16 and 11 respectively, and using varimax (orthogonal) rotation. In both cases this procedure replicated as closely as possible the original factor analysis procedures computed by the test publishers. Where this was not clear due to a lack of detail in the test manual, we used this 'default' procedure of an unrotated initial solution, examination of the eigenvalues and scree test, followed by a Varimax rotated solution, as a generally accepted conservative approach to EFA (Kline, 1986). Nevertheless, for the OPQ this rotated solution explained only 39.6% of total variance and for the BPI a disappointing 26.4% of total variance. In both solutions the loading of items onto primary scales was notably unclear with multiple cross-loadings and a lack of items loading at acceptable levels (Kline, 1986). These solutions are therefore not reported here due to their ambiguous findings (they are available from the authors upon request). Since we failed to replicate the test publishers' factor structures at the item level of analysis, we turned our attention to factor analysis of the primary scales (OPQ and BPI) and HICs (HPI).

Using standardized primary scale and HIC level data, principal components analyses were run by computing an initial solution without rotation or iteration, examination of the scree plot and eigenvalues, and then running a rotated solution with Varimax (orthogonal) rotation limiting the number of components extracted as appropriate. For the BPI the rotated solution accounted for 44.14% of total variance but suggested a four component underlying structure (53.03% of the variance was accounted for by a five component solution), whereas for the OPQ the rotated solution accounted for 52.08% of variance again with a four component solution (57.96% of the variance was accounted for by a five component solution). For the HPI, a principal components analysis of the standardized HIC level data suggested a seven component solution. The rotated solution accounted for 45.91% of total variance (36.17% of the variance was accounted for in a five component solution). Nevertheless, all solutions were ambiguous in terms of several primary scale loadings, multiple cross-loadings were evident. In any event, our solutions did not concur with publisher EFAs reported in the test manuals. We therefore failed in this study to replicate the test publisher EFA solutions at both the item and the primary scale (and HIC) levels of analysis for three personality inventories examined.

This may feasibly have been due to range restriction/enhancement for some scales and low reliabilities reported earlier. The scale descriptives reported earlier in this paper show some notable differences from those reported by the test publishers. This is disconcerting to us because these measures are all intended for precisely the population we sampled—new entrants and student entrants to the British labour market.

We attempted to replicate the test publisher factor structures for the OPQ and BPI using two levels of analysis, using procedures as identical as possible to those used originally by the publisher, and limiting the number of factors extracted to 'fit' with the published underlying structure. As anticipated the results of this aspect of our present study, despite being based on a relatively large sample of British job market entrants, simply did not comply with those reported in the manuals for these two personality inventories (SHL, 1999; Feltham & Woods, 1995). Moreover, the EFAs run at both the item and primary scale levels of analysis failed to generate acceptably clear underlying solutions by established conventions in factor analysis (see e.g. Comrey, 1978; Kline, 1986).

Between-measure construct validity

Pre-analysis checks

Prior to conducting any analyses into construct validity using the five factor framework, it was important to examine inter-rater agreement indices of the ratings of primary scales onto the FFM ($n = 66$). The appropriate indices in this case were Cohen's kappa as modified by Fleiss (1971) in combination with its associated *Z*-value and probability (Shrout & Fleiss, 1979). The kappa coefficient of inter-rater agreement, although in theory ranging between zero (no agreement) and 1.0 (perfect agreement), is in practice a highly conservative statistic with values in excess of 0.5 indicating acceptable levels of inter-rater agreement (Shrout & Fleiss, 1979). For the HPI, kappas ranged from 0.559 to 0.581 (*Z*-values = 119.72–143.31, $p < 0.001$ respectively); for the OPQ, kappas ranged from 0.563 to 0.592 (*Z*-values = 122.01–135.75, $p < 0.001$); and for the BPI, kappas ranged from 0.565 to 0.591 (*Z*-values = 127.83–136.33, $p < 0.001$). These pre-analysis checks therefore indicated sufficient levels of inter-rater agreement for all three measures to permit further more sophisticated investigation at the level of analysis of loadings onto the FFM scales. We should also note that scale classifications into the Big Five, for the most part, agreed with those provided by Hough and Ones (2001). A few differences occurred because Hough and Ones (2001) explicitly acknowledged multiple loadings of some scales (e.g. the ambition scale of the HPI assesses both conscientiousness and extraversion), while in the design of this study we aimed to use strictly the FFM (Goldberg, 1990).

Composite scales were then computed summing primary source scales onto the Five Factor Model framework. Referring back to Tables 1–3, details of the loadings of primary scales onto the FFM are reported in the final column. This procedure was used by the authors to compute the second-order FFM scales in preference to computing these factor analytically by entering all primary scales into a single EFA procedure. This was done on the grounds that (i) inter-rater agreement was high whilst initial EFAs produced markedly inconclusive results and (ii) the task of independent raters in contrast to the original subject sample who completed the measures was to focus at this level of analysis. Construct validity was examined through use of modified MTMM analysis techniques in this case applied to the multiple-measure, multiple-trait matrix generated in this study. First, we examined convergent validity through monotrait–heteromethod (MTHM) correlations within each of the FFM scales. Second, we explored divergent validity through computing mean heterotrait–monomethod (HTMM) correlations within each of the three inventories. Finally, by comparing MTHM correlations against HTMM correlations we were able to ascertain key aspects of construct validity. Results are reported below.

Convergent validity

Table 7 sets out MTHM correlations by FFM dimension. Note that whilst independent raters allocated differing numbers of primary scales to each FFM dimension (ranging between 2 onto Agreeableness to 12 onto Extraversion), correlations within each dimension are notably low. Indeed, only a small minority of statistically significant correlations were found, mostly between primary scales loading onto Extraversion and Openness to Experience, far fewer than might have been expected. Moreover, upon closer inspection even these significant correlations are without exception between primary scales on the OPQ rather than between scales drawn from different inventories. On Extraversion, the OPQ scales Influential, Sociable, and Active correlate significantly (*r*-values range from 0.39 to 0.49, $p < 0.05$), whilst none of the HPI or BPI primary scales correlate significantly even where the scale names are conceptually similar or even

Table 7. Monotrait–heteromethod correlations

	I EMOTIONAL STABILITY	BPI Worrying	HPI: Adjustment	OPQ: Emotionally Controlled	OPQ: Relaxed	OPQ: Optimistic
I EMOTIONAL STABILITY	Mean $r = -0.03$					
BPI Worrying						
HPI: Adjustment		0.01				
OPQ: Emotionally Controlled		−0.05	−0.02			
OPQ: Relaxed		−0.09	0.05	*0.73*		
SOPQ: Optimistic		−0.10	−0.04	0.05	−0.15	

	II EXTRAVERSION	BPI: Competitive	BPI: Warm	BPI: Outgoing	BPI: Risk Taking	BPI: Limelight Seeking	HPI: Ambitious	HPI: Sociable	OPQ: Influential	OPQ: Sociable	OPQ Active	OPQ Out Spoken	OPQ: Competitive
II EXTRAVERSION	Mean												
BPI: Competitive	$r = 0.07$												
BPI: Warm		0.04											
BPI: Outgoing		0.04	−0.03										
BPI: Risk Taking		0.01	0.03	0.02									
BPI: Limelight Seeking		−0.06	0.01	0.09	0.05								
HPI: Ambitious		0.03	0.07	−0.00	0.01	0.02							
HPI: Sociable		0.04	0.07	0.02	−0.01	0.02	*0.46*						
OPQ: Influential		−0.08	0.06	−0.02	0.03	0.00	0.12	0.07					
OPQ: Sociable		0.01	−0.06	0.08	−0.02	0.01	0.09	0.00	*0.46*				
OPQ: Active		−0.11	0.11	0.00	0.08	−0.01	0.13	0.04	*0.39*	*0.44*			
OPQ: Outspoken		0.01	0.06	−0.01	0.00	−0.03	0.05	−0.01	*0.49*	*0.44*	0.40		
OPQ: Competitive		−0.04	0.02	−0.07	0.09	0.02	0.01	0.05	*0.18*	0.00	0.10	0.03	

Continues

Table 7. Continued

	III OPENNESS TO EXPERIENCE	BPI: Stamina	HPI: Intellectance	HPI: School Success	OPQ: Innovative	OPQ: Traditional	OPQ: Data Potential	OPQ: Conceptual
III OPENNESS TO EXPERIENCE	Mean $r = 0.09$							
BPI: Stamina								
HPI: Intellectance		0.06						
HPI: School Success		0.04	*0.38*					
OPQ: Innovative		−0.06	0.06	0.11				
OPQ: Traditional		−0.00	0.06	0.01	*−0.22*			
OPQ: Data Potential		0.01	0.15	0.18	*0.23*	0.10		
OPQ: Conceptual		0.02	0.02	0.02	*0.51*	−0.16	*0.29*	1.00

	IV AGREEABLENESS	HPI: Likability	OPQ Empathic
IV AGREEABLENESS	Mean $r = -0.05$		
HPI: Likability			
OPQ: Empathic		−0.05	

	V CONSCIENTIOUSNESS	BPI: Change Oriented	BPI: Time Managed	BPI: Work Oriented	BPI: Perfectionist	HPI: Prudence	OPQ: Methodical	OPQ: Achieving	OPQ: Decisive
V CONSCIENTIOUSNESS	Mean $r = 0.02$								
BPI: Change Oriented									
BPI: Time Managed		−0.02							
BPI: Work Oriented		0.03	0.09						
BPI: Perfectionist		−0.02	0.00	0.08					
HPI: Prudence		0.02	0.04	−0.04	0.01				
OPQ: Methodical		−0.00	0.07	−0.05	0.01	0.02			
OPQ: Achieving		0.01	0.00	−0.02	−0.05	0.07	*0.26*		
OPQ: Decisive		0.05	0.02	−0.01	−0.12	0.01	0.04	0.07	

Correlations which are statistically significant ($p < 0.05$) are in italics.

identical (e.g. HPI Sociable, BPI Outgoing, BPI Competitive). For instance, OPQ Sociable correlates $r = 0.00$ with HPI Sociable and $r = 0.08$ with BPI Outgoing; OPQ Outspoken correlates $r = -0.01$ with BPI Outgoing; and OPQ Competitive correlates $r = -0.04$ with BPI Competitive. On the FFM dimension Openness to Experience a similar pattern is evident. The OPQ scales Innovative, Traditional, Data Rational, and Conceptual all intercorrelate significantly (Traditional negatively with the other three) with *r*-values varying between -0.22 and 0.51 ($p < 0.05$). Between-measure primary scale correlations are much lower however, with, for instance, OPQ Conceptual correlating $r = 0.02$ with HPI Intellectance. This pattern of intercorrelations clearly suggests a lack of between-measure, single-trait convergent validity.

Within each FFM dimension mean MTHM correlations were then calculated as reported in Table 7. These were as follows: Emotional Stability mean $r = -0.03$; Extraversion 0.07; Openness to Experience 0.09; Agreeableness -0.05; Conscientiousness 0.02. All were therefore near zero and either marginally positive or marginally negative, but in all cases indicative of a troublesome lack of convergent validity across scales perceived to measure the same Big Five traits. The overall mean multitrait–heteromethod correlation for all five dimensions was $r = 0.03$ ($p =$ NS), again so notably close to zero as to raise pertinent concerns over convergent validity.

Divergent validity

Intercorrelations between the FFM dimensions are reported in Table 8. These are similar in magnitude to those reported in earlier meta-analyses of FFM-based measures of normal adult personality (e.g. Ones et al., 1995), suggesting that the underlying structure of relationships between dimensions computed for the present study was not greatly different to that found in earlier larger-scale meta-analytic investigations. Thus, we explored divergent validity through the calculation of heterotrait–monomethod (HTMM) correlations in the usual manner. For the HPI mean HTMM was $r = 0.228$ (7×7 matrix, mean of 21 *r*-values); for the OPQ mean HTMM was $r = 0.105$ (16×16 matrix, mean of 120 *r*-values); and for the BPI mean HTMM was $r = 0.068$ (11×11 matrix, mean of 55 *r*-values). These correlations provide some limited evidence for within-measure divergent validity as they are not particularly high values. However, in comparison with the near-zero MTHM correlations reported in Table 7, it is apparent that these HTMM values are without exception higher than those found for convergent validity. To summarize, it can be stated that MTHM correlations evidential of convergent validity were near zero whereas HTMM correlations indicative of divergent validity were somewhat higher. Both sets of correlations, but especially the comparison between the notably disappointing convergent validity correlations and the low-to-moderate divergent validity correlations, are disconcerting, but to be expected given the range restriction and unreliability issues discussed earlier.

Table 8. Intercorrelations between FFM composite dimensions

	Emotional Stability	Extraversion	Openness to Experience	Agreeableness	Conscientiousness
Extraversion	0.21	—			
Openness to Experience	0.18	0.42	—		
Agreeableness	0.20	0.55	0.47	—	
Conscientiousness	0.15	0.38	0.33	0.07	—

All correlations statistically significant at $p < 0.01$ (two-tailed) except between Agreeableness and Conscientiousness ($n = 497$–480 depending on missing vales).

DISCUSSION

Returning to the four aims specified earlier for this study, we found the following.

(i) Scale descriptive data for the present sample suggested that for some scales there was non-negligible range restriction or range enhancement.
(ii) Scale reliabilities for primary scales were universally lower for this data set compared with alpha coefficients reported in the test manuals. Alphas could not be computed for the HPI, however, due to the publishers' policy of non-release of the item-to-scale scoring key. Overall mean alpha for the OPQ was 0.59 as opposed to 0.81 in the manual, and for the BPI 0.67 as compared with 0.79 in the test manual.
(iii) We failed to replicate the publisher first-order or second-order factor solutions. Although exploratory analyses suggested three to seven factors, depending on inventory and primary scale/HIC *vis-à-vis* item level analyses, the solutions were not unambiguously clean.
(iv) Disconcerting construct validity findings emerged. As regards convergent validity, the between-measure, single-trait MTHM correlations were near zero (see Table 7). With respect to divergent validity, the within-measure, multiple-trait HTMM correlations were higher: HPI mean $r = 0.228$, OPQ mean $r = 0.105$, BPI mean $r = 0.068$.

What accounts for this set of results? As suggested earlier, the vast nomological net provided in personality inventory manuals is only applicable to samples that have similar variability and similar levels of reliability. In samples where variability is markedly reduced, construct validity suffers. Although in this study we focused on factor structure and MTMM analyses, the same phenomenon would also be expected for criterion-related validity. In using personality measures for personnel selection purposes, it is essential to assess variability among job applicant samples. Relatedly, reliability of personality scale scores also has a direct impact on relationships with other constructs. Hence, it may also be essential for IWO psychology practitioners to assess reliability among their job applicant samples in order to compare these against the test publisher reported scale reliabilities. Such independent verification by practitioners will be important to check that scales are performing to the level of internal and test–retest reliability claimed by the commercial test publishers. In this study, we believe the lower levels of observed reliabilities (for OPQ and BPI), compared with those reported by the publishers, affected both our factor analyses and multitrait–multimethod investigations. Again, this could be due in part to sample differences between the original publisher standardization samples and our study sample. Either way, unreliability in measures lowers observed correlations and therefore the lack of convergent validities found in this study can be partially attributed to the lower scale reliabilities observed in this sample. However, low reliabilities of scales cannot explain somewhat higher than anticipated divergent validities. These may point to shared method bias effects.

Convergent validity

For the three entry-level inventories under consideration, the results of this study indicated that the correlations between corresponding Big Five dimension scales were low to negligible. That is, in contrast to our expectations regarding convergent validity, the highest correlations were *not* on the within-category diagonal of the multimethod–multitrait matrix.

Scales from the three different personality inventories tapping the same dimension of the Big Five did not correlate well. Both MTMM and factor analytic work on our data points to the conclusion that, at the primary scale level, it appears that different operationalizations of the Five-Factor Model in the three inventories studied are not psychometrically equivalent. In addition, ratings by an independent sample of study participants of the primary scales onto the FFM (see Tables 1–3) produced intriguing disparities in the number of scales perceived to be loading onto each FFM dimension. This raises questions over the width of coverage (i.e. band-width) and depth of coverage (i.e. fidelity) of the FFM by these three personality inventories. Even though the FFM has become generally accepted amongst personality psychologists as being the most parsimonious representation of the nomological net underlying adult personality, the results of our study would rather suggest that comprehensive band-width and fidelity of coverage cannot be taken for granted in commercially published personality inventories.

Personality measurement and theory construction greatly benefits from the examination of convergent and discriminant validity of personality inventories constructed to measure the same construct domain. This paper has provided one such analysis for three personality inventories widely used in personnel screening and selection in the UK and elsewhere in Europe. Our results have important implications for personality measurement and assessment as well as for using personality inventories in occupational settings. Indeed, the practical implications of multiple-inventory construct validation by independent researchers are varied and numerous. Suffice it to point out here that such studies contribute to professional practice in transparently clear ways, provide practising psychologists with data and results which allow them to examine test properties based upon independent research, and, ultimately, will begin to fuse the major advances in personality psychology over recent years with professional practice in personality testing for employee selection (see e.g. Anderson, Herriot, & Hodgkinson, 2001; Herriot & Anderson, 1997).

The FFM and commercial personality inventories

Putting aside the variability and reliability issues we encountered in this study, it has, now, been known for some time that personality scales purportedly assessing similar (and even the same) personality traits correlate around 0.30–0.40 (observed mean *r*: Ones, unpublished doctoral dissertation; Ones et al., 1995). There is a large specific variance associated with Big Five personality scales. This large specific variance appears to be inventory based. That is, specific personality inventories carry much inventory idiosyncratic variance. From a measurement equivalency perspective, this is problematic. In classical test theory, when the focus is on measuring underlying constructs, specific factor variance is in effect measurement error variance and, clearly, ought to be treated as such. Scale and inventory specific factors cannot be part of the personality trait being measured. An implication is that both test–retest and alpha reliability estimates greatly overestimate the reliability of personality measures. This also causes estimates of correlations between the actual traits that make up the Big Five and other variables to be substantially biased downward (i.e. the correction for the downward bias of measurement error is downwardly biased). In other words, the true score validities that have been reported in the literature for personality dimensions are underestimates (see e.g. Barrick & Mount, 1991; Barrick et al., 2001; Salgado, 1997, 2002).

Reliable measures of the Big Five have, by necessity, to be quite long. We are aware that in many cases personality scales are very long, but we would recommend that they be

longer yet. The internal consistency reliabilities we observed in our sample were uniformly lower than those reported in inventory manuals. This fact added to the issue of large inventory specific variance discussed above points up the importance of having longer measures of personality dimensions. The content validity (i.e. personality trait domain sampling) and therefore construct validity of each personality scale can be improved by having longer measures that sample the trait domain fully and adequately. Conversely, of course, the commercial pressures to market short, quick-to-complete personality inventories may militate against this measurement imperative. Our plea, as researchers in this area, is that such expediency, driven as it undoubtedly is by marketing concerns as opposed to psychometric accuracy, needs to be resisted by professional and responsible test publishers. Fortunately, the three publishers who agreed to take part in the present study did so on the basis of their genuine interest in examining critically their own measures and to obtain useful findings to feed back into their future test development and validation efforts.

There are now many measures that have been constructed to measure personality traits in occupational settings. For the most part, there is some convergence between these measures; however, it is also evident that these various operationalizations differ in the breadth of their various scales (Ones & Viswesvaran, 1996; Ones & Anderson, 2002). Analyses such as those provided in this paper serve to clarify the linkages between different measures and offer insights for future improvements. We wish to make it clear that the results from this study *do not* imply that the measures we examined are in any sense 'poor'. Indeed, such a crude interpretation of our findings would be inaccurate, as each inventory examined here has amassed a large nomological net supporting it. Rather, practically, our findings suggest that when IWO psychology practitioners aim to select/screen job applicants for a particular personality trait, say extraversion, those who are selected will vary depending on (i) which personality inventory is used, (ii) the variability in the sample, and (iii) the reliability of the scales under consideration. As such, this study sounds an important note of caution, which past research using either single-inventory or dual-inventory designs could not feasibly have identified. Indeed, we believe this to be the first multiple-inventory design to have carried out such detailed factor analytic and construct validity analyses using this modified MTMM approach, which, when combined, has offered crucially important insights of both psychometric and pragmatic import.

Limitations and strengths of the present study

There are notably few construct validity investigations that use multiple operationalizations of Big Five constructs based on occupational personality inventories. Thus, the present study aimed to redress this situation by examining the construct validity of three popular personality inventories used in the UK and other European countries for personnel selection and placement. Comparing our findings with the results published in the test manuals, some similar data are evident, but the importance of independent research into this area of concern cannot be overstated. We acknowledge that future research, using large samples, will need to be conducted to examine more fully the construct validity questions we raise in this study, preferably using multi-inventory designs as outlined earlier in this paper. Replication should be an indispensable cornerstone of our science, and in the case of personality inventories used for selection purposes, such replication undoubtedly has major implications for fair and accurate personnel selection. The present study, whilst into an area that has been regrettably under-researched in the past, represents

a first attempt by the authors to examine these important issues empirically in relation to personality tests commonly used in graduate selection. Our findings clearly cannot be interpreted as being definitive or precluding of the need for future research to replicate these findings in terms of both graduate recruitment and recruitment of other levels of job entrants. Further research is therefore called for to replicate and extend the findings of the present study using large sample sizes.

Conclusion

In describing normal, adult personality, the Big Five dimensions have accumulated a massive nomological net, supporting (i) convergent, (ii) divergent, and (iii) criterion-related validity. It is possible to locate specific personality traits as subdimensions or combinations of facets of Big Five dimensions (see appendices provided by Hough & Ones, 2001). In moving forward occupational applications of personality measurement, it would be beneficial for test publishers provide the specific locations of the traits their scales purportedly measure under the Big Five umbrella. Independent research into construct validity questions should be encouraged. We also believe that it will be important for research to begin to uncover cross-cultural effects and the effects of language translations of proprietary personality inventories upon their psychometric properties, construct validity, and criterion-related validity across different cultures. The impact of sample variability, scale unreliability, shared method bias, construct deficiency, and contamination should be systematically assessed. This study provides a constructive first step in that direction for three occupational personality inventories used in the UK and elsewhere in Europe for entry-level selection into organizations. This study also offers methodological guidance on the computation of appropriate factor analytic, MTMM-based, and scale reliability analyses incorporating variance and error variance considerations in multi-inventory construct validation studies. It is thus our hope that it thereby provides a useful, replicable guide for researchers active in this area who are conducting construct validity investigations into personality inventories, or other multi-dimension measures, used for personnel selection.

ACKNOWLEDGEMENTS

An earlier version of this paper was presented at the 1998 annual conference of the American Psychological Association, San Francisco, CA, USA. We wish to thank Jesus Salgado and Chockalingam Viswesvaran for valuable comments on an earlier version of this paper. We also thank Peter H. Dekker and Nathan Kuncel, who provided support and advice on data analysis. We also wish to thank Psychological Consultancy Limited, SHL plc, and ASE for granting permissions to use the personality inventories in this study, and particularly Gillian Hyde at PCL for her assistance in scoring and interpretation of the HPI.

REFERENCES

Anderson, N. (2000). *The future of selection and assessment*. Paper presented to the British Test User Association Conference, London.

Anderson, N., & Herriot, P. (Eds.). (1997). *International handbook of selection and assessment*. Chichester: Wiley.

Anderson, N., Herriot, P., & Hodgkinson, G. P. (2001). The practitioner–researcher divide in Industrial, Work and Organizational (IWO) psychology: Where are we now, and where do we go from here? *Journal of Occupational and Organizational Psychology, 74*, 391–411.

Anderson, N., & West, M. A. (1998). Measuring climate for work group innovation: Development and validation of the team climate inventory. *Journal of Organizational Behavior, 19*, 235–258.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 41*, 1–26.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*(1/2), 9–30.

Bartram, D., Lindley, P. A., Marshall, L., & Foster, J. (1995). The recruitment and selection of young people by small businesses. *Journal of Occupational and Organizational Psychology, 68*, 339–358.

Bond, M. H., Nakazato, H. S., & Shiraishi, D. (1975). Universality and distinctiveness in dimensions of Japanese person perception. *Journal of Cross-Cultural Psychology, 6*, 346–355.

Borkenau, P., & Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: A study of the 5-factor model of personality. *Personality and Individual Differences, 11*, 515–524.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 140–161.

Cellar, D. F., Miller, M. L., Doverspike, D. D., & Klawsky, J. D. (1996). Comparison of factor structures and criterion-related validity coefficients for two measures of personality based on the five factor model. *Journal of Applied Psychology, 81*, 694–704.

Comrey, A. (1978). Common methodological problems in factor analytic studies. *Journal of Clinical and Consulting Psychology, 46*, 648–659.

Conley, J. J. (1985). A personality theory of adulthood and aging. In R. Hogan, & W. H. Jones (Eds.), *Perspectives in Personality* (Vol. 1, pp. 81–115).

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

Costa, P. T., & McCrae, R. R. (1988). From catalog to classification: Murray's needs and the five-factor model. *Journal of Personality and Social Psychology, 55*, 258–265.

Digman, J. M. (1990). Personality structure: Emergence of the Five Factor Model. *Annual Review of Psychology, 41*, 417–440.

Digman, J. M. (1997). Higher order factors of the Big Five. *Journal of Personality and Social Psychology, 73*, 1246–1256.

Digman, J. M., & Inouye, J. (1986). Further specification of the five robust factors of personality. *Annual Review of Personality and Social Psychology, 50*, 116–123.

Digman, J. M., & Takemoto-Chock, N. K. (1981). Factors in the natural language of personality: Re-analysis and comparison of six major studies. *Multivariate Behavioral Research, 16*, 146–170.

Feltham, R., & Woods, J. (1995). *Manual and User Guide for the Business Personality Indicator.* Windsor: ASE.

Ferguson, E., Payne, T., & Anderson, N. (1994). Occupational personality assessment: An evaluation of the psychometric properties of the Occupational Personality Questionnaire (OPQ). *Personality and Individual Differences, 17*, 217–225.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology, 44*, 329–344.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378–382.

Furnham, A. (1994). A content, correlational and factor analytic study of four tolerance of ambiguity questionnaires. *Personality and Individual Differences, 16*, 403–410.

Furnham, A. (1996). The FIRO-B, the Learning Style Questionnaire, and the Five Factor Model. *Journal of Social Behavior and Personality, 11*, 285–299.

Goldberg, L. R. (1990). An alternate 'description of personality', the Big-Five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216–1229.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*, 26–34.
Gorsuch, R. (1983). *Factor Analysis*. London: Erlbaum.
Herriot, P., & Anderson, N. (1997). Selecting for change: How will personnel and selection psychology survive? In N. Anderson, & P. Herriot (Eds.), *International handbook of selection and assessment* (pp. 1–34). Chichester: Wiley.
Hodgkinson, G. P., & Payne, R. L. (1998). Graduate selection in three European countries. *Journal of Occupational and Organizational Psychology, 71*, 359–365.
Hogan, R., & Hogan, J. (1995). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.
Hogan, R. T., & Roberts, B. W. (2001). Introduction: Personality and industrial and organizational psychology. In B. W. Roberts, & R. T. Hogan (Eds.), *Personality psychology in the workplace* (pp. 3–18). Washington, DC: American Psychological Association.
Hough, L. M. (2001). I/Owes its advances to personality. In B. W. Roberts, & R. T. Hogan (Eds.), *Personality psychology in the workplace* (pp. 19–44). Washington, DC: American Psychological Association.
Hough, L. M., & Ones, D. S. (2001). The structure, measurement, validity, and use of personality variables in industrial, work, and organizational psychology. In N. Anderson, D. S. Ones, H. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work, and organizational psychology* (Vol. 1, pp. 233–277). London: Sage.
Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152–194.
Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. London: Methuen.
Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312–320.
Lorr, M., & Youniss, R. P. (1973). An inventory of interpersonal style. *Journal of Personality Assessment, 37*, 165–173.
Matthews, G., Stanton, N., Graham, N. C., & Brimelow, C. (1990). A factor analysis of the scales of the Occupational Personality Questionnaire. *Personality & Individual Differences, 11*(6), 591–596.
McCrae, R. R. (2001). Trait psychology and culture: Exploring intercultural comparisons. *Journal of Personality, 69*, 819–846.
McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81–90.
McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*(5), 509–516.
McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality, 60*, 175–216.
Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66*, 574–582.
Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology, 4*, 681–691.
Ones, D. S. (1993). *The Construct Validity Evidence for Integrity Tests*, copyrighted doctoral dissertation, Iowa City, Iowa: University of Iowa.
Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology, 75*(3), 255–276.
Ones, D. S., Schmidt, F. L., & Viswesvaran, C. (1995, August). *Convergence among three personality inventories measuring the Big Five*. Paper presented at the 103rd annual meeting of the American Psychological Association, New York.
Ones, D. S., & Viswesvaran, C. (1996). Bandwidth–fidelity dilemma in personality measurement for personnel selection. *Journal of Organizational Behavior, 17*, 209–226.
Ones, D. S., & Viswesvaran, C. (1998). Gender, age and race differences on overt integrity tests: Analyses across four large-scale applicant data sets. *Journal of Applied Psychology, 83*, 35–42.

Ones, D. S., & Viswesvaran, C. (2001). Integrity tests and other criterion-focused occupational personality scales (COPS) used in personnel selection. *International Journal of Selection and Assessment, 9*, 31–39.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology (Monograph), 78*, 679–703.

Robertson, I. T., & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational and Organizational Psychology, 66*, 225–244.

Rothmann, S., Meiring, D., & Barrick, M. R. (2002). *Predicting job performance using personality measures in South Africa: A meta-analytic study.* Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Toronto.

Ryan, A. M., McFarland, L., Baron, H., & Page, R. (1999). An international look at selection practices: Nation and culture as explanations for variability in practice. *Personnel Psychology, 52*, 359–392.

Salgado, J. F. (1996). Personality and job competences: A comment on the Robertson and Kinder (1993) study. *Journal of Occupational and Organizational Psychology, 69*(4), 346–351.

Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology, 82*, 30–43.

Salgado, J. F. (2002). The Big Five personality dimensions and counterproductive behaviors. *International Journal of Selection and Assessment, 10*, 117–125.

Salgado, J. F., & Anderson, N. (2002). Cognitive and GMA testing in the European Community: Issues and evidence. *Human Performance, 15*, 75–96.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods and techniques. In N. Anderson, D. S. Ones, H. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology* (Vol. I). (pp. 165–199). London: Sage.

Shackleton, V. J., & Newell, S. (1994). European management selection methods: A comparison of five countries. *International Journal of Selection and Assessment, 2*, 91–102.

SHL. (1999). *OPQ32 Manual and User's Guide*. Surrey: SHL.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Thompson, B., & Borrello, G. M. (1986). Second-order factor structure of the MBTI: A construct validity assessment. *Measurement and Evaluation in Counselling and Development, January*, 148–153.

Watson, D. (1989). Strangers' ratings of the five robust personality factors: Evidence of a surprising convergence with self-report. *Journal of Personality and Social Psychology, 57*, 120–128.

Yang, K., & Bond, M. H. (1990). Exploring implicit personality theories with indigenous or imported constructs: The Chinese case. *Journal of Personality and Social Psychology, 58*, 1087–1095.

Yoo, T. Y., & Min, B. M. (2002). *A meta-analysis of the Big Five and performance in Korea*. Paper presented at the annual conference of the Society for Industrial and Organizational Psychology, Toronto.

