16pf®

Technical Manual

Sixth Edition

Core Personality Insights

powered by psi

Table of Contents

List of Tables and Figures

Acknowledgments

The Sixteen Personality Factor Questionnaire (16pf) holds a special place in modern personality measurement as a foundational instrument in the development of assessment methods in place today. Raymond B. Cattell developed a personality assessment approach that has endured for several decades as a valid and reliable means of assessing human characteristics that are important in the workplace as well as educational and vocational guidance settings. Since the time of Cattell's original work, the 16pf has been refined with the development of several new editions, each one introducing new enhancements and features. At the same time, the essence of the 16pf has remained consistent, building upon the many research and validation studies conducted over the years that have made it one of the most widely researched assessment tools.

This manual describes the technical features and empirical data supporting the Sixth Edition of the 16pf. In keeping with prior editions, the Sixth Edition offers a number of important enhancements and refinements while maintaining consistency and equivalence in measurement. Chapter 4 of the manual describes many of these enhancements.

Similar to prior editions of the 16pf, the development of the Sixth Edition was a collaborative effort involving many psychologists, psychometricians, and assessment consultants representing a diverse range of experience and deep expertise. Acknowledgement is given to the following individuals who contributed to the Sixth Edition revision project. The research and development team included lead researcher on the revision project, Alan Mead, Ph.D., supported by research team members: Joseph Abraham, Ph.D., Dawn Lambert, Ph.D., Ralph Mortensen, Ph.D., Michael Stowers, Psy.D., Scott Stubenrauch, Psy.D., Pamela Becker, M.A., and Amie Lawrence, Ph.D.; John Weiner served as executive sponsor and advisor. Additional acknowledgement is given to the following individuals for item review and/or item writing: Angelina Bennet, Ph.D., Jason Blaik, M.A., Elke Chrystal, M.A., Doug Craig, Psy.D., Anna Crollick, M.S., Craig Gilles, Anne Hennessy, Alan Mead, Ph.D., Ralph Mortensen, Ph.D., Julianna Otremba, M.A., Shefali Sharma, M.A., Kelsey Stephens, M.A., Michael Stowers, Psy.D., Scott Stubenrauch, Psy.D., Nicola Taylor, Ph.D., and Chenxuan, Zhou, Ph.D.. The manual was reviewed by an ad hoc committee, including Paul Sackett, Ph.D., Sheldon Zedeck, Ph.D., and Rick Jacobs, Ph.D..

Lastly, the team recognizes the groundbreaking work of Raymond B. Cattell and the many contributors listed in Appendix A who supported the continued development of the 16pf in subsequent editions leading up to this Sixth Edition.

Chapter 1: Introduction to the 16pf Questionnaire

This chapter briefly introduces the 16pf Questionnaire, including its origins, unique features, and applications.

The remaining chapters provide detailed information, with practical chapters first, followed by more technical chapters. Chapter 2 covers the administration of the 16pf Questionnaire. Chapter 3 covers the interpretation of 16pf scores. Chapters 4 through 6 describe the development of the Sixth Edition. Chapter 7 describes the norm sample, and Chapters 8 through 10 summarize reliability and validity studies. Thus, this manual provides both introductory and detailed technical information including empirical evidence for 16pf users and other interested parties.

Foundations of the 16pf Questionnaire

When the 16pf Questionnaire was originally published in 1949, it was the first test based on systematic, scientific research into the basic dimensions of human personality (Goldberg, 1993). Dissatisfied with the approach of choosing a group of traits a priori and then constructing an inventory to measure them, Raymond B. Cattell (1945) set out to discover the fundamental building blocks of personality using factor analysis. R. B. Cattell convincingly argued the factor-analytically discovered personality factors to be the basic elements of personality.

Factor analysis was a new and laborious technique when R. B. Cattell began applying it in the first half of the 20th century (R. B. Cattell, 1952). Although involving a fair amount of mathematical and statistical sophistication, the basic idea is simple. Self-descriptive words like “inquisitive” and “curious” seem to represent one concept and “fastidious” and “exacting” both seem to represent a second concept, so these four words represent only two concepts and can be mathematically represented by a two-dimensional space. If we considered thousands of self-descriptive words and phrases, how many concepts would be needed to represent this space? This was the overarching question with which R. B. Cattell launched his program of research using factor analysis. It allowed him to estimate the smaller number of unseen (latent) factors that underlay ratings of self-descriptive words and phrases. Basically, factor analysis allows us to “reduce data” by organizing similar information into categories that contain related ideas but differ from the ideas in other categories. We can describe what we see around us in terms of specific colors—azure, royal, and baby or rose, ruby, and cranberry—but we can also simply say blue or red, and convey accurate information. Factor analysis is an empirical way of creating those categories.

R. B. Cattell and his colleagues reasoned that adjectives relating to personality had to correspond to English-language adjectives commonly used to describe people. Therefore, R. B. Cattell began by systematically analyzing the entire range of personality trait descriptors present in the English language, beginning with Allport and Odbert's (1936) 17,953 trait words. Initially, R. B. Cattell and his colleagues asked observers to rate subjects well known to them on the basis of a subset of adjectives, which had been reduced to eliminate similar terms in the Allport and Odbert set. The researchers then subjected the observers' ratings to factor analysis. R. B. Cattell performed this factor analysis with the intent of identifying the "primary" personality traits, or those that could explain the entire personality domain, just as the chemical elements are considered the primary building blocks of all matter and blue, red, and yellow are considered the primary colors.

Factor analyses of the observers' ratings data, termed "Life-data" or "L-data," identified 12 traits that could account for the range of descriptors in the trait lexicon. These traits, called "factors," were named using letters of the alphabet, such as Factors A, B and C, similar to the periodic table in the elements in chemistry. (Within the alphabetical listing of factor names, some letters are skipped over. Factors corresponding to these skipped letters were found in parallel studies of child and adolescent personality but were not found in descriptions of adults.) The adjectives rated for the factors were translated into multiple-choice questionnaire items and were termed "Questionnaire-data" or "Q-data." In a series of studies, responses to the questionnaire items were factor analyzed, and the resultant data were used in constructing the 16 primary scales of the 16pf instrument. Twelve of the scales measure the factors labeled alphabetically and originally identified through analyses of the L-data. The remaining four scales measure factors labeled Q1, Q2, Q3, and Q4 because they originated from analyses of the Q-data.

Through a long series of factor-analytic studies of the behavior ratings and questionnaire data, R. B. Cattell (1946) reduced the myriad of descriptors to 16 basic underlying dimensions that held together as unitary traits—the primary factors of the 16pf instrument. More than 50 published studies have replicated the basic structure of these 16 traits (R. B. Cattell & Krug, 1986). Cattell also developed separate measures of the traits for the early age ranges and provided a detailed analysis of how the traits change and develop throughout the life span (R. B. Cattell, 1979, 1980).

Correspondence to the Big Five Model

In addition to the 16 "primary factors," R. B. Cattell identified five more general human characteristics, which he called "second order" personality traits (R. B. Cattell, 1957) because they were found by factoring the primary scales. His discovery of these "global" factors was the first identification of what would later be called the "Big Five" personality dimensions: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (Costa & McCrae, 1995; Goldberg, 1993).

Although the labels assigned to five factor models have varied, a common way of describing them in general is with the OCEAN acronym:

- Openness (O) reflects a tendency to view new and different experiences with interest and excitement.
- Conscientiousness (C) is a personal orientation towards acting with a sense of duty and the desire to be seen by others as dependable.
- Extraversion (E) characterizes people who consistently gravitate to forming close relationships with other people and gain energy from social interactions.
- Agreeableness (A) is related to forming cooperative social relationships and deferring to others' wishes and ideas.
- Neuroticism (N; reflected) indicates the degree to which individuals are generally emotionally stable and confident.

Five-factor models have become an increasingly popular way for practitioners to consider personality. Even for scales with specific factors, such as the 16pf Questionnaire, Neurotic, Extraversion, Openness Personality Inventory (NEO PI), and Hogan Personality Inventory (HPI), the Big Five provides an organizing principle that makes the complexity of the specific factors easier to understand.

Heather E. P. Cattell (1996) examined how the 16pf global factor model corresponds to the five-factor model of the NEO Personality Inventory (Costa & McCrae, 1985). She administered both instruments to a sample (N=624) and factored the resulting responses. Table 1.1 summarizes her findings.

Table 1.1 Comparison of “Big Five” and Facets for the NEO and 16pf Questionnaire

High level factor	NEO Personality Inventory	16pf Questionnaire
Openness	*Openness*	*Low Tough-Mindedness*
(*r* = 0.60)	O1: Fantasy	A: Warmth
	O2: Aesthetics	I: Sensitivity
	O3: Feelings	M: Abstractedness
	O4: Actions	Q1: Openness to Change
	O5: Ideas	
	O6: Values	
Conscientiousness	*Conscientiousness*	*Self-Control*
(*r* = 0.67)	C1: Competence	F: Liveliness (low)
	C2: Order	G: Rule Consciousness
	C3: Dutifulness	M: Abstractedness (low)
	C4: Need/Achievement	Q3: Perfectionism
	C5: Self-Disclosure	
	C6: Deliberation	
Extraversion	*Extraversion*	*Extraversion*
(*r* = 0.67)	E1: Warmth	A: Warmth
	E2: Gregariousness	F: Liveliness
	E3: Assertiveness	H: Social Boldness
	E4: Activity	N: Privateness (low)
	E5: Excitement Seeking	Q2: Self-Reliance (low)
	E6: Positive Emotion	
Agreeableness	*Agreeableness*	*Low Independence*
(*r* = 0.30)	A1: Trust	E: Dominance (low)
	A2: Straightforward	H: Social Boldness (low)
	A3: Altruism	L: Vigilance (low)
	A4: Compliance	Q1: Openness to Change (low)
	A5: Modesty	
	A6: Tender-Mindedness	
Neuroticism	*Neuroticism*	*Anxiety*
(*r* = 0.72)	N1: Anxiety	C: Emotional Stability (low)
	N2: Angry Hostility	L: Vigilance
	N3: Depression	O: Apprehension
	N4: Self-Conscious	Q4: Tension
	N5: Impulsiveness	
	N6: Vulnerability	

Note: 16pf Questionnaire primary factors labeled “(low)” indicate that low scores contribute to their global factor. For example, high scores on Liveliness/F contribute to lower Conscientiousness scores. *r* was the correlation between the global dimension from the two questionnaires. Reproduced from H. E. P. Cattell (1996).

As shown in Table 1.1, correlations between the factors of the NEO five-factor model and the corresponding five 16pf global factors ranged from 0.30 (between Agreeableness and Low Independence) to 0.72 (between Neuroticism and Anxiety).

Overall, this indicates a good degree of alignment between these two perspectives. However, examination of the 16pf primary personality factors compared to the lower level personality facets of the NEO highlighted important differences in how they are defined and measured. For example, whereas the NEO Agreeableness factor and the 16pf Independence factor (reflected/low Independence) both measured aspects of this trait, the 16pf places greater emphasis on being a dominant, forceful person as well as open-mindedness and fewer neurotic tendencies such as anxiety, depression and vulnerability. Similarly, Openness as measured by the NEO contained other cognitive traits, for instance openness to ideas, imagination, and values, whereas lower 16pf Tough-Mindedness (reflected/low Tough-Mindedness) also incorporated being receptive to other people and to new experiences.

Other researchers such as Goldberg (1992) and Hogan and Hogan (1992) have proposed their own schema for capturing broader personality dimensions of interest. In general, each author has made either their own theoretical assumptions about the nature of personality or applied unique decision rules to deducing what their research indicates about its structure (Block, 1995; H. E. P. Cattell, 1996). For example, Hogan and Hogan (1992) have added factors by dividing Openness into the two factors of Inquisitive and Learning Approach. Test authors also have chosen different routes to the construction of individual test items such as only gathering self-assessments or the use of true/false responses.

Although it cannot be said that a single, definitive model of high-level personality dimensions exists, the various Big FIve frameworks are useful tools to guide research into commonly seen behavior patterns. Users are cautioned, though, to carefully examine the details before reaching the conclusion that profiles obtained from different inventories are truly identical. The average correlation of .59 between NEO factors and corresponding broad 16PF factors is far short of the value that would be needed to view these as alternate measures of the same underlying constructs.

These five broad personality characteristics are consistent predictors of many behaviors and outcomes such as academic success (Trapmann, Hell, Hirn, & Schuler, 2007), job performance (e.g., Barrick & Mount, 1991; Hurtz & Donovan, 2000; Salgado, 1997), leadership effectiveness (Judge, Bono, Ilies, & Gerhardt, 2002) and career interests (Barrick, Mount, & Gupta, 2003). However, careful examination of narrower traits such as the 16pf primary factors can provide a more nuanced and precise picture of the individual test taker (e.g.Ashton, Paunonen, & Lee, 2013; Hermann, 2009).

Applications of the 16pf Questionnaire

Since its inception, the 16pf Questionnaire has provided psychologists, trained human resources managers, educators, consultants, and public safety and security professionals with valuable insights into the personality profiles of their clients, job candidates, employees, and students. The following are a few examples of how individuals in each occupation have leveraged the rich, in-depth information offered by the 16pf Questionnaire. (An in-depth guide to interpreting the global and primary personality factors can be found in Chapter 3.)

Human resources professionals and consultants can gain a greater appreciation of both job candidates and employees. Along with employment interviews, work sample assessments, application forms, and background checks, 16pf results will inform judgments about applicants' personality- and competency-related job fit. After selection decisions are made, 16pf profiles can offer useful ideas about how to best provide new employee onboarding experiences. Knowledge of personality patterns also is helpful when weighing employee development tactics and the most promising career directions. A key employee's or leader's personality profile can be a valuable baseline for providing relevant job performance coaching. At a higher level, collecting and feeding back the collective profile of work teams can facilitate member and leader appreciation of individual differences, likely areas of compatibility and conflict, and shared strengths and weaknesses or potential blind spots.

Psychologists and counselors will find that their respondents' 16pf profiles provide them with greater understanding of individuals' unique social styles, emotional dynamics, ways of viewing the world, motives, and thinking patterns. These nuanced portraits enrich respondent self-insight, facilitate dialogue, and inform practitioners about the best approaches to counseling and therapy. Reviewing the 16pf profiles of couples also assists psychologists with understanding how partners are likely to see each other, communicate, and react emotionally when issues arise in a relationship or marriage.

Educators will find that 16pf findings are particularly useful for both student career advice and personal counseling. The scores can be combined with other student information to pinpoint individuals' likely career interests and contribute to discussions of the most promising degree specialties or academic focus and career choices. Teachers and professors will find that appreciation of their distinctive personality traits will allow students to better grasp their own problem solving and decision making styles, study habits and interests, relationships with classmates and faculty, and responses to stress and pressure. Finally, understanding of their personal 16pf profiles can assist faculty with understanding their own special strengths and developmental needs.

Public safety and security professionals will discover that the 16pf Questionnaire's personality insights are essential intelligence when screening candidates for dangerous work and high-risk assignments. Tailored 16pf reports present both standard personality profiles as well as new windows into the personality dimensions specifically related to success in these important and difficult roles. Users with advanced mental health credentials can have access to all this information and also supplemental, key indicators of possible clinical issues that could affect job fit and performance. The findings of 16pf questionnaires have aided users in local, state, national, and international law enforcement and security roles as well as in private security organizations and government contractors.

Next Steps

Users of the 16pf Questionnaire are encouraged to study this manual carefully. Doing so will take advantage of its many resources as they explore the workings of the inventory and its multiple personality dimensions, the different report options available and learn the most useful ways to put these powerful personality insights to use. If needed, additional professional assistance and training is available from PSI Services LLC via www.16pf.com.

References

Allport, G. W., & Odbert, H. S. (1936). Trait-names, a psycholexical study. *Psychological Monographs, 47*, 171.

Ashton, M. A., Paunonen, S. V. & Lee, K. (2013). On the validity of narrow and broad personality traits: A response to Salgado, Moscoso and Berges (2013). *Personality and Individual Differences*, *56*, 24-28.

Barrick, M. R., & Mount, P. K. (1991). The Big FIve personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*, 1-26.

Barrick, M. R., Mount, P. K., & Gupta, R. (2003). Meta-analysis of the relationship between the five-factor model of personality and Holland's occupational types. *Personnel Psychology*, *56*, 45-74.

Block, J. (1995). A contrarian view of the five factor approach to personality description. *Psychological Bulletin, 2,* 187-215.

Cattell, H. E. P. (1996). The original Big Five: A historical perspective. *European Review of Applied Psychology, 46,* 5-14.

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

Cattell, R. B. (1946). *The description and measurement of personality*. New York, NY: Harcourt, Brace & World.

Cattell, R. B. (1952). *Factor analysis*. New York, NY: Harper and Brothers.

Cattell, R. B. (1957). *Personality and motivation structure and measurement*. New York, NY: World Book.

Cattell, R. B. (1979). *Personality and learning theory: The structure of personality in its environment, Vol. 1*. New York, NY: Springer.

Cattell, R. B. (1980). *Personality and learning theory: A systems theory of maturation and structured learning, Vol. 2*. New York, NY: Springer.

Cattell, R. B., & Krug, S. E. (1986). The number of factors in the 16pf: A review of the evidence with special emphasis on methodological problems. *Educational and Psychological Measurement, 46*, 509–522.

Costa, P. T. Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment, 64*, 21-50.

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

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

Hermann, A. (2009). *Using broad or narrow personality measures to predict leadership success: Does keeping it simple have an impact on predictive power and utility*? [White paper]. Retrieved from https://www.16pf.com/wp-content/uploads/Leadership-Success-Broad-vs-Narrow-Measures-WHITE-PAPER.pdf

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology*, *85*, 869-879.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*, *87*, 765-780.

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

Trapmann, S., Hell, B., Hirn, J. W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Journal of Psychology*, *215*, 132-151.

Chapter 2: Administration and Scoring

Introduction

This chapter presents information on administering and scoring the 16pf Sixth Edition Questionnaire. The questionnaire is designed to be administered to individuals aged 16 and older. The questionnaire is administered online and has an overall readability estimated at the sixth-grade level (roughly ages 11 to 12). Users with a need for alternative administration methods should contact PSI Services LLC via www.16pf.com.

The 16pf questionnaire was normed on a sample with ages from 16 to over 75 years of age. Whether the questionnaire is appropriate for an individual younger than 16 is a decision that should be based on professional consideration of the respondent's maturity level.

Proctored Versus Unproctored 16pf Assessment

Proctoring refers to an assessment professional supervising and observing the completion of an assessment. The 16pf Questionnaire may be administered with or without proctoring, but the 16pf user must factor the administration context into interpretation of the results. For example, responses to unproctored assessments may not reflect the intended respondent (someone else might have completed the 16pf Questionnaire), may include the assistance of others, or may reflect the unauthorized use of assessment aids on the Reasoning/B items. Regardless of the type of proctoring, respondents should be provided with a means by which they may ask questions or seek assistance with the assessment.

Although caution is advised, there is evidence that scores across proctored and unproctored assessments may be interchangeable. Jones and Newhouse (2005) reported no differences in personality scores between test takers who completed the 16pf in proctored versus unproctored online administration. In mental health settings, reviews generally have indicated high levels of user satisfaction and acceptance of remotely provided assessments and services (Luxton, Pruitt, & Osenbach, 2014).

Users should avoid comparing candidates assessed using proctored and those assessed without a proctor unless evidence is available showing that these scores can be compared.

Preparing for Assessment

Although the 16pf Sixth Edition questionnaire can be completed without proctoring, the administrator is advised to take time to establish a comfortable rapport with respondents since the creation of a favorable attitude toward the questionnaire is important to facilitate receiving accurate assessment data. With this in mind, the administrator should give thoughtful attention to respondents' questions and should reinforce the assessment objectives by explaining to respondents that, in the long run, the results will be most accurate by being frank and honest in their self-descriptions.

Response Format

The personality statements have a five-choice Likert agreement response format: strongly disagree, disagree, neutral, agree, and strongly agree. The instructions encourage respondents to decide whether they agree or disagree with a statement and then decided on the strength of that agreement or disagreement. They are instructed to choose the middle response if they cannot agree or disagree with a statement. (A Likert format is a change from previous editions; see Chapter 4 for complete information.)

The Reasoning/B scale items each have three multiple-choice options with one correct response (a small number of items require the respondent to enter a numeric value). The Reasoning scale is administered after the personality items and uses an adaptive administration format with a variable length stopping rule that administers at least 10 items and no more than 20 items. Details of the adaptive administration format are explained in Chapter 5 (although most users will prefer the adaptive format because it is more efficient, a user with a need for a fixed form of the scale may inquire about the availability of 20-item fixed forms). The questions are designed to be answered without the aid of outside devices and respondents are advised to avoid use of such aids (calculators, dictionaries, Internet searches, etc.).

Scoring and Reporting

The 16pf questionnaire is automatically scored by the PSI True Talent platform once the respondent completes all the questions. Assessment reports are available online through the 16pf administrator account. In the unlikely event that a 16pf user is unable to access their electronic copy of a report, or has other customer support issues, assistance can be obtained by contacting PSI Services LLC via www.16pf.com.

Answers to the 16pf Sixth Edition Questionnaire can be used to generate a variety of reports for different applications, including employee selection, development, and promotion; public safety and protective services assessment; as well as clinical or counseling applications in which a broad assessment of normal adult personality is needed. If multiple reports are needed for a single respondent, there is no need to answer the questionnaire again. The additional reports can be ordered from PSI Services LLC using the person's original profile. Please reference www.16pf.com for information on available reports.

Administration and Completion Time

The assessment is untimed, but respondents should be encouraged to work at a steady pace. The administrator may want to discourage respondents from agonizing over possible responses by reiterating this caution included in the assessment directions: "Remember, don't spend too much time thinking over any one question. Give the first, natural answer as it comes to you." Average assessment completion time is 20-30 minutes.

The 16pf Sixth Edition can be administered from any device with reasonably fast Internet access. The administration system, PSI True Talent, may offer respondents a choice of language and then administer instructions and the questions. The Likert personality items are presented in a matrix format with several items on one screen, and respondents may change their previous responses; on the adaptive Reasoning/B scale, items are administered individually, and previous responses cannot be viewed or changed. It is extremely important that respondents complete the items and then also click the final button labeled "finish" (in English). This is the trigger for the PSI True Talent system to immediately score the assessment and make reports available after the assessment is complete. Additional information regarding PSI True Talent can be obtained by contacting PSI Services LLC via https://www.16pf.com.

Research has demonstrated that scores obtained from computerized administration are typically equivalent to scores obtained via paper-and-pencil administration for untimed assessments (Mead & Drasgow, 1993) and that the measurement qualities of traditional personality scales are comparable across web-based devices (e.g., laptop, smart phone, tablet; for a review, see Dadey, Lyons, & DePascale, 2018). Users should be cautious about interpreting Factor B/Reasoning scores obtained on different kinds of devices, because some research suggests that assessments of cognitive ability are harder when completed on devices with small screens (Schroeders & Wilhelm, 2010). Users may avoid this issue by requiring respondents to use large-screen devices such as desktops, laptops, and large tablets.

When should a respondent be retested?

Personality traits should be relatively stable over time (Ferguson, 2010; Terracciano, Costa, & McCrae, 2006). However, the possibility of fluctuations in scores always exists, particularly due to the individual's current psychological state (H. E. P. Cattell & Schuerger, 2003) or due to maturation. Under normal circumstances, the recommendation is that a respondent be retested after a 6-month to 1-year period. If the respondent has experienced a major life event that could be expected to influence his or her psychological state, retesting after a shorter time interval is strongly advised.

Is it possible to stop answering the questionnaire and restart it later?

If the respondent is interrupted, the assessment software automatically records the completed answers. The individual can resume answering the remaining questions once available again. However, we recommend that respondents reserve the necessary time to maintain their concentration and minimize the inconvenience of signing on to the web site multiple times.

Why don't the Reasoning items have numbers?

The Reasoning items are administered adaptively and can vary from 10 to 20 items. As a result, they are not numbered.

Why does the progress bar show 100% during Reasoning?

The Reasoning items are administered adaptively and can vary from 10 to 20 items. As a result, the platform displays the progress bar as 100% complete. Most respondents do not notice this, and it is not a problem for the assessment.

My respondent finished but the assessment is still shown as "in progress"

Typically, this is because the respondent answered all the items but failed to click the "Finish" button at the very end. As a result, the system keeps the record active and does not score the record. This can be resolved by having the candidate log back into the system and clicking the button or administratively through the administrator's interface. Please contact technical support for additional help with this issue.

References

Cattell, H. E. P., & Schuerger, J. M. (2003). *Essentials of 16pf assessment*. New York, NY: John Wiley & Sons.

Dadey, N., Lyons, S., & DePascale, C. (2018) The comparability of scores from different digital devices: A literature review and synthesis with recommendations for Practice. *Applied Measurement in Education, 3*1, 30-50.

Ferguson, C. J. (2010). A meta-analysis of normal and disordered personality across the life span. *Journal of Personality and Social Psychology, 98*, 659-667.

Jones, J. W., & Newhouse, N. (2005). *Unproctored online testing programs: Managing risks to benefit from a flexible and low-cost assessment model*. Working paper, Institute for Personality and Ability Testing.

Luxton, D. D., Pruitt, L. D., & Osenbach, J. E. (2014). Best practices for remote psychological assessment via telehealth technologies. *Professional Psychology: Research and Practice*, *45*, 27-35.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458.

Schroeders, U. & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment, 26*, 284-292.

Terracciano, A., Costa, P. T. Jr., & McCrae, R. R. (2006). Personality plasticity after age 30. *Personality and Social Psychology Bulletin, 32*, 999-1009.

Chapter 3: Score Interpretation

Introduction

Intended as a user-friendly guide for interpreting 16pf Sixth Edition results, this chapter provides general interpretive information, a profile interpretation strategy, and specific scale descriptions. The content synthesizes findings from a number of different studies described in Chapters 8 and 9 of this manual.

General Interpretive Information

The interpretive information that follows is based on the current body of evidence available for the Sixth Edition. As users continue to develop a database on the latest edition, interpretation guidance will be refined to reflect the incoming data.

Factor Analysis

This chapter focuses on interpreting the 16pf scores, which are the result of R. B. Cattell's use of the factor-analytic approach in identifying the basic structure of human personality (see Chapter 1 for a brief overview of factor analysis). Users of the 16pf are well-advised to familiarize themselves with the technical and methodological details of the instrument. Chapters 4 to 6 provide details on how factor analytic and related methods were used to develop the current edition of the questionnaire. Chapters 7 through 10 describe reliability and validity evidence. Chapter 11 describes research and practice applying the 16pf scores to the study of organizational leadership.

Primary Factor Scales

Historically, the basic scales of the 16pf Questionnaire have been labeled with letters (e.g., Factor A, Factor B, etc., through Factor Q4). The Sixth Edition continues the tradition of using factor letters and also provides "common-language" names for each scale (see Table 3.1).

Table 3.1 Primary Factor Scale Descriptors

Factor		Left-side meaning	Right-side meaning
A	Warmth	Reserved, impersonal, distant	Warm, outgoing, attentive to others
B	Reasoning	Lower general mental capacity, less intelligent, concrete thinking	Higher general mental capacity, more intelligent, bright, analytical
C	Emotional Stability	Reactive, emotionally changeable	Emotionally stable, adaptive, mature
E	Dominance	Deferential, cooperative, avoids conflict	Dominant, forceful, assertive
F	Liveliness	Serious, restrained, careful	Lively, animated, spontaneous
G	Rule-Consciousness	Expedient, nonconforming	Rule conscious, dutiful
H	Social Boldness	Shy, threat sensitive, timid	Socially bold, venturesome, thick skinned
I	Sensitivity	Utilitarian, objective, unsentimental	Sensitive, aesthetic, sentimental
L	Vigilance	Trusting, unsuspecting, accepting	Vigilant, suspicious, skeptical, wary
M	Abstractedness	Grounded, practical, solution oriented	Abstracted, imaginative, idea oriented
N	Privateness	Forthright, genuine, artless	Private, discreet, nondisclosing
O	Apprehension	Self-assured, unworried, complacent	Apprehensive, self-doubting, worried
Q1	Openness to Change	Traditional, attached to familiar	Open to change, experimenting
Q2	Self-Reliance	Group oriented, affiliative	Self-reliant, solitary, individualistic
Q3	Perfectionism	Tolerates disorder, unexacting, flexible	Perfectionistic, organized, self-disciplined
Q4	Tension	Relaxed, placid, patient	Tense, high energy, impatient, driven

Bipolar Scales

As shown in Table 3.1, the 16pf scales are bipolar in nature; that is, both high and low scores have meaning. Generally, professionals should not assume that high scores are "good" and that low scores are "bad." For example, high scorers on Factor A tend to be warm interpersonally, whereas low scorers tend to be more reserved interpersonally. In some situations, being reserved might be quite fitting or useful. In other situations, being warm might be more suitable.

Throughout this chapter, the right-side pole, or high-score range, of a factor is described as the plus (+) pole. The left-side pole, or low-score range, is the minus (-) pole. For example, high scorers on Factor A are described as Warm (A+); low-scorers are described as Reserved (A-).

Usually, the correlation of one 16pf scale with another is framed in terms of the positive correlation. For example, Warmth (A+) is positively correlated with the Extraversion Global Factor. That is, being high on Warmth (A+) contributes to being high on Extraversion. On the other hand, Sensitivity (I+) is negatively correlated with the Tough-Mindedness Global Factor; that is, being high on Sensitivity (I+) contributes to being low on Tough-Mindedness. Thus, Sensitivity (I+) could be said to be negatively correlated with Tough-Mindedness or positively correlated with Receptivity, the minus pole of Tough-Mindedness. In most such cases (negative poles relating to one another), the correlation is described in the positive manner (e.g., being Sensitive [I+] contributes to Receptivity).

Global Factors

In addition to the primary scales, the 16pf instrument contains a set of five scales that combine related primary scales into Global Factors of personality. (The Global Factors have historically been called "second-order factors" in 16pf literature and result from a factor analysis of the test's primary scales.) Table 3.2 lists the Global Factors and gives brief descriptors of each factor pole. As described in Chapter 1, these Global Factors correspond in many ways to the Big Five model of personality structure. Each of the descriptions of the Global Factors below includes a comparison to the similar Big Five factor.

Table 3.2 Global Factor Scale Descriptors

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

Sten Scales

The 16pf instrument uses "standardized ten" (sten) score scales. Sten scores are usually calculated in a norm sample (see Chapter 7 for more details on Sten scores and Figure 7.1 in that chapter for a visual of sten scores), which represents the population of test takers for whom the test is intended. Sten scores range from 1–10, with a mean of 5.5 and a standard deviation of 2. Scores that fall farther from the mean (either in the high or the low direction) are considered more extreme. The more extreme a score is toward a given factor pole, the more likely that the descriptors for the scale's pole will apply for that score and that the trait will be apparent in the test taker's behavior.

Historically, 16pf stens of 4–7 have been considered within the average range; stens of 1–3, in the low range; and stens of 8–10, in the high range (see Figure 3.1) with reference to the norm sample. These same ranges continue to be used for the Sixth Edition, with a sten score of 4 being described as "low-average" and a sten score of 7, as "high-average." In a sten distribution, most people are expected to score in the middle (theoretically, about 68% obtain a score within plus or minus one standard deviation from the mean). About 16% score at the low end and another 16% score at the high end. The actual percentages may vary somewhat, depending on the shape of the distribution for any given factor scale.

Because most people tend to score towards the middle of the sten scale for a given factor, extreme 16pf scores represent more distinctive individual behavior patterns than the norm. These high and low scores are likely to represent the person's unique or "signature" personality characteristics compared to other people. They tend to be more stable behaviors across situations. In contrast, profiles in the average range can represent a personal style that varies more, depending on the circumstances.

Measurement Limits

Professionals need to integrate an understanding of measurement limits when interpreting 16pf Sixth Edition profiles. Because the scales are relatively short (approximately 10 items each), they necessarily are an estimate of a person's true score on any given personality factor. Theoretically, a person's true score falls, 68% of the time, in a band of plus or minus one standard error unit. The 16pf scales have a standard error of measurement (SEM) that is slightly below 1 sten score point. (See Table 7.3 for standard errors of measurement for the scales.) Assuming approximate normality of the scores, slightly more than 68% of the time, the true score for a person falls within the score range of plus or minus 1 sten score point around his or her obtained score. That is, the true score for a sten score of 8 on a factor would be expected to fall, 68% of the time, within a sten score range of 7–9. For a 95% confidence interval, the score band expands to plus or minus two standard error units; that is, for a sten of 8, the true score falls, 95% of the time, within a sten range of 6–10.

Professionals should be careful not to overinterpret sten score differences. This caution especially applies to interpreting scores at the extremes of the distribution where small differences in raw scores can shift sten scores. See Table 7.2 in Chapter 7 for data concerning how raw scores are converted into sten scores.

As mentioned previously, scores of 4 and 7 are termed “low-average” and “high-average” respectively. Professionals should realize that a test taker's true score might fall outside the average range because it is on the line between “average” and “distinctive” scores, and because the scales are not perfect measures of traits. For example, a test taker's sten of 4 might shift down a sten-score point, thus falling outside the average range, if he or she were to be retested. Similarly, scores of 3 and 8, which fall outside the average line but along the line between average and extreme, should not be overinterpreted as extreme because true scores might fall in the average range.

Interpretive Strategy: Approach to a 16pf Profile

Recommended Strategy

The recommended strategy for 16pf profile interpretation involves evaluating the following in the sequence indicated:

1. Response Style Indices
2. Global Factor scales
3. Contributing Primary Factor scales
4. Reasoning Factor and Related Primary Scales
5. General Trends – Extreme Scores

Each of the interpretive steps is described in the sections that follow. In general, Response Style Indices are evaluated first as a check for atypical test-response styles. The Global Factors are examined next because they provide a broad picture of the person. Next, the contributing Primary Factor scales are evaluated to obtain more nuanced details of the personality picture. Subsequently, information about the individual's cognitive ability and problem-solving style can be obtained by evaluating the reasoning factor score and by interpreting the score in association with other primary factors. Finally, general trends in primary factors are evaluated to provide overall sense of the profile.

Step 1: Evaluate Response Style Indices

The Sixth Edition has three response style indices: Impression Management (IM), Infrequency (INF), and Acquiescence (ACQ). For full details on the development and use of these scales, professionals can consult Chapter 6 of this manual.

Reviewing all three scales provides data about test-taking response styles. If a test taker's score on any of the indices is extreme, the professional should generate hypotheses about the test taker's attitude and, if possible, review information about the test taker (e.g., background data, other test results, notes from previous interactions, and discussions after the testing). In some cases, retesting may be necessary.

16pf Sixth Edition computer-based interpretive reports automatically score all three scales.

Interpretive information for each of the response style indices is given in the sections that follow. This information is based on the body of evidence available for the Sixth Edition.

Impression Management (IM) Scale

This bipolar scale consists of six items. The items are scored only on the IM scale and do not contribute to any of the primary personality scales.

General Scale Meaning

IM is essentially a social-desirability scale, with high scores reflecting socially desirable responses and low scores reflecting willingness to admit undesirable attributes or behaviors. The item content reflects both socially desirable and undesirable behaviors or qualities.

Social-desirability response sets include elements of self-deception as well as elements of other deception. Thus, high scores can reflect impression management (presenting oneself to others as tending to behave in desirable ways), or they can reflect a test taker's self-image as a person who behaves in desirable ways. In both cases, the possibility exists that the socially desirable responses might be more positive than the test taker's actual behavior (i.e., conscious or unconscious distortion) or that the test taker really might behave in socially desirable ways (i.e., the response choices accurately reflect the person's behavior).

Item Content/Typical Self-Report

The IM scale includes items such as "I am always willing to help people." Answering "agree" or "strongly agree" to such items contributes to a higher score on IM, indicating a socially desirable response set, whereas answering "disagree" or "strongly disagree" indicates a willingness to admit to less socially desirable behaviors.

Correlations With Other 16pf Factors

The IM scale correlates with several Sixth Edition primary personality scales. Correlations between the IM scale and the 16pf primary factors are presented in Chapter 6. The IM scale's main relationships are to primary scales that contribute to the Anxiety Global Factor and some components of the remaining Global Factors (see Table 6.3). IM correlates most highly with Emotional Stability (C+), lack of apprehension (O-) and Relaxedness (Q4-). In fact, high IM scorers may tend to score in the nonanxious direction on all scales related to the Anxiety Global Factor, including the remaining factor, Trust (L-). Moreover, high IM scorers also may tend to score in the Extraverted direction on some scales related to this Global Factor. The highest correlations are with Warmth (A+) and Social Boldness (H+). Further, high IM scorers may tend to score in the positive direction on Rule-Consciousness (G+), and lower on Groundedness (M-), two primary scales related to the Self-Control Global Factor. Conversely, low scores on IM tend to correlate with the same primaries but in the direction of admitting Anxiety, Introversion, and less Self-Control. Finally, higher (A+) and Abstractedness (M+) scores combined with high IM may contribute to lower Tough-Mindedness scores because both of these primary factors are inversely related to Tough-Mindedness.

Use of the IM Scale

Full elaboration of the use of the IM scale is presented in Chapter 6. Briefly, if a test taker's score exceeds a certain level (usually the 95th percentile for the high end of the IM scale and the 5th percentile for the low end), the professional should consider possible explanations for the extreme response set (e.g., job applicants generally have elevated IM scores). For the Sixth Edition, raw scores of 26 or higher fall at or above the 95th percentile and raw scores of 13 or lower fall at or below the 5th percentile compared to the norm sample. (See Chapter 6, Table 6.4 for the set of possible raw scores and their corresponding percentile values.) Depending on the reasons for testing, the professional might consider retesting, especially if deliberate distortion is suspected.

Infrequency (INF) Scale

The INF score consists of five independent items chosen because most respondents will agree or disagree when responding attentively. Unexpected answers (answers which research sample participants chose rarely) increase the INF score.

General Scale Meaning

Infrequent test taker responses indicate one of two possibilities. The first is that the individual test taker was rushed and inattentive to the actual content of the questions (or, similarly, had extreme reading issues interfering with understanding). As a result, the person chose a higher number of unusual answers than is typical. The second possibility is that the individual believed that a particular rare answer was actually descriptive of them. In either case, it is important for the test giver to try to discuss the results with the test taker and learn more about the situation. Their overall pattern of answers departs noticeably from how most people respond to these relatively simple and straightforward questions.

Sample Item Content

Answers such as disagreeing with the statement “I want my loved ones to be well” are representative of items contributing to a higher INF score. In most cases, 16pf respondents are more likely to indicate agreement than disagreement. The individual test taker may not have been paying close attention to the item. Bear in mind that the questionnaire’s items have deliberately been written at a primary school reading level. Native English speakers should have little difficulty with understanding or answering them.

Use of the INF Scale

The total raw score on the INF scale is converted to a percentile that compares the test taker to the normative sample for the Sixth Edition. The lowest score possible is 5, and the content of the items is such that attentive respondents can sometimes receive scores higher than 5, but very high scores may indicate inattention, language barriers, or extremely idiosyncratic responding. Percentiles of the standardization sample are shown in Table 6.6 of Chapter 6. Raw scores of 16 or greater fall at or above the 95th percentile relative to the norm sample and are considered to be high. If a test taker’s INF score is at or above the 95th percentile (or another designated cut-off), the professional should try to determine why the individual chose unusual answers. Possibly, they were not attentive to the items, were using a pattern of responses such as an excessive number of middle responses (which affects Sixth Edition INF but to a lesser extent than was the case in the Fifth Edition), or purposefully chose a series of unusual responses.

The importance of correctly identifying invalid protocols varies in different situations. Professionals may choose to set their own cut offs for classifying protocols as invalid in accordance with the information presented here and relative to individual respondent cases. Base rate and hit rate considerations are discussed in Chapter 6.

Note that attentiveness can wax and wane during assessment. The INF items are administered roughly evenly spaced through the sequence of the personality items (when the 16pf questionnaire is administered in its standard format). Although elevated INF scores indicate an unusual response pattern, a low score cannot rule out unusual or inattentive responses to some non-INF items. Like the other response style indices, INF does not directly address responses to the Reasoning/B scale.

Acquiescence (ACQ) Scale

The Acquiescence (ACQ) scale measures the tendency to answer "agree" or "strongly agree" to an item, no matter what its content.

General Scale Meaning

An acquiescent response set reflects a test taker's tendency to answer "true" to incongruous items such as both of these: "I'm a 'take charge' kind of person" and "I feel uncomfortable telling other people what to do." This response set may denote a misunderstanding of item content, random responding, difficulty in attending to self-evaluative questions, or inability to choose a self-descriptive response. An acquiescent response set might also indicates an unclear self-image or a high need for approval from the testing professional, or people in general.

Item Content/Typical Responses

All items on the ACQ scale are agree–disagree items. Thus, a high score indicates an overall pattern of tending to respond "agree" or "strongly agree" to items rather than choosing answers based on the item content.

Use of the ACQ Scale

As with the other response style scales, scores above the 95th percentile on the ACQ scale signify the possibility of an acquiescent response set. Raw scores of 112 or higher are considered high (at or above the 95th percentile; see Table 6.5 in Chapter 6 for the set of raw scores and their corresponding percentile values). The testing professional should try to determine whether the high score reflects random, inconsistent, or indecisive responding, or a high need for approval. Protocols with extreme acquiescence bias should be invalidated.

Step 2 and 3: Evaluate Global Factor Scales and Their Contributing Primary Factors

Table 4.4 in Chapter 4 presents the factor pattern for the five Global Factors around which the primary scales cluster: Extraversion, Independence, Tough-Mindedness, Self-Control, and Anxiety. Descriptions of both poles of each Global Factor are listed in Table 3.2. Readers may recognize links between the 16pf Global Factors and the “Big Five” model of personality that is discussed in personality literature. Interested readers can find more technical details about the Global Factors in Chapter 4.

For each Global Factor, a set of primary scales “load on” the global construct; that is, the scale set contributes to, or makes up, the global construct. For example, Warmth (A-), Liveliness (F+), Social Boldness (H+), Forthrightness (N-), and Group-Orientation (Q2-) compose the scale set that contributes to the Extraversion Global Factor.

An understanding of the Primary Factor scales is critical to understanding the Global Factor scales. Therefore, users of the 16pf instrument should become familiar with such test characteristics as scale reliabilities, score distributions and standard errors of measurement (SEM), and correlations with other measures. Evidence for these characteristics is presented in this manual’s text and tables.

The sections that follow discuss how to evaluate broad trends evident at the Global Factor level in a 16pf profile. Each Global Factor is described in terms of the primary scales that contribute to it and its meaning. The pole of the bipolar primary scale that contributes to the Global Factor will be identified by a plus (+) or minus (-) following the factor name. For example, scoring high (+) on Warmth (Factor A), Liveliness (Factor F), and Social Boldness (Factor H) contributes to being Extraverted on the Global Factor. Scoring low (-) on Privateness (Factor N) and Self-Reliance (Factor Q2) also contributes to the Extraversion Global Factor score. Equations for calculating Global Factor scores are presented in Table 4.4 in Chapter 4.

Broad Trends

Before examining the specific global scale scores in a 16pf profile, testing professionals are encouraged to look at broad trends within the profile.

Evaluate Number of Extreme Scores

As noted in prior explanations of the sten distribution, the extreme scores in a profile usually indicate a test taker's most distinctive traits. Thus, the greater the numbers of extreme scores, the more distinctive the personality expression is likely to be.

Table 3.3 presents the number of extreme Global Factor sten scores (scores that fall outside the average range of 4–7). About half of the test takers (47.1%) obtain all average scores at the Global Factor level or are extreme on only one or two Global Factors. That a test taker would have extreme scores on all five Global Factors is rare. Only about 4.7% of the Sixth Edition standardization sample had Global Factors scores that were so distinctive.

Table 3.3 Number of Extreme Global Factor Scores on 16pf Profiles

Number of extreme scores	Percent of sample	Percentile
0	12.5	6.3
1	22.0	23.5
2	25.1	47.1
3	21.6	70.4
4	14.0	88.3
5	4.7	97.6

Note: Standardization sample, N=2528.

Examine Global Factor Clusters

A review of patterns presented by the five Global Factors taken together can present an initial high-level picture of the person. H. E. P. Cattell and Schuerger (2003) suggest examining three clusters of global factors to gain an impression of the individual and their profile.

First, consider Extraversion and Independence together. These two Global Factors offer possible views of the test taker from an interpersonal perspective. Is the individual more outgoing and prone to approach others in an interested, friendly way (high Extraversion), or do they tend to be socially reserved and shy in most cases (low Extraversion)? Do they take the initiative in social interactions and try to exert influence (higher Independence) or do they tend to follow the other people and defer to them (lower Independence)?

Second, a review of the Tough-Mindedness and Self-Control factors provides early clues about how the person makes decisions and approaches activities. Is this individual open to new ideas, people, and experiences (low Tough-Mindedness), or do they lean towards more familiar practices, people, and situations (high Tough-Mindedness)? Further, does the person tend to approach life and work in a more structured, methodical way (high Self-Control), or are they likely to be more spontaneous and make decisions and improvise in the moment (low Self-Control)?

Third, what does the Anxiety profile indicate about the test taker's emotional life in general? Do they appear to struggle with life's demands and challenges (higher Anxiety scores)? Or, do they usually react to uncertainties and unexpected events with composure (lower Anxiety scores)?

Once the professional has formed a preliminary impression of a test taker based on these clusters, a closer examination of the Global Factor scores and their components will enrich the picture.

Remember the Primary Factor Scale Relationships

When interpreting a Global Factor score, the testing professional should identify (a) contributing primary scale scores that are in the expected direction for the Global Factor, and (b) primary scale scores that are in the opposite direction. With a knowledge that certain scales are expected to contribute to a given Global Factor, the professional can begin to identify unusual factor combinations and can form hypotheses about possible ways that conflicting scores might be expressed in a test takers' life.

For example, if a test taker is extraverted and all the related primary scale scores are in the extraverted direction, he or she probably moves toward other people in a consistent manner. On the other hand, if a test taker is extraverted on some relevant primary scales and introverted on others, he or she may experience conflict. That is, the test taker may be extraverted in some situations or ways but not in others or may be ambivalent about how to or whether to move toward others.

Another example involves an overall global Extraversion score that is low-average. Such a score can reflect various combinations of the primary scales because several primary scales contribute to the Global Factor score.

For instance, one person with this score might be Reserved (A-), average on Liveliness (Factor F) and Social Boldness (Factor H), and high on Self-Reliance (Q2+). This person could be expected to be reserved, serious, and self-sufficient but not timid. If there is no sign of anxiousness or lack of self-confidence, the person may be comfortable with his or her introversion. Another person with a low-average Extraversion score might be average on Warmth (Factor A) and Liveliness (Factor F), but also Shy (H-) and Group-Oriented (Q2-) This introvert shows more timidity and dependency on other people and less orientation away from people than the introvert in the previous example. A fair hypothesis would be that the second introvert might enjoy being around people but that his or her reticence and shyness intrude. Another possibility is that this person would like to be group oriented so that he or she can get lost in a crowd (Q2-) as a way to deal with the evident timidity.

The Global Factor interpretive information that follows is based on the body of evidence available for the 16pf Fifth Edition. As test users continue to develop a database, interpretation guidance will be refined to reflect the incoming data. Each Global Factor is described, followed by an explanation of the primary factors that contribute to that higher-level factor.

Extraversion

Table 3.4 Extraversion (Extraverted Versus Introverted)

Introversion	Weight in Scoring equation	Extraversion
Reserved (A-)	0.2	Warm (A+)
Serious (F-)	0.4	Lively (F+)
Shy (H-)	0.2	Socially Bold (H+)
Private (N+)	0.3	Forthright (N-)
Self-Reliant (Q2+)	0.3	Group-Oriented (Q2-)

General Factor Meaning

Extraversion has been included in even the earliest descriptions of personality. The construct is largely attributed to Jung (1971) but has been found and described in many subsequent studies such as those by Eysenck (1960) and R. B. Cattell (1957), and is one of the "Big Five" factors (Goldberg, 1992). In the original 16pf Handbook, Extraversion was said to orient around a general social participation (R. B. Cattell, Eber, & Tatsuoka, 1970). Extraverts tend to be people oriented and to seek out relationships with others. Introverts tend to be less outgoing and sociable; they tend to spend more time in their own company than in that of others. Extraversion has several contributing aspects, as reflected in the Primary Factor scales that play a role in the overall Global Factor.

As shown in Table 3.4, Extraversion includes interpersonal Warmth (A+), a stimulation-seeking type of sociability called Liveliness (F+), Social Boldness (H+), a tendency for self-disclosure portrayed by Forthrightness (N-), and the need to affiliate with other people, especially in groups, called Group-Orientation (Q2-).

As mentioned, a consistent relationship exists between social desirability and the Extraversion Global Factor. Several of the Extraversion-related primary factors are correlated with the Sixth Edition Impression Management (IM) scale. (See individual primary scale descriptions for further evidence.) Even though Introversion is seen as less desirable than Extraversion, it may be associated with independence of thought and a tendency to think and deliberate.

Comparison to the Big Five Extraversion: The 16pf Extraversion is quite similar to the Big Five factor of the same name, but most Big Five models view Dominance (E) as a facet of Extraversion. 16pf users already familiar with the Big Five should avoid direct inferences about leading or dominating derived from a Big Five Extraversion perspective, although Social Boldness (H) does share some characteristics with aspects of Dominance (E).

Contributing Primary Factors

Factor A (Warmth): Warm Versus Reserved

General Factor Meaning

Factor A addresses the tendency to be warmly involved with people versus the tendency to be more reserved socially and interpersonally; both poles are normal. Reserved (A-) people tend to be more cautious in involvement and attachments. They tend to like working alone, often on mechanical, intellectual, or artistic pursuits. Warm (A+) people tend to have more interest in people and to prefer occupations dealing with people. Warm individuals tend to be comfortable in situations that call for closeness with other people.

Warm (A+) behavior tends to be more socially desirable, and, in fact, Factor A correlates positively with the Impression Management (IM) scale. However, extremely high scores can indicate that the desirable aspect of Warmth represents an extreme need for people and for close relating. Extremely Warm (A+) people may be uncomfortable in situations where the close relationships they seek are inaccessible. Low scorers, on the other hand, can be quite uncomfortable in situations that call for extensive interaction or for emotional closeness. In previous editions of the 16pf Questionnaire, Karson and O’Dell (1976) point out that Reserved (A-) people can be quite effective (e.g., famous researchers are often reserved). Karson and O’Dell also state that, in some cases, an extremely low Warmth score may indicate a history of unsatisfactory or disappointing interpersonal relationships.

Item Content/Typical Responses

High scorers might agree or strongly agree with the statement that their friends describe them as warm and comforting.

Correlations With Other 16pf Factors

Warmth (A+) is correlated with Liveliness (F+), Social Boldness (H+), Forthrightness (N-), and Group-Orientation (Q2-) as other components of Extraversion. It is related to seeking closeness to people, clearly a component of the general orientation to people that typifies Extraversion. It also is correlated with greater Sensitivity (I+). Reserved (A-) scores also contribute to Vigilance (L+) and Tension (Q4+), or general irritability and impatience with others. This latter combination of factors suggests a tough, unemotional pattern with which the lower Warmth of A- is consistent.

Factor F (Liveliness): Lively Versus Serious

General Factor Meaning

In The 16pf: Personality in Depth, Factor F’s exuberance is compared to the natural self-expression and spontaneity exhibited by children before they learn self-control (H. B. Cattell, 1989). High scorers are enthusiastic, spontaneous, and attention seeking; they are lively and drawn to stimulating social situations. Extreme scores may reflect a flighty quality that is seen as unreliable or immature. The attention seeking and liveliness of F+ people can be inappropriate for certain situations, especially those that call for restraint or decorum. In contrast, low scorers on Factor F tend to take life more seriously. They are quieter, more cautious, and less playful. They tend to inhibit their spontaneity, sometimes to the point of appearing constricted or serious. Although they may be regarded as mature, they may not be perceived as fun or entertaining.

Item Content/Typical Responses

High scorers are likely to say that they like being in the middle of excitement and activity.

Correlations With Other 16pf Factors

Factor F is correlated with Warmth (A+), Social Boldness (H+), Forthrightness (N-), and Group Orientation (Q2-). Factor F’s social exuberance has a more lively, impulsive, high-spirited flavor than other Extraversion-related primary scales. This may explain the contribution of Liveliness (F+) to the Unrestrained pole of the Self-Control Global Factor in the Sixth Edition.

Factor H (Social Boldness): Socially Bold Versus Shy

General Factor Meaning

High scorers consider themselves to be bold and adventurous in social groups, and show little fear of social situations. They tend to initiate social contacts and aren't shy in the face of new social settings. A large element of need for self-exhibition is evident at the high pole, with a flavor of Dominance (E) more prevalent than in other Extraversion-related factors. Low scorers tend to be socially timid, cautious, and shy; they find speaking in front of a group to be a difficult experience. The possibility of subjective experience of discomfort may relate to shyness (H-) as well as to some lack of self-esteem and discomfort in new settings, particularly interpersonal settings.

Item Content/Typical Responses

High scorers tend to say that starting conversations with strangers never gives them trouble.

Correlations With Other 16pf Factors

Social Boldness is correlated with Warmth (A+), Liveliness (F+), Forthrightness (N-), and Group-Orientation (Q2-). Factor H's contribution to Extraversion seems to relate more to boldness, status, and self- exhibition in comparison to the contributions of the other four primary scales. Social Boldness (H+) also contributes positively to the Independence Global Factor, along with Dominance (E+) and Openness to Change (Q1+). The ability to interact boldly with others plays a part in Independence, which involves elements of persuasion and self-expression. Further, Social Boldness (H+) relates to greater Emotional Stability (C+) and Openness to Change (Q1+) as well as to a relaxed, nontense manner (Q4-), indicating a high degree of ease in new circumstances and social situations. Last, Social Boldness can be affected by elevated Impression Management scores. Describing one's self as more interpersonally courageous can be socially desirable in some situations, such as being a job candidate for a customer-facing position.

Factor N (Privateness): Private Versus Forthright

General Factor Meaning

This factor addresses the tendency to be Forthright (N-) and personally open versus being Private (N+) and non-disclosing. Related to the Extraversion Global Factor in the Sixth Edition, Factor N content addresses whether self-disclosure is part of one's orientation to people. Low scorers tend to talk about themselves readily; they are genuine, self-revealing, and forthright. At the extreme, low scorers may be Forthright (N-) in situations where doing so may not be to their advantage. High scorers, on the other hand, tend to be personally guarded. High scorers seem to indicate that they "play their hand close to their chest," whereas low scorers "put all their cards on the table." At the extreme, high scorers may maintain their privacy at the expense of developing close relationships with others. This may reflect disinterest in or fear of closeness, as suggested by correlations with other measures. Factor N shows a modest correlation (r = -.21) with the Impression Management (IM) scale, with Forthrightness (N-) being the socially desirable pole.

Item Content/Typical Responses

Low scorers (more Extraverted individuals) may say that they tend to talk about their feelings readily when other people seem interested.

Correlations With Other 16pf Factors

Correlations with Reserve (A-), Seriousness (F-), Shyness (H-), and Self-Reliance (Q2+) and the negative loading of Factor N on the Extraversion Global Factor support the link between Privateness (N+) and Introversion, especially with Introversion's components of timidity, reserve, and self-reliance.

Factor Q2 (Self-Reliance): Self-Reliant Versus Group Oriented

General Factor Meaning

This factor tends to be about maintaining contact with or proximity to others. Low scorers are Group-Oriented (Q2-); they prefer to be around people and like to do things with others. High scorers are Self-Reliant (Q2+); they enjoy time alone and prefer to make decisions for themselves. It appears to be more socially favorable to present oneself as scoring in the Extraverted, Group-Oriented (Q2-) direction rather than in the Self-Reliant (Q2+) direction, as possibly reflected by the moderate but significant negative correlation (r = -0.29) of Self-Reliance with the Impression Management (IM) scale.

Being extremely Group-Oriented (Q2-) may not be optimally effective in situations where help is unavailable or where others are providing poor direction or advice. On the other hand, excessively Self-Reliant (Q2+) people may have difficulty in working alongside others, and they also may find it hard to ask for help when necessary. Although Self-Reliant people can act autonomously when the need arises, those having extremely high scores may neglect interpersonal aspects and consequences of their actions.

Item Content/Typical Responses

Low scorers (N-), who are more Extraverted, often say that they like it best when they have people around them.

Correlations With Other 16pf Factors

In the 16pf Sixth Edition, Group Orientation (Q2-) is correlated with the other Extraversion primary factors of with Warmth (A+), Liveliness (F+), Social Boldness (H+), and Forthrightness (N-).

Independence

Table 3.5 Independence (Independent Versus Accommodating)

Accommodating	Weight in scoring equation	Independent
Deferential (E-)	0.6	Dominant (E+)
Timid (H-)	0.3	Bold (H+)
Trusting (L-)	0.1	Vigilant (L+)
Traditional (Q1-)	0.4	Open to Change (Q1+)

General Factor Meaning

Independence revolves around the tendency to be actively and forcefully self-determined in one's thinking and actions. Independence has several contributing aspects, as reflected in its Primary Factor scales. As shown in Table 3.5, this Global Factor includes tendencies to be Dominant (E+), Socially Bold (H+), Vigilant (L+), and Open to Change (Q1+).

Independent people tend to enjoy trying new things and exhibit an intellectual curiosity. A strong element of social forcefulness also is evident in Independence. Independent people tend to form and to express their own opinions. They often are persuasive and forceful, willing to challenge the status quo, and suspicious of interference from others. Extreme Independence— especially when not tempered with Self-Control or the sociability of Extraversion, or the sensitivity of Receptivity—can assume a certain amount of disagreeableness. In the Sixth Edition, Independence may have flavors of inflexibility and domination. Independent people may be uncomfortable or ineffective in situations that involve accommodating other people.

In contrast to Independent, Accommodating people tend to be Deferential (E-), Shy (H-), Trusting (L-), and Traditional (Q1-). They tend not to question; instead, they value Agreeableness and accommodation more than self-determination or getting their own way. External situations and other people tend to influence them, both in terms of forming opinions and shaping behavior. They may be very uncomfortable or ineffective in situations that call for self-expression, assertiveness, or persuasion. Accommodation may be linked with the wish to avoid harm or with anxiousness, as suggested by the correlations.

Comparison to the Big Five Agreeableness: Low scores on Independence share characteristics with the Big Five concept of Agreeableness, but Independence includes Dominance and high scores are colored by this influence and are less about being mean, contrary, or disagreeable. In contrast, most Big Five models view Dominance as a facet of Extraversion. This difference, which has been attributed to technical choices in factor analysis (oblique vs. orthogonal rotations; Child, 1998), is one of the biggest interpretational differences between the 16pf Global Factors model and the Big Five.

Contributing Primary Factors

Factor E (Dominance): Dominant Versus Deferential

General Factor Meaning

This factor involves the tendency to exert one's will over others (Dominance) versus accommodating others' wishes (Deference). Factor E is more about dominance than about simple assertiveness. Whereas assertiveness serves to protect one's rights, wishes, and personal boundaries, dominance serves to subjugate others' wishes to one's own (H. B. Cattell, 1989). A high score does not eliminate the possibility that a test taker can be assertive rather than aggressive. However, most high scorers tend to be forceful, vocal in expressing their wishes and opinions even when not invited to do so, and pushy about obtaining what they want. They feel free to criticize others and to try controlling others' behavior. Whereas Dominance can lend a certain amount of commanding social presence, extreme Dominance can alienate people who do not wish to be subjugated.

Low scorers tend to avoid conflict by acquiescing to the wishes of others. They are self-effacing and willing to set aside their wishes and feelings. Extreme deference can be alienating to those who wish for a more forceful or participating response.

Item Content/Typical Responses

High scorers say that they are "take charge" persons.

Correlations With Other 16pf Factors

Dominance (E+) is the strongest contributor to the Independence Global Factor, along with Social Boldness (H+), Vigilance (L+), and Openness to Change (Q1+). In being Independent, one is forcefully self-determined and attempts to influence others. The Dominance element and the willingness to assert oneself contribute to this Independent stance. Dominance also is correlated with Self Assured (O-) profiles, indicating a tendency to experience little self-doubt about one’s words and actions.

High Factor H (Social Boldness)

As described in the Extraversion section, Socially Bold individuals are socially adventurous and will take the initiative in social situations. That orientation can lead to a more Independent, expressive orientation toward others rather than being deferential or shy in pursuing one’s own wishes and personal agenda.

Factor L (Vigilance): Vigilant Versus Trusting

General Factor Meaning

This factor relates to the tendency to trust versus being vigilant about others’ motives and intentions. This vigilance leads high scorers to expectations that they will be misunderstood or taken advantage of by others, and they also experience themselves as separate from other people. In the context of global Independence, Vigilant people may be competitive and alert for signs of unfairness. High scorers may be unable to relax their Vigilance (L+) when it might be advantageous to do so. At the extreme, high scorers’ mistrust may have an aspect of animosity. Sometimes a Vigilant stance is in response to life circumstances. For example, members of oppressed minority groups have historically tended to score higher on Vigilance [L+], although this tendency was more pronounced in the Fifth Edition (Cohen’s d = -.50 for White/Black comparison, d = -.27 for White/Hispanic comparison) than in the Sixth Edition (d = -.12 and -.19, respectively).

Low scorers tend to expect fair treatment, loyalty, and good intentions from others. Trust (L-) tends to be related to a sense of well-being and satisfactory relationships However, extremely low scorers may be taken advantage of because they do not give enough thought to others’ motivations.

Factor L is correlated -0.46 with the Impression Management (IM) scale; Trust is the socially desirable pole for Factor L.

Item Content/Typical Responses

High scorers say that they are suspicious of others' actions.

Correlations With Other 16pf Factors

In addition to its contribution to the Independence global factor, Vigilance (L+) is correlated with the Anxiety primary factors of Privateness (N+) and Apprehension (O+), and with Self Reliance (Q2+). The picture is of someone who guards personal information, may fret about their actions, and is prone to keep their own company. As stated earlier, Vigilance is also related to Impression Management. Higher IM scores can result in lower Vigilance scores because this trait can be viewed as less desirable in many people.

Factor Q1 (Openness to Change): Open to Change Versus Traditional

General Factor Meaning

High scorers tend to be curious and to think of ways to improve things and to enjoy experimenting. If they perceive the status quo as unsatisfactory or dull, they are inclined to change it. Low scorers tend to prefer traditional ways of looking at things. They don't question the way things are done. They prefer life to be predictable and familiar, even if life is not ideal.

Item Content/Typical Responses

High scorers tend to say that they like to think about ways the world could be improved.

Correlations With Other 16pf Factors

Openness to Change (Q1+) is correlated with the Dominance (E+) and Social Boldness (H+) components of the Independence Global Factor. Q1+ also contributes to the Receptive pole of the Tough-Mindedness Global Factor, along with Warmth (A+), Sensitivity (I+), and Abstractedness (M+). Factor Q1's elements of nonconformity and openness to new ideas are reflected in its correlations with Liveliness (F+) and with Abstractedness (M+).

Tough-Mindedness

Table 3.6 Tough-Mindedness (Tough-Minded Versus Receptive)

Receptive	Weight in scoring equation	Tough-Minded
Warm (A+)	0.3	Reserved (A-)
Sensitive (I+)	0.5	Utilitarian (I-)
Abstracted (M+)	0.3	Grounded (M-)
Open to Change (Q1+)	0.4	Traditional (Q1-)

General Factor Meaning

R. B. Cattell originally called this Global Factor "Cortertia," an abbreviation for "cortical alertness" (R. B. Cattell et al., 1970). High scorers on Cortertia were described as alert and tending to deal with problems at a dry, cognitive level. The factor later assumed the more popularized term "Tough Poise." In the Fifth Edition, this Global Factor was called Tough-Mindedness, and it has several contributing aspects, as reflected in its related Primary Factor scales. Tough-Minded people tend to be Reserved (A-), Utilitarian (I-), Grounded (M-), and Traditional (Q1-) (see Table 3.6). In addition to operating at a dry, cognitive level, extremely Tough-Minded people may portray a sense of being "established," possibly to the degree of being set or fixed in their thinking. They may not be open to other points of view, to unusual people, or to new experiences. Receptive people, on the other hand, are Warm (A+), Sensitive (I+), Abstracted (M+), and Open to Change (Q1+). Although they may be more open than their Tough-Minded counterparts, Receptive people may overlook the practical or objective aspects of a situation.

Prior to the Fifth Edition, the Tough-Minded label had been given to Factor I, one of the main primaries that contributed to this Global Factor. For the Sixth Edition, Tough-Mindedness is the name of the high pole of the Global Factor because it represents the overriding thread that runs through all the contributing primary scales. Factor I's contribution is more specific to sensitivity and aesthetic values on the high end and to utilitarian values and objectivity on the low end. Hence, the primary scale Factor I was renamed "Sensitivity" for the Fifth and Sixth Editions. The low pole of the Tough-Mindedness Global Factor is named Receptive in the Sixth Edition. Receptive people tend to deal with problems in a cultured, refined, or sensitive way. They also tend to be open to interpersonal involvement (Warmth, A+), to sensitive perceptions (Sensitivity, I+), to ideas and fantasy (Abstractedness, M+), and to change (Openness to Change, Q1+).

A certain inflexibility and lack of openness may be apparent in Tough-Mindedness. In fact, toughness and resoluteness can border on inflexibility and entrenchment. Tough-Minded people may have difficulty in accepting new viewpoints, including those that involve emotions. In contrast, Receptive people can be more open to experiencing feelings, possibly even negative affective states. As a result, Receptive people may experience difficulty in setting aside their feeling reactions to attain objectivity and, consequently, may overlook the practical aspects of situations. Historically, gender stereotypes have sometimes been associated with Tough-Mindedness and Receptivity, the former being more "masculine" and the latter being more "feminine." However, gender differences in Tough-Mindedness in the Sixth Edition are only moderate, with standardized group differences favoring men (Cohen's *d* = -.52, see Chapter 7).

Comparison to the Big Five Openness: Arguably, Openness (i.e., Openness to Experience) is the least well-defined Big Five factor, with some significant differences in how this factor is conceptualized in different Big Five models. Indeed, some formulations of the Big Five view this factor primarily as Culture or Intellect rather than as Openness to Experience. As a result, 16pf users familiar with the Big Five should verify any Openness (or Culture/Intellect) inferences when interpreting Tough-Mindedness, depending upon the Big Five model that is being utilized as a frame of reference.

However, high scores on Tough-Mindedness reflect many of the characteristics of the low pole the Big Five Openness, a lack of receptivity and traditionalism. Tough-Mindedness, however, does not address academic interest and/or success, which are encompassed in some Big Five models of Openness that focus primarily on Intellect. Rather, Reasoning/B may be helpful in inferring school success.

Contributing Primary Factors

Low Factor A (Warmth): Reserved

Tough-Minded individuals tend to score in the Reserved direction (A-) of the Warmth factor. This indicates a more distant, impersonal view of the world and of other people. It suggests lower receptivity to becoming personally and emotionally involved with others.

See the primary factor description under the Extraversion section for more detailed information.

Factor I (Sensitivity): Sensitive Versus Utilitarian

General Factor Meaning

Relative to the Tough-Mindedness Global Factor, low scorers, or Utilitarian (I-) people evince little sentimentality, attending more to how things operate or work than to personal values or tastes. These individuals tend to have a more utilitarian focus. The content of the Sixth Edition Factor I scale focuses on people's sensitivities and sensibilities; high scorers tend to base judgments on personal tastes and aesthetic values. Sensitive (I+) people rely on empathy and sensitivity in their considerations.

Utilitarians (I-) tend to be concerned with utility and objectivity and may exclude people's feelings from consideration. Because they don't tend to indulge vulnerability, people with extreme I- scores may have trouble dealing with situations that demand sensitivity. In contrast, Sensitive (I+) people tend to be more refined in their interests and tastes and more sentimental than their Utilitarian (I-) counterparts. At the extreme, I+ people may be so focused on the subjective aspects of situations that they overlook more functional aspects. In previous editions of the 16pf Questionnaire, the Sensitivity factor is linked to the Jungian concept of judging functions: Thinking versus Feeling (H. B. Cattell, 1989). This interconnection is supported by correlations with other measures.

One of the changes made during revision of the 16pf Sixth Edition was to reduce the relationship of Factor I to gender stereotypes and gender score differences are thus much reduced but not eliminated. As with any group difference result, it is important to recognize that there is considerable within-gender variation in Sensitivity/I scores; however, as a group, women tend to score moderately higher on this factor than do men (Cohen's d = -.52; see Chapter 7 for statistics on group differences).

Item Content/Typical Responses

Low scorers say that in school they preferred math to English in school. High scorers endorse an attraction to aesthetic or artistic pursuits. Low scorers exhibit practical, utilitarian tendencies.

Correlations With Other 16pf Factors

The Utilitarian (I-) pole has the highest correlations with Privateness (N+) and is correlated with a Reserved (A-) and Serious (F-) orientation. In addition, it is related to Traditional (Q1-) and Grounded (M-) profiles. These patterns portray someone who is prone to be socially restrained, deliberate, and unlikely to speculate or easily embrace new ideas or possibilities.

Factor M (Abstractedness): Abstracted Versus Grounded

General Factor Meaning

Factor M addresses the type of things to which people give thought and attention. Tough-Minded people often score low on this primary factor. Grounded (M-) people focus on their senses, observable data, and the outer realities of their environment in forming their perceptions. On the other hand, Abstracted people (M+) are more oriented to internal mental processes and ideas rather than to practicalities. In previous editions, this factor is linked to the Jungian perceiving functions, Sensation versus Intuition (H. B. Cattell, 1989).

In previous editions, high scores, reflecting an intense inner life rather than a focus on outer environment, are associated with the absent-minded professor image (Krug, 1981). High scorers are Abstracted (M+); that is, they are occupied with thinking, imagination, and fantasy, and they often get lost in thought. In contrast, low scorers are Grounded (M-); that is, they focus more on the environment and its demands.

Although low scorers may think in a practical and down-to-earth manner, they may not be able to generate possible solutions to problems. In fact, extremely Grounded (M-) people may be so overly concrete or literal that they “miss the forest for the trees.” Abstracted (M+) thinking, on the other hand, often leads to plentiful idea generation. However, high scorers may generate ideas without considering the practical realities of people, processes, and situations. It is important to keep in mind that Abstractness/M reflects thinking and problem solving style or approach, rather than ability, which would be assessed using Reasoning/B scores and/or supplemental ability measures.

Extremely Abstracted (M+) people sometimes seem less in control of their attention or of situations, and sometimes report that they have mishaps or accidents because they are preoccupied. In fact, Factor M loads negatively on the Self-Control Global Factor, with Abstracted (M+) people being less self-controlled.

Item Content/Typical Responses

Low scorers tend to say that they rarely get lost in their thoughts.

Correlations With Other 16pf Factors

Groundedness (M-) contributes to a Tough-Minded stance; that is, to say one is Grounded (M-) means the individual has a more practical and here-and-now perspective rather than being imaginative or idea-focused. A Grounded (M-) person is more likely to seek and be observant of rules and moral codes (Rule Conscious, or G+).

The reverse is also true. Factor M's negative correlation with Factor G suggests a link between Abstractedness (M+) and Expedience (G-), a tendency to observe rules and even laws when personally convenient. In addition, Factor M correlates negatively with Reactivity (C-) and with Apprehension (O+), suggesting somewhat greater emotionality and self-criticism. The remaining correlations suggest a link between Abstractedness (M+) and lowered Self-Control: It correlates with Tolerance of Disorder (Q3-).

Finally, test takers with inflated Impression Management scores (IM+) tend to score lower on Abstractedness (M-). Like the Anxiety primary factors, the behavior associated with higher scores may appear to be less desirable to some individuals.

<u>Low Factor Q1 (Openness to Change): Traditional</u>

The Tough-Minded person is often found to have a more Traditional (Q1-) or conservative outlook. He or she will seek predictability and stability and to tend rarely think about how the world could be different. This person is unlikely to seek new experiences or to relish surprises.

See the description under the Independence Global Factor for further information.

Self-Control

Table 3.7 Self-Control (Self-Controlled Versus Unrestrained)

Unrestrained	Weight in scoring equation	Self-Controlled
Lively (F+)	0.3	Serious (F-)
Expedient (G-)	0.4	Rule-Conscious (G+)
Abstracted (M+)	0.5	Grounded (M-)
Tolerates Disorder (Q3-)	0.3	Perfectionistic (Q3+)

General Factor Meaning

Self-Control concerns curbing one's urges. High scorers tend to be able to inhibit their impulses and may do so in several ways, depending on the pattern of scores on the related Primary Factor scales. For example, Self-Controlled people can be Serious (F-), Rule-Conscious (G+), practical and Grounded (M-), and possibly Perfectionistic (Q3+) as a means to Self-Control (see Table 3.7). Either Self-Controlled people simply do not value flexibility or spontaneity, or they may have acquired self-control at the expense of these qualities. A link also exists between Self-Control and social desirability, with higher control being more socially desirable.

In contrast to Self-Controlled people, Unrestrained people tend to follow their urges more. This Unrestrained behavior can be reflected in several ways: in spontaneity and Liveliness (F+), in Expedience versus rigor in following rules and duty (G-), in being imaginative or higher in Abstractedness (M+), and potentially in a Tolerance of Disorder (Q3-). Unrestrained people may be flexible in their responses; however, in situations that call for self-control, they may find it difficult to restrain themselves.

They may be perceived as self-indulgent, disorganized, irrepressible, or irresponsible, depending on whether they can muster resources for self-control when doing so is important.

Comparison to the Big Five Conscientiousness: There is a general similarity of inferences based on 16pf Self-Control and Big Five Conscientiousness, but 16pf users familiar with the Big Five should verify specific inferences. Some Conscientiousness scales include elements like loyalty and achievement that would be seen as combinations of 16pf primary scales.

Contributing Primary Factors

<u>Low Factor F (Liveliness): Serious</u>

As previously described, a Serious (F-) person tends to act in a careful and deliberate manner. They think before they speak and act, which can contribute to a more measured and self-disciplined approach to work, relationships, and life. These individuals are often seen by others as mature because of their paced, more methodical style. Further description of this factor is found in the Extraversion section.

<u>Factor G (Rule-Consciousness): Rule-Conscious Versus Expedient</u>

General Factor Meaning

This factor addresses the extent to which cultural standards of right and wrong are internalized and used to govern behavior (R. B. Cattell et al., 1970). It has been associated with the psychoanalytic concept of superego, in which moral ideals from the culture and environment are internalized and used to control the id impulses of self-gratification.

High scorers tend to perceive themselves as strict followers of rules, principles, and manners. In previous 16pf editions, high scorers are described as those who endorse conventional cultural values in their responses to Factor G items (H. B. Cattell, 1989). Rule-Conscious people emphasize the importance of conformance to regulations, depicting themselves as rule-bound, conscientious, and persevering.

They can be perceived as staid, inflexible, or self-righteous because of their dogmatism. Low scorers tend to place little value on rules and regulations, doing so either because they have a poorly developed sense of right and wrong (e.g., lacking internalized moral values) or because they ascribe to values that are not solely based on conventional mores in deciding which rules and principles should govern their actions. Expedient (G-) behaviors seem to have elements of need for autonomy, need for play, and need for flexibility, as suggested by correlations with other measures. Low scorers might have difficulty in conforming to strict rules and regulations. It is important for the testing professional to gather additional information to help evaluate whether low scorers have failed to develop moral standards or whether they simply follow unconventional standards. In either case, their behaviors may be perceived as unpredictable unless their guiding principles and motivations are known. Other Primary Factor scales can indicate resources that might influence the Expedient (G-) person's Self-Control, especially those scales with which this factor correlates. For example, paired with an elevated Liveliness (F+) score, a G- profile could indicate the person is prone to act impulsively and ignore consideration of conventional rules and guidelines. Similarly, an individual with a higher Abstractedness/M result and an Expedient (G-) score might interpret the meaning of a policy in an unusual way.

A correlation of 0.33 exists between the cultural values endorsed by Rule-Conscious (G+) people and social desirability. Factor G shows a significant positive correlation with social desirability as measured by the Impression Management (IM) scale. That is, saying that one follows the rules is more socially desirable than admitting that one does not conform.

Item Content/Typical Responses

High scorers agree that it is important to be respectful of rules, laws, and moral codes.

Correlations With Other 16pf Factors

Factor G contributes positively to the Self-Control Global Factor, and it has a positive correlation with Perfectionism/Q3, which also contributes positively to the Self-Control Global Factor. It also is related to being more Concrete (M-). The picture presented is one of a rule-oriented, detail conscious person who favors planning and is unlikely to be extraordinarily imaginative or freewheeling in their thinking.

Low Factor M (Abstractedness): Concrete

As described in the Tough-Mindedness section above, a Concrete (M-) profile indicates that the person is practical and well-attuned to the external environment instead of to the inner world of ideas. Such people are likely to maintain their attention to immediate surroundings and to the demands of the current situation, qualities that help them be productive and present a conscientiousness, task-conscious image to others.

Factor Q3 (Perfectionism): Perfectionistic Versus Tolerates Disorder

General Factor Meaning

High scorers want to do things right. They tend to be organized, to keep things in their proper places, and to plan ahead. Perfectionistic (Q3+) people are likely to be most comfortable in highly organized and predictable situations, and may find it hard to deal with unpredictability. At the extreme, they may be seen as inflexible.

In contrast to high scorers, low scorers leave more things to chance and tend to be more comfortable in a disorganized setting. However, low scorers may be perceived as lackadaisical, unorganized, or unprepared. They may not be able to muster a clear motivation for behaving in planful or organized ways, especially if these behaviors are unimportant to them.

Item Content/Typical Responses

High scorers say that they prefer to plan ahead even if it takes longer to do a task.

Correlations With Other 16pf Factors

Perfectionism (Q3+) contributes to the Self-Control Global Factor, along with Rule-Consciousness (G+), and Groundedness (M-). Viewed together, these factor scores suggest an individual who is attuned to following guidelines and rules, doing so in a practical and carefully planned way.

Anxiety

Table 3.8 Anxiety (Anxious Versus Unperturbed)

Low anxiety	Weight in scoring equation	High anxiety
Emotionally Stable (C+)	0.4	Reactive (C-)
Trusting (L-)	0.2	Vigilant (L+)
Self-Assured (O-)	0.4	Apprehensive (O+)
Relaxed (Q4-)	0.3	Tense (Q4-)

General Factor Meaning

Like Extraversion, Anxiety has been described since early studies of personality and continues to be described in studies of the "Big-Five" dimensions of personality (Goldberg, 1992). Anxiety has several contributing aspects, as reflected in its related Primary Factor scales. As shown in Table 3.8, Anxiety includes a tendency to be Reactive (C-) to events rather than adaptive, distrustful of others and Vigilant (L+), worry-prone and Apprehensive (O+), and impatient or Tense (Q4+).

Anxiety can be aroused in response to external events, or it can be internally generated. Anxiousness may be an activation of the "fight-or-flight" state associated with perceived or actual threat, as suggested by known correlations. Low-anxious people tend to be unperturbed; however, they may minimize negative affect or be unmotivated to change because they are comfortable. Evidence from known correlations also suggests that anxious people often experience more negative affect; they may have difficulty controlling their emotions or reactions and may act in counterproductive ways.

As previously described, strong relationship exists between social desirability and Anxiety; several of the Anxiety-related primary factors are strongly correlated with the Sixth Edition Impression Management (IM) scale.

Comparison to the Big Five Neuroticism: There is a general similarity of inferences based on 16pf Anxiety and Big Five Neuroticism but 16pf users familiar with the Big Five should verify specific inferences. Both the 16pf Anxiety factor and most conceptualizations of Big Five Neuroticism tend to incorporate facets involving tension, worry, and lack of composure.

Contributing Primary Factors

Factor C (Emotional Stability): Emotionally Stable Versus Reactive

General Factor Meaning

This factor largely concerns feelings about coping with day-to-day life and its challenges. High scorers tend to take life in stride and to manage events and emotions in a balanced, adaptive way. Low scorers feel a certain lack of control over life. Low scorers tend to react to life, whereas high scorers make adaptive or proactive choices in managing their lives. This factor has an element of emotional well-being. However, an extremely high score on this scale can indicate that a test taker may be strongly disinclined to report, or even to experience, so-called "negative" feelings.

Factor C shows a very strong correlation with the Impression Management (IM) scale. Presenting oneself as able to cope with life is socially desirable; admitting that one feels unable to manage feelings or adapt to life is socially undesirable. Whenever a test taker obtains an extremely low score, he or she is admitting undesirable feelings. In previous editions of the 16pf Questionnaire, Karson and O'Dell (1976) suggest that a test taker should always be questioned about reported experiences of distress and reactivity. They also advise that interpretation of a high Emotional Stability (C+) score, especially when it is accompanied by a high score on the IM scale, should address whether the test taker denied any problems in order to present himself or herself favorably.

Item Content/Typical Responses

High scorers tend to say that their emotions are well-balanced most of the time. Low scorers report that their mood is susceptible to change.

Correlations With Other 16pf Factors

Reactivity (C-) is a strong contributor to the Anxiety Global Factor, having strong correlations with Vigilance (L+), Apprehension (O+), and Tension (Q4+). Self-perceptions of feeling unable to adapt to life and its demands contribute to general anxiousness. Emotional Stability (C+) is also related to Dominance (E+) and to Social Boldness (H+), which indicates a degree of social fearlessness and willingness to state and pursue one's wishes and beliefs.

Factor L (Vigilance)

This factor was described in detail under the Independence Global Factor. Relative to Anxiety, high Vigilance (L+) indicates the tendency to distrust others and avoid taking their words and actions at face value. Although not suggestive of outright paranoia, higher scores suggest that the individual is predisposed to view other people as dishonest and possibly manipulative. As previously stated, this quality may be habitual or a result of repeated life experiences such as being in a social minority (H. E. P. Cattell & Schuerger, 2003). It also may be a temporary reaction to recent events such as being dramatically deceived or cheated, which raise the person's suspicions that the people around them are less benevolent than they appear.

Factor O (Apprehension): Apprehensive Versus Self-Assured

General Factor Meaning

High scorers tend to worry about things and to feel apprehensive and insecure. Sometimes, these feelings are in response to a particular life situation. In other cases, these feelings are part of a characteristic response pattern, appearing across situations in a person's life. Worrying can have positive results, in that a person can anticipate dangers in a situation and can see how actions might have consequences, including interpersonal effects. However, Apprehensive (O+) people can make a poor social presence.

In contrast to high scorers, low scorers tend to be more self-assured, neither prone to apprehensiveness nor troubled about their sense of adequacy. Low scorers present themselves as confident and self-satisfied. If a person's score is extremely low, his or her confidence may be unshaken, even in situations that provide opportunities for self-evaluation and self-improvement. In such instances, the person's self-assurance may result from blocking out awareness of negative elements of self.

There also is an element of social desirability in Factor O, with Self-Assured (O-) response choices being the socially desirable pole. Individuals with elevated Impression Management scores (IM+) may also have somewhat lowered Apprehension scores.

Item Content/Typical Responses

High scorers tend to say that they sometimes feel they've done something wrong even if they haven't.

Correlations With Other 16pf Factors

Apprehension (O+) contributes to the Anxiety Global Factor, along with Reactivity (C-), Vigilance (L+), and Tension (Q4+). Thus, Apprehension (O+) seems to contribute to a general anxiousness. It also is related to the Deferential (E-) pole of Independence and to the Extraversion primary factors of Seriousness (F-) and Shyness (H-). These findings suggest that Apprehensive people may be more withdrawn, timid, and serious.

Factor Q4 (Tension): Tense Versus Relaxed

General Factor Meaning

This scale is associated with nervous tension. High scorers tend to have a restless energy and to be fidgety when made to wait. Although a certain amount of tension can be focused effectively and can motivate action, extremely high tension can lead to impatience and irritability. High tension may sometimes get in the way of self-control or may impede effective action. Professionals may want to address the source of tension whenever high scores occur in a profile because such scores may reflect either tension that is characteristic of a person or tension that is specific to a person's present life situation.

Low scorers tend to feel relaxed and tranquil. They are patient and slow to become frustrated. At the extreme, their low level of arousal can make them unmotivated or complacent. That is, because they are comfortable, they may be disinclined to change or push themselves.

Social desirability can affect Factor Q4 results. In fact, the correlation between Factor Q4 and the Impression Management (IM) scale is among the highest in the 16pf Sixth Edition (r = -0.62).

Item Content/Typical Responses

High scorers say they tend to appear to become agitated quite quickly.

Correlations With Other 16pf Factors

Factor Q4 is a contributor to the Anxiety Global Factor, along with Reactivity (C-), Vigilance (L+), and Apprehension (O+). High tension (Q4+) tends to be related to lower Extraversion-related scores such as being Reserved (A-), Serious (F-), Shy (H-), and Private (N+) along with greater Self-Reliance (Q2+).

Step 4: Evaluate Reasoning Scale and Related Primary Factors

Factor B (Reasoning): Analytical Versus Concrete

About the Scale

The Factor B scale is composed of items which tap the ability to solve problems using reasoning. In the Sixth Edition, this scale is longer and expanded with a broader range of item types such as interpreting graphs and deductive reasoning in addition to prior editions' verbal and mathematical problems. The development and construction of the B factor is described in greater detail in Chapter 5. Even though Reasoning is not a personality trait, it is included in the 16pf Questionnaire because cognitive style has been observed to moderate the expression of many personality traits.

Item Content/Typical Responses

The scale represents nine different types of items (see Table 5.1 in Chapter 5). Descriptions of item types and development of the scale are presented in Chapter 5. An example Factor B item is "Which word does NOT belong with the other two? (a) cat, (b) dog, (c) house." The Sixth Edition Factor B items are entirely new and incorporate a variety of common reasoning assessment tasks (a greater variety than the previous edition). A small number of items have numeric responses that are entered directly by the respondent (rather than choosing a response).

Because of the many advantages of adaptive administration (see Chapter 5), this form is administered “adaptively.” As a result, different respondents are likely to get different items but the resulting sten score can be compared among respondents just like other 16pf scores.

Score Meaning

High scorers solve more of the reasoning problems correctly; low scorers choose a higher number of incorrect answers. In general, better results may indicate greater ability to think in more rational, integrated ways, whereas lower scores indicate difficulty with problem solving or logical thinking.

However, in previous editions of the 16pf Questionnaire, H. B. Cattell (1989) suggests that high scores frequently reflect higher reasoning ability because people are unlikely to obtain high scores by chance. At times, however, average or low scores may not accurately reflect people’s reasoning ability. These instances might occur in test takers who are educationally disadvantaged or who are experiencing emotions that interfere with their test performance. They might also occur when test takers are distracted by environmental stimuli, are wrong in their interpretations of the instructions, or are, for various reasons, not motivated to spend the time figuring out the correct answers.

A lower-than-expected score can result when a test taker has extreme reading difficulties or speaks English as a second language. A low score also may indicate that a test taker did not pay full attention to the questions. A review of the Infrequency (INF) scale score may support this possibility.

Correlations With Other 16pf Factors

Because reasoning is typically seen as a separate domain from personality, Factor B has only very low correlations with the other 16pf factors. Its strongest correlation in the standardization sample is with Perfectionism (Q3; r = -.17).

Interpretation of Thinking Style Using Other Factors

Reasoning scale results may be helpful in considering broader intellectual style in combination with other factors such as Liveliness (F), Sensitivity (I), and Abstractedness (M). In particular, individuals who score in the Serious (F-) direction may work deliberately with new information, using a methodical and step-by-step method. In contrast, high scorers (F+) may be more impulsive and prone to seize their first thoughts in evaluating situations and reaching decisions. Additionally, a Utilitarian score (I-), which indicates a more logical and factual thinking style, can suggest a dry, unemotional analytical orientation, whereas higher Sensitivity scorers (I+) may employ their emotional reactions and intuitive and subjective information in evaluating information and making decisions. Finally, a Grounded (M-) profile combined with Reasoning (B) could indicate a tendency to focus on the immediate situation and the information available. High Abstractedness scorers, (M+), however, often can engage in far-ranging interpretation of information, perceiving possible connections to other ideas and situations. They may generalize about the meaning of a problem and possibly consider multiple interpretations. Such generalizations and interpretations may be more accurate among people with high Reasoning (B) scores. These higher-level combinations of factors can substantially enrich 16pf profile interpretation.

Step 5 Evaluate Primary Factor General Trends

To fully understand the 16pf Fifth Edition primary scales, testing professionals should not only study this chapter but also scale information presented elsewhere in this manual. For example, professionals should understand scale reliabilities, score distributions and standard errors of measurement (SEM), intercorrelations among the scales, as well as correlations with other measures. These data, which are presented in tables throughout this manual and in the appendices, are synthesized in the sections that follow; see Chapter 9 for more information about the construct validity results.

The interpretive information that follows is based on the body of evidence available for the Sixth Edition.

Broad Trends

In addition to examining the specific primary scale scores in a 16pf profile, testing professionals are encouraged to look at broad trends within the profile. One important consideration is evaluating the number of extreme scores.

It’s helpful to remember that the 16pf factors are bipolar. Individuals who score in the middle range are likely to be flexible and exhibit behaviors that represent either pole of the factor, depending on the situation. More extreme scorers’ actions are likely to be consistent and stable across time and circumstances. For example, an individual who scores in the average Warmth range (stens 5-6) mostly likely has a mixture of both personally close and impersonal and distant relationships. High scorers (A+ individuals) will tend to consistently seek out warm and emotionally intimate relationships with people. In contrast, low scorers (A- profiles) are prone to be selective and form close personal connections with only a few people.

Evaluate Number of Extreme Scores

As noted previously, extreme scores in a profile usually indicate a test taker’s most distinctive traits. Therefore, greater numbers of extreme scores are likely to indicate a more distinctive personality expression. Remember that fewer people in general tend to score in the low (stens 1-3) or high (stens 8-10) range than towards the middle.

Table 3.9 Number of Extreme Primary Factor Scores on 16pf Profiles

Number of extremes	Percent of sample	Percentile
0	4.4	2.2
1	9.7	9.3
2	11.6	19.9
3	11.9	31.6
4	11.6	43.4
5	11.7	55.0
6	9.0	65.3
7	8.0	73.8
8	6.9	81.3
9	5.4	87.4
10	4.1	92.2
11	2.9	95.7
12	1.5	97.8
13	0.9	99.0
14	0.3	99.6
15	0.2	99.9

Note: Standardization sample, N=2528.

Table 3.9 presents the number of extreme sten scores (those outside the 4–7 average range) obtained by the norm sample for the Sixth Edition. A test taker not having at least one extreme score is a rare occurrence. Most profiles show extreme scores on two to nine primary scales. If the number of extremes is ten or more, the test taker is among only about 10% of people whose profiles are this distinctive. If the number of extremes is below two, the test taker is among only about 14% of people whose profiles are flat. No profile in the norm sample had extreme scores on all 16 scales.

If the profile shows few extreme scores, the test taker possibly chose a large number of middle responses, indicating uncertainty about which response choice better described him or her. If the number of middle responses is not elevated, the test taker may have answered a given scale's questions inconsistently. In either case, the reasons for the "flat" profile can be pursued by the qualified user.

Remember the Primary Factor Scale Relationships–They Are CRITICAL to Effective Interpretation of Results

Because the 16pf instrument uses oblique factors (i.e., R. B. Cattell assumed that the primaries would be related), the structure of the 16pf tool shows that the scales are indeed intercorrelated. These intercorrelations are predictable: the primary scales cluster along the five Global Factors of Extraversion, Anxiety, Tough-Mindedness, Independence, and Self-Control.

With a knowledge of how certain scales are expected to intercorrelate, the testing professional can identify unexpected factor combinations, thus adding a richness beyond an evaluation that involves only a single factor at a time. In general, Primary Factor scale scores that cluster on a given Global Factor tend to be consistent; that is, a person who scores in the introverted direction on the Global Factor often tends to score in the introverted direction on the Primary Factor scales that make up Introversion (Reserved [A-], Seriousness [F-], Shyness [H-], Privateness [N+], and Self-Reliance [Q2+]). However, it is not uncommon that one of the primary scale scores will be in the extraverted direction, even when the person's score on the Global Factor falls in the introverted direction. For example, a generally introverted person might be Reserved (A-), Shy (H-), and Private (N+), but Group-Oriented (Q2-). (The latter is a score in the extraverted direction.) This person might be reserved and timid but wishing for more group contact, or the person might rely on group interactions to get "lost in the crowd" because of his or her reserve and timidity. Given the likelihood that this person experiences a conflict between the urge to be in groups and the tendency to be timid, the testing professional can generate a number of hypotheses about the person's orientation to people.

In evaluating a profile, then, how conflicting tendencies are played out should be considered, and hypotheses should be generated. Comparing the findings with other data about the test taker also can be helpful. Finally, in cases where findings are shared with the test taker, a discussion of conflicting patterns could be valuable.

References

Cattell, H. B. (1989). *The 16pf: Personality in depth*. Champaign, IL: Institute for Personality and Ability Testing.

Cattell, H. E. P., & Schuerger, J. M. (2003). *Essentials of 16pf assessment*. New York, NY: John Wiley and Sons.

Cattell, R. B. (1957). The conceptual and test distinction of neuroticism and anxiety. *Journal of Clinical Psychology, 13*, 221–233.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the 16pf*. Champaign, IL: Institute for Personality and Ability Testing, Inc.

Child, D. (1998). Some technical problems in the use of personality measures in occupational settings illustrated using the “Big-Five.” In S. Shorrocks-Taylor (Ed.), *Directions in educational psychology*. London, UK: Whurr Publishing.

Eysenck, H. J. (1960). *Handbook of abnormal psychology*. London, UK: Pitman.

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

Jung, C. G. (1971). Psychological types (H. G. Baynes, Trans. revised by R. F. C. Hull). In *The collected works of C. G. Jung (Volume 6).* Princeton, NJ: Princeton University Press. (Original work published in 1921).

Karson, S., & O’Dell, J. W. (1976). *A guide to the clinical use of the 16pf*. Champaign, IL: Institute for Personality and Ability Testing.

Krug, S. E. (1981). *Interpreting 16pf profile patterns*. Champaign IL: Institute for Personality and Ability Testing.

Chapter 4: Development of the 16pf Sixth Edition

Introduction

Since its initial publication in 1949, the 16pf Questionnaire has undergone five revisions (1956, 1962, 1967-1969, 1993, 2018). The most recent revision, which produced the 16pf Sixth Edition, had several goals:

- Update the item content to reflect modern language usage;
- Improve psychometric score properties with the shortest possible scales;
- Improve response efficiency by using a standardized response scale for all 16pf items (except Reasoning/B);
- Improve psychometric score properties and test security of Reasoning/B scale using computerized adaptive testing (CAT);
- Update normative data;
- Improve the Infrequency/INF response style index to better reflect inattentive responding; and
- Maintain compatibility with the existing factor structure, ideally sufficient to reuse existing Fifth Edition predictive equations with updated cut scores.

This chapter describes the development of the scales for the 16pf Sixth Edition, including the 16 Primary Factor scales, five global scales (i.e., second-order scales), similar to the "Big Five," and the response style indices.

Overview of Updates in the 16pf Sixth Edition

The 16 Personality Factor Questionnaire (16pf) represents R. B. Cattell's endeavor to identify the primary components of personality by factor analyzing all English-language adjectives describing human behavior. The 16pf Sixth Edition, although updated and revised, continues to measure the same 16 primary personality factor scales identified by R. B. Cattell nearly 70 years ago. Factor scales remain denoted by letters as assigned by R. B. Cattell, such as "Factor A," but they are also designated by more descriptive labels, such as "Warmth."

The broad personality domains are called "Global Factors" and indicate factors later popularized as the "Big Five" personality dimensions often summarized with the OCEAN acronym: Open-Mindedness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. In fact, authors of later personality inventories such as the NEO questionnaire have often used the 16pf Global Factors as reference variables in creating their own assessments (H. E. P. Cattell, 1996).

The 16pf Sixth Edition contains 155 items measuring the 15 primary personality factors and independent measures of Impression Management (IM), which assesses socially desirable responding, and Infrequency (INF), which measures inattentive responding. Each Primary Factor scale contains approximately 10 items. The IM and INF scales consist of six and five items. The 16th factor, Reasoning, is independently assessed by 10–20 adaptively administered items drawn from a large item pool. As in the previous edition, the Reasoning/B items are administered together at the end of the questionnaire.

The 16pf Sixth Edition can be administered individually or in a group setting and takes approximately 20–30 minutes to complete online. Readability statistics suggest that the questionnaire is easy to read; the Flesch-Kincaid grade level was 6.1 and the Flesch reading ease was 71.3%. The easy readability, shortened scales, and reduced administration time should ensure that the 16pf Sixth Edition is accessible to a wide range of populations and useful in settings where testing time is limited.

Like its predecessors, the Sixth Edition is computer scored by the publisher, PSI Services LLC. Computer-generated reports, as well as numerous source books and articles, are available to enrich the interpretation of assessment results.

A summary of new features and updates included in the Sixth Edition are presented below:

1. Item content has been revised to reflect modern language usage and to remove ambiguity.
2. Item content has also undergone statistical analysis and independent review to ensure freedom from gender, race, and cultural differences that might lead to score bias.
3. Psychometric properties have been improved. Internal consistency reliability estimates for the primary scales average .83, with a range from .72 to .90. Test–retest reliability estimates average about .85 for a 2-week interval and .83 for a 3-month interval. Reliability is discussed further in Chapter 8.
4. Scales have been shortened by about 9% and administration time by about 10%.
5. Response choices are consistent for all personality items, with a 5-choice Likert format of Strongly Disagree/Disagree/Neither Agree nor Disagree/Agree/Strongly Agree, thus providing a uniform response choice. The previous 16pf edition contained two answer choices that varied across questions plus a middle

response that was labeled "?" and thus could reflect several different reasons for not selecting either the "agree" or "disagree" alternative. Preliminary empirical analyses suggested that a Likert response scale provided considerably more information than the traditional response format.

6. The Reasoning/B scale is available in a Computer Adaptive Test (CAT) administration format that achieves higher reliability and more uniform information while being shorter. Traditional, static forms have also been prepared and are available from the publisher upon request. Development of the new Reasoning scale is discussed in Chapter 5.
7. Normative data have been updated to reflect the U.S. demographics based on recent Census information (the American Community Survey [ACS] 2015). The normative sample is discussed in Chapter 7.
8. Response style indices provide the same information as in the previous edition; however, Infrequency (INF) has been improved beyond merely measuring excessive middle responses to more directly measuring attentiveness. Development of these Response Style Indices is discussed in Chapter 6.
9. The 16pf factor structure had been maintained, and the sten scores of the 16pf Sixth Edition are compatible with existing equations. Validity information collected on previous editions should generalize to this form due to construct equivalence (but cut scores may need to be updated). Equivalency is discussed in Chapter 8, and validity evidence is presented in Chapters 9 and 10.

In addition to the above improvements, the Global Factor scales are defined in terms of the same primaries.

As a broad measure of personality, the 16pf Questionnaire is used to generate an assortment of reports that are useful in a variety of settings to predict a wide range of life behaviors. Human resources personnel consider the test a useful component of selection batteries and essential for personal and professional development planning. Vocational counselors find the links to occupational and other interests helpful in guiding clientele.

Development of the Primary Factor Scales

Development of the Primary Factor scales occurred in two parallel tracks. The first track developed the items of the personality and response style scales, whereas the item pool for the Reasoning/B items were developed in a parallel track.

Likert Item Pool

The plan for the revision selected and updated the "best items" from the 16pf Fifth Edition Questionnaire and then combined these with new items to create an updated form. Items had to meet the following criteria.

Items were indicative of the intended psychological construct.

Items from the Fifth Edition were adapted to a Likert format using clear, modern wording while changing the item meaning as little as possible. The adaptation to a Likert format required that the items be written as statements; for example, the hypothetical item "I would rather: a. be rich; b. be happy" would have been adapted as "I would rather be wealthy than happy" or possibly "I would rather be happy than wealthy." As an example of updating the language of the items, the phrase "I quite enjoy…" was changed to "I really enjoy…" These changes were informed by the experience of the team members and by reviews of the items by people unconnected to the revision, who flagged items that seemed odd in their content or wording. The words "minister" (i.e., member of the clergy) and "slapstick" (i.e., humorously embarrassing comedy) were examples of words flagged in these reviews. An item involving a choice between exercising by either dancing or fencing was identified as having overly narrow content.

The revision team also arranged for new items to be written. As a reference for item writers, a 62-page item-writing guide was developed to cover these topics: an introduction, information about the revision goals, 39 general and specific rules for Likert and ability items, detailed definitions of each scale, a chapter on distinguishing related primary factors, a Frequently Asked Questions list, and a glossary. Item writers were trained in groups of 2–6 authors, including practice writing and critiquing items.

A total of 19 authors participated in writing English-language items. Authors had diversity with respect to location (8 U.S., 5 U.K., 3 South Africa, 2 Australia, 1 China), identified gender (74% identified as women; 36% as men), age (ranged 23 to 64; mean 40.1), education (5 doctoral, 10 masters, 4 bachelors), identified "race" (16 identified as White, 2 as Asian, 1 as multiracial), experience using the 16pf (1 month to 25 years; mean 9.5 years), and practice area (12 in Selection and Development, 3 in Research, 4 in Other).

During item selection, all items selected had high corrected item-total correlations and most items correlated with their intended scale better than with other scales. In factor analysis of bundles (i.e., parcels of 3-5 items), all bundles loaded highest on their intended factor.

In an effort to ensure content validity, the content of each scale was broken into 3–5 "content clusters" and item writers were instructed to target specific content clusters. Special attention was given to ensuring that all important content areas for a factor were sampled. For example, several Warmth/A items about vocational preferences were replaced by new items that covered more directly the caring, interpersonal aspects represented by this content cluster. The aim was to construct cohesive factor scales that still tapped diverse content.

Table 4.1 presents a summary of modifications made to the final Sixth Edition items. Approximately 20% of the items are unchanged from the Fifth Edition, 17% contain minor changes, 1% have substantial changes, and 61% are new items.

Table 4.1 Source of 16pf Sixth Edition Items

Primary factor	Unchanged	Minor changes	Substantial changes	New
Warmth/A	--	3	1	6
Emotional Stability/C	1	1	--	8
Dominance/E	1	--	--	9
Liveliness/F	2	1	--	8
Rule-Consciousness/G	2	1	--	8
Social Boldness/H	4	2	--	2
Sensitivity/I	3	3	--	6
Vigilance/L	--	--	--	8
Abstractedness/M	3	1	--	6
Privateness/N	2	2	--	5
Apprehension/O	3	--	--	5
Openness to Change/Q1	1	1	1	8
Self-Reliance/Q2	2	5	--	1
Perfectionism/Q3	1	2	--	6
Tension/Q4	5	--	--	4
Impression Management/IM	1	5	--	--
Infrequency/INF	--	--	--	5
Total number	**31**	**27**	**2**	**95**
Percentage	**20%**	**17%**	**1%**	**61%**

Note: Unchanged includes one item where a comma was dropped. Minor changes were generally incidental wording or adaptation to a Likert format. Substantial changes addressed the same idea using different words. Percentages do not sum to 100 because of rounding error.

Items were short, simple, and unambiguous.

To achieve this criterion, wording was simplified, and sentences were shortened whenever possible. Items with awkward sentence structure were either rewritten or dropped. For example, the item "I like to join in with people who are doing something together such as going to a park or to a museum" (which had a true–false response) was shortened to "I like to join in with people who are doing something together."

The goal was to lower the overall reading level requirement of the test and to shorten test-taking time. The Flesch-Kincaid grade level of 6.1 and Flesch reading ease score of 71.3% indicate that the Sixth Edition is not difficult to read.

Dated or datable content should be removed and avoided.

Items with words or ideas that were outdated or that might become dated were rewritten or dropped. For example, the following Liveliness/F item was eliminated because of the out-of-date phrasing and reference to television in a video streaming era: "I greatly enjoy the racy and slapstick humor of some television shows." More generic terms, such as "video" or "news media," were used instead of specific terms like "television" or "newspapers," and we eliminated an item about being lost in a car because of the prevalence of mapping technology and the possibility that cars may be autonomous in the future.

Content that might suggest gender, race, or disability differences were avoided.

All gender-specific content (and content that might tap differential experience by gender) was avoided or removed from items to ensure that items worked well for all individuals. For example, the following Sensitivity/I item was removed: "I'm always interested in mechanical things and am pretty good at fixing them." In addition, reporting was modified to avoid gendered language, removing the need to know (or assume) a specific recipient gender. Importantly, items were reviewed to ensure that content associated with a wide variety of disabilities were eliminated.

In addition to the reviews described above, statistical analyses were performed to empirically investigate certain group differences. Standardized mean differences between protected group status (based on sex, age, and race) were calculated so that items could be chosen to minimize the possibility of group differences.

Items that are not easily translatable into other languages or cultures were avoided.

Because the 16pf Questionnaire has been translated into multiple languages, avoiding slang or content that was not easily translatable was important. For example, the following item was removed because of the colloquial phrasing: "In dealing with people, it's better to: a. 'put all your cards on the table;' b. 'play your hand close to your chest.'"

Material that might be considered intrusive, offensive, or otherwise unacceptable in a work setting was avoided.

The revision team was predominantly trained as organizational psychologists and avoided items about sexual, religious, or political behavior. For example, the following item was dropped on the basis of this criterion: "Teachers, ministers, and others spend too much time trying to stop us from what we want to do."

Content that is socially desirable or undesirable was avoided to reduce motivational distortion.

To meet this criterion, an intuitive approach was used in the original writing and rewriting of items, which led to dropping items such as the following: "I make smart, sarcastic remarks to people if I think they deserve it."

Toward the end of the revision, the intuitive process was checked empirically by calculating the correlation of each item with the Impression Management (IM) scale. Items that correlated unusually strongly with IM were subsequently dropped.

Item Evaluation

The entire development team collaborated in revising the existing Fifth Edition items into Likert form (all items had to be written as statements) and implemented wording changes. For eight items, variants were produced to evaluate alternative phrasing. Five items were dropped.

A total of 704 personality items were eventually written in two rounds of item writing. To double check that the items had been classified into primary factors correctly, after each round of item writing, three experienced 16pf users reclassified all items into factors. All items where these three individuals disagreed were discussed and either a group consensus was reached or the item was discarded. This process also afforded another chance to correct minor wording issues.

The first pilot test included 409 personality items and 32 response style items. The 409 items included most of the 158 prior items, rewritten to fit a Likert format, as well as over 250 new or rephrased items. Between 19 and 42 items (mean of 27.3) were included on each scale, including revised versions of all Fifth Edition items (except for Dominance/E, which only had seven original items) and all nine variant items. A total of N=477 participants were recruited from Amazon Mechanical Turk (MTurk) to complete the form. After a review of the data using excessive INF items and administration time extremes as decision points, N=407 individuals remained.

Items were screened by computing the corrected item-total correlation (CITC) of the item with the remaining Fifth Edition items. For example, form Warmth/A, 12 rewritten Fifth Edition Warmth/A items and 28 new items were correlated with the total defined by the original Fifth Edition items. The correction was that, for the original items, the total was computed leaving out the target item. Using a total defined by the original items helped ensure that the new items were being evaluated against the same constructs as measured by the Fifth Edition scales. In addition, the "cross-correlations" (the correlation of each item with the remaining 15 scales) were computed for all items. Items were selected by meeting statistical and content considerations. Statistically, items needed to have a high CITC that exceeded all cross-correlations with all other factors. Content considerations were evaluated during the item selection meeting by the entire development team and included length, wording, face validity, and redundancy with the other items. To accurately compare Fifth Edition reliabilities with the new values, a correction for the longer lengths of the pilot scales was applied. Although these corrections tended to slightly reduce the observed values of the pilot scales, mean scale reliability did increase from 0.76 for the Fifth Edition to 0.81 for the pilot forms.

The best items of the first pilot form were augmented with additional new items for a second round of pilot testing. A total of 440 items were included, with pilot scale lengths ranging from 21 to 39 items with a mean of 29 items. A total of N=562 participants were recruited from MTurk. After data cleaning, N=477 cases remained for analysis. The same item analysis and review steps were repeated, but the goal of this final review was to select 15-item scales for Form S, the assessment used in the standardization (norming) sample. Form S was designed with longer scales to allow for final item selecting using the normative sample. Although items derived from the original Fifth Edition were favored where they were statistically and conceptually equivalent to new items, there was no requirement to include all original items. Thus, many original items were replaced in this step simply because there were better new items. Additionally, care was taken to include approximately the same number of positively and negatively worded items on the standardization form. Length-corrected mean scale reliability rose from 0.81 for the first pilot form to 0.82 for the standardization form.

The final step was a sensitivity review by a diverse group of 12 consulting professionals including American, British, and Australian reviewers who represented African American, Asian, Hispanic/Latino, and White/Caucasian ethnicities; one reviewer was consulted on LGBTQ issues. Each item was rated on six principles of fairness (Treating People with Respect; Minimizing Irrelevant Knowledge; Sensitive Topics; Avoidance of Stereotypes; Appropriate Labels for Groups; and Representation of Diversity) and provided written notes. A total of six items were removed due to this sensitivity review. An example was a Vigilance/L item: "A lot of people will 'stab you in the back' in order to get ahead," which was seen as violating the principles of Minimizing Irrelevant Knowledge and Sensitive Topics. Additionally, items were flagged if even a single member of the sensitivity panel disliked the item in any way and these flagged items were avoided during final item selection.

Final Item Selection

The standardization form, Form S, was administered to a normative sample (N=2,528; described in Chapter 7), and final item selection for the operational 16pf Sixth Edition occurred in a series of team meetings, again balancing statistical and content considerations. The selection criteria were:

- No fewer than eight items in a scale;
- Maximize corrected item-total correlations (CITC);
- Minimize the highest cross-loading;
- Select a mixture of content clusters;
- Minimize redundancy within a scale (avoid having substantially the same item repeated within a scale);
- Maximize reliabilities (ideally reliability > 0.80);
- Shorten scales (but no fewer than eight items on any scale);
- No offensive items; minimize use of items flagged in the sensitivity review;
- Maximize equivalency with the corresponding Fifth Edition scale; and
- Approximately balance positively and negatively worded items.

Final item selection was based on these 10 criteria. Scales were reduced in length about 9% from 170 personality items to 155 items. The Impression management (IM) scale was reduced from 12 items to 6, but five new Infrequency (INF) items were added for a net savings of 8%. In addition, Reasoning (B) contains a variable number of adaptively administered items, as described in Chapter 5.

IRT Analyses of the Likert Responses

Likert's (1932) scoring simply assigns integers to the response points. For positively worded items, "Strongly Disagree" = 1, "Disagree" = 2, "Neutral" = 3, "Agree" = 4, and "Strongly Agree" = 5; for negatively worded items, scoring is reversed: "Strongly Agree"=1, "Agree" = 2, and so on. Using consecutive integers implies that the "psychological meaning" of each response is ordered (the responses exhibit "ordinality") and that the "psychological distance" between each response point is equal (the responses exhibit "interval" properties or "intervalness"). In contrast, Thurstone's method (Thurstone & Chave, 1929) involved estimating weights for each response, which might not be ordinal or interval. Some have questioned the ordinality of the traditional 16pf responses (scored 0/1/2), arguing that the middle response should always be scored as being in between the other two responses (i.e., 1 is not always an appropriate score for the middle, "?," response; Hernandez, Drasgow, & Gonzalez-Roma, 2004; Murray, Booth, & Molenaar, 2016).

The purpose of item response theory (IRT) analyses was to scale the Likert items using modern latent trait theory. IRT estimates of item location and loading are useful in their own right and demonstrate the quality of the scales in terms of modern latent trait theory. The model used in these analyses also assumes ordinality (that "Strongly Disagree" < "Disagree" < "Neutral" < …) and thus provide evidence of the Likert scoring model.

A generalized form of Samejima's (1969) Graded Response Model (GRM) was fit to the Likert item responses of each personality scale using Parscale (Muraki & Bock, 2002) with default settings. This model provides a "slope" and a "difficulty" or "location" parameter for each item, as well as locations of the "thresholds" of 2PL curves distinguishing each response. Because we collected data on a 5-point Likert scale, four threshold locations are estimated. The first threshold, t1, distinguishes between Strongly Disagree and Disagree (or above), the second threshold, t2, distinguishes between Disagree and Neutral (or above), and so forth.

Note that in this generalized model, Samejima's GRM thresholds can be obtained by summing these generalized thresholds with the overall location. That is, Samejima's b1 for item Warmth/A item L1 is -0.36 - 2.30 = -2.66. Note that this implies that these generalized thresholds have the *opposite* sign as compared to Samejima's thresholds; negative thresholds indicate "high" or "right" standing on the latent trait and positive thresholds represent "low" or "left" standing.

All items of each scale were fit in such a way that all parameters (the slope, overall location, and the threshold locations) were free to vary (in Parscale terminology, each item had its own estimation "block"). These parameter estimates are shown in Appendix B. Model-data fit was adequate.

Note that the GRM assumes (and enforces) ordinality (thresholds are estimated in such a way that t1 > t2 > t3 > t4), so this analysis is not a test of ordinality. But the fact that this model (which assumes ordinality) fits well is fairly strong evidence that ordinality holds. Also, the obtained estimates of the thresholds suggest ordinality (there is no instance where two thresholds are very close to each other; i.e., for all estimates t1 >> t2 >> t3 >> t4).

In summary, this IRT analysis provides detailed item-level statistics, including the "slope" for each item and "locations" corresponding to the points on the 5-point Likert scale. An examination of these results reinforces the other analyses that suggest that the items are of good quality or above. Also, the fit of the GRM IRT Model provides evidence that the 5-point Likert responses have ordinal properties.

Factor Analyses and Development of the Global Factor Scales

Factor analysis was always a primary tool in 16pf research. Beginning with R. B. Cattell's original model, the 16pf primary factors were intercorrelated. These relationships led to the exploration of a higher-order factor structure and to the discovery that small clusters of the primary scales comprise "second-order" factors of personality, similar to the "Big Five." In the Fifth Edition, these factors were first described as global to better reflect the broad personality domains that they represent. This section describes analyses of the factors and the development of the global factor scores. The revision was successful in avoiding changes to the second order factor structure (as demonstrated by the analyses described in this section), and thus the same primary scales indicate the same second-order, "Global Factor" scales.

Exploratory Factor Analysis of Parcels

A Maximum Likelihood Exploratory Factor Analysis was performed after the final item selection. Parcels, or item bundles, which are small groupings of the items within a scale, were factor analyzed instead of individual items because parcels have been found to be more reliable (Bernstein & Teng, 1989; R. B. Cattell & Burdsal, 1975; Gorsuch, 1983). Three parcels were created for each primary scale by first ranking items by their ITC (item-test correlation) and applying a spiral assignment method. A total of 48 item bundles were generated, and most of them consists of 3 or 4 items. The scores of the items composing each parcel were averaged to achieve a parcel score.

Sixteen factors were extracted, explaining 77.64% of the variance. These factors were rotated using the Direct Oblimin method with Delta=0. Table 4.2 presents the factor pattern from the analysis of the standardization test form. The resulting factor solution showed a hyperplane count of 79.8% for loadings of .05 or less (absolute value) and 90.1% for loadings of .10 or less (absolute value). The hyperplane count represents the number of factor loadings that are close to zero. In this case, the percentages of loadings between −.05 and .05 as well as between −.10 and .10 are given. R. B. Cattell (1952, 1966) stressed the use of hyperplane count as an analytical criterion when evaluating a factor pattern.

Overall, the pattern showed a good simple structure for the 16pf primary factors, with all of the 48 parcels having a loading of .40 or greater on one and only one factor. The highest cross-loadings were shown in three C parcels that they loaded relatively high on Factor O but not as strongly as on Factor C. Overall, these results provided strong support for the basic factor structure of the 16pf Sixth Edition.

Table 4.2 Rotated Factor Pattern Loadings of 16pf Primary Factors

Parcel	*16pf primary factor*															
	A	**B**	**C**	**E**	**F**	**G**	**H**	**I**	**L**	**M**	**N**	**O**	**Q1**	**Q2**	**Q3**	**Q4**
A1	**-72**	01	-03	05	04	02	03	08	03	03	-09	-05	-09	05	01	-03
A2	**-64**	-02	08	00	14	06	-01	-01	05	-04	-08	06	01	03	00	-12
A3	**-61**	00	-04	01	-06	06	05	06	01	00	02	-01	23	08	04	-02
B1	00	**71**	00	-02	07	02	-02	01	01	01	-02	01	01	-06	-02	03
B2	-02	**70**	-04	-01	-03	-05	03	01	-02	00	02	-01	00	03	02	-01
B3	01	**63**	03	02	-05	01	-01	-03	01	-02	00	00	-01	03	-02	-01
C1	-03	04	**46**	03	07	02	07	-04	02	-13	-01	-29	12	-04	02	-07
C2	02	00	**52**	05	05	09	03	-03	05	-12	00	-21	07	03	00	-10
C3	02	01	**51**	04	01	04	03	-04	06	-10	03	-20	02	06	03	-15
E1	-04	00	-01	**78**	01	01	05	-01	-02	-07	-01	-03	02	-01	02	04
E2	01	-02	03	**79**	-04	00	01	04	-02	01	01	02	-01	03	-01	-03
E3	-01	01	-02	**78**	08	-03	-01	-04	03	02	-03	-03	06	-01	02	04
F1	-01	01	06	05	**65**	-02	08	11	02	03	-04	00	00	08	00	-01
F2	-07	-05	-01	-01	**65**	-03	06	01	02	01	01	-02	12	13	00	00
F3	-05	-01	-01	14	**46**	-05	16	-02	02	06	-10	01	03	10	00	01
G1	-09	00	06	-02	-07	**78**	00	00	-01	-01	02	07	01	04	00	03
G2	04	-05	-04	01	00	**86**	02	-02	-01	01	00	-04	01	00	01	00
G3	01	04	00	-01	05	**76**	00	00	05	-03	-03	-01	-07	-04	02	-03
H1	00	00	00	-03	-03	00	**96**	02	-01	-01	-02	-02	03	02	01	01
H2	-03	-03	01	01	11	01	**66**	-02	06	-06	-05	01	03	05	01	-05
H3	01	01	02	27	11	02	**54**	-04	01	04	-06	-04	-05	-02	-01	-01
I1	-10	-07	05	-01	-04	-05	02	**66**	03	00	-01	04	-08	-05	-05	03
I2	01	00	-04	02	05	04	02	**69**	00	01	02	-01	15	-02	05	-02
I3	05	07	-04	01	05	01	-04	**84**	-03	04	-03	-03	03	07	01	-03
L1	-02	01	01	-02	-01	-01	01	-01	**-79**	03	-01	03	02	00	-01	-01
L2	01	00	-02	00	05	-01	-03	00	**-88**	-02	-05	01	-02	-03	-02	00
L3	01	-02	02	04	-04	00	01	01	**-71**	-01	04	-03	-01	00	03	02
M1	-01	-01	-08	00	04	-03	-07	-02	-05	**72**	00	-03	02	-04	-03	01
M2	02	-05	01	-04	02	-02	00	04	00	**72**	-01	-02	01	05	-04	-01
M3	-01	05	02	01	-03	00	03	-01	02	**81**	01	05	00	-04	02	01
N1	17	05	08	02	-09	00	-05	-04	-03	11	**62**	-03	01	04	-02	04
N2	-01	-03	-03	01	00	-01	01	02	00	-06	**84**	00	-03	-04	-01	-06
N3	-03	02	-01	-05	05	01	-04	-01	00	01	**80**	03	00	-04	01	05
O1	06	01	00	01	-02	01	00	04	-06	01	-01	**84**	-03	00	00	-01
O2	-04	-03	-16	-12	03	-02	01	-02	-12	09	04	**54**	07	-01	-01	04
O3	-04	00	-07	-02	00	01	-08	-03	04	02	01	**70**	03	-03	01	06
Q11	02	09	05	12	01	-02	00	05	05	03	-05	06	**67**	00	01	-07
Q12	-09	-07	02	02	02	-06	03	-01	-01	03	-02	-03	**78**	-01	00	03
Q13	02	00	-02	-02	07	-02	04	07	-02	02	01	-01	**70**	04	-03	-03
Q21	00	-03	-02	03	-05	02	01	01	00	04	00	-02	04	**-81**	02	-01
Q22	00	03	01	-04	00	-03	00	01	-02	-03	05	01	00	**-78**	-04	05
Q23	05	02	00	-02	-01	-01	-05	-02	-03	-01	01	03	-05	**-73**	01	-02
Q31	00	01	04	01	03	-04	02	-02	02	03	-01	-02	02	-02	**85**	-04
Q32	-04	00	05	-01	-06	00	-01	00	01	-03	00	06	-06	01	**76**	07
Q33	03	-03	-09	01	04	07	00	02	-03	-03	01	-04	02	01	**73**	-04
Q41	12	02	06	-02	-01	-02	-01	-02	01	00	-03	-01	-09	00	-01	**71**
Q42	-08	02	00	02	-02	00	-01	04	-06	04	06	15	01	-06	03	**63**
Q43	01	-02	-11	02	03	00	-01	-01	-04	00	01	-04	05	00	-02	**83**

Note: Standardization sample, N=2,528. Values shown to two decimal places; decimal point omitted. A=Warmth, B=Reasoning, C=Emotional Stability, E=Dominance, F=Liveliness, G=Rule-Consciousness, H=Social Boldness, I=Sensitivity, L=Vigilance, M=Abstractedness, N=Privateness, O=Apprehension, Q1=Openness to Change, Q2=Self-Reliance, Q3=Perfectionism, Q4=Tension.

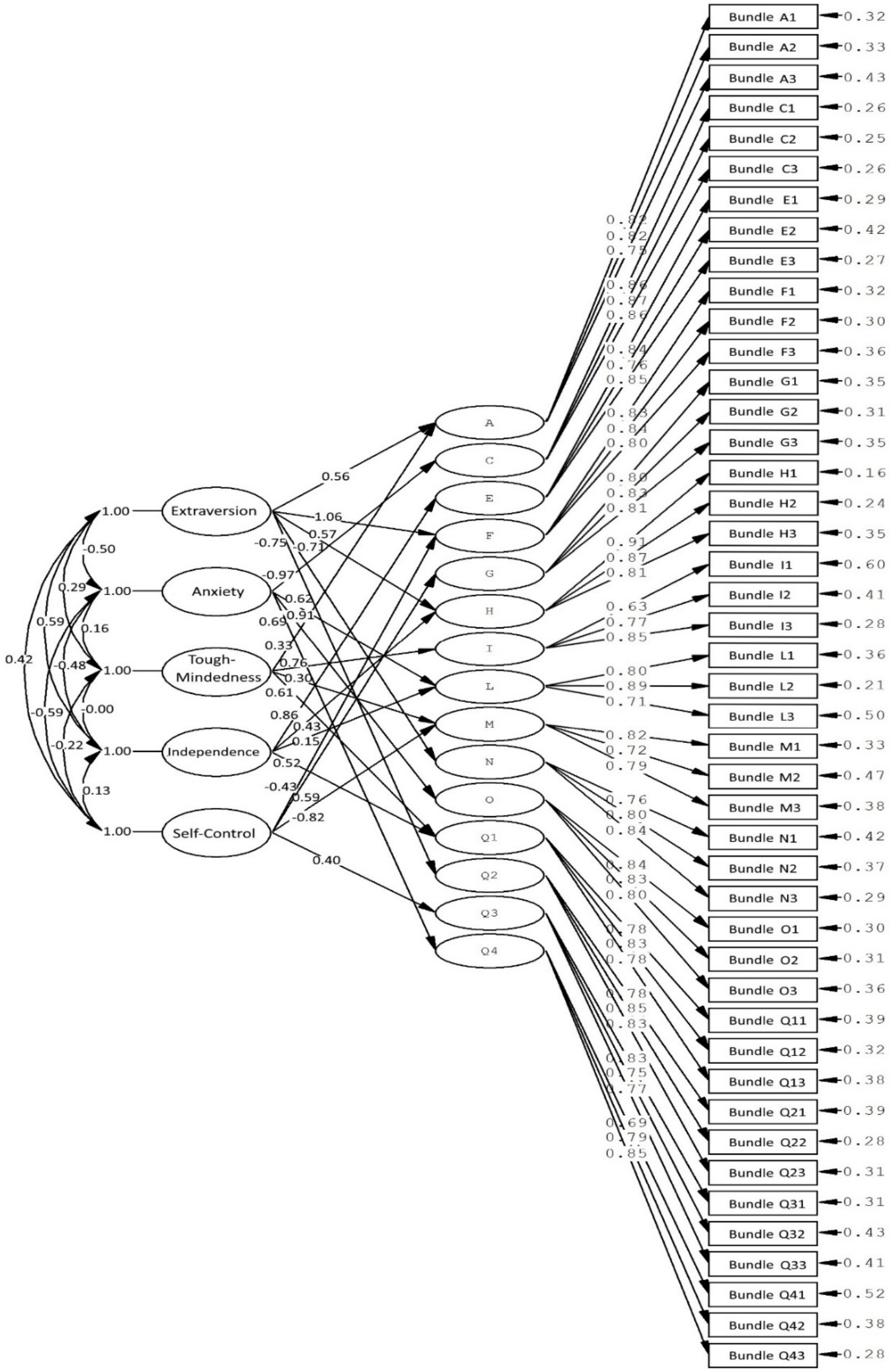
Confirmatory Factor Analyses

To define the Global Factor scales for the Sixth Edition, two confirmatory factor analysis models were tested using the standardization sample (N=2,528). Primary scale Reasoning/B does not load on any global factor, and thus was not included in these CFA models.

In the first model, shown in Figure 4.1, the same parcels used in previous EFA were re-used as indicators of primary factors with global factors being the higher-order factors. Model fit for the Fifth Edition second-order factor structure, evaluated using RMSEA and CFI, was good (RMSEA = 0.06, CFI=0.97). These results replicated the traditional second-order factor structure. For example, Warmth/A, Liveliness/F, Social Boldness/H, Privateness/N (negatively), and Self-Reliance/Q2 (negatively) loaded on the global factor Extraversion. The current model showed very good fit to the data, supporting the factor structure found for the earlier 16pf editions. The path coefficients shown in Figure 4.1 are fully standardized coefficients.

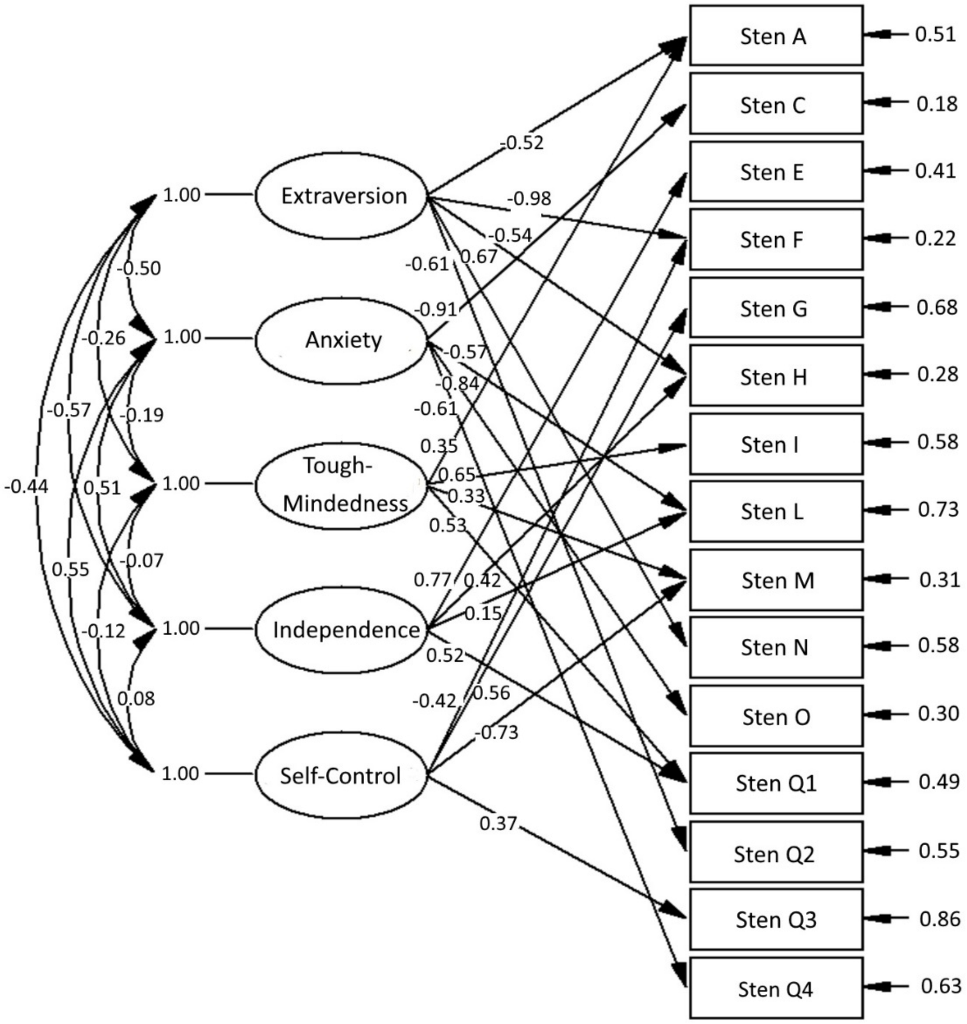
To better define the global factor scores, a second model (shown in Figure 4.2) was fit using sten scores as the indicators of the global factors using the traditional factor structure. This model fit adequately (RMSEA = 0.12, CF = 0.90). Figure 4.2 shows the fully standardized path coefficient estimates. Note that as shown by the sign of coefficients estimates between primary scales and Global Factors, Extraversion, Anxiety, and Tough-Mindedness came out in the opposite direction as theoretically defined. For example, higher Extraversion/EX scores indicate introversion and lower scores indicate extroversion. Similarly, Factors A, F, and H have negative loadings and N and Q2 have positive loadings. This is due to the factor score indeterminacy, which does not affect fit or the definition of the global factor scores (the “direction” of these bipolar scales is arbitrary).

Figure 4.1 Path Diagram for Item Bundles Predicting Primary and Global Factors



Note: A=Warmth, C=Emotional Stability, E=Dominance, F=Liveliness, G=Rule-Consciousness, H=Social Boldness, I=Sensitivity, L=Vigilance, M=Abstractedness, N=Privateness, O=Apprehension, Q1=Openness to Change, Q2=Self-Reliance, Q3=Perfectionism, Q4=Tension. Reasoning item bundles are not included in the model. Note Tough-Mindedness is actually Accommodation.

Figure 4.2 Path Diagram for Primary Factors Predicting Global Factors



Note: A=Warmth, C=Emotional Stability, E=Dominance, F=Liveliness, G=Rule-Consciousness, H=Social Boldness, I=Sensitivity, L=Vigilance, M=Abstractedness, N=Privateness, O=Apprehension, Q1=Openness to Change, Q2=Self-Reliance, Q3=Perfectionism, Q4=Tension. Reasoning scale was not included in the model. Extraversion, Anxiety and Tough-Mindedness in the model came out in the opposite direction.

Traditionally the global factor scores are weighted combinations of the primary scale sten scores. In addition, the global factor equations need to be standardized to produce scores with a sten distribution (mean 5.5 and standard deviation 2.0). The raw (not standardized) coefficients of the second CFA model, shown in Table 4.3, were used as provisional and unstandardized equations. The scoring of Extraversion, Anxiety, and Tough-Mindedness was reversed (so that these equations calculated the appropriate scores).

Table 4.3 CFA Raw Coefficient Estimates of 16pf Primary Factors Predicting Global Factors

	Global Factor				
Primary Factor	**Extraversion**	**Anxiety**	**Tough-Mindedness**	**Independence**	**Self-Control**
Warmth/A	1.03	--	-0.69	--	--
Reasoning/B	--	--	--	--	--
Emotional Stability/C	--	-1.79	--	--	--
Dominance/E	--	--	--	1.51	--
Liveliness/F	1.95	--	--	--	-0.83
Rule-Consciousness/G	--	--	--	--	1.13
Social Boldness/H	1.09	--	--	0.84	--
Sensitivity/I	--	--	-1.29	--	--
Vigilance/L	--	1.10	--	0.28	--
Abstractedness/M	--	--	-0.63	--	-1.39
Privateness/N	-1.27	--	--	--	--
Apprehension/O	--	1.65	--	--	--
Openness to Change/Q1	--	--	-1.05	1.02	--
Self-Reliance/Q2	-1.29	--	--	--	--
Perfectionism/Q3	--	--	--	--	0.71
Tension/Q4	--	1.21	--	--	--

Note: Standardization sample, N=2,528. The sign of the loadings of Extraversion, Anxiety, and Tough-Mindedness are reversed so that the scales indicate the traditional global scale.

Scores for each of the five global factor scales were calculated for each case in the N=2,528 standardization sample and a linear transformation was calculated to transform those scores to the sten metric. This transformation was then applied to the provisional equation to produce an equation that produces sten scores with a mean of 5.5 and a standard deviation of 2.0. Traditionally, global factor sten scores are rounded to the nearest tenth and truncated to range from 1 to 10. The final, standardized global factor equations are shown in Table 4.4. The Global Factor scale equations of the Fifth Edition are also shown in Table 4.4 for comparison and show a fairly high degree of similarity.

Table 4.4 Global Factor Scale Equations

Scale	Scaled Equation					
(Sixth Edition)						
Extraversion	3.87	+0.20A	+0.38F	+0.21H	-0.25N	-0.25Q2
Anxiety	2.90	-0.39C	+0.24L	+0.36O	+0.26Q4	
Tough-Mindedness	13.54	-0.28A	-0.52I	-0.25M	-0.42Q1	
Independence	-2.03	+0.57E	+0.32H	+0.11L	+0.38Q1	
Self-Control	6.29	-0.32F	+0.44G	-0.54M	+0.27Q3	
(Fifth Edition)						
Extraversion	4.40	+0.30A	+0.30F	+0.20H	-0.30N	-0.30Q2
Anxiety	1.60	-0.40C	+0.30L	+0.40O	+0.40Q4	
Tough-Mindedness	13.80	-0.20A	-0.50I	-0.30M	-0.50Q1	
Independence	-2.20	+0.60E	+0.30H	+0.20L	+0.30Q1	
Self-Control	3.80	-0.20F	+0.40G	-0.30M	+0.40Q3	

Note: Global Factor scale equations were scaled using the Sixth Edition Standardization sample (N=2,528).

These Sixth Edition global factor equations are quite similar to those found for earlier 16pf editions (R. B. Cattell, Eber, & Tatsuoka, 1970; Krug & Johns, 1986; IPAT, 1991, 2009). Descriptions of the five Global Factor scales and their contributing Primary Factor scales are presented below. Descriptions of the scales are also shown in Table 4.5.

Extraversion

Primary Factor scales having high loadings on the Extraversion Global Factor are Warmth (A), Liveliness (F), Social Boldness (H), Privateness (N), and Self-Reliance (Q2). Both Privateness (N) and Self-Reliance (Q2) are negatively weighted to represent the more personally open and group-oriented aspects of Extraversion.

Anxiety

The Sixth Edition Anxiety Global Factor contains the same combination of primary scales shown in earlier factor analyses: Emotional Stability (C), Vigilance (L), Apprehension (O), and Tension (Q4). Emotional Stability is negatively weighted; less emotional stability is characteristic of more anxious individuals.

Tough-Mindedness

Primary scales having high loadings on the Tough-Mindedness Global Factor are Warmth (A), Sensitivity (I), Abstractedness (M), and Openness to Change (Q1). In Second through Fourth 16pf editions, this scale was called Tough Poise. Its renaming in the Fifth Edition reduced the confusion and awkwardness in interpreting the concept of Tough Poise. The title of Tough-Mindedness reflects the prominent contribution of

Sensitivity (I), which was defined as "tough-minded" at the low end on the Fourth Edition. This indicates reliance in logic and fact versus feelings and intuitions in making decisions.

Table 4.5 16pf Factor Names and Descriptors

Descriptors of Low Range	*Factor*	*Descriptors of High Range*
Introverted, socially inhibited	**Extraversion**	**Extraverted, socially participating**
Reserved, impersonal, distant	Warmth (A)	Warm, Outgoing, Attentive to Others
Serious, restrained, careful	Liveliness (F)	Lively, animated, spontaneous
Shy, threat sensitive, timid	Social Boldness (H)	Socially bold, venturesome, thick skinned
Private, discreet, nondisclosing	Privateness (-N)	Forthright, genuine, artless
Self-reliant, solitary, individualistic	Self-Reliance (-Q2)	Group oriented, affiliative
Low anxiety, unperturbable	**Anxiety**	**High anxiety, perturbable**
Emotionally stable, adaptive, mature	Emotional Stability (-C)	Reactive, emotionally changeable
Trusting, unsuspecting, accepting	Vigilance (L)	Vigilant, suspicious, skeptical, wary
Self-assured, unworried, complacent	Apprehension (O)	Apprehensive, self-doubting, worried
Relaxed, placid, patient	Tension (Q4)	Tense, high energy, impatient, driven
Receptive, open-minded, intuitive	**Tough-Mindedness**	**Tough-minded, resolute, unempathic**
Warm, Outgoing, Attentive to Others	Warmth (-A)	Reserved, impersonal, distant
Sensitive, aesthetic, sentimental	Sensitivity (-I)	Utilitarian, objective, unsentimental
Abstracted, imaginative, idea oriented	Abstractedness (-M)	Grounded, practical, solution oriented
Open to change, experimenting	Openness to Change(-Q1)	Traditional, Attached to Familiar
Accommodating, agreeable, selfless	**Independence**	**Independence, persuasive, willful**
Deferential, cooperative, avoids conflict	Dominance (E)	Dominant, forceful, assertive
Shy, threat sensitive, timid	Social Boldness (H)	Socially bold, venturesome, thick-skinned
Trusting, unsuspecting, accepting	Vigilance (L)	Vigilant, suspicious, skeptical, wary
Traditional, Attached to Familiar	Openness to Change (Q1)	Open to change, experimenting
Unrestrained, follows urges	**Self-Control**	**Self-controlled, inhibits urges**
Lively, animated, spontaneous	Liveliness (-F)	Serious, restrained, careful
Expedient, nonconforming	Rule-Consciousness (G)	Rule conscious, dutiful
Abstracted, imaginative, idea oriented	Abstractedness (-M)	Grounded, practical, solution oriented
Tolerates disorder, unexacting, flexible	Perfectionism (Q3)	Perfectionistic, organized, self-disciplined
Lower general mental capacity, less intelligent, concrete thinking	Reasoning (B)	Higher general mental capacity, more intelligent, bright, abstract-thinking

Note: Global Factors are in Bold font. Primary scales are shown along with corresponding global factor(s). (-) indicates the reversed primary scales. Reasoning/B does not load on any global factor.

Independence

The primary scales having the highest loadings on Independence in the Sixth Edition are the same as those in the Fifth Edition. These primaries are Dominance (E), Social Boldness (H), Vigilance (L), and Openness to Change (Q1). All primary factors are scored in the positive direction.

Self-Control

The Self-Control Global Factor denotes this scale's focus on the control of one's own thoughts, feelings, and behaviors rather than by responding to others. Primary scales having high loadings on Self-Control are Liveliness (F), Rule-Consciousness (G), Abstractedness (M), and Perfectionism (Q3). The Liveliness and Abstractedness primary factors are negatively weighted to indicate that high Self-Control individuals tend to be deliberate and cautious as well as prone to focus on practical and solution-oriented matters.

References

American Community Survey (2015). http://proximityone.com/acs2015.htm

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 467–477.

Cattell, H. E. P. (1996). The original big five: A historical perspective. *European Review of Applied Psychology, 46(*1), 5-14.

Cattell, R. B. (1952). *Factor analysis*. New York, NY: Harper and Brothers.

Cattell, R. B. (1966). *Handbook of multivariate experimental psychology*. Chicago, IL: Rand McNally.

Cattell, R. B., & Burdsal, C. A. (1975). The radial parcel double factoring design: A solution to the item-vs.-parcel controversy*. Multivariate Behavioral Research, 10*, 165–179.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the 16pf*. Champaign, IL: Institute for Personality and Ability Testing, Inc.

Gorsuch, R. L. (1983). *Factor Analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hernandez, A., Drasgow, F., & Gonzalez-Roma, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology, 89*, 687-699.

IPAT (1991). *Administrator's manual for the 16pf*. Champaign, IL: Institute for Personality and Ability Testing.

IPAT (2009). *16pf Fifth Edition Questionnaire Manual*. Champaign, IL: Institute for Personality and Ability Testing.

Krug, S. E., & Johns, E. F. (1986). A large-scale cross-validation of second-order personality structure defined by the 16pf. *Psychological Reports, 59*, 683–693.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 5-55.

Muraki, E., & Bock, D. (2002) *PARSCLE 4.1 Computer program*. Chicago, IL: Scientific Software International, Inc.

Murray, A. L., Booth, T., & Molenaar, D. (2016). When middle really means "top" or "bottom": An analysis of the 16pf5 using bock's nominal response model. *Journal of Personality Assessment, 98*, 319-331.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society.

Thurstone, L. L., & Chave, E. J. (1929). The measurement of attitude. Chicago, IL: University of Chicago Press.

Chapter 5: Development and Validation of the Adaptive Reasoning Scale

Introduction

This chapter describes the development of a new and improved Reasoning/B scale, including the nature of the Reasoning/B scale, the goals of the revision, the item writing process, the psychometric work, design of the Computerized Adaptive Testing (CAT), and norming. The result is a significantly improved Reasoning/B scale.

Nature of the Reasoning Scale

As a brief measure of mental ability, the Reasoning (Factor B) scale is unique among the scales of the 16pf instrument. Factor B is included as a supplementary measure to the 16pf personality scales because reasoning ability is such an important dimension in individual differences (R. B. Cattell, Eber, & Tatsuoka, 1970). R. B. Cattell developed the Factor B scale because "the undoubted influence of general ability upon some personality variables... invites a variety of hypotheses" (R. B. Cattell, 1957, p. 873). For instance, he noted that intelligence directly aids certain personality development, such as the growth of conscientiousness.

The Reasoning/B scale measures both fluid and crystallized intelligence (R. B. Cattell, 1963, 1971; Horn & R. B. Cattell, 1966). Fluid intelligence is a general ability to recognize patterns and learn novel stimuli. Many measures of fluid intelligence are nonverbal (although reasoning on verbal items may also tap fluid intelligence). In contrast, crystallized intelligence refers to reasoning using accumulated knowledge and acquired skills.

In practice, the Reasoning scale has been interpreted as a brief measure of mental ability. H. B. Cattell (1989) defines the ability measured by Reasoning as "the capacity to discern relationships in terms of how things stand, relative to one another" (p. 30). She suggests that, in a clinical setting, a Reasoning score is useful in determining whether a respondent would benefit from insight therapy or another form of treatment. H. B. Cattell also notes that a Reasoning score can help predict how an individual's traits are likely to be expressed; that is, a Reasoning score can serve to moderate the interpretation of a given personality trait score. An example of this is how a Reasoning score could moderate the interpretation of an individual's Warmth (A) score: A lower Reasoning score and a high Warmth score indicate that the individual would be more likely to be "duped" by a con artist than if reasoning ability was higher.

The other major interpretation of Reasoning/B is in terms of the diagnostic value of low scores. For instance, low scores may indicate reading difficulties, lack of attention, misunderstanding of instructions, or test sabotage. Karson and O'Dell (1976) state that the Reasoning score can also be a good indicator of an respondent's attentiveness while completing the 16pf Questionnaire. For example, a low score for a college student would likely indicate inattentiveness. H. B. Cattell also notes that in many instances, average or low scores might not reflect an respondent's actual intellectual ability. Such scores might be obtained by respondents who are educationally disadvantaged, who are depressed or anxious, who did not allocate enough time, or who are completing the assessment in a non-native language.

Revision Goals

The overarching goal of the revision was to improve the Reasoning/B scale to make it more useful for practitioners and researchers. The Reasoning/B scale was different from other 16pf scales in three ways. First, as with any assessment of cognitive ability, the items had correct and incorrect responses, and were scored dichotomously as correct and incorrect. Correct responses indicated both knowledge of the correct response and guessing. As a result of the dichotomous scoring and the presence of guessing, each item produces less information than a personality item and this is manifest in a lower degree of reliability. The Fifth Edition Reasoning/B scale was the longest on the questionnaire at 15 items, but it had modest reliability.

In addition, the Reasoning/B scale could be compromised. Whereas the personality scales did not have "correct" answers, the items and correct answers to this scale could be shared. This became a significant issue with the rise of the Internet. Searching for the text of Fifth Edition items often produced "hits" showing the item and its correct answer. In some cases, the item could be compromised because it relied upon common information (e.g., series items based on common numerical progressions could be compromised by educational mathematics websites that present these progressions). The risk of compromise was exacerbated by the short, static nature of the 16pf Fifth Edition Reasoning/B form.

The increasing use of unproctored, online testing also raised the risk that respondents would use other resources, such as dictionaries or calculators. Clearly, items based on the difficult vocabulary would be severely affected if the respondent has access to a dictionary or thesaurus. The difficulty of series items and other items depending on mathematical computations might be seriously compromised by the use of calculators. Both of these kinds of resources (dictionary/thesaurus and calculators) are commonly available on smartphones.

Finally, the Reasoning/B scale also required an inordinate amount of time. Estimates varied but completing a single Reasoning/B item most likely takes four to five times longer than a personality item. Thus, albeit "short," the Reasoning/B scale probably required 26% to 31% of total testing time, which is disproportionate to its compromising only 6% of the 16pf scores (i.e., 1 out of 16 primary scale scores).

These concerns cannot wholly be abated. State of the art assessments of cognitive ability have time-consuming, dichotomously scored items that are subject to compromise, but adaptive testing was chosen because it would improve reliability while shortening scale length and would select items from a large pool of items that could more easily be replenished. In addition, some of the new items are graphical, which may be more robust to compromise.

Some applications, such as paper-and-pencil testing and possibly some translations (e.g., where it is infeasible to translate the entire CAT pool), will still use static forms. To accommodate such uses, 20-item forms composed of pool items have been prepared.

Development of the Revised Reasoning Scale

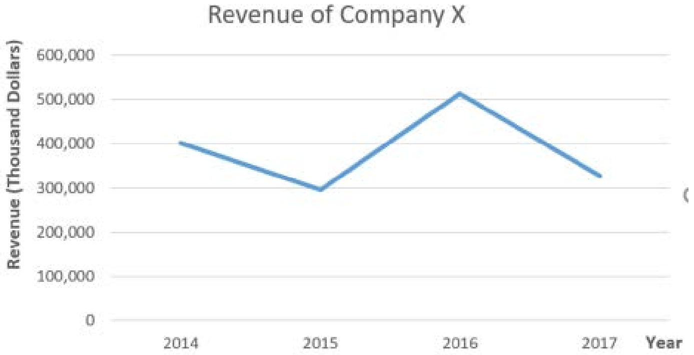
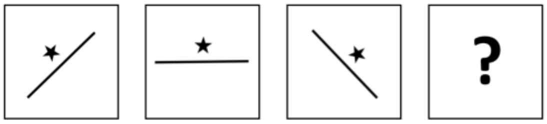
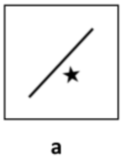
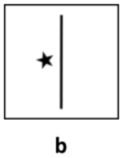
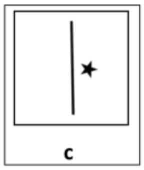
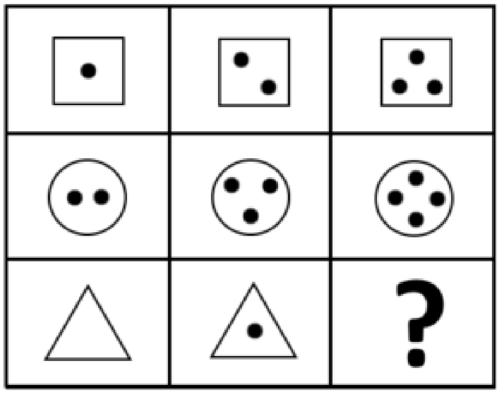
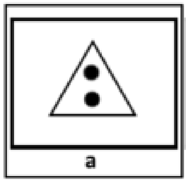
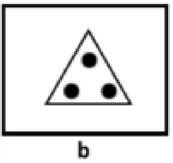
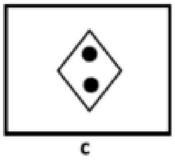
The revised Reasoning scale of the 16pf Sixth Edition was developed via a series of studies separate from those conducted with the 15 personality scales. Items were written to fit into a wider array of content clusters (all measuring reasoning ability), pilot tested, and analyzed. The adaptive test was then designed, and normative considerations were addressed.

Types of Reasoning Items

Table 5.1 shows the nine types of items. Series, synonyms, verbal analogies, and verbal reasoning were items types used in the Fifth Edition scale. Graphs, linear progression, logical reasoning, matrix, and numerical reasoning are new item types. All item types were chosen to measure crystallized and/or fluid intelligence, and all item types are common tasks on many of the widely used intelligence tests. Together, these items are intended to measure a general intelligence factor.

Following tradition, all of the items used a three-option, multiple-choice format (except a small number of numerical fill-in-the-blank items). Research suggests that three options are optimal in terms of balancing time required with minimizing guessing (Bruno & Dirkzwager, 1995; Lord, 1980). Sixteen of the items used a short answer response format. One of these items was later eliminated over concern about scoring; the other 15 items require numerical responses that are easily scored.

Table 5.1. Types of Reasoning/B Items

Item type	Description	Example
Graph	Graphs and simple math	 In which year did Company X have the highest revenue? a. 2015; **b. 2016;** c. 2017
Linear progression	Figural patterns	Which figure should be next, following the pattern?    
Logical reasoning	Logic problems	To disarm a bomb, the red and blue wires must not be cut consecutively. The red wire is cut after the green. Which order disarms the bomb? **a. blue, green, red**; b. green, red, blue; c. red, green, blue
Matrix	Matrix figural patterns	Which figure below completes the matrix?    

Numerical reasoning	Math problems and knowledge	Which number is the smallest? a. 1/2; **b. 0.3**; c. 5/12
Series	Patterns in numbers and letters	Which number should come next at the end of this series: 2, 4, 6, 8, _? a. 9; **b. 10**; c 12
Synonyms/ antonyms	Vocabulary synonyms and antonyms	The word that means the same as big is: **a. ample**; b. slight; c. untrue
Verbal analogies	Verbal analogies	Carton is to milk as pen is to: a. cap; **b. ink**; c. paper
Verbal reasoning	Which word doesn't belong	Which word does NOT belong with the other two? a. cat; b. dog; **c. house**

Item Writing

Items were written by the item writers described in Chapter 4 to conform to the nine item types. To avoid introducing bias against individuals with lower acuity for color discrimination, graphical items were required to be solvable using a grayscale version of the image. Item writers were instructed to avoid content that would cause an item to become much easier if the respondent uses aids such as a dictionary, web search, or calculator (such as difficult vocabulary words or complex calculations). Items were then reviewed by a second team member before being finalized for pilot testing.

Table 5.2 shows the number of items written for each of the nine item types. All item types had at least 20 items except graphs, for which there were only 16. About 91% of the items survived pilot testing and IRT calibration to be included in the final item pool. Mean IRT slope and difficulty indicate that overall the items are fairly high-quality and fairly easy. The similarity of the means across item content areas indicates that all item types indicate general reasoning ability fairly well and were of somewhat similar difficulty, although verbal reasoning items were noticeably lower in quality and easier, synonyms were generally easier, and logical reasoning items were harder.

Table 5.2. Summary of Item Development

Item type	Items written	Items in pool	Yield	Mean slope	Mean difficulty
Graph	16	16	100.0	0.73	-0.76
Linear progression	20	19	95.0	0.92	-0.98
Logical reasoning	32	27	84.4	0.93	-0.28
Matrix	20	20	100.0	0.77	-0.62
Numerical reasoning	35	29	82.9	1.04	-0.93
series	33	31	93.9	0.89	-0.64
Synonyms/antonyms	33	32	97.0	0.75	-1.80
Verbal analogies	33	29	87.9	0.79	-1.32
Verbal reasoning	26	22	84.6	0.57	-1.90
Total	**248**	**225**	**90.7**	**0.83**	**-1.05**

Note: Reasoning pilot study sample, N = 951; All items written were included in the pilot test. The “Items in Pool” indicate the items that survived item analysis. Yield is the percentage of written items that were included in the final item pool. Mean Slope and Mean Difficulty are average slope (IRT *a* parameter) and difficulty (IRT *b* parameter) estimated in the pilot study and transformed to the operational metric.

Pilot Testing

Amazon Mechanical Turk (MTurk) was used to recruit a total of 951 participants. MTurk was instructed to offer this survey only to US citizens who had at least 5,000 accepted work assignments with an average acceptance rate of 98% (i.e., who had a track record of successfully completing short MTurk tasks). Participants were paid $0.70 for each form completed (there were 10 forms; see below) and, as an incentive to do well and to complete as many forms as possible bonuses of $2 were paid to the 100 top-scoring participants. After data cleaning, the final sample size was 796 and sample size per item varied from 354 to 412. Fourteen percent of participants completed all forms (see below), 39% of participants completed a single form, and the remaining 47% completed between two and nine forms. The sample was more diverse than a university subject pool but not as diverse as the US general population; a majority of the participants identified themselves as White (72.4%; 9.0% were missing or not reported; 8.2% identified as African-American/Black; 4.8% as Asian or Pacific Islander, 3.8% as Hispanic, 1.8% as multiracial, 0.1% as Native American, 0.1% as “other”), but the sample had diversity in terms of gender (63.7% female), age (ranged from 18 to 77, with a mean of 39.7 years), and education (13% no college, 33% some college/associates degree, 32% college degree, 13% graduate degree, 9% missing). Most participants described English as their native language (89%) and being employed (69%).

A total of 293 items were pilot tested: 248 new Reasoning items, 15 Fifth Edition reasoning items, and 30 very easy items designed to detect inattention. This large number of items could not be administered on a single form, therefore they were arranged onto 10 content-balanced, randomly equivalent pilot forms and administered using a design which ensured that the samples completing each form were randomly equivalent. Each form contained 24 or 25 new items and three attention check items; in addition, an "anchor block" composed of the 15 items of the Fifth Edition Reasoning/B scale were injected into the first form completed by each participant. Each time a participant accessed the testing platform, a form was chosen at random from the available forms not previously completed by that participant. Forms were retired when a sample size of about 400 was collected.

Data cleaning was performed on 3,725 individual forms, of which 416 (about 11%) were discarded because of evidence that the participant was unmotivated or otherwise responded improperly. Including such cases could cause the IRT analyses to set a metric that made items to appear to be too hard (when compared to a motivated sample). The data cleaning steps included screens for excessive missing responses, short administration time, evidence of inattention, and low percent correct. Three hundred and sixty records were discarded for missing most of the responses (this is common in Internet data collection). An examination of administration times suggested that administration time shorter than 3 minutes was extremely unlikely; 31 cases were discarded for this reason. Each form had three very easy items (median percent correct >99%) near the end of the form designed to assess the attentiveness of responding; 29 records were removed because more than one item was answered incorrectly or was missing. Finally, 13 records with percent of correct responses lower than chance (33.3%) were removed.

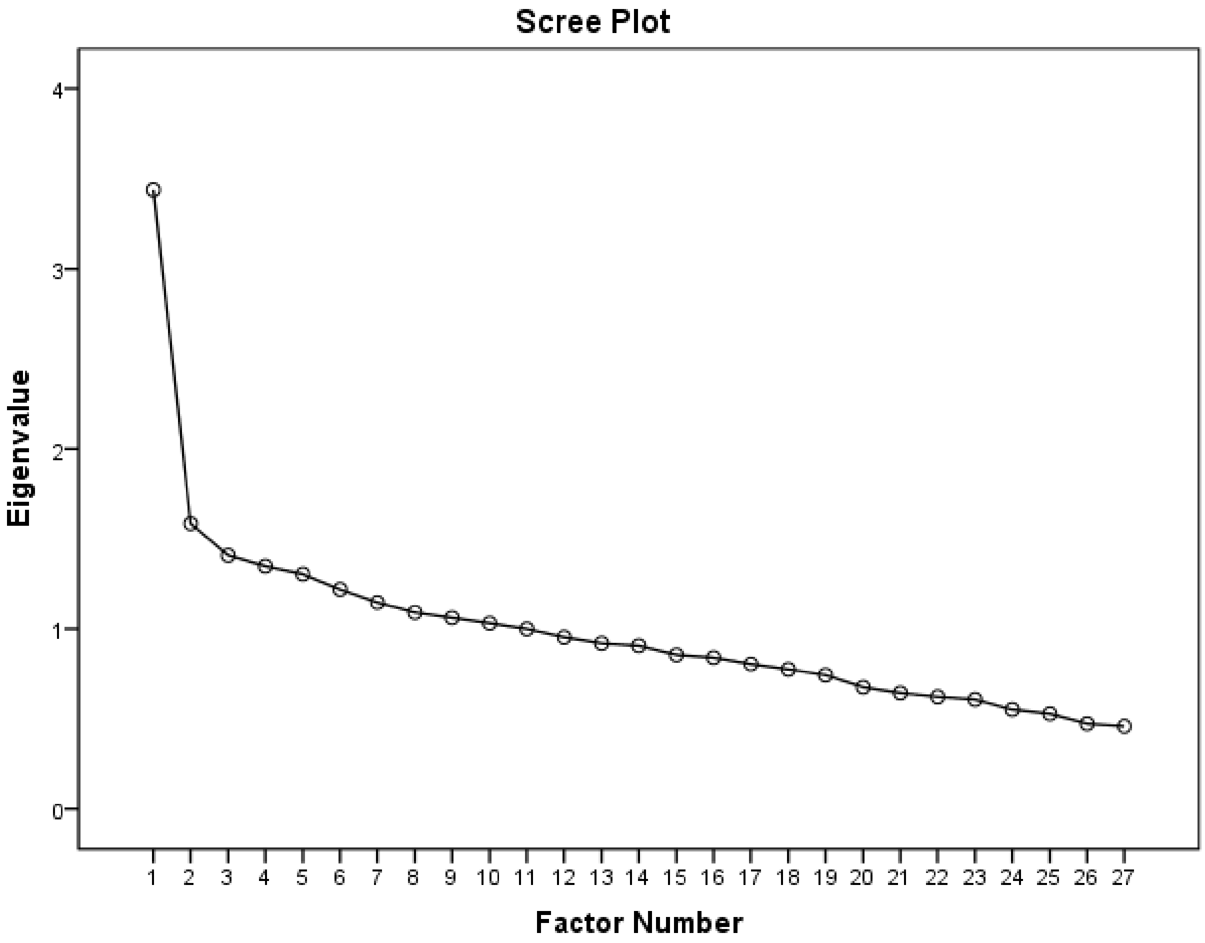
Psychometric Analyses

A total of 3,309 individual forms were collapsed into a single analysis dataset with 796 records, each containing all scored responses of a respondent with missing data coded as "not administered." Preliminary analyses suggested that two items were mis-keyed, and these items were rescored with the correct key. Subsequently we assessed the unidimensionality of the item responses, eliminated items based on poor classical test theory (CTT) item statistics, and fit the item response theory (IRT) three-parameter logistic (3PL) model to all items.

Unidimensionality analysis

The IRT models assume that the items measure a single trait. To test this assumption, we factored the phi matrix of scored item responses. Missing data in our block-sparse data matrix precluded factoring all items. Because the forms were designed to be randomly equivalent, the items of Form 1 were factored using a 1-factor model with maximum likelihood extraction. The scree plot shown in Figure 5.1 shows a strong first factor and weak second and subsequent factors. This shape closely resembles the figure presented by Lord (1980) as a "reasonably unidimensional" test. Velicer's MAP test (O'Connor, 2000; Velicer, 1976) also suggested that a 1-factor solution fit best (using both Velicer's square criterion and O'Connor's fourth-power criterion). Thus, this analysis suggests that Reasoning/B is sufficiently unidimensional for IRT analysis.

Figure 5.1 Scree Plot for Reasoning/B Form 1



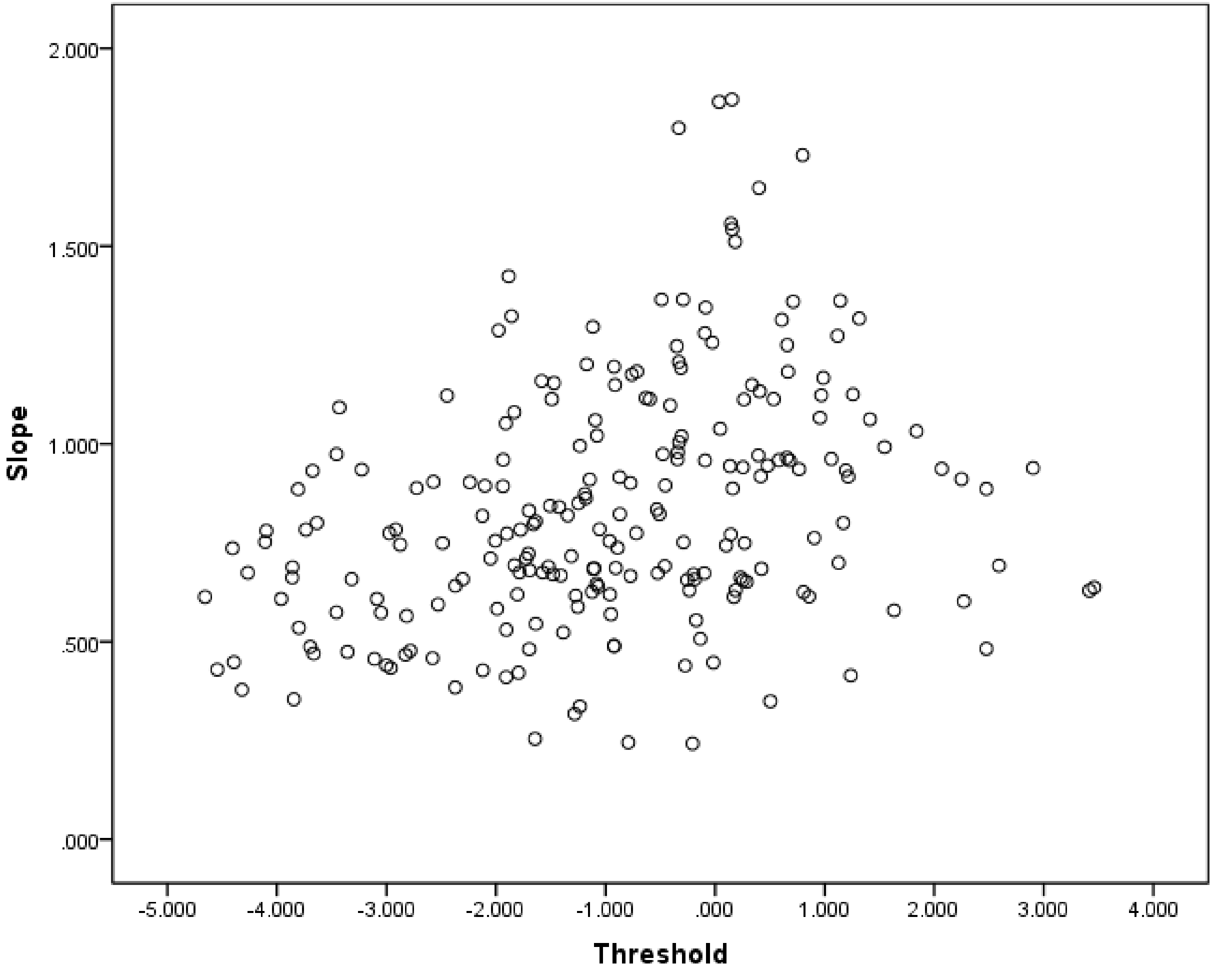
Psychometric CTT and IRT analyses

This section describes the psychometric analyses of the items that make up the pool of CAT items. Proper psychometric analysis is a key foundation for CAT because the CAT algorithm needs to know the difficulty (and potentially other characteristics) of the items, and psychometric analyses estimate the difficulty (and potentially other characteristics) of items through an analysis of responses to the items.

Classical test theory (CTT) and item response theory (IRT) analyses of 263 items (248 new items and 15 items of the Fifth Edition Reasoning/B scale) were conducted using BILOG-MG 3.0 (Zimowski, Muraki, Mislevy, & Bock, 1996). Eight new items were removed due to negative corrected item-total point-biserial correlations and two additional new items were removed due to being extremely difficult (2.5% and 6.6% correct). IRT 3PL models were fit to the remaining 253 items using default settings except: (a) default, diffuse priors were requested for the threshold/b parameter; (b) 30 EM cycles were allowed; and (c) “fit plots” were requested for all items. Both EM and “Newton” cycles converged. Examination of BILOG “fit-plots” suggested that all items fit the data.

Table 5.2 shows the mean IRT slope and difficulty for the items in the final CAT pool. The overall mean of 0.83 represents fairly high-quality items and the overall difficulty of -1.05 shows that the average item is fairly easy (i.e., a mean level of difficulty about one standard deviation below the mean of the examinees. Figure 5.2 shows a scatterplot of these values. The triangle shape of the scatterplot indicates that there was a slight tendency for the highest quality items to be of moderate difficulty, although items with slopes exceeding 1.0 had a wide range of difficulties (from below -3 to almost +2).

Figure 5.2 Scatterplot of Slope (IRT *a*) and Threshold/Difficulty (IRT *b*)



CAT Design and Norming

This section describes the design of the adaptive Reasoning/B scale.

What is a CAT?

A computerized adaptive test (CAT) is an assessment employing a narrow form of artificial intelligence to improve the efficiency of an assessment. In a static administration, all respondents complete all items. Because respondents of a wide range of ability may complete the form, it must include easy, medium, and difficult items, but easy items provide little information for high-ability respondents, and hard items provide little information about low-ability respondents (because items provides the most information when their item difficulty is matched to respondent ability). CAT achieves greater efficiency by performing this matching during administration. A CAT requires that all items in a pool are precalibrated (i.e., have known difficulty) and begins by administering a medium difficulty item. After each item response, the CAT engine updates its estimate of the respondent's ability and picks a new item to maximize information about the ability of the respondent. Simulation studies of the Sixth Edition CAT Reasoning/B scale suggest that most respondents can be assessed more reliably using about 11 adaptively administered items than the 15 static items of the Fifth Edition scale.

IRT scale scores are usually assumed to be standard scores with mean 0 and standard deviation 1.0. The IRT parameter for ability is generally called theta, and estimated theta scores (i.e., IRT CAT scores) are thus called theta-hat. Theta-hat is estimated (Bock & Mislevy, 1982), rather than being computed, but the correlation of theta-hat and number correct is generally very high (>0.90). However, theta-hat has the advantage that comparable scores can be computed on different sets of items, even if the items differ in difficulty. In contrast, number correct scores are specific to the set of items and must be equated across forms.

Design of the Reasoning CAT

Just as two reasoning assessments may differ in design, different adaptive assessments also differ in design, including differences in item selection, updating, stopping criteria, and length and exposure control algorithm. The Sixth Edition Reasoning/B CAT uses maximum item information with "randomesque" item exposure (i.e., each item is selected at random from among the 20 most informative unused items in the pool; Kingsbury & Zara, 1989; Revuelta & Ponsoda, 1998). The theta-hat is updated using EAP theta-hat estimation with a standard normal prior (Bock & Mislevy, 1982). The exam is ended when the number of items administered falls between 10 and 20 items, and the estimated standard error falls below 0.5. These values were chosen to achieve the desired degree of psychometric rigor on the basis of several preliminary simulations.

Simulations of the CAT Reasoning/B scale using a standard normal theta density suggest a CTT reliability around 0.80 with about 11 items administered on average. Simulations suggest that about 22% of the item pool will rarely be used.

CAT Metric

The CAT Reasoning/B scale was not available during standardization. Therefore two steps were taken to ensure that Reasoning/B sten scores had a useful metric. First, the metric of the item pool was equated to that of the operational 16pf Fifth Edition Reasoning/B scale. Subsequently a simulation was conducted to establish a conversion from raw scores (theta-hat values) to sten scores.

Just as temperature can be measured in different scales (Celsius, Fahrenheit, Kelvin, etc.) so the metric of the CAT scores must be established. Furthermore, just as Celsius and Fahrenheit are both equally “valid” ways of measuring temperature, no metric in inherently better for CAT scores; test developers typically chose a convenient, meaningful metric.

The default metric used for latent traits (e.g., Reasoning ability) is the standard distribution with the mean of zero and standard deviation of one derived from the sample used for IRT analysis. Thus, the metric established in the IRT analysis was based on the sample that completed the Reasoning pilot, which might not exactly match the intended operational population.

Note also that with a mean of 0 and a standard deviation of 1.0, about half of scores would be negative and virtually all scores would be in the range -3 to +3. Although this metric is commonly accepted by psychometricians employing IRT, it might be unusual for practitioners unaccustomed to Reasoning scores like +0.3 or -1.4. Therefore, to remove the possibility of negative scores, when reported on the raw score summary page, 5.0 is added to the raw score. Therefore, raw scores on the summary page are likely to have a range of about 2.0 to 7.0 with a mean of about 5.0 and a standard deviation of 1.0. This transformation only affects the presentation of this raw Reasoning score. Practitioners wishing to convert that raw score back to the standard IRT metric (e.g., to look up a sten in Table 5-3) should subtract 5.0 from the raw score.

The metric of the Sixth Edition Reasoning/B CAT scores was chosen to match that of the existing Fifth Edition scores. That is, the mean and standard deviation of the CAT Reasoning/B sten scores should match that of the operational Fifth Edition Reasoning/B sten scores and practitioners should see little change in mean or dispersion when moving from the old to the new edition.

A large (N=5,000) sample (63.9% male; mean age was 37.4 years with an SD of 12.1; 72.6% White, 11.2% Asian, 8.0% Black, 6.9% multiple or others; 7.7% identified themselves as Hispanic; 96.0% were high school graduate or above, 58% had bachelor's degree or above) of operational Fifth Edition Reasoning/B cases were used to estimate IRT parameters in the current, operational scale metric. Mean–mean equating was used to link the pilot sample IRT parameters of these items (which were on the same metric as the new Reasoning items, because they had been administered together and analyzed simultaneously). The obtained linking parameters were A=0.999 and K=-0.417, indicating that the dispersion was virtually unchanged, but the metric of the Fifth Edition items was about 42% of a standard deviation higher than that of the CAT pool (i.e., the items of the CAT pool needed to be made about 0.42 units easier to adhere to the metric of the Fifth Edition items, which were harder). The difference of 0.42 standard deviation units reflects item difficulty, but item difficulty is partially a matter of respondent motivation, and thus this difference may ultimately reflect the difference between a motivated operational sample and a research sample. These linking parameters were used to transform the IRT parameters of the new Reasoning/B items to the same metric as the operational Fifth Edition items.

Simulation

The operational CAT was not a part of the norming study, so CAT sten scores needed to be created using a different methodology. We therefore conducted a Monte Carlo simulation study using commonly accepted simulation methods (Harwell, Stone, Hsu, & Kirisci, 1996).

A total of 10,000 simulees were selected from a standard normal distribution and a CAT administration was simulated using the actual pool and IRT item parameter estimates. Theta-hat values were rounded to the nearest tenth and then a sten score conversion table was computed in the standard fashion (see Chapter 7). The thresholds between sten scores were computed as the midpoints of the 0.1 score range. That is, theta-hat values of -1.8 or less should be converted into a sten of 1, and values between -1.7 and -1.4 should be converted into a sten of 2. Thus, the theta-hat cut-score distinguishing stens of 1 and 2 should be midway between -1.8 and -1.7, or -1.75. Table 5.3 presents the resulting raw-score to sten-score conversion table.

Table 5.3 Estimated Theta (CAT Raw Score) to Sten Score Conversions

	Sten 1	Sten 2	Sten 3	Sten 4	Sten 5	Sten 6	Sten 7	Sten 8	Sten 9	Sten 10
Lower bound	-	-1.7500	-1.3500	-0.9500	-0.5500	-0.0500	0.4500	0.9500	1.3500	1.7500
Upper bound	-1.7501	-1.3501	-0.9501	-0.5501	-0.0501	0.4499	0.9499	1.3499	1.7499	-

Note: Theta to sten score conversions were based on simulation results (*N* =10,000).

To test whether the sten score conversions were useful, the conversions in Table 5.3 were applied to the test–retest sample, in which the CAT Reasoning/B scale was used. Although this sample is not perfectly representative of the general population (see retest sample in Table 8.1), it represents a large diverse sample and the distribution of scores in this sample gives practitioners an idea of sten score distributions that they might observe in practice. Table 5.4 shows the distribution of sten scores in the test–retest sample (N=233); the mean is 5.50, the standard deviation is 1.69, the sten scores span the range from 1 to 10 and are approximately distributed according to the target percentages (e.g., about 1.2% are expected for a sten of 1; see Figure 3.1). The observed standard deviation is slightly low and the percentages in each sten only approximate the sten score distribution, but these results in a (real-data) research sample strongly suggest that the sten score conversion achieved the objective of producing useful scale scores for the CAT Reasoning/B scale.

Table 5.4 Distribution of CAT Reasoning/B Sten Scores in Research Sample

Sten	Frequency	Percent
1	3	1.3
2	8	3.6
3	19	8.5
4	23	10.3
5	53	23.8
6	57	25.6
7	36	16.1
8	18	8.1
9	4	1.8
10	2	0.9
Total	**223**	**100.0**

Note: Test–retest study sample, N=233. Mean = 5.50, SD = 1.69.

Construct-Related Validity of the Revised Reasoning Scale

Several types of validity evidence were collected. The Reasoning/B scores were correlated across Fifth and Sixth editions. Also, the Sixth Edition scores were correlated with other measures of reasoning.

The correlation between the Reasoning/B scales of the Fifth and Sixth editions was calculated in the equivalency sample (N=488; see Chapter 8). The raw score correlation was 0.67. Using the standard formula for disattenuation (Allen & Yen, 2001), the estimated true-score correlation was 0.88, which indicates that the scores of the Fifth and Sixth Editions Reasoning/B scales measure essentially the same construct.

To further establish its construct validity, the revised Reasoning/B scale was correlated with additional measures of general ability and with self-reported grade point average (GPA). Table 5.5 presents these correlations. The correlations in the lower diagonal are observed values (i.e., the uncorrected correlations), whereas those in the upper diagonal are disattenuated (i.e., corrected for unreliability using the standard disattenuation formula; Allen & Yen, 2001). The pattern of correlations with the Employee Aptitude Survey (EAS) scales suggests that Reasoning/B measures a general reasoning ability shared by these scales. The true-score correlation is 0.61 with EAS05 (spatial) and 0.73 with EAS06, EAS07 and EAS10 (numerical, verbal and symbolic/logical). The estimated true-score correlation with a unit-weighted EAS composite was 0.79. Reasoning/B also had significant correlations with the Dupuy measure of vocabulary (Dupuy, 1974; r=0.21, p<0.05 one-tailed) and self-reported GPA (r=0.21, p<0.05 one-tailed). The EAS composite (of four EAS scales) had similar observed and estimated true-score correlations with the Dupuy and GPA, indicating that the Reasoning/B scale scores were about as predictive of these criteria as the EAS composite scores.

Table 5.5 Observed (Lower) and Disattenuated (Upper) Construct Validity Correlations

	N	$\rho_{xx'}$	B	EAS 05	EAS 06	EAS 07	EAS 10	EAS Comp.	Dupuy	GPA
Reasoning/B	223	0.69	--	0.61	0.73	0.73	0.73	0.79	0.30	0.25
EAS05 Spatial	212	0.89	0.48	--	0.55	0.66	0.63	0.99	0.23	0.19
EAS06 Numerical	211	0.81	0.55	0.47	--	0.59	0.68	0.80	0.24	0.12
EAS07 Verbal Reasoning	209	0.82	0.55	0.57	0.48	--	0.65	0.93	0.34	0.28
EAS10 Symbolic Reasoning	206	0.82	0.55	0.54	0.56	0.53	--	0.91	0.27	0.18
EAS Composite	212	0.90	0.62	0.89	0.68	0.8	0.78	--	0.31	0.24
Dupuy	129	0.76	0.22	0.19	0.19	0.27	0.21	0.25	--	0.06
GPA	166	--	0.21	0.18	0.11	0.25	0.16	0.23	0.05	--

Note: $\rho_{xx'}$ = reliability. Observed correlations are shown in the lower triangle, whereas disattenuated correlations are shown in the higher triangle. Reasoning/B was sten scores of CAT administration; reliability was the average of test–retest reliabilities. EAS = Employee Aptitude Survey. EAS05 = Space Visualization; EAS06 = Numerical Reasoning; EAS07 = Verbal Reasoning; and EAS10 = Symbolic Reasoning. EAS composite was the unit-weighted mean of four EAS tests. GPA = grade point average. Reliabilities for EAS scales were alpha coefficients reported in the EAS manual and the EAS composite reliability was estimated using stratified alpha. Coefficient alpha estimate of reliability for Dupuy scores was calculated in the current sample.

The correlations in Table 5.5 suggest that the Reasoning/B scores measure a general intelligence dimension common among the EAS scales and indicate that despite being brief, the Reasoning/B scale is useful as a quick, general measure of reasoning ability. Practitioners are cautioned that unexpectedly low scores warrant the attention of those interpreting the test scores. For instance, low scores may indicate reading difficulties, lack of attention, misunderstanding of instructions, or test sabotage. In such cases, Reasoning/B scores should be discounted or the candidate should be retested.

References

Allen, M. J., & Yen, W. M. (2001). *Introduction to Measurement Theory.* Long Grove, IL: Waveland Press.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Bruno, J. E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55*, 959-966.

Cattell, H. B. (1989). *The 16pf: Personality in depth*. Champaign, IL: Institute for Personality and Ability Testing.

Cattell, R. B. (1957). *Personality and motivation structure and measurement*. New York, NY: World Book.

Cattell, R. B. (1963). The theory of fluid and crystallized intelligence: A crucial experiment. *Journal of Educational Psychology, 57*, 253–270.

Cattell, R. B. (1971). *Abilities: Their structure, growth and action*. Boston, MA: Houghton Mifflin.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the 16pf*. Champaign, IL: Institute for Personality and Ability Testing, Inc.

Dupuy, H. J. (1974). The rationale, development, and standardization of a basic word vocabulary test. DHEW Publication No. (HRA) 74-1334. Rockville, MD: National Center for Health Statistics.

Harwell, M. R., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*, 101-125.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 54*, 1–22.

Karson, S., & O'Dell, J. W. (1976). *A guide to the clinical use of the 16pf*. Champaign, IL: Institute for Personality and Ability Testing.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*, 359-375.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32*, 396-402.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321-327.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.

Chapter 6: Response Style Indices

Introduction

Among the long-standing concerns in the history of objectively scored psychological test development is how to identify and measure known components of test variance (Loevinger, 1957). One such component, response style, may have a significant effect on the variance of test items (Meade & Craig, 2012; Wiggins, 1973). Response style refers to how a respondent reacts to a test and the test-taking setting. Examples of different response styles include tendencies to give socially desirable, acquiescent, critical, extreme, or random answers, regardless of item content.

The development and use of response style measures–particularly social desirability–has generated some controversy concerning the degree to which response tendencies confound assessment scores. In his review of the response style literature, Furnham (1986) points out that although responses to some questionnaires can be faked, it does not mean that the measure has no validity. In fact, he noted that empirical evidence suggests a conceptual relationship between mental health and responding in a socially desirable manner.

The 16pf Sixth Edition addresses certain response tendencies in the same way as the previous edition, via three Response Style Indices: Impression Management (IM), Acquiescence (ACQ), and Infrequency (INF). The IM and ACQ scores presented on the 16pf Sixth Edition are very similar to those scores presented for the 16pf Fifth Edition. As described below, the INF score is conceptually identical but is constructed in a somewhat different manner. That is, IM, ACQ, and an inattention interpretation of these scores remains the same (but the raw score scales of these indices have changed).

Fifth Edition users should note that INF no longer provides any inference about the use of middle responses. Users of the Sixth Edition may review the number of middle responses on the score summary page if such information is desired..

Impression Management

In personality measurement, the primary goal is to provide the most accurate assessment of a person's characteristics or attributes. Assessment with self-reports can be contaminated by misrepresentation or a bias in responding to a psychological measure (Paulhus, 1986). The issue of response style has concerned researchers throughout the early history of self-report psychological test development (e.g., Goldstein, 1945; Gough, 1947; Meehl & Hathaway, 1946; Mischel, 1968) and has continued to be debated more recently (e.g., Edwards, 1990; Furnham, 1986; Hough &

Oswald, 2008; McCrae & Costa, 1983; Morgenson, et al., 2007; Nicholson & Hogan, 1990; Ones, Dilchert, Viswesvaran, & Judge, 2007; Paulhus, 1990; Rothstein & Goffin, 2006; Walsh, 1990).

Over the years, a number of terms have been used to describe the general concept of response style, such as social desirability, lying, dissimulating, faking good, and faking bad. The term faking was defined by Furnham (1986) as referring "specifically to those occasions when a respondent is deliberately giving false responses in order to create a specific impression—he or she is ill or merits a job" (p. 385). Social desirability was defined by Nederhof (1985) as the tendency "to deny socially undesirable traits and to claim socially desirable ones, and the tendency to say things which place the speaker in a favorable light" (p. 269).

Furnham (1986) distinguished between faking and social desirability. To paraphrase, he defined faking as a general sort of dishonest self-presentation and social desirability as referring more specifically to the tendency of presenting oneself in a positive light. Paulhus (1990) also makes this distinction. On the other hand, Jackson (1989a) discussed a low score on the Social Desirability scale of the Personality Research Form (PRF) as possibly reflecting the tendency toward malingering (p. 26). R. B. Cattell (1973) referred to response style as "motivational distortion," a systematic error in responding to questionnaires by which "the subject either consciously or unconsciously presents a misleading set of responses" (p. 390).

Development of an Impression Management Scale

For users of an instrument like the 16pf Questionnaire, the need is to provide a measure of motivational distortion that is simultaneously brief and useful (sensitive to changes). Early editions of the questionnaire used an empirical approach to rescore existing items to produce two indexes, "Fake Good" (FG) and "Fake Bad" (FB; Winder, O'Dell, & Karson, 1975).

Rescoring existing items had the obvious advantage of not adding to the length of the questionnaire but had two significant disadvantages. First, beginning with the 16pf Fifth Edition and continuing into the current, Sixth Edition, the items of the 16pf questionnaire are written and selected to minimize sensitivity to motivational distortion (see Chapter 4), which makes them poorly suited as measures of impression management. Second, an ipsative relationship is introduced between the FG/FB scores and the primary scales containing the rescored items (i.e., certain primary scale scores would be numerically impossible for extremes of FG/FB and this causes inflated correlations between FG/FB and the primary scales). Preliminary analyses during the development of the 16pf Sixth Edition showed that if IM were based on existing items, such correlations could increase beyond 0.80 due to the inflation inherent in re-use of the same responses for both measures.

For these reasons, the authors of the previous edition devised a 12-item scale of independent items intended to measure socially desirable behaviors, values, and feelings. Research on the Fifth Edition IM scale suggested that the scale was reliable (estimates ranged from 0.63 to 0.70) and had strong, statistically significant ($p < 0.01$, two tailed) empirical correlations with other measures of social desirability (0.48 to 0.55; Conn & Rieke, 1994).

The Sixth Edition IM scale consists of six of the original 12 items. Because of the Likert response scale, these six items provide 25 possible raw score levels (i.e., raw scores range from 6 to 30), which matches the range of the 12-item Fifth Edition scale (raw scores ranged 0-24). Because the items on the 16pf sixth Edition are a subset of the 16pf Fifth Edition IM scale, users may expect scores to behave in a similar manner (as shown in Table 6.1, the correlation between IM on the Fifth and Sixth Editions is 0.78, which is about as high as the reliability allows). Additional information about interpreting IM is presented in Chapter 3.

Table 6.1 presents the corrected item-total correlations of the Sixth Edition IM scale, which ranged from 0.22 to 0.44 with a median of 0.34 and a mean of 0.33. Cronbach's alpha in the standardization sample was 0.60. These results likely reflect the heterogeneity of item content, which would tend to reduce interitem correlations and coefficient alpha. Conversely, the mean test–retest reliability estimate was 0.82 (see Table 6.2), which is almost as high as the reliabilities of the 16pf primary scale scores.

Table 6.1 Corrected Item-Scale Correlations of the Impression Management (IM) Scale

IM item	Corrected item-total correlation
16	0.22
75	0.44
85	0.39
95	0.29
101	0.40
153	0.23
Average	0.33

Note: Standardization sample, *N*=2528, Cronbach's Alpha = 0.598. IM item numbers correspond with the position of the items in the Sixth Edition administrative sequence. Item-total correlations were corrected by omitting the studied items (i.e., the "total" used in each item's analysis omits that item and computes the total score from the remaining items).

Table 6.2 presents correlational evidence of validity. The test–retest reliability estimate is the average of six pairwise estimates based on four surveys (an initial survey and retesting after 2 weeks, 3 months, and 3.5 months; see Chapter 8). The Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1990) is a 40-item measure of motivational distortion comprising 20 items measuring self-deception (labeled "Self-deception enhancement") and 20 items measuring other deception (labeled "impression management"). Sixth Edition IM has substantial correlations with both the self-deception and other-deception scores of the BIDR. Finally, the highest correlation was between the Fifth and Sixth Edition IM scores. The estimated true-score correlation is around 1.0, which implies that IM measures the same construct across the two editions. Again, this is hardly surprising, given that the Sixth Edition items are a subset of those used in the Fifth Edition. These results speak strongly to the construct validity of the Sixth Edition IM scale.

Table 6.2 Correlations Between Impression Management (IM) and Other Measures of Social Desirability

Social Desirability Scale	IM
Mean Test-Retest Reliability	0.82*
BIDR Self-Deception Enhancement	0.64*
BIDR Impression Management	0.69*
Impression Management (16pf, Fifth edition)	0.78*

Note: BIDR=Balanced Inventory of Desirable Responding. *p<.01

Should Scores Be Corrected for Distortion?

Score corrections are sometimes made to personality profiles based on elevated faking good and faking bad scores. These corrections represent an attempt to compensate for social desirability; that is, the profile is changed to reflect the effects of a respondent's high social desirability score(s). Score corrections made on the basis of a respondent's faking scores reflect the direction and magnitude of the correlations found in research (Krug, 1978). For example, if a respondent's Faking Good (FG) score were 10 on the 16pf Fourth Edition (Form A), the following score corrections would be made: two sten scores added to Q4; two sten scores subtracted from C; one sten score added to F, L, and O; and one sten score subtracted from A, G, H, and Q3. If a respondent's Faking Bad (FB) score were 10, these corrections would be made: two sten scores added to C; one sten score subtracted from L, O, and Q4; and one sten score added to A, H, I, and Q3. In the rare event that a respondent scored a 10 on both FG and FB, his or her profile would be essentially unchanged because the corrections would cancel each other.

Walsh (1990) argued that social desirability scores reflect contamination in personality measurement, suggesting that controlling for social desirability results could produce differential validity. In contrast, Nicholson and Hogan (1990) asserted that the correlation between social desirability scales and personality scales represents content overlap, and therefore, that controlling for social desirability results in lower validity coefficients.

To be useful, corrections of personality scores for distortion should be situation-specific; that is, different corrections are needed for different testing situations, such as in a job application or personal counseling setting (R. B. Cattell, 1973). On the other hand, corrections on the basis of a single universal faking good or faking bad cut-off score would inevitably partial out some real personality variance (R. B. Cattell, 1973; R. B. Cattell, Eber, & Tatsuoka, 1970). Other researchers also have discouraged the use of corrected scores (Costa & McCrae, 1992; Ellingson, Sackett, & Hough, 1999; Hogan & Nicholson, 1988; Nicholson & Hogan, 1990; Ones, Viswesvaran, & Reiss, 1996). Thus, such corrections should not be used.

Table 6.3 presents the correlations of IM with the 16pf primary scale scores. IM has substantial correlations with the low Anxiety poles of the AX primary factors Emotional Stability/C, Apprehension/O, Tension/Q4, and, to a slightly lesser extent, Vigilance/L. Abstractedness/M was also found to be negatively correlated with IM. Other correlations were modest or low.

One implication of the correlations in Table 6.3 is that correcting (i.e., partialling out), IM would reduce the variability of the Anxiety primary scales and Abstractedness considerably. Changing these primary factor scores on the basis of a high IM score might result in a distorted measure of these primary factor scores and of Anxiety, and would probably reduce predictive validity coefficients considerably.

Table 6.3 Correlations of the 16pf Primary Scale Sten Scores With the Response Style Indices

Primary factor	IM	ACQ	INF
Warmth/A	0.30	-0.22	0.32
Reasoning/B	-0.09	-0.08	-0.08
Emotional Stability/C	0.64	-0.05	-0.07
Dominance/E	0.21	-0.09	0.11
Liveliness/F	0.21	0.04	0.24
Rule-Orientation/G	0.33	-0.20	0.01
Social Boldness/H	0.36	-0.01	0.15
Sensitivity/I	-0.02	-0.12	0.23
Vigilance/L	-0.46	0.04	-0.01
Abstractedness/M	-0.39	0.13	0.01
Privateness/N	-0.21	0.03	-0.12
Apprehension/O	-0.61	-0.05	0.11
Openness to Change/Q1	0.15	-0.06	0.24
Self-Reliance/Q2	-0.28	-0.06	0.01
Perfectionism/Q3	0.16	-0.07	0.07
Tension/Q4	-0.61	-0.02	0.04

Note: Standardization sample, N = 2528. Correlations of magnitude 0.04 or larger are statistically significantly different from zero, $p < 0.05$ (two-tailed). IM, ACQ, and INF are raw scores of the three response indices: Impression Management, Acquiescence, and Infrequency. All 16pf primary scale scores are sten score.

Using the IM Scale

The IM scale is meant to be used as one of several checks on the overall validity of a 16pf protocol. One way to use IM as a validity check involves choosing high and low IM cut-off scores. A respondent's IM score greater than the high cut off or less than the low cut off would signal a possible problem with the 16pf protocol.

Table 6.4 presents the percentile ranks of the IM raw scores. Using a traditional cut of the 5th and 95th percentiles, scores of 13 or less would be low and scores of 26 or more would be high. Other cuts could be devised from this distribution.

Table 6.4 Impression Management (IM): Raw Score to Percentile Conversions

Raw score	n	Percentile
6	1	< 0.1
7	3	0.1
8	4	0.2
9	6	0.4
10	12	0.8
11	21	1
12	35	3
13	72	5
14	117	8
15	190	14
16	232	23
17	266	33
18	275	43
19	292	55
20	231	65
21	211	74
22	176	81
23	105	87
24	97	91
25	63	94
26	58	96
27	28	98
28	11	98.9
29	6	99.2
30	16	99.7
Total	2528	

Note: Standardization sample, N = 2528.

An unusually high IM score may suggest that a respondent has exaggerated his or her socially desirable qualities while denying undesirable characteristics. An unusually low IM score may indicate excessive malingering, self-criticism, or reading difficulties. Positive or negative self-presentation will likely manifest itself throughout the 16pf scale scores. When a high or low IM score occurs, different actions can be taken. For example, the professional could determine possible motives for the exaggerated self-presentation by reviewing other encounters with the respondent and possibly other test results. Alternatively, retesting may be necessary.

In interpreting IM, it is important to take into consideration the context. For example, job applicants generally have a strong incentive to present in more socially desirable directions (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001), but this does not seem to fundamentally affect the factor structure of the 16pf scores (Ellingson, Smith, & Sackett, 2001).

Acquiescence

An acquiescent response style can be defined as a tendency "to agree to personality items as self-descriptive, independently of the particular content of the items" according to Wiggins (1973, p. 440). Although research on acquiescence has been slight, some self-report personality instruments contain measures to screen for this response style (e.g., Hathaway, et al., 1989; Costa & McCrae, 1992). Moreover, several personality instruments have been constructed to minimize the effects of acquiescence by including a balance of true-keyed and false-keyed items (e.g., Jackson, 1989a, 1989b).

Development of the Acquiescence (ACQ) Index

Development of the Acquiescence (ACQ) index for the Sixth Edition followed the empirical approach used in the prior edition. Because the items were Likert items, all items could be used, but following the tradition of the Fifth Edition, the items of Infrequency (INF) scale were excluded. For each item, a response of "Strongly Agree" or "Agree" was scored as one and any other response was scored as zero.

A partially cleaned version of the general population norm sample (N = 3056) was used to compute percentiles. The final norm sample may have had direct range restriction on ACQ because we removed people who selected too many identical responses (which would have removed some people with high ACQ scores). The "partial" cleaning applied to this sample removed only individuals with quick responding or excessive missing responses (see Chapter 7). Table 6.5 presents the percentile rank of each raw score, which can be used to determine cut scores.

Table 6.5 Acquiescence (ACQ): Raw Score to Percentile Conversions

Raw score	n	Percentile	Raw score	n	Percentile
0	19	<1	70	70	51
1-12	27	1	71	65	53
13-22	32	2	72	71	55
23-28	29	3	73	70	58
29-32	35	4	74	59	60
33-35	27	5	75	90	62
36-38	33	6	76	59	65
39	21	7	77	63	67
40-41	30	8	78	53	69
42-43	45	9	79	67	71
44	22	10	80	61	73
45	35	11	81	48	74
46	20	12	82	45	76
47-48	46	13	83	52	77
49	19	14	84	43	79
50	34	15	85	41	80
51	34	16	86	38	82
52	30	17	87	37	83
53	43	18	88	31	84
54	36	20	89	27	85
55	41	21	90	28	86
56	44	22	91	26	87
57	42	24	92-93	39	88
58	43	25	94-95	44	89
59	68	27	96	17	90
60	61	29	97-98	29	91
61	59	31	99-101	35	92
62	57	33	102-105	29	93
63	70	35	106-111	33	94
64	67	37	112-119	25	95
65	86	40	120-127	35	96
66	62	42	128-134	30	97
67	67	44	135-144	30	98
68	52	46	145-149	19	99
69	85	48	150	26	100
			Total	**3056**	

Note: Partially cleaned standardization sample, *N* =3056.

The correlations between ACQ and the primary scales in Table 6.3 demonstrate that most primary scale scores have small correlations with ACQ. No validity evidence was collected for ACQ because extreme scores are obviously concerning; scores of 95th percentile and higher indicate that about 75% or more of the items have been positively endorsed and this is obviously different from any typical response pattern. Like any other response style index, ACQ can be used by the professional to generate hypotheses about a respondent's approach to the test. For example, a high ACQ value might indicate a high need for approval or acceptance and recognition by the professional. Such hypothesized relationships need to be fully explored. Results from other tests and interviews with the respondent may be helpful in interpreting an unusually high ACQ value.

Note that when ACQ is very high, scores on the primary scales are constrained, and the profile tends to be flattened because the items of the primary scales are fairly well balanced in terms of wording. Thus, individuals with very high ACQ tend to have many contradictory responses that cancel and tend to produce mid-range raw scores. For this reason, high ACQ scores are likely to distort the primary scale scores. Thus, profiles with high ACQ scores should be interpreted with caution.

If a person agrees with all the Infrequency (INF) items, he or she will have a raw score in the range 14-17, which is elevated (> 99TH percentile). Because the INF items are independent of the ACQ items, INF could be low when ACQ is very high, which would indicate that the individual distinguished the INF items and answered them differently from the remaining items. However, ACQ and Impression Management (IM) share items, so a person with a high ACQ score is likely to have agreed to all IM items and have a raw score in the range 13-16, which is not elevated.

Infrequency

Unlike items on ability or achievement tests, those on a self-report personality measure have no truly correct responses. Consequently, no means external to the test, such as a scoring key, can determine whether a respondent has attended to the item content. Therefore, a means of determining inattentive responding must be built into the instrument. Two popular methods of accomplishing this goal are a rational-intuitive approach and an empirical approach.

The rational-intuitive approach involves developing a separate scale composed of items reflecting content for which most people agree (i.e., the probability of endorsement is extremely high or extremely low; Meade & Craig, 2012; Millon, 1987; Jackson, 1989a). Such items describe behaviors that are thought to be highly plausible or implausible to virtually everyone. An example is "All my friends say I would make a great poodle" (Meade & Craig, 2012, p. 441). The assumption is that those who endorse such an item as true are inattentive to item content.

The empirical approach to checking for random responding involves determining the endorsement frequency of items in a self-report measure and then combining into a scale those items for which a response alternative was chosen very infrequently (Karson & O'Dell, 1976). If many of the infrequent response alternatives are endorsed by an respondent, random (or very unusual) responding is suspected.

Development of the Infrequency (INF) Index

The Fifth Edition INF scale used the empirical approach, identifying response options endorsed by 5% or fewer of norm sample participants. However, this construction method leveraged the Fifth Edition response format where the middle, '?' response was rare, and all the responses scored as part of the INF scale on that edition were middle responses. This approach was not directly transferrable to the Sixth Edition Likert response scale where essentially no responses were endorsed by 5% or fewer of the normative sample participants. Therefore, a mixed approach was used in which a set of new items was written to assess inattentive responding and then scored to create an INF score. This section describes that development.

Initially, a set of 15 inattention items were developed, and the set of items was refined over rounds of pilot testing. The initial set consisted of five blatant attention check items (e.g., "Please choose 'Disagree' (D)"), six fairly obvious items (e.g., "I believe that one plus two equals three"), and four more surreptitious items (e.g., "At work, I spend all my time sleeping").

The blatant attention check items had the advantage of having a single "correct" response (failing to pick "Disagree" to the above item would be inattentive) but were found to be almost without benefit in research samples. One individual was observed with zero item latencies to all items (probably through the use of a browser plugin to choose random responses) except these obvious inattention items, where he or she manually selected the correct response. There was almost no variance in the response to these items, even where other measures, like response to the remaining items and response time, indicated that the respondent was answering at random. These items were subsequently excluded from the pool of INF candidate items.

The six fairly obvious inattention items worked better in terms of assessing attention, but they had other issues. One disadvantage is that respondents may have idiosyncratic interpretations that seek to rationalize the items as personality items. For example, Curran and Hauser (2015) describe a participant who endorsed the Meade and Craig (2012) "[making] a great poodle" inattention item because "friends have said my hair looks like a poodle." Another Meade and Craig item about sleeping "less than one hour per night" was endorsed too frequently by respondents who otherwise seemed attentive.

There is also concern that such items may elicit negative reactions in some attentive respondents. Fleischer (2016) found that participants in a condition with obvious attention-check items were much more likely to drop out of his study. Comments from attentive respondents indicated that many of these items were obviously inattention items. One commented "I disagreed with that item about having visited every country in the world, although through my military service, personal travels, and missionary work, I have visited every country." Whether this comment is factually correct, this individual obviously understood that this item was intended as an attention check.

To avoid this problem, INF items used on the 16pf are all plausible personality items that have little variance in typical responses. A hypothetical example might be: "I never notice when people are talking to me." This hypothetical item is related to introversion, and a few extremely introverted individuals might endorse it (translating "never" to "rarely"). During rounds of testing, additional items were written and retired if they failed to have the expected extreme directional responding. The final five items all had 6% or fewer respondents responding in the unexpected way in cleaned research samples (e.g., 94% or more of respondents provided the expected "agree" or "disagree" answer).

Both dichotomous and polytomous scoring was considered. In dichotomous scoring, a score of 1 was produced for a neutral response or an answer in the unexpected (atypical) direction with an item to which most people agreed and zero otherwise. For example, most people should disagree with the item "I never notice when people are talking to me," so a "neutral," "agree," or "strongly agree" response would be scored 1 and a "disagree" or "strongly disagree" would be scored 0. These scores were then summed to produce raw scores in the range 0-5. In polytomous scoring, this item would be scored "strongly disagree" = 1, "disagree" = 2, "neutral" = 3, "agree" = 4, and "strongly agree" = 5. Summing over five items produces a raw score in the range 5-25. Parallel analyses suggested that polytomous scoring was more sensitive to inattention and this scoring scheme was adopted for operational use.

Table 6.6 contains the frequency distribution and percentiles of INF scores in the previously described, partially cleaned dataset (N=3,056) in the cleaned 16pf Sixth Edition norm sample (N = 2,528), and in a simulated dataset (N=2,528; described in the next section). The percentiles computed in the partially cleaned dataset should be used to determine the extremity of data (because the fully cleaned dataset had direct range restriction on INF). In that distribution, a raw score value of 16 indicates a high score (a score exceeding the 95th percentile).

Table 6.6 Infrequency (INF): Raw Score to Percentile Conversions

	Partially cleaned standardization sample		Cleaned standardization sample		Simulation sample	
Raw score	N	Percentile	N	Percentile	N	Percentile
5	634	10	633	12.5	1	0.0
6	606	31	605	37.0	1	0.1
7	443	48	439	57.7	15	0.4
8	306	60	303	72.3	34	1.3
9	239	69	223	82.7	55	3.1
10	192	76	175	90.6	103	6.2
11	121	81	76	95.6	144	11.1
12	87	85	44	97.9	197	17.9
13	85	87	24	99.3	255	26.8
14	88	90	6	99.9	255	36.9
15	101	93	0	100	310	48.1
16	66	96			308	60.3
17	57	98			265	71.6
18	24	99.4			215	81.1
19	5	99.85			153	88.4
20	0	99.93			103	93.5
21	2	99.97			68	96.8
22	0	100			32	98.8
23					10	99.6
24					4	99.9
25					0	100
Total	3056		2528		2528	

The correlations in Table 6.3 show that INF has modest or trivial correlations with the primary scale scores. The largest correlations are in the .23 to .32 range for primaries associated with low scores on Tough-Mindedness (e.g., receptive, open-minded individuals) and high scores on Liveliness/F, which probably indicates that very open-minded individuals and those who are fun loving and sensation seeking may be slightly more likely to provide idiosyncratic answers or possibly to be more distracted during assessment. However, these are fairly modest correlations.

Validation and Use of INF Index

To determine how well different cut-off scores for INF would identify random responding, a Monte Carlo simulation study was conducted to simulate completely random responding for the five INF scale items. The objective was to discover the "hit rate" (the percentage of correctly identified valid and invalid test protocols) that could be expected at different INF values.

For the Monte Carlo study, five random Likert item responses were generated for a simulated sample of 2528 (the size of the 16pf Sixth Edition norm sample; described in Chapter 7). To simulate each INF item response, a random integer [1-5] was sampled from a uniform distribution. Like the actual INF, the five simulated item responses were totaled to yield a simulated INF value. Table 6.6 contains percentile rankings for the simulated INF values as well as rankings for the actual sample. A value of 16, which was above the 95th percentile in the partially cleaned sample, was at the 60th percentile in this distribution, indicating that about 40% of random responders would be removed with a cut score of 16 without identifying any norm sample participants as inattentive. A cut score of 12 would identify 80% of the random responders while mislabeling fewer than 3% of the cases in the cleaned norm sample. The next step was to systematically assess "hit rate" for different cut scores.

Note that this analysis assumes that all norm sample participants in the cleaned sample were completely attentive, which is probably an optimistic assumption. As such, this simulation represents a conservative perspective on the hit rate (because some norm group participants, labeled as attentive, were probably at least a little inattentive).

To calculate the overall hit rate for various INF values, the percentage of correctly identified valid protocols was compared with the percentage of correctly identified random protocols. Prior to conducting the comparisons between the two conditions (valid versus simulated random), a base rate was chosen. Base rate refers to the percentage of the population expected to respond randomly. For example, choosing a base rate of 5% means that 5% of the population can be expected to respond randomly. The term population also can refer to a "special" population, such as a clinical population, a job applicant population, a student population, or a prison population.

Using a 5% base rate, which is reasonable in most settings, the overall hit rates were calculated for selected INF values (Table 6.7).

When deciding which cut off to use, two issues must be considered: (a) How many random protocols can be expected in a particular test setting; and (b) which is the greater risk: that of not detecting a truly random profile or that of misclassifying a truly valid profile as random (see Wiggins, 1973).

Table 6.7 Overall Hit Rates for Various INF Cut Scores

Cut-off scores	Valid group	Random group	Overall hit rate
5	25.0	100.0	28.8
6	49.0	99.9	51.5
7	66.3	99.3	68.0
8	78.3	98.0	79.3
9	87.1	95.8	87.6
10	94.1	91.7	93.9
11	97.1	86.0	96.5
12	98.8	78.2	97.8
13	99.8	68.2	98.2
14	100.0	58.1	97.9
15	100.0	45.8	97.3
16	100.0	33.6	96.7
17	100.0	23.1	96.2
18	100.0	14.6	95.7
19	100.0	8.6	95.4
20	100.0	4.5	95.2

Note: Valid group = cleaned standardization sample, N=2528. Random group = simulated random responders. Base rate = 5% random responders. See text for explanations of cut scores and hit rates.

For example, the overall hit rates presented in Table 6.7 represent a 5% base rate. Choosing an INF cut-off score of 9 would result in 87.1% of the valid protocols being classified as valid and 12.9% being misclassified as random. In addition, 95.8% of the truly random protocols would be classified as random and 4.2% misclassified as valid. Therefore, choosing a cut-off score of 6 would result in an overall hit rate of 87.6% (.871 x 95 + .958 x 5 = .876). Because risk is determined by the percentages of misclassifications, the risk in deciding whether a protocol is random would involve a 12.9% chance of a false positive (classifying a valid protocol as random) and a 4.2% chance of a false negative (classifying a random protocol as valid).

Although choosing an INF cut-off score can be complex, two rules of thumb may be used: (a) Choose the cut-score with the highest allowable Type I error rate, or (b) select a cut-off score that yields the smallest difference in rates of correct classification (see Berry et al., 1991). The first approach is useful to minimize misclassifying attentive respondents. The second approach will result in the chance of either kind of misclassification being as equal as possible. Thus, for a base rate of 5%, the best choice of cut off would be 10 (see Table 6.7) because the difference between the valid score hit rate (94.1%) and random score hit rate (91.7%) is smallest at 2.4%. At this cut off, the expected overall hit rate is 93.9%. Note that these results are dependent on the chosen base rate of 5% (i.e., this example assumes 5% of respondents are inattentive; if more or fewer were inattentive, the hit rates would change).

If misclassifying a valid protocol as random is more critical than accepting a truly random protocol as valid, the cut off should be increased. Such a decision might be made if the hypothesis is that respondents are unlikely to respond in a random fashion (e.g., in a job selection setting). On the other hand, the INF cut off might be lowered if the hypothesis is that random responding is likely to occur (e.g., in a court-ordered psychological assessment).

In the Fifth Edition, high values of INF implied the likelihood that the respondent had selected excessive middle responses; this is not necessarily true for the Sixth Edition. Users who wish to evaluate the number of middle responses for the Sixth Edition protocols may consult the score summary page of the report, where the number of middle responses are shown, and compare the number to Table 6.8, which shows the percentiles of various numbers of middle responses. In the partially clean normative sample, 95% of respondents picked 89 or fewer middle responses (out of 150 Likert responses to non-INF items).

Table 6.8 Percentile Ranks for Number of Middle Responses

Raw score	n	Percentile
0	65	1
1	37	3
2-3	33	4
4-5	35	5
6	29	6
7	26	7
8	27	8
9	39	9
10-11	42	10
12	34	11
13	25	12
14	45	14
15	38	15
16	31	16
17	39	17
18	55	19
19	36	20
20	48	22
21	31	23
22	48	24
23	40	26
24	45	27
25	58	29
26	52	30
27	47	32
28	55	34
29	41	35
30	57	37
31	39	39
32	55	40
33	48	42
34	55	43
35	52	45
36	52	47
37	61	49
38	62	51
39	53	53
40	59	54
41	52	56
42	44	58
43	44	59
44	46	61
45	53	62
46	52	64
47	51	66
48	46	67
49	31	69
50	32	70
51	38	71
52	41	72
53	48	74
54	31	75
55	34	76
56	27	77
57	38	78
58	37	79
59	32	80
60	33	81
61	31	82
62	20	83
63	19	84
64	18	85
65	25	85
66-67	35	86
68	15	87
69-70	38	88
71-73	33	89
74-75	26	90
76-78	40	91
79-80	19	92
81-84	35	93
85-89	32	94
90-94	27	95
95-103	31	96
104-113	32	97
114-123	30	98
124-149	31	99
150	15	>99

Note: Partially cleaned standardization sample, *N*=3056.

Summary

The Sixth Edition continues the tradition of three response style indices and these indices are interpreted in very similar ways on the Sixth Edition, as compared to the Fifth Edition scores, although Infrequency (INF) no longer implies the selection of many middle responses.

Each of the three Response Style Indices reveals different aspects of a respondent's approach to the 16pf Sixth Edition. The revelations may be specific to the testing situation, such as the respondent's mood at the time of testing, or may be indications of more enduring traits or personal problems. The professional needs to identify possible reasons for an elevated response style index. Asking questions and raising issues or hypotheses about the respondent's attitude toward the testing is recommended. For example, is a high Impression Management (IM) score due to conscious or unconscious motives? Does it indicate that an overly positive self-presentation was specific to the purpose for testing, such as a job application or promotion? Was the self-presentation a more enduring, yet naive, characteristic of the respondent's self-image, or does the respondent truly possess the attributes identified by the IM items? On the other hand, is an unusually low IM score an indication of low self-esteem, excessive self-criticism, malingering, or a misunderstanding of test instructions? Unusually high or low IM scores warrant attention to such issues.

In the case of an elevated Acquiescence (ACQ) score, other questions may arise. For instance, does the respondent exhibit a high need for approval by the testing professional or by people in general? Did the respondent fully understand the instructions? Did he or she give full attention to the test?

A high Infrequency (INF) score indicates that the respondent endorsed several items in unusual pattern that may indicate a lack of attention or extremely idiosyncratic interpretation of items. Possible explanations include a lack of commitment due to psychological stressors that need attention; a hostile attitude toward the testing or toward the reason for being tested (e.g., a court-ordered assessment); inattentive responding; an attempt to sabotage test results; or a respondent reading problem or learning disability that hampers understanding of item content. The professional is encouraged to use his or her training and experience to explore possible reasons for an elevated score. Additional testing and face-to-face inquiries are among the recommended strategies. Retesting may be a necessary option. A therapeutic or counseling referral may be another.

References

Berry, D. T. R., Wetter, M. W., Beer, R. A., Widiger, T. A., Sumpter, J. C., Reynolds, S. K., & Hallam, R. A. (1991). Detection of random responding on the MMPI-2: Utility of F, Back F, and VRIN scales. *Psychological Assessment, 3*, 418–423.

Cattell, R. B. (1973). *Personality and mood by questionnaire*. San Francisco, CA: Jossey-Bass.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the 16pf*. Champaign, IL: Institute for Personality and Ability Testing, Inc.

Conn, S. R., & Rieke, M. L. (1994). *16pf fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.

Costa, P. T. Jr., & McCrae, R. R. (1992). *Professional manual for the revised NEO Personality Inventory*. Odessa, FL: Psychological Assessment Resources.

Curran, P. G., & Hauser, K. A. (2015, April). *Understanding responses to check items: A verbal protocol analysis*. Paper presented at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

Edwards, A. L. (1990). Construct validity and social desirability. *American Psychologist, 45*, 287–289.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*(2), 155.

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology, 86*, 122-133.

Fleischer, A. (2016). A comparison of different methods of detecting inattentive responding on self-report personality measures. (Doctoral dissertation). Illinois Institute of Technology, Chicago, IL.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7*, 385–400.

Goldstein, H. (1945). A malingering key for mental tests. *Psychological Bulletin, 42*, 104–118.

Gough, H. G. (1947). Simulated patterns on the MMPI. *Journal of Abnormal Social Psychology, 42*, 215–225.

Hathaway, S. R., McKinley, J. C., Butcher, J. N., Dahlstrom, W. G., Graham, J. R., & Tellegen, A. (1989). *Manual for administration and scoring for the MMPI-2*. Minneapolis, MN: University of Minnesota Press.

Hogan, R., & Nicholson, R. A. (1988). The meaning of personality test scores. *American Psychologist, 43,* 621–626.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and Industrial-Organizational Psychology: Reflections, progress and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1(*3), 272–290.

Jackson, D. N. (1989a). *Personality Research Form manual*. Port Huron, MI: Sigma Assessment Systems.

Jackson, D. N. (1989b). *Basic Personality Inventory manual*. Port Huron, MI: Sigma Assessment Systems.

Karson, S., & O'Dell, J. W. (1976). *A guide to the clinical use of the 16pf*. Champaign, IL: Institute for Personality and Ability Testing.

Krug, S. E. (1978). Further evidence on 16pf distortion scales. *Journal of Personality Assessment, 42*, 513–518.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

McCrae, R. R., & Costa, P. T., Jr. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51*, 882–888.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437-455.

Meehl, P., & Hathaway, S. (1946). The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology, 30,* 525–564.

Millon, T. (1987). *Manual for the MCMI-II (2nd ed.).* Minneapolis, MN: National Computer Systems.

Mischel, W. (1968). *Personality and assessment*. New York, NY: Wiley and Sons.

Morgenson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: a review. *European Journal of Social Psychology, 15*, 263–280.

Nicholson, R. A., & Hogan, R. (1990). The construct validity of social desirability. *American Psychologist, 45*, 290–292.

Ones, D.S., Dilchert, S., Viswesvaran, C., & Judge, T. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995–1027.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*(6), 660.

Paulhus, D. L. (1986). Self-deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), *Personality assessment via questionnaires: Current issues in theory and measurement*. New York, NY: Springer-Verlag.

Paulhus, D.L. (1990). *Assessing self-deception and impression management in self-reports: The balanced inventory of desirable responding*. Unpublished paper, University of British Columbia Department of Psychology, Vancouver.

Rothstein, M. G., & Goffin, R.D. (2006). The use of personality measures in personnel selection: What does the current research support? *Human Resource Management Review, 16*, 155–180.

Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86*, 943-953.

Walsh, J. A. (1990). Comment on social desirability. *American Psychologist, 45*, 289–290.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.

Winder, P., O'Dell, J. W., & Karson, S. (1975). New motivational distortion scales for the 16pf. *Journal of Personality Assessment, 39*, 532–537.

Chapter 7: Norms and Standardization

Introduction

This chapter describes the collection of a normative sample and the production of sten score conversion tables. The final norm sample after data cleaning was N=2,528. (A partially cleaned sample of N = 3,056 was used in some analyses of the response style indices; see Chapter 6). The final sample generally showed good alignment with census targets for sex, race, age, educational, and geographic variables reported in 2015 American Community Survey (ACS) published by the U.S. Census. Because this sample was large and representative, analyses were conducted to investigate the relationship between 16pf scores and demographic variables including sex, race, age, and education level.

Sample

A survey research firm recruited participants from their panel to complete the standardization form, Form S, of the 16pf Sixth Edition Questionnaire on the operational administration platform. Participants were recruited via email to match the required demographic groups and completed Form S in return for “points” redeemable for modest rewards by the survey research firm. The six demographic variables and their levels are shown in Table 7.1. During data collection, data cleaning was performed regularly, and the survey closed when the cleaned sample size was greater than the target of 2,500 participants and the demographics were well aligned with the census targets (> 80% and < 125% of target). Considerably more than 2,500 cases were eventually collected (5,995 cases were initially collected), primarily for two reasons. First, because the data cleaning removed cases from the sample. Second, because recruitment was mainly limited to targeting a single demographic category (e.g., men or specific age range) while the sample was evaluated simultaneously on multiple demographic categories (e.g., men in a specific age group, identified race, employment, education, and region categories).

A total of 2,504 cases were removed due to missing data. The primary objective of data cleaning for the remaining 3,491 cases was to remove individuals who responded with insufficient effort and attention (Meade & Craig, 2012). Data cleaning used three indications of inattention: unrealistically fast total assessment time, high Infrequency/INF score, and low variability of responses. Cut scores on these three indicators were determined empirically with the constraint that as few cases as possible be eliminated. Most deleted cases were flagged on more than one of these issues. Assessment time was the total time from beginning to end; individuals with times shorter than 20 minutes were eliminated. The Infrequency score used at this step was an early version of the one documented in Chapter 6, scored dichotomously (like the Fifth Edition INF score); individuals who answered more than two (of seven) questions in an unexpected way were removed. Individuals who answered many or all items the same were suspected of not being attentive. Response variability was measured two ways. First, the standard deviation of the numeric values of the Likert responses (prior to reverse scoring) was calculated; values at or close to zero indicate little variability in responding. We also counted largest number of "same responses." For each of the five Likert responses, we counted the number of responses and recorded the largest number for each individual (i.e., a person who picked "Strongly Agree" most often at 52 times would have a score of 52). Median standard deviation was 0.95 and we flagged individuals with SD <= 0.10. Median number of same responses was about 110 (representing 47% of all Likert responses), and cases were flagged if more than 85% of the items (more than 200 out of 233 items) had the same response.

Table 7.1 shows the demographic breakdown of the final sample of N = 2,528. For each of the six demographic stratification variables, the table presents the frequency (i.e., the count of individuals with a given demographic group), the percent of the sample, the frequency expected based on census figures, the expected percentage based on census figures, and the percentage match to the target (i.e., Frequency/Census Target).

- Size of the norm sample is 2,528: 1,211 men and 1,317 women (47.9% male, 52.1% female).
- The sample is 62.1% White, 16.2% Hispanic, 11.0% Black/African American, 5.0% Asian American, 0.6% Native American or Alaska Native, 0.2% Native Hawaiian or other Pacific Islander, and 4.9% other race.
- Ages range from 16 to 75, with a mean age of 45.4 years.
- Years of education range from "less than high school" to "having a doctorate," with the majority having at least some college course work (63.9%).
- Approximately 21.6% of those in the sample reside in Middle West states, 18.5% in Northeastern states, 35.7% in Southern states, and 23.7% in Western states.

The "Target %" values are almost all in the range 80% to 125%. Thus, the normative sample approximated the census targets (ACS, 2015) reasonably well. The counts closely matched the census figures for sex, age, and race. The norm group is slightly overeducated, as compared to the census targets, and individuals aged 16-17 were represented at lower rates than in the U.S. general population. The group's composition reflects the kind of person who routinely takes the 16pf Questionnaire.

Table 7.1 Demographic Characteristics of the US Standardization Sample (N=2528)

Sex	Frequency	Percent	Census target	Census %	Target %
Male	1211	47.9	1229	48.6	98.5
Female	1317	52.1	1299	51.4	101.4
Age	**Frequency**	**Percent**	**Census target**	**Census %**	**Target %**
16-17	55	2.2	86	3.4	64.0
18-19	81	3.2	86	3.4	94.2
20-24	220	8.7	229	9.0	96.1
25-34	452	17.9	434	17.2	104.1
35-44	421	16.7	411	16.2	102.4
45-54	452	17.9	444	17.6	101.8
55-59	207	8.2	211	8.4	98.1
60-64	186	7.4	186	7.4	100.0
65+	454	18.0	442	17.5	102.7
Race	**Frequency**	**Percent**	**Census target**	**Census %**	**Target %**
African-American/ Black	279	11.0	301	11.9	92.7
Asian	127	5.0	124	4.9	102.4
Hispanic	409	16.2	420	16.6	97.4
Native American or Alaska Native	14	0.6	16	0.6	87.5
Native Hawaiian or Other Pacific Islander	5	0.2	4	0.2	125.0
White	1570	62.1	1528	60.5	102.7
Other	124	4.9	134	5.3	92.5
Employment status	**Frequency**	**Percent**	**Census target**	**Census %**	**Target %**
Employed	1490	58.9	1497	59.2	99.5
Unemployed	142	5.6	139	5.5	102.2
Not in workforce	896	35.4	890	35.2	100.7
Education	**Frequency**	**Percent**	**Census target**	**Census %**	**Target %**
High school	913	36.1	986	39.0	92.6
Some college	786	31.1	657	26.0	119.6
College degree	546	21.6	531	21.0	102.8
Graduate degree	283	11.2	329	13.0	86.0
Region	**Frequency**	**Percent**	**Census target**	**Census %**	**Target %**
Middle West	545	21.6	556	22.0	98.0
Northeastern	467	18.5	455	18.0	102.6
South	902	35.7	935	37.0	96.5
West	598	23.7	581	23.0	102.9
(Missing)	16	0.6	--	--	--

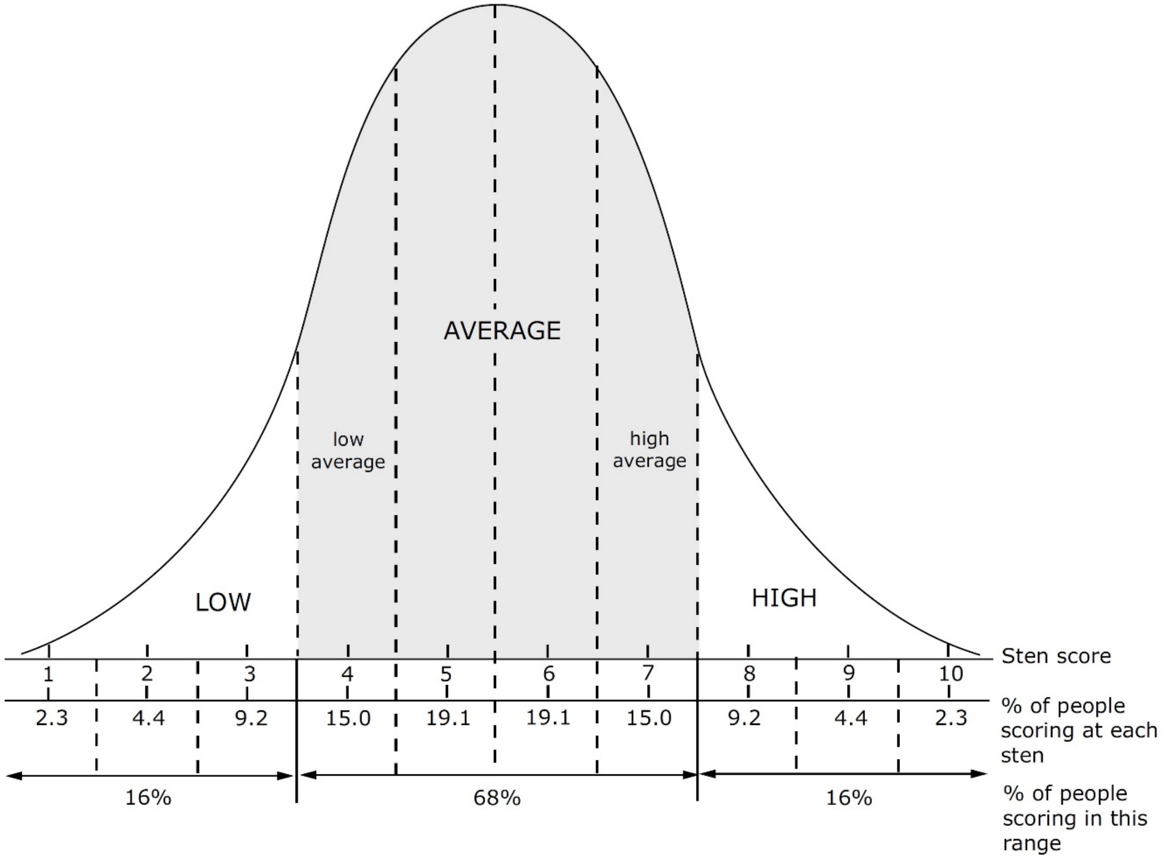
Note. For employment status, the category "Not in workforce" represents students, homemakers, and retired people.

Norms

The items of the 16pf Questionnaire are reverse-scored as needed and then summed to form raw scores. Because the Likert responses range from 1 to 5, the smallest possible raw score is equal to the number of items on the scale and the largest possible score is five times the number of items. As such, the interpretation of a raw score depends on the number of items on a scale (as well as the distribution of raw scores).

To make 16pf scores easy to interpret, raw scores of the primary scales are converted to “sten” scores (short for “standard ten”) using Table 7.2, prepared from the normative sample. Sten scores range from 1 to 10, with a mean of 5.5 and a standard deviation of 2. Because stens are standardized across scales, an individual obtaining the same sten score on two different factor scales will fall at approximately the same percentile rank on both scales, relative to the normative group. This simplifies the comparison of an individual’s scores across different factor scales. Figure 7.1 shows the sten score distribution, including the percentage of the total distributed into each sten score.

Figure 7.1 Sten Distribution



To produce the raw score conversion table (Table 7.2), raw scores are grouped (starting at the extremes and moving into the middle of the distribution) to match the percentages defined in Figure 7.1. For example, as shown in Table 7.2, Warmth/A raw scores of 10 through 21 were grouped because the total percent of the norm sample obtaining a raw score of 10 to 21 was about 2.3%. Then raw scores 22, 23, 24, and 25 were grouped because about 4.4% of the norm sample obtained these scores. This process was repeated for raw scores through 35; then the larger scores were grouped by moving down the raw score distribution, starting with raw scores 50, 49 and 48, which were assigned a sten of 10 (because about 2.3% of the sample obtained these three raw scores).

This process illustrates that the sten conversion is a normalizing process; even if the raw scores were not normally distributed, sten scores are approximately normally distributed (in the population). More scores fall into the middle of the distribution (the average-score range, 4-7) than into the more extreme stens (1-3 and 8-10). This process also illustrates that norms are intertwined with standard scores (e.g., stens of 1 and 10 represent the most extreme 2.3% of the population).

In Table 7.2, the Reasoning/B scale is shown separately. The B scale uses a computer adaptive administration methodology that produces raw scores on a “z-score” metric (mean of about 0, standard deviation of about 1.0). The CAT B scores are continuous, so the conversion includes an upper and lower bound for all stens (except stens 1 and 10). The process used to compute these bounds is identical to the process described above but used simulated CAT B scores rounded to the nearest 0.1 (i.e., about 2.3% of the simulated respondents had CAT B score below -1.7 and were assigned a sten of 1, about 4.4% had scores between -1.7 and -1.3 and were assigned a sten of 2, etc.; see Chapter 5 for more information about the adaptive Reasoning scale and this simulation).

The global scales are not shown on Table 7.2 because these scores are computed from the sten scores of specific primary scales. Thus, there are no global scale raw scores. Chapter 4 describes the factor analytic research and scaling conducted to produce the global factor sten scores.

Percentiles for the raw scores of the response style indices are presented in Chapter 6: Impression Management (IM; Table 6.4), Acquiescence (ACQ; Table 6.5), and Infrequency (INF; Table 6.6). As a general rule, response style indices produce flags when the response style is extreme (at or above the 95th percentile for all indices, or at or below the 5th percentile for IM; although other cut-scores could be used). Flagged cases suggest the possibility that test-taker response style might influence the respondent’s 16pf profile of primary scale scores. Some 16pf users may wish to adjust these "cut offs" for the unique needs of their testing situation. See Chapter 6 for complete descriptions of these scores, their development, and these percentiles.

Table 7.2 16pf Sixth Edition Raw to Sten Conversion

Factor	Sten 1	Sten 2	Sten 3	Sten 4	Sten 5	Sten 6	Sten 7	Sten 8	Sten 9	Sten 10
A	10-21	22-25	26-29	30-32	33-35	36-38	39-42	43-45	46-47	48-50
C	10-16	17-21	22-25	26-30	31-34	35-37	38-41	42-44	45-47	48-50
E	10-19	20-22	23-26	27-29	30-32	33-36	37-39	40-42	43-46	47-50
F	11-16	17-20	21-24	25-28	29-32	33-36	37-40	41-43	44-47	48-55
G	11-23	24-28	29-31	32-35	36-38	39-41	42-44	45-48	49-51	52-55
H	8-9	10-12	13-16	17-20	21-24	25-28	29-31	32-34	35-37	38-40
I	12-22	23-27	28-31	32-35	36-38	39-42	43-46	47-50	51-54	55-60
L	8-13	14-15	16-17	18-20	21-23	24-26	27-29	30-32	33-35	36-40
M	10-14	15-17	18-20	21-23	24-26	27-29	30-32	33-36	37-40	41-50
N	9-15	16-18	19-22	23-25	26-29	30-32	33-35	36-39	40-42	43-45
O	8-11	12-13	14-16	17-20	21-23	24-27	28-30	31-34	35-37	38-40
Q1	11-23	24-27	28-30	31-34	35-37	38-40	41-43	44-46	47-48	49-55
Q2	8-14	15-17	18-20	21-23	24-26	27-29	30-32	33-35	36-38	39-40
Q3	9-17	18-19	20-22	23-25	26-29	30-32	33-35	36-38	39-41	42-45
Q4	9-14	15-17	18-20	21-23	24-26	27-29	30-33	34-36	37-39	40-45
B										
Lower Bound	-	-1.7500	-1.3500	-0.9500	-0.5500	-0.0500	0.4500	0.9500	1.3500	1.7500
Upper Bound	-1.7501	-1.3501	-0.9501	-0.5501	-0.0501	0.4499	0.9499	1.3499	1.7499	-

Note: A=Warmth, B=Reasoning, C=Emotional Stability, E=Dominance, F=Liveliness, G=Rule-Consciousness, H=Social Boldness, I=Sensitivity, L=Vigilance, M=Abstractedness, N=Privateness, O=Apprehension, Q1=Openness to Change, Q2=Self-Reliance, Q3=Perfectionism, Q4=Tension. Conversion for nonreasoning scales were based on Standardization Sample, N=2528; Reasoning/B scale conversion was based on simulation results, N=10000; see Chapter 5.

Descriptive Statistics

Table 7.3 provides means, standard deviations, and standard error of measurement (SEM) for raw and sten scores of the primary and global scales, and for the response style indices. The global scores do not have raw scores and the response style indices do not have sten scores. The sten scores all have a mean close to 5.5 and a standard deviation close to 2.0.

Table 7.3 Means, Standard Deviations, and Standard Errors of Measurement (SEM)

	Raw score			Sten score		
Primary Factor	**Mean**	***SD***	**SEM**	**Mean**	***SD***	**SEM**
Warmth/A	35.58	6.52	2.61	5.54	1.96	0.79
Emotional Stability/C	33.71	7.60	2.63	5.51	1.96	0.68
Dominance/E	32.85	6.65	2.66	5.51	1.97	0.79
Liveliness/F	32.23	7.80	3.02	5.45	1.99	0.77
Rule-Consciousness/G	38.29	6.77	2.71	5.50	2.00	0.80
Social Boldness/H	23.97	7.29	2.30	5.44	2.01	0.64
Sensitivity/I	38.78	7.70	3.61	5.49	1.98	0.93
Vigilance/L	23.86	5.59	2.23	5.59	1.91	0.77
Abstractedness/M	26.50	6.05	2.71	5.46	1.92	0.86
Privateness/N	28.95	6.62	2.65	5.45	1.96	0.78
Apprehension/O	23.74	6.65	2.58	5.48	1.97	0.76
Openness to Change/Q1	37.00	6.36	2.77	5.44	1.97	0.86
Self-Reliance/Q2	26.13	5.83	2.33	5.39	1.93	0.77
Perfectionism/Q3	29.03	6.11	2.73	5.48	1.91	0.86
Tension/Q4	26.53	6.33	2.61	5.44	2.00	0.82
Global Factor						
Extraversion				5.50	1.96	0.28
Anxiety				5.50	1.97	0.34
Tough-Mindedness				5.50	1.96	0.52
Independence				5.50	1.97	0.48
Self-Control				5.50	1.95	0.48
Validity Index						
IM	18.79	3.70				
INF	7.04	1.93				
ACQ	69.39	16.76				

Note: Standardization sample, N=2528. Internal consistency reliabilities (Alpha) of primary scales (presented in Table 8.3) were used to calculate SEMs for raw scores and stens. Stratified Alpha estimates of global scales (presented in Table 8.3) were used to calculate SEMs for Global Factor Stens. The CAT Reasoning/B scale was not available in the standardization sample; for the 20-item fixed Reasoning/B scale in the standardization, raw score mean, SD, and SEM were 10.82, 3.77, and 1.99, and sten score mean, SD, and SEM were 5.57, 1.92, and 1.02. CAT Reasoning/B has an SEM of about 0.50 in the CAT standard score metric; the sten score SEM is about 1.0.

The standard error of measurement (SEM) is often used to build a confidence interval around an obtained score. The SEM can be thought of as the standard deviation of the error in an individual's score. Adding plus-or-minus (±) 1 SEM to the obtained score provides an approximate 68% confidence interval for that individual's true score. For example, because most 16pf scales have SEM of about 1.0, if an individual had an observed sten score of 3, an approximate 68% confidence interval would be [2-4]. That is, 68% of the confidence intervals constructed in this way will contain the respondent's true score (Allen & Yen, 2001). Approximate 95% confidence intervals can be constructed by adding and subtracting 2 SEMs. For Warmth/A with an SEM of 0.79, the 68% confidence interval for a score of 3 is [2.2, 3.8] and the 95% confidence interval is [1.6, 4.6]. Scores different by 2 of more SEM's are likely to be "significantly different" from each other.

Tables 7.4 and 7.5 present the intercorrelation of the global scale sten scores and the primary scale sten scores, respectively. The global scales show generally small to moderate correlations, but Extraversion and Independence are more strongly correlated at 0.65. The primary scale scores have intercorrelations consistent with their relationships to global factors. For example, Warmth/A is more highly positively correlated with Liveliness/F, Social Boldness/H than other scales and negatively correlated with Privateness/N and Self-Reliance/Q2. However, Warmth/A also has only slightly weaker strong correlations with Sensitivity/I, Vigilance/L (negatively), Openness to Change/Q1, and Tension/Q4 (negatively).

Table 7.4 Global Scale Intercorrelations (Sten Scores)

Global Factor	**EX**	**AX**	**TM**	**IN**
Extraversion (EX)	-			
Anxiety (AX)	-39	-		
Tough-Mindedness (TM)	-47	-00	-	
Independence (IN)	65	-31	-49	-
Self-Control (SC)	-18	-28	45	-18

Note: Standardization Sample, N=2528. Values shown to two decimal places; decimal point omitted.

Table 7.5 Primary Scale Intercorrelations (Sten Scores)

	A	B	C	E	F	G	H	I	L	M	N	O	Q1	Q2	Q3
A	-														
B	-13	-													
C	16	04	-												
E	24	-01	34	-											
F	45	-10	23	43	-										
G	26	-09	22	-4	-09	-									
H	41	-12	41	59	66	06	-								
I	38	-08	-10	09	27	-08	11	-							
L	-30	-03	-42	-06	-21	-26	-28	-03	-						
M	-09	01	-50	-15	06	-40	-20	25	28	-					
N	-48	06	-16	-34	-51	-05	-55	-18	29	07	-				
O	-05	-01	-77	-34	-23	-11	-41	13	41	43	16	-			
Q1	37	02	18	39	48	-17	36	35	-06	22	-21	-12	-		
Q2	-40	07	-25	-24	-58	-12	-48	-08	34	12	48	25	-23	-	
Q3	15	-17	11	15	02	30	14	-02	-02	-30	-07	-04	01	-03	-
Q4	-36	04	-55	-09	-24	-22	-30	-07	45	26	22	48	-25	33	-06

Note: Standardization Sample, N=2528. Values shown to two decimal places; decimal point omitted. A=Warmth, B=Reasoning, C=Emotional Stability, E=Dominance, F=Liveliness, G=Rule-Consciousness, H=Social Boldness, I=Sensitivity, L=Vigilance, M=Abstractedness, N=Privateness, O=Apprehension, Q1=Openness to Change, Q2=Self-Reliance, Q3=Perfectionism, Q4=Tension.

Analyses of Demographic Group Differences

To better understand how demographic characteristics might impact the interpretation of 16pf Sixth Edition scores, group differences were examined for several demographic variables, including sex (male; female), self-identified race (Asian, Black, Hispanic, and White), age (under 40; 40 and older), and Education (see Table 7.1). For each analysis, a measure of “effect size” shows the degree that the demographic variable is related to each 16pf scale score.

Effect Size

In nontechnical terms, an effect size is a standard index of the strength of a phenomenon (Cohen, 1988). The phenomenon investigated in these analyses was the difference in means for various demographic groups. A small effect size indicates that there is little difference between demographic groups, whereas a large effect size indicates that there are large differences between the demographic groups (although the largest effect sizes observed in these analyses were only moderate).

The effect size most often used in these analyses was the standardized mean difference (denoted as Cohen’s *d*; Cohen, 1988) between two groups (e.g., the standardized mean difference of a score between men and women). The analysis of education used an effect size measure called omega-squared.

Standardized mean differences always involve two groups and are calculated in two steps. First, the pooled standard deviation of the two groups is calculated (by taking the weighted average of the two groups' variances). Then *d* is calculated as the difference between the group means is divided by the pooled standard deviation. Thus, these *d* values can be interpreted as the "number of standard deviation units" by which the two group means are separated. In practice, this is easier than it sounds. Cohen (1988) suggested, strictly as a rough rule of thumb, that *d* values of 0.20 and smaller are "small;" values of 0.50 are "medium;" and values of 0.80 are "large." As a result, effect sizes much smaller than 0.50 were not interpreted. Because difference exist regarding the context in which 16PF scores are used, interpretation of score difference should be relative to the user's situation.

The direction of the difference matters to the interpretation. The *d* values in this analysis were calculated so that negative numbers indicate that women, minorities, and older individuals have higher scores. Of course, higher 16pf scores are not better (nor worse) except on the Reasoning/B scale where higher scores imply more items correct.

Education had four levels and was analyzed using ANOVA. Omega-squared is the appropriate effect size to use in this case, which indicates the proportion of variation in 16pf score that is attributable to the four educational levels. Values of 0.01 are considered "small;" values of 0.06 are considered "medium;" and values of 0.14 are considered "large" (Cohen, 1988).

When evaluating the results presented in this section, readers are advised that mean differences between two groups do not indicate test bias, only that differences are indicated. These differences may be expected based on experience and prior research. For information about the distinction, see Reynolds (1995).

Group Differences Analyses for Sex

In previous 16pf editions, mean differences between sexes were noted for primary factor scales such as Warmth/A, Sensitivity/I and Apprehension/O (Conn & Rieke, 1994). Table 7.4 presents the results of the analysis of the Sixth Edition normative data. An effect size greater in magnitude than .50 was found for Sensitivity/I (*d* = -0.52), and the effect size for Warmth/A was -0.40. Women as a group scored higher than men on both of these primary factors, although the sizes of those differences were reduced from the values previously found for the Fifth Edition.

Table 7.4 Female-Male Standardized Mean Differences

	Male (N=1211)		*Female (N=1317)*		Cohen's *d*
	Mean	***SD***	**Mean**	***SD***	
Primary Factor					
Warmth/A	5.14	1.95	5.91	1.90	-0.40
Reasoning/B	5.80	1.94	5.36	1.88	0.23
Emotional Stability/C	5.71	1.89	5.32	2.02	0.20
Dominance/E	5.67	1.89	5.37	2.03	0.16
Liveliness/F	5.59	1.97	5.31	2.01	0.14
Rule-Consciousness/G	5.20	1.95	5.77	2.01	-0.29
Social Boldness/H	5.58	1.93	5.31	2.08	0.14
Sensitivity/I	4.97	1.98	5.97	1.86	**-0.52**
Vigilance/L	5.55	1.82	5.63	2.00	-0.04
Abstractedness/M	5.50	1.87	5.41	1.96	0.05
Privateness/N	5.56	1.89	5.34	2.01	0.11
Apprehension/O	5.14	1.88	5.78	2.01	-0.33
Openness to Change/Q1	5.65	1.96	5.24	1.95	0.21
Self-Reliance/Q2	5.25	1.91	5.52	1.94	-0.14
Perfectionism/Q3	5.23	1.80	5.70	1.99	-0.25
Tension/Q4	5.32	1.93	5.54	2.05	-0.11
Global Factor					
Extraversion	5.51	1.95	5.49	1.97	0.01
Anxiety	5.27	1.85	5.72	2.04	-0.23
Tough-Mindedness	5.78	1.99	5.24	1.90	0.27
Independence	5.71	1.92	5.31	2.00	0.21
Self-Control	5.24	1.88	5.74	1.98	-0.26
Validity Index					
IM	18.89	3.70	18.71	3.71	0.05
INF	7.34	2.05	6.77	1.77	0.30
ACQ	69.37	17.40	69.42	16.15	-0.00

Note: Standardization sample, N=2528. Primary scale and Global scales are Sten scores; Validity Indices are raw scores. Negative value of Cohen's d indicates that female mean was higher than the male mean. Cohen's d meeting or exceeding the cut off of .50, a moderate effect, are shown in bold type.

Group Differences Analyses for Race

Race is a social construct that can represent important differences for many people in North America (AAA Executive Board, 1998). During standardization, participants were invited to self-identify as one of the seven "racial categories" shown in Table 7.1, and analyses were conducted on "racial" groups based on self-identification. Including "Hispanic" as a "racial category" is not uncommon but departs from the treatment of

Hispanic as an orthogonal ethnicity on the U.S Census. However, press reports suggest that the U.S. Census is considering adopting this “unidimensional” format for the 2020 census because many respondents are confused by the existing, more complex, “two dimensional” treatment. The unidimensional treatment was ultimately chosen because it represented the simplest choice for respondents and ensured accurate representation of self-identified Hispanics in these analyses.

Tables 7.5, 7.6, and 7.7 present analyses for Whites compared to Blacks, Hispanics, and Asians, respectively. Comparisons with other racial groups could not be made due to the small sample size of the subgroups (i.e., Native Americans or Alaska Natives, Native Hawaiians or Other Pacific Islanders, and others).

Table 7.5 shows effect sizes slightly exceeding 0.50 for Reasoning/B (*d*=0.53) and Openness to Change/Q1 (*d*=-0.54). Also, the global factors Tough-Mindedness and Independence had d values almost reaching 0.50 (d=0.42 and d=-0.49, respectively). White participants had higher Reasoning/B scores than Black participants by about half a standard deviation. This result is about half the magnitude typically found for general ability tests, in which the mean for Whites is often one standard deviation above the mean for Blacks (Neisser, et al., 1996; Sackett, Schmitt, Ellingson, & Kabin, 2001). For Openness to Change/Q1, the mean score was higher for Blacks than Whites, indicating that Black participants have a moderate tendency towards being more open to change, more innovative, and possibly more unconventional. The two global factor score differences are probably due to the Q1 difference, because Q1 is a component of both these global factors.

Table 7.6 presents effect sizes for Hispanics as compared to Whites. No effect sizes exceeded 0.50, although there was a moderate tendency for Hispanics to score higher on Liveliness/F (*d*=-0.41) and Openness to Change/Q1 (*d*=-0.43).

Table 7.7 presents group differences for Asians compared to Whites. The two medium effect sizes for Reasoning/B, where Asians outperformed whites (*d*=-0.52) and on the Infrequency response style index (*d*=-.67). Prior research has noted a tendency for Asians to outperform Whites on reasoning assessments (Sackett, et al., 2001). Asian’s higher mean score on INF may be because of a tendency to avoid extremes. Note too that the actual difference is about 1.2 score points. That’s approximately the difference between selecting “Strongly Agree” and “Agree” on one out of five items.

Table 7.5 Black–White Standardized Mean Differences

	White (N=1570)		*African American/Black (N=279)*		**Cohen's *d***
	Mean	***SD***	**Mean**	***SD***	
Primary Factor					
Warmth/A	5.48	1.95	6.06	1.93	-0.30
Reasoning/B	5.67	1.90	4.67	1.74	**0.53**
Emotional Stability/C	5.49	1.97	5.70	1.95	-0.10
Dominance/E	5.39	2.00	5.95	1.82	-0.29
Liveliness/F	5.22	2.00	5.77	2.01	-0.27
Rule-Consciousness/G	5.64	2.05	5.61	1.93	0.02
Social Boldness/H	5.30	2.01	6.00	1.97	-0.35
Sensitivity/I	5.39	2.07	5.82	1.75	-0.21
Vigilance/L	5.48	1.98	5.71	1.83	-0.12
Abstractedness/M	5.32	1.95	5.39	1.69	-0.03
Privateness/N	5.42	2.02	5.54	1.84	-0.06
Apprehension/O	5.50	1.99	5.16	1.98	0.17
Openness to Change/Q1	5.12	1.97	6.18	1.88	**-0.54**
Self-Reliance/Q2	5.51	1.93	5.29	1.98	0.11
Perfectionism/Q3	5.42	1.93	5.86	1.82	-0.23
Tension/Q4	5.58	1.97	4.91	1.97	0.34
Global Factor					
Extraversion	5.35	1.98	5.84	1.96	-0.25
Anxiety	5.53	2.00	5.21	1.95	0.16
Tough-Mindedness	5.73	2.01	4.90	1.72	0.42
Independence	5.26	1.97	6.20	1.83	-0.49
Self-Control	5.68	1.99	5.58	1.78	0.05
Validity Index					
IM	18.66	3.53	19.47	4.27	-0.22
INF	6.84	1.79	7.14	2.14	-0.16
ACQ	68.29	15.55	72.77	17.95	-0.28

Note: Standardization sample, N=2528. Primary scale and Global scales are sten scores; Validity Indices are raw scores. Negative value of Cohen’s d indicates that the Blacks’ mean was higher than the Whites’ mean. Cohen’s d meeting or exceeding the cut off of .50, a moderate effect, are shown in bold type.

Table 7.6 Hispanic–White Standardized Mean Differences

	White (N=1570)		*Hispanic (N=409)*		Cohen's *d*
	Mean	*SD*	Mean	*SD*	
Primary Factor					
Warmth/A	5.48	1.95	5.59	1.96	-0.06
Reasoning/B	5.67	1.90	5.29	1.78	0.20
Emotional Stability/C	5.49	1.97	5.45	1.97	0.02
Dominance/E	5.39	2.00	5.78	1.94	-0.20
Liveliness/F	5.22	2.00	6.04	1.91	-0.41
Rule-Consciousness/G	5.64	2.05	5.18	1.93	0.23
Social Boldness/H	5.30	2.01	5.66	2.00	-0.18
Sensitivity/I	5.39	2.07	5.70	1.80	-0.16
Vigilance/L	5.48	1.98	5.85	1.81	-0.19
Abstractedness/M	5.32	1.95	5.79	1.87	-0.24
Privateness/N	5.42	2.02	5.37	1.89	0.02
Apprehension/O	5.50	1.99	5.61	1.90	-0.06
Openness to Change/Q1	5.12	1.97	5.96	1.82	-0.43
Self-Reliance/Q2	5.51	1.93	5.07	1.87	0.23
Perfectionism/Q3	5.42	1.93	5.50	1.95	-0.04
Tension/Q4	5.58	1.97	5.30	2.06	0.14
Global Factor					
Extraversion	5.35	1.98	5.87	1.88	-0.26
Anxiety	5.53	2.00	5.60	1.92	-0.04
Tough-Mindedness	5.73	2.01	5.08	1.78	0.33
Independence	5.26	1.97	5.93	1.90	-0.35
Self-Control	5.68	1.99	5.00	1.86	0.35
Validity Index					
IM	18.66	3.53	19.08	3.91	-0.12
INF	6.84	1.79	7.47	2.06	-0.34
ACQ	68.29	15.55	71.77	17.94	-0.22

Note: Standardization sample, N=2528. Primary scale and Global scales are sten scores; Validity Indices are raw scores. Negative value of Cohen’s d indicates that the Hispanics’ mean was higher than the Whites’ mean.

Table 7.7 Asian–White Standardized Mean Differences

	White (N=1570)		***Asian (N=127)***		**Cohen's**
	Mean	***SD***	**Mean**	***SD***	***d***
Primary Factor					
Warmth/A	5.48	1.95	5.22	1.64	0.13
Reasoning/B	5.67	1.90	6.66	1.92	**-0.52**
Emotional Stability/C	5.49	1.97	5.28	1.60	0.11
Dominance/E	5.39	2.00	4.98	1.62	0.21
Liveliness/F	5.22	2.00	5.50	1.60	-0.14
Rule-Consciousness/G	5.64	2.05	5.13	1.66	0.25
Social Boldness/H	5.30	2.01	5.17	1.78	0.06
Sensitivity/I	5.39	2.07	4.99	1.62	0.20
Vigilance/L	5.48	1.98	5.43	1.58	0.02
Abstractedness/M	5.32	1.95	5.55	1.69	-0.12
Privateness/N	5.42	2.02	5.40	1.57	0.01
Apprehension/O	5.50	1.99	5.51	1.64	-0.01
Openness to Change/Q1	5.12	1.97	5.23	1.60	-0.05
Self-Reliance/Q2	5.51	1.93	4.99	1.73	0.27
Perfectionism/Q3	5.42	1.93	5.43	1.63	-0.00
Tension/Q4	5.58	1.97	5.32	1.68	0.13
Global Factor					
Extraversion	5.35	1.98	5.52	1.70	-0.09
Anxiety	5.53	2.00	5.54	1.49	-0.01
Tough-Mindedness	5.73	2.01	5.91	1.67	-0.09
Independence	5.26	1.97	5.01	1.65	0.13
Self-Control	5.68	1.99	5.26	1.65	0.21
Validity Index					
IM	18.66	3.53	18.43	3.18	0.06
INF	6.84	1.79	8.06	2.25	**-0.67**
ACQ	68.29	15.55	69.69	20.98	-0.09

Note: Standardization sample, N=2528. Primary scale and Global scales are sten scores; Validity Indices are raw scores. Negative value of Cohen's d indicates that the Asians' mean was higher than the Whites' mean. Cohen's d meeting or exceeding the cut off of .50, a moderate effect, are shown in bold type.

Group Differences Analyses for Age

According to research literature on personality traits such as those measured within a Big Five framework, both consistency and change are shown across the life span (McCrae & Costa, 1999, 2008). Additionally, cohort effects also contribute to differences in personalities between older and younger individuals (Milojev & Sibley, 2017). Table 7.8 presents mean differences analyses comparing those younger than 40 to those 40 and older. The cut off of 40 years of age was determined because these individuals are considered to be a protected class according to the Age Discrimination in Employment Act of 1967 (ADEA). Two primaries have medium effect sizes. On average, older individuals scored lower on Abstractedness/M (d=0.69) and on Openness to Change/Q1 (d=0.53). These primaries contributed to medium effect sizes for the globals Tough-Mindedness (d=-.62) and Self-Control (d=-0.66); older individuals were less receptive and more self-controlled.

Table 7.8 Age Standardized Mean Differences

	Under 40 (N=1048)		***40+ (N=1480)***		**Cohen's *d***
	Mean	***SD***	**Mean**	***SD***	
Primary Factor					
Warmth/A	5.65	2.03	5.46	1.91	0.10
Reasoning/B	5.71	1.90	5.47	1.93	0.12
Emotional Stability/C	5.10	2.05	5.80	1.85	-0.36
Dominance/E	5.53	2.03	5.50	1.93	0.01
Liveliness/F	5.88	2.05	5.15	1.89	0.37
Rule-Consciousness/G	5.08	1.99	5.80	1.96	-0.36
Social Boldness/H	5.26	2.12	5.57	1.93	-0.15
Sensitivity/I	5.94	1.92	5.18	1.97	0.39
Vigilance/L	5.91	1.97	5.36	1.84	0.29
Abstractedness/M	6.19	1.91	4.94	1.74	0.69
Privateness/N	5.56	1.94	5.37	1.96	0.10
Apprehension/O	5.90	2.03	5.18	1.87	0.37
Openness to Change/Q1	6.03	1.87	5.02	1.92	0.53
Self-Reliance/Q2	5.29	1.99	5.46	1.89	-0.09
Perfectionism/Q3	5.42	1.91	5.52	1.92	-0.05
Tension/Q4	5.50	2.11	5.39	1.91	0.05
Global Factor					
Extraversion	5.64	2.02	5.40	1.92	0.12
Anxiety	5.90	2.04	5.22	1.86	0.35
Tough-Mindedness	4.82	1.85	5.98	1.89	-0.62
Independence	5.70	2.01	5.36	1.93	0.18
Self-Control	4.78	1.90	6.00	1.82	-0.66
Validity Index					
IM	18.25	3.87	19.17	3.54	-0.25
INF	7.22	2.05	6.92	1.83	0.16
ACQ	71.47	17.58	67.92	15.99	0.21

Note: Standardization sample, N=2528. Primary scale and Global scales are sten scores; Validity Indices are raw scores. Negative value of Cohen's d indicates that the mean for individuals 40 years of age and older was higher than the mean for individuals under 40 years of age. Cohen's d meeting or exceeding the cut off of .50, a moderate effect, are shown in bold type.

Group Differences Analyses for Educational Level

To investigate the relationship between educational level and 16pf Fifth Edition primary factors, an ANOVA was performed for each raw and sten score. Table 7.9 shows the results of these analyses. Although the results differed, the conclusions were identical. The omega-squared effect sizes are all small except for Reasoning/B, which had a large effect size. In other words, individuals with higher educational attainment scored higher on B, which is to be expected. In part, this may be due to the result of self-selection in that people with more reasoning skills are more likely to seek higher education (Neisser et al., 1996). It may also be due to the result of education itself, in that more education may improve reasoning skills and ensure greater experience with reasoning types of test questions.

Table 7.9 Relationship Between Education Level and the 16pf Scores

	Raw Scores		*Sten Scores*	
	F-Value	**Effect size**	**F-Value**	**Effect Size**
Primary Factor				
Warmth/A	2.93	0.003	3.21	0.004
Reasoning/B	120.47*	0.125	120.17*	0.125
Emotional Stability/C	9.51*	0.011	8.82*	0.010
Dominance/E	11.88*	0.014	11.56*	0.014
Liveliness/F	1.79	0.002	1.64	0.002
Rule-Consciousness/G	1.17	0.001	1.15	0.001
Social Boldness/H	4.95*	0.006	5.31*	0.006
Sensitivity/I	1.47	0.002	1.47	0.002
Vigilance/L	9.22*	0.011	9.05*	0.011
Abstractedness/M	6.72*	0.008	7.24*	0.009
Privateness/N	3.99	0.005	3.87	0.005
Apprehension/O	1.59	0.002	1.51	0.002
Openness to Change/Q1	9.33*	0.011	9.31*	0.011
Self-Reliance/Q2	4.80*	0.006	4.63*	0.005
Perfectionism/Q3	0.72	0.001	0.65	0.001
Tension/Q4	0.36	0.000	0.55	0.001
Global Factor				
Extraversion			4.12	0.005
Anxiety			6.04*	0.007
Tough-Mindedness			3.18	0.004
Independence			12.19*	0.014
Self-Control			0.49	0.001
Validity Index				
IM	0.30	0.000		
INF	0.64	0.001		
ACQ	1.50	0.002		

Note: Standardization sample, N=2528. Degrees of freedom for each F-test were 3 and 2524. In an effort to minimize the capitalization on chance, a p-value less than 0.003 (0.05/16) was used to determine statistical significance (denoted by *). Effect size was defined as the amount of variance accounted for by Education level (high school graduate or less, some college, BA, and advanced degrees).

Summary

The 16pf Sixth Edition normative sample was large (N=2,528) and closely matched the demographic composition of the U.S. general population. Analyses of demographic group differences found a small number of mainly expected differences (9 medium effect sizes in 120 comparisons). The vast majority of 16pf scales have small or trivial differences for men and women, people self-identifying differently in terms of "race," and older and younger individuals. As expected, Reasoning/B was found to be quite strongly related to educational attainment. The norms were used to provide sten scores that allow for easy interpretation and comparison of the primary scale scores. Practitioners may use these results to predict which scales have greater potential to demonstrate group differences and possible outcomes depending upon the assessment application. For example, Factor B is strongly related to educational attainment. Selecting job candidates using B is likely to favor more educated candidates.

References

AAA Executive Board. (1998). AAA statement on race. *American Anthropologist, 100*, 712-713.

Age Discrimination in Employment Act of 1967. 29 U.S.C. Sec. 621 et seq. (1967).

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

American Community Survey (2015). http://proximityone.com/acs2015.htm

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Conn, S. R., & Rieke, M. L. (1994). *16pf fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.

McCrae, R. R., & Costa, P. T., Jr. (1999). A five-factor theory of personality. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality theory and research* (Vol. 2, pp. 139-153). New York, NY: Guilford Press.

McCrae, R. R., & Costa, P. T., Jr. (2008). The five-factor theory of personality. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 159-181). New York, NY: Guilford Press.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437-455.

Milojev, P., & Sibley, C. G. (2017). Normative personality trait development in Adulthood: A 6-year cohort-sequential growth model. *Journal of Personality and Social Psychology, 112*(3), 510-526.

Neisser, U., Boodoo, G., Bouchard, T. J. Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101.

Reynolds, C. R. (1995). Test bias and the assessment of intelligence and personality. In D. H. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence*. New York, NY: Plenum Press.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist, 56*, 302–318.

Chapter 8: Reliability and Equivalency

Introduction

Although many aspects of an assessment may be important, the primary technical issues addressed during test development are that the assessment has measurement precision and measures constructs that are practically useful. These two related requirements are called reliability and validity. Reliability is the topic of this chapter. Validity is discussed in Chapters 9 and 10. This chapter also discusses the equivalence of the editions.

Reliability

All psychological measures suffer from measurement error. In classical test theory, measurement error represents the discrepancy between the observed test scores and the true scores. Smaller errors make the observed scores closer numerically to true scores. Highly reliable measures have small measurement errors. The primary goal of psychometrics is to characterize measurement error and to help test developers to improve measurement precision.

Because true and error scores are unobserved, psychometric theory has developed methods to estimate the reliability of measures. Test scores are termed *observed* scores to emphasize the fact that they are known (in contrast to unobservable true and error scores). The correlation between observed and true scores is referred to as the index of reliability, and the square of this correlation is called the reliability coefficient (Allen & Yen, 2001). The reliability coefficient provides an estimate of the proportion of the variation in observed test scores that is attributable to true-score variance. Therefore, reliability coefficients range from a low of .00 to a high of 1.00 (technically, as a correlation coefficient, reliability could be negative, but such values are not found in practice); the higher the reliability coefficient, the smaller the error variance.

Reliability is also related to scale length: longer scales tend to be more reliable than shorter scales. For this reason, the global scales are generally found to have higher reliability than their constituent primary scales. However, longer scales have the obvious disadvantage that they require more respondent time. One goal of test development is to produce scores with maximal reliability while keeping scales as short as possible.

A reliability coefficient has important implications in regard to how a particular measure correlates with another. Because errors scores are, by definition, random, the true score represents the only predictable variation in scores and the correlation of the true and observed scores represent an absolute maximum of any possible validity correlations of assessment scores. Therefore, reliability provides an upper bound on any criterion-related validity correlations (Allen & Yen, 2001). Consequently, a low correlation between two measures may reflect a low true-score correlation or it may reflect the low reliability of one measure or both, and high levels of reliability are essential for valid uses of assessments.

There is no one best method for estimating reliability. The method used depends on the purpose and meaning that one wishes to attach to the reliability coefficient. Two commonly used methods in studies of psychological and educational instruments are test–retest reliability and internal consistency reliability. Test–retest reliability is important in examining consistency of scores over time or in conducting longitudinal studies. Internal consistency reliability is appropriate to investigating the homogeneity of the test content.

Test–Retest Reliability Estimation

The test–retest reliability estimate is the correlation between the scores obtained from two different administrations of a single instrument. As such, test-retest reliability estimates indicate the “reproducibility” of scores over time. The length of time between test administrations is called the “retest period” and commonly varies from short (as short as a day) to months or even years. There is no single correct retest period; short intervals may lead to memory effects, whereas long intervals may cause reliability estimates to reflect respondent change or maturation effects. It is a best practice to use different retest periods, to give an idea of the changes in scores over time. Generally, estimated reliability is higher for short retest period and lower for longer retest periods (Schuerger, Tait, & Tavernelli, 1982; Schuerger, Zarrella, & Hotz, 1989).

Internal Consistency Reliability Estimates

An internal consistency estimate of reliability (coefficient alpha, in these analyses) uses statistical properties of the item scores to estimate the true and error score variation. Heterogeneity of content can cause decreased internal consistency estimates because the essential source of internal consistency is the magnitude of intercorrelations among the items; that is, the larger the item intercorrelations, the greater the internal consistency. For this reason, internal consistency is the appropriate form of reliability to use with scales designed to measure unitary constructs, or traits, such as the 16pf Primary Scales.

Reliability of 16pf Sixth Edition

This section describes the analyses conducted to determine the internal consistency and test–retest reliability estimates for the 16pf Sixth Edition.

Two samples were available for estimating score reliability using internal consistency, the N=2528 normative sample and an N=488 equivalency sample. Participants for the equivalency sample were recruited from two sources: N=305 volunteers from Amazon Mechanical Turk (MTurk) and N =183 volunteers recruited from a temporary staffing agency. The normative sample was described in Chapter 7 and intended to match the US general population. The sample demographics for these two samples, shown in Table 8.1, show a close match for the normative sample to the US population. The equivalency sample demographics were reasonably diverse in terms of sex, age, ethnic background, and education, although this sample tended to be younger and more highly educated than the US population and disproportionately composed of women.

Retest data were collected at three time points, Waves 1, 2 and 3. The “Wave 1” data was the N=305 MTurk sample described above who completed the Fifth and Sixth Edition forms and then were resampled at later dates to retest on the Sixth Edition and to complete measures used to establish construct validity (described in Chapter 9). The time between Waves 1 and 2 was approximately 2 weeks and approximately 3 months elapsed between Waves 2 and 3. Thus 2-week, 3-month, and 3.5 month retest estimates were available from this sample. Because fewer individuals retested, the retest sample sizes are all less than N=305, and the demographics were recomputed on the sample that completed Wave 3. As shown in Table 8.1, these participants were reasonably diverse but had disproportionately more women and fewer people of color. Most participants reported being employed, students, or retired.

Table 8.1. Demographics of the Retest Sample

	Standardization sample N=2528	Equivalency sample N=488	Retest sample N=233
Sex			
Male	47.9	34.8	33.9
Female	52.1	64.8	59.2
(missing)	-	0.4	6.9
Age			
19 or younger	5.4	0.2	-
20-24	8.7	9.6	5.6
25-34	17.9	36.3	36.1
35-44	16.7	20.7	22.3
45-54	17.9	16.0	13.3
55-59	8.2	8.4	6.9
60-64	7.4	3.3	3.9
65+	18.0	5.1	5.2
Missing	-	0.4	6.9
Average age (*SD*)	45.4 (17.0)	39.3 (13.3)	39.8 (12.5)
Race			
African American/Black	11.0	8.0	3.4
Asian	5.0	5.9	8.2
Hispanic	16.2	11.5	6.0
Native American or Alaska Native	0.6	0.4	0.4
Native Hawaiian or Other Pacific Islander	0.2	0.4	0.4
White	62.1	69.3	72.1
Other/two or more races	4.9	4.1	2.6
(missing)	-	0.4	6.9
Education			
Less than high school graduate	5.0	0.6	-
High school graduate	31.1	7.6	8.6
Some college	21.1	23.4	15.5
Associates degree/vocational school	10.0	12.9	13.7
Bachelor's degree	21.6	43.4	40.8
Advanced degree (MA/professional degree/doctorate)	11.2	11.6	14.6
(missing)	-	0.4	6.9

Table 8.2 presents the sten score means and standard deviations for each scale. The means and standard deviations of the scores tended to be about 5.5 and slightly below 2.0 for the standardization sample (where the standardization occurred), while being somewhat more varied in the equivalency and retest samples.

Table 8.2 Means and *SD*s of Stens for Three Samples

	Standardization N=2,528		Equivalency N=488		Test–Retest N=233	
	Mean	***SD***	**Mean**	***SD***	**Mean**	***SD***
Primary Factor						
Warmth/A	5.54	1.96	6.09	2.07	5.72	2.13
Reasoning/B	5.57	1.92	5.09	1.61	5.50	1.70
Emotional Stability/C	5.51	1.96	5.82	2.13	5.92	2.12
Dominance/E	5.51	1.97	5.58	2.20	5.60	2.27
Liveliness/F	5.45	1.99	5.87	2.21	5.27	2.15
Rule-Orientation/G	5.50	2.00	5.34	2.12	5.48	2.06
Social Boldness/H	5.44	2.01	5.52	2.28	5.06	2.22
Sensitivity/I	5.49	1.98	6.49	2.00	5.88	1.84
Vigilance/L	5.59	1.91	5.27	2.23	5.55	2.48
Abstractedness/M	5.46	1.92	5.67	2.11	5.24	2.00
Privateness/N	5.45	1.96	5.30	2.25	5.59	2.25
Apprehension/O	5.48	1.97	5.54	2.27	5.41	2.32
Openness to Change/Q1	5.44	1.97	6.20	2.13	5.85	1.96
Self-Reliance/Q2	5.39	1.93	5.25	2.17	5.64	2.16
Perfectionism/Q3	5.48	1.91	5.76	2.04	5.87	2.18
Tension/Q4	5.44	2.00	5.11	2.03	5.21	2.19
Global Factor						
Extraversion	5.50	1.96	5.84	2.25	5.31	2.10
Anxiety	5.50	1.97	5.25	2.20	5.26	2.21
Tough-Mindedness	5.50	1.96	4.47	2.09	5.13	1.97
Independence	5.50	1.97	5.81	2.22	5.57	2.14
Self-Control	5.50	1.95	5.23	2.10	5.74	2.10

Note: N=223 for Reasoning/B scale in test-retest sample.

Coefficient alpha and test–retest reliability estimates for the Sixth Edition scales are shown in Table 8.3. Excluding Reasoning/B, the internal consistency estimates for the primary scales ranged from .78 to .93 with a mean of .85 and a median of .84. Reasoning/B had lower estimates of 0.71 and 0.72, but those were for a 20-item static form of new Reasoning items constructed for research purposes. Only the Wave 3 data collection included the computer adaptive (CAT) measure of Reasoning/B. In Monte Carlo simulations, reliability was estimated approximately 0.80 (see Chapter 5). Internal consistency estimates were computed for the global scales using stratified alpha. The global factor scales show high levels of estimated reliability.

Table 8.3. Reliability Estimates for the 16pf Sixth Edition Personality Scales

	Internal consistency (Alpha)		Test–Retest				
	Form S (N=2528)	Equivalency sample (N=488)	2-week (N=219)	3-month (N=131)	3.5-month (N=169)	Weighted average	SEMs
Primary Factor							
Warmth/A	0.84	0.87	0.85	0.86	0.85	0.85	0.77
Emotional Stability/C	0.88	0.90	0.90	0.85	0.85	0.88	0.67
Dominance/E	0.84	0.88	0.85	0.85	0.83	0.85	0.77
Liveliness/F	0.85	0.88	0.85	0.86	0.85	0.85	0.76
Rule-Orientation/G	0.84	0.86	0.86	0.80	0.76	0.84	0.80
Social Boldness/H	0.90	0.93	0.87	0.90	0.85	0.90	0.64
Sensitivity/I	0.78	0.79	0.86	0.85	0.79	0.79	0.91
Vigilance/L	0.84	0.90	0.87	0.84	0.83	0.85	0.74
Abstractedness/M	0.80	0.84	0.83	0.76	0.80	0.81	0.85
Privateness/N	0.84	0.89	0.86	0.88	0.88	0.85	0.76
Apprehension/O	0.85	0.91	0.89	0.84	0.87	0.86	0.73
Openness to Change/Q1	0.81	0.84	0.86	0.87	0.84	0.82	0.83
Self-Reliance/Q2	0.84	0.87	0.82	0.82	0.80	0.84	0.77
Perfectionism/Q3	0.80	0.82	0.87	0.87	0.83	0.81	0.83
Tension/Q4	0.83	0.84	0.83	0.82	0.81	0.83	0.82
Global Factor							
Extraversion	0.98	0.99	0.91	0.91	0.90	0.97	0.34
Anxiety	0.97	0.96	0.93	0.88	0.90	0.96	0.40
Tough-Mindedness	0.93	0.95	0.90	0.89	0.87	0.93	0.53
Independence	0.94	0.95	0.89	0.90	0.88	0.93	0.51
Self-Control	0.94	0.97	0.88	0.86	0.86	0.93	0.50

Note: Internal consistency estimates were calculated using scored items and test–retest reliabilities are calculated using stens. Stratified alpha is shown for the internal consistency estimates for the global scales. Except for the administration at 3.5 months, Reasoning/B CAT scores have a reliability of approximately 0.80 (see Chapter 5) and a standard error of about 1.0. For calculating standard error of measurement (SEM), SDs of the primary scale stens from the Standardization sample (as reported in Table 8.2) were used, and reliabilities were weighted reliability average for all scales.

Two-week test–retest estimates for the primary scales ranged from .82 to .90 with a mean and median of .86. Three-month test–retest estimates for the primary scales ranged from .76 to .90 with a mean of 0.84 and median of .85. Finally, 3.5-month test–retest estimates for the primary scales ranged from .76 to .88 with a mean and median of 0.83. Comparing the 2-week, 3-month, and 3.5-month estimates, the reliability decline over these time spans appears to be modest. Scores are generally sufficiently reliable after 3.5 months.

No test–retest data were available for the CAT Reasoning/B scale, but two 20-item static B forms had a 2-week test–retest estimate of 0.79 and the correlation between a static form and the CAT form over several weeks was about 0.68, which may suggest that motivation plays a significant part in the retest reliability of B.

Test-retest reliability estimates for the global scales were higher, as expected due to the longer length of these composite scales; estimates ranged from 0.86 to 0.93 with mean reliability falling slightly from 0.90 for the 2-week retest to 0.89 for the 3-month retest and 0.88 for the 3.5-month retest.

The weighted average in Table 8.3 is computed using the three estimates that are least affected by time (the two internal consistency estimates and the 2-week test–retest estimate), to get an indication of the reliability of the scale scores at a single point in time. Using this weighted average and the standard deviations of the scales in the standardization sample, standard error of measurement (SEM) values were calculated for all scores. These SEM values ranged from 0.64 to 0.91 with an average of 0.78 and a median of 0.77. These values indicate that most respondents will score within about 1.5 stens of their true sten score and can be used for other confidence interval calculations. Because the weighted average includes estimates of two types of error, time and item sampling, those requiring a strict interpretation of SEM are encouraged to use the columns of Table 8.3 that they feel are most applicable to their application. SEM values can easily be calculated for any reliability estimate. For example, the SEM shown in Table 8.3 for Factor A using an average reliability is 0.77 but can be calculated using the 3-month test-retest reliability estimate as 0.73 $(1.96\sqrt{1-0.86})$. It should be noted that due to the homogeneity of values in any scale (single row), resulting SEM's will be highly similar for given scale.

Equivalency Between Fifth and Sixth Editions

The issue of equivalency between the scores of a revised assessment and its predecessor is important for users of the assessment scores because the level of equivalency will indicate the degree to which interpretations of those scores remain the same.

Note that whenever possible, users are advised to administer the same edition to all individuals to be compared. Equivalency is primarily a topic when it is not possible to administer the same edition, as when a Sixth Edition profile is being compared to a Fifth Edition profile.

The analyses described in this section addressed the issue of equivalency by comparing scores obtained from administrations of the 16pf Sixth Edition and its predecessor, the 16pf Fifth Edition. Correlational and regression analyses were conducted on the equivalency sample, described in detail below. The 16pf Fifth and Sixth Edition questionnaires were administered to a sample of volunteers recruited from two sources: N=305 individuals were recruited from MTurk (and who form Wave 1 of the retest/construct validity sample) and N=183 respondents recruited by a temporary staffing agency. The order in which participants completed the two inventories was counterbalanced, and no appreciable time interval elapsed between the completion of the two tests. The assessments were administered on the online platform that will be used for operational assessment and with instructions similar to the operational instructions. Intercorrelations of the scales in the equivalency study are presented in Tables 8.4 and 8.5 and the intercorrelation across editions is shown in Tables 8.6 and 8.7.

Table 8.4. Intercorrelations of the 16pf Sixth Edition Global Scales

Global Factor	**EX**	**AX**	**TM**	**IN**
Sixth Edition				
Extraversion (EX)	-			
Anxiety (AX)	-46	-		
Tough-Mindedness (TM)	-54	09	-	
Independence (IN)	65	-39	-49	-
Self-Control (SC)	-25	-20	55	-29
Fifth Edition				
Extraversion (EX)	-			
Anxiety (AX)	-46	-		
Tough-Mindedness (TM)	-44	01	-	
Independence (IN)	45	-23	-41	-
Self-Control (SC)	-16	-21	52	-20

Note: Equivalency sample (N=488).

Table 8.5. Intercorrelations of the 16pf Sixth Edition Primary Scales

Sixth Edition

	A	B	C	E	F	G	H	I	L	M	N	O	Q1	Q2	Q3
A	-														
B	-07	-													
C	16	-07	-												
E	18	-06	38	-											
F	45	-11	30	49	-										
G	21	-15	20	-10	-07	-									
H	39	-13	42	64	72	01	-								
I	43	-04	-06	08	35	-05	19	-							
L	-39	01	-38	-08	-33	-20	-34	-12	-						
M	04	21	-46	-07	08	-36	-07	29	17	-					
N	-52	09	-13	-33	-57	-12	-59	-31	40	-05	-				
O	-08	12	-77	-44	-34	-05	-47	07	35	39	16	-			
Q1	36	18	21	40	47	-21	38	30	-09	33	-26	-20	-		
Q2	-44	11	-29	-31	-63	-13	-52	-19	40	10	53	26	-30	-	
Q3	03	-16	22	09	-02	25	02	-09	04	-39	09	-14	-09	02	-
Q4	-39	08	-60	-20	-33	-20	-36	-11	44	28	29	52	-31	42	-09

Fifth Edition

	A	B	C	E	F	G	H	I	L	M	N	O	Q1	Q2	Q3
A	-														
B	-18	-													
C	26	-14	-												
E	18	-03	24	-											
F	51	-10	33	27	-										
G	18	-22	23	-05	-11	-									
H	45	-10	46	43	54	08	-								
I	38	-01	-11	-10	17	-08	01	-							
L	-35	03	-42	05	-18	-13	-30	-07	-						
M	-04	15	-37	02	07	-34	-09	14	21	-					
N	-56	-00	-21	-10	-37	-10	-42	-16	37	-05	-				
O	-05	10	-63	-33	-18	-11	-44	22	28	31	05	-			
Q1	19	13	15	30	32	-34	28	14	-09	34	-17	-13	-		
Q2	-58	09	-36	-18	-55	-12	-48	-09	36	15	45	18	-22	-	
Q3	-01	-12	20	10	-11	32	00	-10	03	-32	10	-15	-14	00	-
Q4	-32	15	-49	-05	-23	-20	-33	05	36	20	23	42	-21	38	-07

Note: Equivalency sample (N=488). Values shown to two decimal places; decimal point omitted. A=Warmth, B=Reasoning, C=Emotional Stability, E=Dominance, F=Liveliness, G=Rule-Orientation, H=Social Boldness, I=Sensitivity, L=Vigilance, M=Abstractedness, N=Privateness, O=Apprehension, Q1=Openness to Change, Q2=Self-Reliance, Q3=Perfectionism, Q4=Tension.

Table 8.6 Correlations of the 16pf Fifth and Sixth Edition Global Scales

	Fifth Edition Global Scale				
Sixth Edition Global Scale	**EX**	**AX**	**TM**	**IN**	**SC**
Extraversion (EX)	90	-47	-41	49	-13
Anxiety (AX)	-41	86	-01	-25	-21
Tough-Mindedness (TM)	-50	09	78	-36	41
Independence (IN)	54	-37	-39	82	-15
Self-Control (SC)	-23	-15	55	-29	80

Note: Equivalency sample, N=488. Values shown to two decimal places; decimal point omitted. See Table 8.8 for an analysis of the equivalency coefficients.

Table 8.7 Correlations of the 16pf Fifth and Sixth Edition Primary Scales

Sixth Edition Primary	Fifth Edition Primary Scale														
	A	**B**	**C**	**E**	**F**	**G**	**H**	**I**	**L**	**M**	**N**	**O**	**Q1**	**Q2**	**Q3**	**Q4**
A	72	-15	23	06	41	16	33	34	-33	-01	-43	03	31	-45	06	-28
B	-16	66	-09	-03	-08	-21	-07	-04	-02	18	-01	11	24	06	-11	09
C	13	-11	78	24	19	18	41	-14	-30	-38	-10	-63	10	-23	23	-47
E	26	-08	36	72	33	00	62	-09	-06	-03	-23	-42	34	-29	13	-18
F	52	-14	40	35	81	-03	67	10	-27	05	-44	-28	36	-60	-05	-29
G	20	-20	18	-12	-12	74	03	00	-13	-35	-08	-04	-35	-13	25	-16
H	47	-18	44	45	55	08	86	00	-28	-06	-44	-41	31	-47	03	-33
I	41	-05	-01	05	36	-14	17	70	-12	23	-23	12	28	-23	-10	-04
L	-39	04	-42	03	-25	-16	-32	-07	75	24	38	27	-06	37	03	36
M	02	20	-39	-06	12	-33	-07	19	11	76	-08	35	31	07	-37	20
N	-63	12	-26	-20	-47	-13	-55	-18	34	00	79	11	-20	52	09	26
O	-12	13	-69	-32	-24	-10	-45	19	29	30	08	79	-14	22	-16	45
Q1	21	08	17	31	38	-18	31	04	-12	30	-22	-14	69	-29	-07	-28
Q2	-56	13	-38	-18	-53	-13	-50	-06	36	17	42	20	-19	84	01	35
Q3	00	-14	17	05	-05	26	-02	-06	04	-32	09	-12	-15	01	76	-04
Q4	-33	14	-53	-07	-26	-16	-34	03	40	26	22	43	-20	40	-12	81

Note: Equivalency sample Sten scores, N=488. Values shown to two decimal places; decimal point omitted. A=Warmth, B=Reasoning, C=Emotional Stability, E=Dominance, F=Liveliness, G=Rule-Orientation, H=Social Boldness, I=Sensitivity, L=Vigilance, M=Abstractedness, N=Privateness, O=Apprehension, Q1=Openness to Change, Q2=Self-Reliance, Q3=Perfectionism, Q4=Tension. See Table 8.8 for an analysis of the equivalency coefficients.

Correlational Analysis

The first analysis was to correlate the corresponding scale scores (e.g., Fifth and Sixth Edition Factor A scores were correlated). The two editions are interchangeable to the extent that these correlations are high. As shown in Table 8.8, the observed correlations between the forms are quite high, generally above 0.70 and often above 0.80, for both sten scores and raw scores (Global Factors are only defined for sten scores). The two exceptions are Reasoning/B and Openness to Change/Q1, which had sten score correlations of 0.66 and 0.69. An additional question is whether these correlations suggest equivalence of the constructs being measured. To examine this question, true-score correlations were estimated (using the standard disattenuation formula; see Allen & Yen, 2001; labeled “Disattenuated” in Table 8.8), which show the estimated correlations of the constructs being measured. These correlations are all close to 1.0 (above 0.88) indicating close correspondence between the constructs measured by the Fifth and Sixth Edition scales. The lowest scales were Reasoning/B (true score correlation of 0.86) and Openness to Change/Q1 (true-score correlation of 0.88). These true-score correlations are high enough that there is little practical difference between the constructs measured by the Fifth and Sixth Edition scales. Three of the estimated true-score correlations in Table 8.8 exceed 1.0, which is to be expected when estimates are made of a value close to 1.0 (as is the case here). As seen in Table 8.8, the correlations are slightly higher for the raw scores, although probably not practically important. To the extent that the scales have true-score correlations below 1.0, these differences may be partially the result of adding new and revised items to the Sixth Edition scales (see Table 4.1) and the Likert response scale.

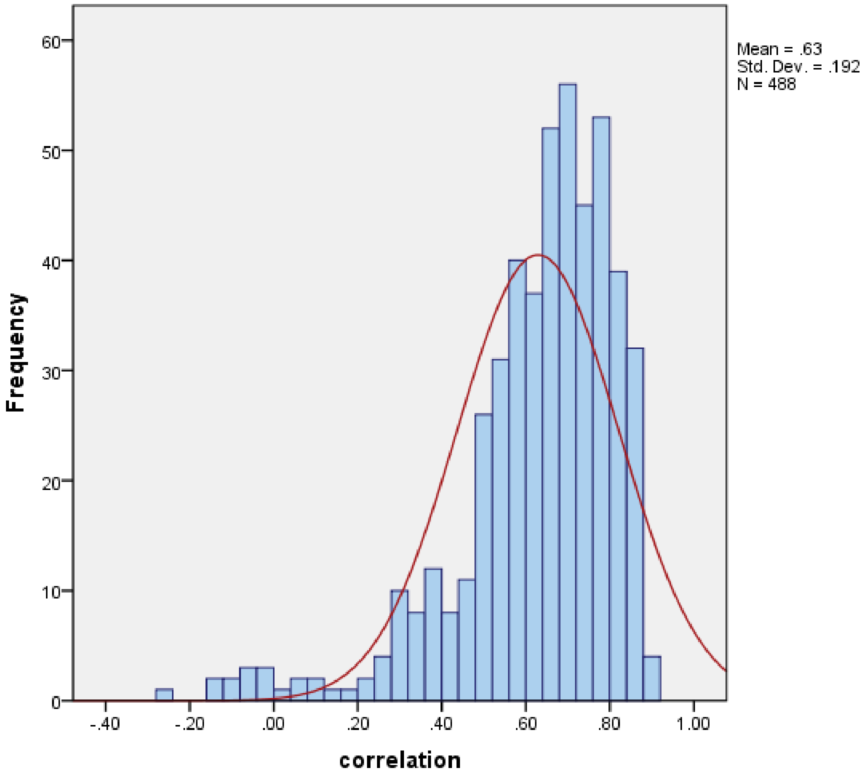
Table 8.8 Observed and Disattenuated Correlations of Fifth and Sixth Edition Scores

	Raw score		Sten score	
	Observed	Disattenuated	Observed	Disattenuated
Primary factors				
Warmth/A	0.74	0.91	0.72	0.89
Reasoning/B	0.67	0.88	0.66	0.86
Emotional Stability/C	0.84	0.96	0.78	0.89
Dominance/E	0.75	0.94	0.72	0.91
Liveliness/F	0.83	1.00	0.81	0.99
Rule-Orientation/G	0.76	0.94	0.74	0.91
Social Boldness/H	0.90	0.98	0.86	0.94
Sensitivity/I	0.74	0.97	0.70	0.93
Vigilance/L	0.75	0.90	0.75	0.90
Abstractedness/M	0.80	0.98	0.76	0.93
Privateness/N	0.82	0.96	0.79	0.93
Apprehension/O	0.81	0.94	0.79	0.91
Openness to Change/Q1	0.71	0.90	0.69	0.88
Self-Reliance/Q2	0.87	1.00	0.84	0.98
Perfectionism/Q3	0.78	1.02	0.76	0.99
Tension/Q4	0.83	1.00	0.81	0.98
Global Factors				
Extraversion			0.90	1.02
Anxiety			0.86	1.02
Tough-Mindedness			0.78	0.90
Independence			0.82	0.96
Self-Control			0.80	0.94

Note: Equivalency sample (N=488). Observed refers to the observed correlations. Disattenuated shows estimated true-score correlations. Raw score correlations were disattenuated using internal consistency calculated in the equivalency sample. For sten correlations among primary scales, Fifth Edition scale reliability was internal consistency in the equivalency sample, and Sixth Edition scale reliability was sample size-weighted test-retest reliability. Operational adaptive Reasoning/B reliability was estimated to be approximately 0.80 in Monte Carlo simulations (see Chapter 5) and was used for disattenuating both raw and sten B scores. For global scale stens, reliabilities were as reported in the Fifth Edition norm supplement (2002) and Sixth Edition sample size-weighted test–retest reliability in Table 8.3.

Figure 8.1 presents the within-person profile correlations of the equivalency sample participants (as a histogram). That is, for each participant, the sixteen Fifth and Sixth edition scores were correlated. We would like to see all very high correlations, but correlations based on 16 data points would have large standard errors. We observe a highly skewed distribution where most values are large and positive, but some values are close to zero and a few are negative. The mean of this distribution is 0.63 and the median is 0.65, indicating that the shape of the profiles on the Fifth and Sixth Editions will be similar for the vast majority of respondents. Ninety-four percent of the sample has a strong positive correlation. Eleven individuals have slightly negative correlations, and 19 individuals have small positive correlations (r < 0.30). There is evidence that sampling or measurement error accounts for at least some of these 30 individuals because they have flatter profiles than the sample as a whole. A completely flat profile cannot have any correlation (these correlations index covariability across primary scales; by definition, profiles without any variability have zero covariance). The mean within-person standard deviation of scores for the entire samples was 2.09. For the 19 individuals with small positive correlations, and 11 individuals with small negative correlations, the mean within-person standard deviations were 1.38 and 1.70, respectively. So, in summary: profiles on the Fifth and Sixth Editions are affected by sampling and measurement error, but the vast majority of individual had strong positive correlations across the primary scales within an individual (i.e., their profiles would have similar shapes). For the 6% with different shapes, there is evidence that they tended to have a flatter profile, which would depress this measure of similarity.

Figure 8.1 Distribution of Within-Person Correlations Across Fifth and Sixth Edition Profiles



These correlation analyses provide strong evidence that the Fifth and Sixth Edition scales measure the same, or almost indistinguishable, constructs.

Analysis of Sten Score Differences

The next analysis was a comparison of stens obtained by the same sample. Because the Fifth and Sixth Editions measure the same constructs and are scaled equivalently, we can expect individuals to have approximately the same true scores on the two editions (i.e., same score except for measurement error), and we can compare the stens obtained by a large group of individuals who completed both editions to demonstrate this equivalency. This kind of equivalency is very relevant to practitioners; in practice, however, the equivalence will be imperfect due to measurement error and due to differences in the normative samples of the two editions, as well as the effects of any other changes (item content, response scale, etc.). Table 8.9 shows the descriptive statistics for the Fifth and Sixth Edition scales in the equivalency sample. Note that several sten scores are missing from the Fifth Edition distributions (e.g., Sten scores of 10 were not possible on the Fifth Edition Warmth/A scale) whereas all 10 stens are possible on all Sixth Edition scales (in Table 8.9, no respondent obtained a sten of 10 on Reasoning/B but it is possible to do so). At the far right of Table 8.9 are the Cohen's d values for each scale. These values show the standardized mean differences between the sten scores of the scales across editions. Values with magnitude 0.20 or lower are small effect sizes, values of 0.50 are considered "medium" effect sizes, and values of 0.80 would be considered large. The only value in Table 8.9 that approaches large is for Self-Reliance/Q2, which was 0.76. These differences likely reflect differences in the standardization samples between the Fifth Edition, which used operational data, and the current, Sixth Edition, which came from research data. Users who are comparing Fifth Edition scores to score from the Sixth Edition are encouraged to review Table 8.9 and decide the importance of the observed differences listed.

Table 8.9 Descriptives of Primary Scale Sten Scores in Combined Equivalency Sample

	Fifth Edition				Sixth Edition				Cohen's
Primary	Min	Max	Mean	*SD*	Min	Max	Mean	*SD*	*d*
Warmth/A	1	9	4.93	1.91	1	10	6.09	2.07	-0.58
Reasoning/B	1	9	5.17	1.82	1	9	5.09	1.61	0.05
Emotional Stability/C	1	8	4.84	1.94	1	10	5.82	2.13	-0.48
Dominance/E	1	9	4.71	1.72	1	10	5.58	2.20	-0.45
Liveliness/F	2	9	5.19	1.85	1	10	5.87	2.21	-0.34
Rule-Orientation/G	1	9	5.13	1.73	1	10	5.34	2.12	-0.11
Social Boldness/H	2	9	4.91	2.11	1	10	5.52	2.28	-0.28
Sensitivity/I	1	9	6.13	1.49	1	10	6.49	2.00	-0.21
Vigilance/L	1	10	6.13	1.97	1	10	5.27	2.23	0.41
Abstractedness/M	2	10	5.77	1.64	1	10	5.67	2.11	0.05
Privateness/N	1	9	6.11	1.88	1	10	5.30	2.25	0.39
Apprehension/O	2	9	5.83	1.83	1	10	5.54	2.27	0.14
Openness to Change/Q1	1	10	5.54	1.81	1	10	6.20	2.13	-0.33
Self-Reliance/Q2	2	10	6.82	2.02	1	10	5.25	2.17	0.75
Perfectionism/Q3	1	9	6.08	1.73	1	10	5.76	2.04	0.17
Tension/Q4	2	9	5.38	1.62	1	10	5.11	2.03	0.15

Note: Negative values for Cohen's *d* indicate the Sixth Edition scores are higher.

Table 8.10 presents an analysis of the sten scores obtained on the two editions. If there were no measurement errors and if the sten scales were the same and the constructs measured identical, then each respondent would be expected to have the same sten score. However, measurement error will result in some differences, even if the two scales were otherwise identical. Therefore, the score similarities were compared to expected ranges. For each scale, a comparison was made between the percent of individuals with the same sten (e.g., a sten of 5 on both the Fifth and Sixth Edition Warmth/A scales); the percent within one sten (e.g., a sten of 5 on Fifth Edition Warmth/A scale and a sten of 6 on the Sixth Edition); and the percent within two stens (e.g., a sten of 5 on Fifth Edition Warmth/A scale and a sten of 7 on the Sixth Edition). The remaining percentage of respondents had larger differences (e.g., 16.6% of respondents had scores discrepant by 3 or more stens on Warmth/A). On average about a quarter of people obtained the same sten on both editions. Based on the Fifth Edition SEM of 1.0 sten scores and assuming normality, we might expect 68% of people to score within 1 sten score and about 96% to score within two stens. The actual values are slightly lower (66% and 89%), suggesting that there are very slight differences in the expected sten scores across the two editions.

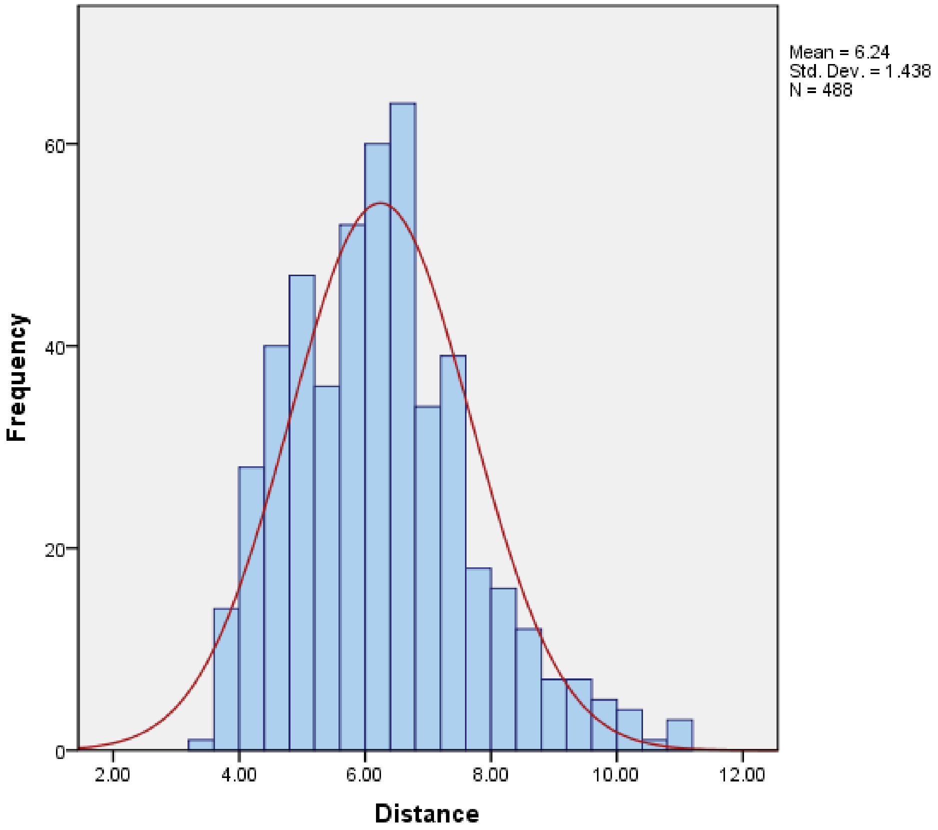
Table 8.10 Percent Within Various Degrees of Comparability

Scale	% Same sten	% Within 1 sten	% Within 2 stens
Warmth/A	18.2	56.4	83.4
Reasoning/B	16.2	50.2	79.7
Emotional Stability/C	24.2	64.5	86.7
Dominance/E	22.1	61.1	84.6
Liveliness/F	23.8	69.7	93.6
Rule-Orientation/G	28.1	71.3	90.8
Social Boldness/H	27.7	76.2	95.5
Sensitivity/I	28.9	69.2	90.8
Vigilance/L	20.1	61.1	87.1
Abstractedness/M	30.3	77.0	92.4
Privateness/N	24.0	67.0	89.1
Apprehension/O	28.9	72.3	93.2
Openness to Change/Q1	21.1	59.0	86.9
Self-Reliance/Q2	13.9	48.4	79.3
Perfectionism/Q3	28.3	75.0	93.0
Tension/Q4	34.6	78.7	96.1
Mean	**24.4**	**66.1**	**88.9**

Note: Equivalency sample, N=488. From left to right, percentages are for cumulative ranges.

Similarly, Figure 8.2 shows the "Euclidean distance" distribution of profiles. The Euclidean distance has many interpretations, but is perhaps most easily imagined as the total "distance" (in sten scores) of the Fifth and Sixth Edition profiles. For example, a value of 4 means an average of one sten point across each of the 16 scores between the Fifth and Sixth Edition profiles. The concept of "distance" does not allow for cancellation (e.g., being 1 sten higher on one scale, and 1 sten lower on another would contribute square-root of 2 sten units to distance) and overweights larger differences (e.g., being 1 sten higher on one scale and 3 stens lower on another would contribute square root of 10 sten units to distance). The normal curve is overplotted and shows that the distribution deviates only slightly from normality, being slightly more peaked and slightly skewed (slightly fewer small differences, slightly more large differences). The mean of 6.11 indicates that the average distances is far less than 1 sten score point.

Figure 8.2 Distribution of Euclidean Distances Between Fifth and Sixth Edition Profiles



These analyses shows that differences in the scores obtained on the two editions are only slightly greater than would be expected by measurement error, and any differences between the scales of the two editions affect fewer than 10% of the respondents (because less than 10% deviated from the expected SEM bands). In other words, if an individual were assessed using the Fifth and Sixth Editions, practitioners might expect most scores to be within one sten score and almost 90% to be within two sten scores.

Nevertheless, users are advised to exercise caution when comparing profiles from individuals completing different editions and to administer the same edition of the questionnaire across subjects where practical.

Regression Analysis

The final equivalency analysis concerned predicted scores. There is a large library of predicted scores that are used in reporting 16pf results. To demonstrate the similarity of predictions from the two editions, a selection of 14 Fifth Edition predictive equation scores were calculated in three ways (described below). The 14 equations were chosen to represent a sample of the approximately 300 equations used in various reports. The first six equations predict Holland's (CITE) RIASEC scores. Also, four leadership equations were included (Leadership Potential and three styles: Assertive, Facilitative, and Permissive). Four predictions developed during the Fifth Edition and documented in that manual (Conn & Rieke, 1994; Self-Esteem, Empathy, Creative Potential, and Creative Achievement).

A comparison was made of between the Fifth Edition calculations (using the Fifth Edition equations with Fifth Edition primary sten scores) and two possible calculations using Sixth Edition sten scores; (A) using the (unmodified) Fifth Edition equations with Sixth Edition primary sten scores; and (B) using regression of the predicted score upon the Sixth Edition primary sten scores (i.e., using an updated equation with the sixth edition). The "Model A" predictions correlated with the Fifth Edition scores shows the degree of equivalence for reusing the predictive equations without modification (simply dropping Sixth Edition sten scores into equations designed for the Fifth Edition), whereas the "Model B" predictions describes the gains possible by using optimal prediction methods with Sixth Edition scores. If the second and third calculations produce similar correlations with the first calculation, then little is to be gained by revising existing equations.

These analyses represent a strong test of the equivalence of the Fifth and Sixth Edition scores because prediction depends on the equivalency of all the intercorrelations of the stens scores across the Fifth and Sixth Editions. These analyses also provide evidence about the degree of similarity that can be obtained by practitioners who have developed predictive equations using Fifth Edition sten scores and now wish to use Sixth Edition sten scores.

As shown in Table 8.11, the Sixth Edition sten scores replicate the Fifth Edition results well in terms of correlations, although there were a few modest mean differences. The Fifth Edition means and variability may be compared to the Sixth Edition using Fifth Edition equations (a) and using revised equations (b). The average correlation is 0.82 between the predictive scores created using Fifth Edition scores and those using Sixth Edition scores. Furthermore, on average this correlation is only 0.03 higher (0.85) when optimal (regression) predictions are used. The improvement as a percentage ranges from 1.1% for equation "HRL3" to 9.2% for "HTEH" but the mean and median percentages are 3.9% and 2.5%. Although it would be optimal to recreate Sixth Edition equations using Sixth Edition scores, these results show that the differences are usually trivial and never even modestly large.

Thus, predictive equations formed using Fifth Edition scores can be used interchangeably with Sixth Edition scores with small differences in prediction. New equations intended for the Sixth Edition should be created using optimal methods with the Sixth Edition stens as predictors but for most existing applications, existing Fifth Edition equations can be used with Sixth Edition scores.

However, differences observed in mean levels suggest that cut scores may need to be revised in some instances. For example, the mean of equation HTRL was 5.30 using the Fifth Edition and 4.90 using the Sixth Edition sten scores. This difference in mean might suggest a practical advantage for practitioners to revise cut scores for predictions using Sixth Edition scores. For example, if the "passing rate" was known for Fifth Edition scores, a sample of participants could be tested using the Sixth Edition and a new cut score set using the new scores based on the same percentage exceeding the hurdle (e.g., if 50% "passed" the old cut score of 4.3 and 50% of Sixth Edition respondents exceed a predicted score of 5.1, then 5.1 should be used as the cut score for that predictive equation when calculated using Sixth Edition scores).

Table 8.11 Equivalence of Predicted Scores

Code	Description	Fifth Edition		Model A			Model B		
		Mean	*SD*	Mean	*SD*	r	Mean	*SD*	r
HTRL	Holland Realistic	5.30	1.73	4.90	2.01	0.80	4.93	2.19	0.82
HTIH	Holland Investigative	5.52	1.73	4.96	1.74	0.73	4.93	2.07	0.79
HTAH	Holland Artistic	5.56	1.70	6.15	2.26	0.82	6.20	2.30	0.84
HTSH	Holland Social	4.81	1.91	5.82	2.16	0.81	5.82	2.25	0.87
HTEH	Holland Enterprising	4.48	1.66	5.28	1.77	0.76	5.45	2.25	0.83
HTCH	Holland Conventional	5.45	1.60	5.27	2.10	0.81	5.15	2.21	0.83
SESE	Self-Esteem	4.74	1.95	5.66	2.26	0.85	5.67	2.27	0.88
EMEM	Empathy	4.65	2.22	6.02	2.62	0.89	5.97	2.32	0.91
LDLD	Leadership Potential	4.62	2.08	5.52	2.52	0.89	5.60	2.25	0.91
CRCP	Creative Potential	5.11	2.01	5.80	2.47	0.85	5.81	2.28	0.87
CRCA	Creative Achievement	6.02	1.75	5.95	2.15	0.73	6.05	2.19	0.79
HRL1	Assertive Leadership Style	4.95	1.78	5.30	2.18	0.81	5.50	2.08	0.82
HRL2	Facilitative Leadership Style	5.14	1.79	5.80	2.13	0.85	5.71	2.15	0.86
HRL3	Permissive Leadership Style	6.49	2.15	5.29	2.60	0.90	5.23	2.31	0.91

Note: Fifth Edition = descriptives and correlation for Fifth Edition scores inserted into Fifth edition equations. Model A = descriptives and correlation for Sixth Edition scores inserted into Fifth Edition equations (r is the correlation of "Fifth Edition" and "Model A"). Model B = descriptives and correlation for Sixth Edition scores inserted into equations estimated by predicting Fifth Edition predicted scores (r is the correlation of "Fifth Edition" and "Model B").

Discussion

This chapter described the reliability and equivalence of the 16pf Sixth Edition scales. Both internal consistency and test–retest reliability estimates were found to be consistently high and generally higher for Sixth Edition primary scales than for Fifth Edition primary scales.

The second investigation described in this chapter concerned the equivalency of the Fifth and Sixth Edition scores. Equivalence was high in terms of the correlations, but some differences were observed in the means. In other words, the Sixth Edition scores measure the same constructs as the Fifth Edition, but sometimes the distributions of scores were different. In practice, this means that correlational relationships (correlations, validity coefficients, regression equations) created for the Fifth Edition sten scores tend to hold for Sixth Edition sten scores. However, cut scores and other criteria that depend on a numerical score value may need to be adjusted for Sixth Edition use, and users are advised, when practical, to avoid comparing profiles from individuals completing different editions. Users using raw scores will need to adjust all cut scores for Sixth Edition raw scores because the raw score values changed dramatically due to the Likert scoring (compare the sten score conversions in Table 7.2 to that of the Fifth Edition).

In summary, the 16pf Sixth Edition measures the same fundamental trait characteristics as the 16pf Fifth Edition and same personality domain is being measured, but by more reliable, more robust factor scales, as demonstrated by the increased reliability of the Sixth Edition scales scores.

References

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

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

Schuerger, J. M., Tait, E., & Tavernelli, M. (1982). Temporal stability of personality by questionnaire. *Journal of Personality and Social Psychology, 43*, 176-182.

Schuerger, J. M., Zarrella, K. L., & Hotz, A. S. (1989). Factors that influence the temporal stability of personality by questionnaire. *Journal of Personality and Social Psychology, 56*, 777-783.

Chapter 9: Construct-Related Validity of the 16pf® Sixth Edition

Introduction

Construct validation evidence of psychological test scores is a standard by which test developers establish the relationship between a test score and the theoretical construct, or trait, which the test is designed to measure (Joint Committee on the Standards for Educational and Psychological Testing, 2014). A common procedure for establishing this relationship is to correlate test scores with other measures and assessments of behavior reflecting the same underlying construct, for "it is only through the empirical investigation of the relationships of test scores to other external data that we can discover what a test measures" (Anastasi, 1988, p. 162). This chapter describes construct validity analyses of the 16pf Sixth Edition and three other well-established personality inventories. These analyses weave the 16pf scores into a broader nomological network of personality constructs and support its use.

Methodology

Several surveys were administered to a sample of participants recruited through Amazon Mechanical Turk (MTurk) in five "waves" between October 2017 and January 2018. Participants were asked to respond in one or more waves to provide within subject relationships between the surveys. Table 9.1 shows the measures included in each wave of surveys. After data cleaning, a total of 379 participants completed at least one survey, but sample sizes for particular instrument pairs were smaller. The demographics of the sample are indicated in Table 9.2, which describes the Wave 3 sample.

Table 9.1. Five Waves of Longitudinal Research Surveys

Wave	Dates	Assessments
1	July 24-28 2017	16pf Sixth Edition Form S; 16pf Fifth Edition; Demographics
2	August 8-9, 2017	16pf Sixth Edition Form S
3	November 3-19, 2017	16pf Sixth Edition; Hogan Personality Inventory
4	December 12-17, 2017	IPIP Marker Scales
5	January 12-19, 2018	Global Personality Survey

Table 9.2. Demographics of the Construct Validity Sample

	Percentages
Sex	
Man	33.9
Woman	59.2
(missing)	6.9
Age	
20-29	18.0
30-39	37.3
40-49	16.7
50-59	12.0
60-69	8.2
>=70	0.9
(missing)	6.9
Average age (*SD*)	39.8 (12.5)
Race	
African American/Black	3.4
Asian	8.2
Hispanic	6.0
Native American or Alaska Native	0.4
Native Hawaiian or Other Pacific Islander	0.4
White	72.1
2 or more races	2.6
(missing)	6.9
Education	
High school graduate	8.6
Some college	15.5
Associates degree/vocational school	13.7
Bachelor’s degree	40.8
Advanced degree (MA/professional degree/doctorate)	14.6
(missing)	6.9

Note: N = 233; Wave 3 sample.

Measures

As previously discussed, the 16pf is a measure of normal personality. As such, its personality factors should converge with other normal personality measures. Three well-known and well-established measures were chosen to establish construct convergence and divergence. To demonstrate construct validity, the 16pf Sixth Edition questionnaire was administered alongside the 50-item IPIP “Big Five” marker scales (Goldberg, 1992);

the Hogan Personality Inventory (HPI; Hogan & Hogan, 2007); and the ViewPoint General Personality Survey (GPS; Abraham & Morrison, 2010).

Because the 16pf was not developed to specifically measure the Big Five factors, identifying construct validity requires additional steps. Results of past studies and rational judgment can be used to hypothesize how the 16pf primary factors should relate to scales on other inventories. Results from a previous study between the 16pf and the NEO-PI (H. E. P. Cattell, 1996) and a comparison of the factor definitions/measurements by a team of psychologists yielded a set of hypothesized relationships. Because the 16pf measures narrow traits, several 16pf factors are expected to be significantly related to the Big Five factor of interest. Within the pattern of convergence, there are certain 16pf factors that are more central to the Big Five measurement than others, they are noted with asterisks. Those central factors should have the highest magnitude correlations when looking at the convergence pattern. For two Big Five measures, Neuroticism and Extraversion, several 16pf factors are marked as central, whereas the other factors have one main 16pf factor. Table 9.3 is a summary of the hypothesized relationships between the 16pf primary factors and the Big Five factors. The plus sign suggests a strong, positive relationship and the minus sign purports a strong, negative relationship. To establish construct validity, it is important to show convergence (relationships between similar measures) as well as divergence (little or no relationships with dissimilar measures). We hypothesize divergence by theorizing that the cognitive component of the 16pf (Factor B: Reasoning) is not related (zero correlation) to any of the personality factors in the other measures. Additionally, low to no correlation between 16pf Factors and Big Five scales provides additional evidence of divergence. This table is the basis for establishing construct convergence and divergence.

Table 9.3

16 pf Global Primary Factors	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
Factor A: Warmth		+		+*	
Factor B: Reasoning	0	0	0	0	0
Factor C: Emotional Stability	-*				
Factor E: Dominance		+*			
Factor F: Liveliness		+*			
Factor G: Rule Consciousness				+	+
Factor H: Social Boldness		+*			
Factor I: Sensitivity			+	+	
Factor L: Vigilance	+			-	
Factor M: Abstractedness			+		-
Factor N: Privateness		-		-	
Factor O: Apprehension	+*	-			
Factor Q1: Openness to Change			+*	+	
Factor Q2: Self-Reliance		-		-	
Factor Q3: Perfectionism					+*
Factor Q4: Tension	+*				

***** ***Indicates a central relationship; expected to be strongest***

The first of the Big Five factors is Neuroticism. This factor is also known as Emotional Stability. Individuals who score high in this factor are typically considered emotionally unstable and not particularly well-adjusted. These individuals are often plagued with worry and tend to have lower self-confidence. As such, the 16pf primary factors proposed to correlate are Factor C (Emotional Stability) in a negative direction and Factors L (Vigilance), O (Apprehension), and Q4 (Tension) in a positive direction.

Extraversion is characterized by traits related to being social and outgoing. Extraverted individuals enjoy the company of others and seek out experiences with other people. They are often lively and energetic, and feel confident in social situations. As such, the 16pf primary factors proposed to positively relate to this factor are Factor A (Warmth), Factor E (Dominance), Factor F (Liveliness). On the opposite end, Factors N (Privateness), O (Apprehension), and Q2 (Self-Reliance) are expected to be negatively related to this factor.

Openness is a personality factor characterized by traits related to seeking out new and different experiences and perspectives. Individuals high in this factor tend to enjoy trying new things, discussing topics from all perspectives and generating new ideas. The 16pf factors proposed to relate to this factor are Factor I (Sensitivity), Factor M (Abstractedness), and Factor Q1 (Openness to Change).

Agreeableness is characterized by friendliness and cooperation. Highly agreeable individuals tend to work to maintain peaceful and collaborative relationships with others. They do not create conflict and are often willing to defer to the wishes of others. They enjoy the company of others and are concerned about the welfare and well-being of those around them. The 16pf factors expected to positively relate to this Big Five factor are Factor A (Warmth), Factor G (Rule Consciousness), Factor I (Sensitivity), and Factor Q1 (Openness to Change). It is also expected that Factors L (Vigilance), N (Privateness), and Q2 (Self-Reliance) are negatively related to this factor.

Last, Conscientiousness is a factor related to working hard and being dependable. Individuals high in Conscientiousness typically follow the rules, do the right thing and follow through on their responsibilities. They are reliable and often have a high attention to detail, which they apply to their tasks and projects. Given this, the following 16pf factors are expected to be positively relate to this factor, Factor G (Rule Consciousness) and Factor Q3 (Perfectionism), whereas Factor M (Abstractedness) is expected to be negatively correlated. We expect to see these general relationships when comparing the 16pf Sixth Edition Primary Factor scales to each of the Big Five factor measures used in this construct validation study. Each one is described in more detail below.

International Personality Item Pool (IPIP) Big Five Marker Scales

The IPIP marker scales are well-known scales developed by Goldberg (1992) to serve as markers of the “Big Five” personality traits. They are freely available and have become common in studies involving the five-factor model. This survey measures Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Intellect.

The IPIP website reports coefficient alpha estimates of reliability that range from 0.79 to 0.87 with a mean of 0.84. Each scale consists of 10 short items (e.g., item H34: “Am the life of the party”) designed for a Likert scale of agreement or accuracy. These items are approximately balanced in terms of positive and negative wording. They were administered with a five-point Likert agreement response scale (“Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree”) and the items were amended to be complete statements (by prepending “I” to the beginning of the item; e.g., H34: “I am the life of the party”). Item H34 indicates Factor I (Surgency or Extraversion). Other example items include: H21: “I am interested in people” indicating Factor II (Agreeableness), X87: “I am always prepared” indicating Factor III (Conscientiousness), E141: “I am relaxed most of the time” indicating Factor IV (Emotional Stability), and H1276: “I have a rich vocabulary” indicating Factor V (Intellect or Imagination).

Hogan Personality Inventory (HPI)

The Hogan Personality Inventory (HPI; Hogan & Hogan, 2007) is a well-known and widely used measure of normal adult personality. The HPI is particularly well-known among assessment professionals working with organizations. The measure is related to the Big Five but predates the wide-spread acceptance of the “Big Five” and has seven higher-order scales and over 40 homogeneous item clusters (HICS). The seven higher order constructs are Adjustment (N), Ambition, Sociability (E), Interpersonal Sensitivity (A), Prudence (C), Inquisitiveness (O), and Learning Approach. The HICS are clusters of a small number of items (3-6) on a similar topic, such as empathy, anxiousness, guilt, and so forth. Because the HICS are short, they tend to be less reliable, 0.34 to 0.86 with most having an estimated reliability above 0.50.

ViewPoint General Personality Survey (GPS)

The ViewPoint General Personality Survey (GPS) is a 155-item inventory that measures the Big Five factors of personality using 22 subscales (Abraham & Morrison, 2002). Validity of the instrument is supported by strong convergent correlations with instruments measuring similar constructs. For example, in college student samples all IPIP Big Five scales correlated highest with corresponding scales from the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991). Uncorrected convergent correlations with BFI self-ratings ranged from .62 (Agreeableness) to .87 (Stability); and uncorrected convergent correlations with BFI observer ratings ranged from .24 (Agreeableness) to .61

(Extraversion; Abraham & Morrison, 2002). Additionally, the criterion-related validity of the GPS has been supported in a number of contexts (e.g., Abraham & Morrison, 2003; Morrison & Abraham, 2003; Morrison, Abraham, & Dennis, 2004; Skyrme, Wilkinson, Abraham, & Morrison, 2005; Robson, Abraham, & Weiner, 2010). Each item is rated on a 5-point scale ranging from very inaccurate to very accurate. The GPS demonstrates acceptable internal consistency with alpha coefficients ranging from .75 for Agreeableness to .85 for Conscientiousness and Emotional Stability.

Data Analysis

Means and standard deviations for each of the measures appear in Appendix C, Tables 1 to 6. The 16pf sten scores are all close to having a mean of 5.5 and standard deviation 2.0. The statistics for the other measures depend on their score metric. For example, the IPIP scales are 5-point Likert raw scores (that range 5-50) with means of 25.3 to 39.7. Construct validity was analyzed by calculating the bivariate correlations between the 16pf primary factors and the general factors of the IPIP, HPI, and GPS and comparing the pattern of relationships (convergence and divergence) relative to expectations as outlined in Table 9.3.

Results

Table 9.4 presents the correlations of the 16pf primary factor scores with the Big Five factor scores of the comparison instruments. Refer to Table 9.3 for a review of the hypothesized relationships relative to Table 9.4.

	Neuroticism			Extraversion			Openness			Agreeableness			Conscientiousness		
16pf primary factor	IPIP	HPI	GPS	IPIP	HPI	GPS	IPIP	HPI	GPS	IPIP	HPI	GPS	IPIP	HPI	GPS
Factor A: Warmth	-.157*	.254**	.188*	.312**	.245**	.346**	.205**	0.121	.229**	.760**	.640**	.653**	0.085	.279**	0.124
Factor B: Reasoning	-0.016	0.044	-0.033	-0.104	-0.032	-0.050	0.115	.146*	0.112	-0.134	-0.089	-0.062	-0.109	0.025	0.044
Factor C: Emotional Stability	-.789**	.702**	.775**	.355**	.140*	.398**	.346**	.231**	.486**	.244**	.377**	.250**	.441**	.322**	.532**
Factor E: Dominance	-.369**	.137*	.475**	.601**	.440**	.601**	.438**	.354**	.477**	.157*	.219**	0.070	.234**	-0.092	.481**
Factor F: Liveliness	-.276**	.281**	.375**	.676**	.815**	.674**	.319**	.370**	.375**	.379**	.585**	.291**	-0.018	-0.085	0.117
Factor G: Rule-Consciousness	-0.117	.201**	.180*	-0.076	-.187**	0.052	-0.121	-.285**	0.017	.175*	.172**	.476**	.259**	.479**	.348**
Factor H: Social Boldness	-.442**	.362**	.512**	.833**	.608**	.789**	.322**	.326**	.403**	.345**	.540**	.284**	.142*	0.016	.332**
Factor I: Sensitivity	-0.027	0.081	0.108	.259**	.254**	.187*	.354**	.239**	.211**	.407**	.362**	.241**	0.064	0.051	0.147
Factor L: Vigilance	.303**	-.556**	-.319**	-.294**	-.147*	-.284**	0.019	-0.101	-0.150	-.338**	-.568**	-.568**	-0.012	-.407**	-0.136
Factor M: Abstractedness	.378**	-.365**	-.348**	0.013	.217**	-0.133	.203**	.166*	0.022	-0.062	-0.067	-.225**	-.469**	-.399**	-.375**
Factor N: Privateness	0.106	-.159*	-.164*	-.528**	-.325**	-.415**	-0.117	-0.107	-0.080	-.422**	-.520**	-.421**	0.026	-0.096	-0.015
Factor O: Apprehension	.735**	-.681**	-.743**	-.440**	-.201**	-.455**	-.283**	-.231**	-.472**	-.152*	-.323**	-.181*	-.329**	-.187**	-.419**
Factor Q1: Openness to Change	-.194**	.141*	.272**	.431**	.509**	.359**	.616**	.707**	.657**	.311**	.301**	0.149	0.002	-0.122	.212**
Factor Q2: Self-Reliance	.169*	-.276**	-.202*	-.459**	-.424**	-.405**	-0.047	-.143*	-.169*	-.348**	-.570**	-.389**	0.093	-.135*	-0.001
Factor Q3: Perfectionism	-.142*	0.093	0.139	0.004	-0.090	0.093	-0.032	-0.093	0.029	0.075	-0.003	.205*	.665**	.314**	.425**
Factor Q4: Tension	.616**	-.686**	-.603**	-.277**	-.159*	-.298**	-.336**	-.314**	-.392**	-.379**	-.501**	-.382**	-.284**	-.353**	-.338**

Table 9.4: Summary of Correlations Between 16pf Primary Factors and Three Big Five Measures

Nonsignificant, zero (or close to zero) correlations between reasoning and personality measures indicate one assessment of divergence. When identifying convergence, statistical significance was used along with magnitude of correlation and the pattern of relationships between the factors of the two measures. Due to the sample size and common method variance, most of the correlations are significant. Only relationships at .30 or above were considered large enough to suggest convergence. In some cases, the magnitudes were low, lower than the threshold, but the highest magnitude observed was with the expected factor, indicating some level of convergence and partial support for the hypothesized relationship. The results of the analysis are described below, in detail, for each of the comparison personality measures and their factors.

IPIP

Neuroticism (Emotional Stability)

The Goldberg IPIP's measure of Emotional Stability is meant to measure the Big Five measure of Neuroticism. As shown in Table 9.3, several 16pf factors are expected to relate to this factor. The 16pf's Factor C (Emotional Stability), Factor O (Apprehension), and Factor Q4 (Tension) are the central factors, and they did show the highest magnitude correlations with r=.78, r=.74 and r=.62, respectively. Emotional Stability convergence was also observed for and Factor L (Vigilance) (r=.30). Some unexpected relationships also emerged. There were strong negative relationships between the IPIP Emotional Stability scale and the 16pf factors H (Social Boldness) and E (Dominance). This finding suggests that individuals who are less emotionally unstable, as measured by the IPIP, are shy and not likely to voluntarily dominate social situations and become the center of attention. Although not hypothesized, these additional relationships make logical sense and support the pattern of relationships between the surveys. Last, a strong positive relationship was observed between IPIP Emotional Stability and the 16pf Factor M (Abstractedness) suggesting that individuals who are higher in neuroticism, according to the IPIP, are more abstract in their thinking style as measured by the 16pf. Factor B (Reasoning) was not significantly related to the IPIP factor, showing evidence of appropriate divergence. In sum, the pattern of findings for the IPIP Emotional Stability factor show strong convergence and divergence with the 16pf.

Extraversion

Several 16pf primary factors were hypothesized to relate to the IPIP Big Five factor of Extraversion: Factor A (Warmth), Factor E (Dominance), Factor F (Liveliness), Factor H (Social Boldness), Factor N (Privateness), Factor O (Apprehension), and Factor Q2 (Self-Reliance), where the latter three were expected to be negatively related. The central factors E, F, and H were expected to have the highest magnitudes. The observed pattern of relationships supports all of these hypotheses. They all related significantly in the expected direction. Two additional primary factors, Emotional Stability and Openness to Change, were significantly positively related to the IPIP Extraversion above the threshold (r=.36 and r=.43). While unexpected, it is logical to expect individuals who enjoy social interaction and attention to also be emotionally stable and open to experiences and people around them. No illogical or unexpected relationships were observed. Factor B (Reasoning) was not significantly related to the IPIP factor, showing evidence of divergence. In sum, the pattern of findings for the IPIP Extraversion factor shows strong convergence and divergence with the 16pf.

Openness

The 16pf central factor for relating to the IPIP's measure of Openness is factor Q1 (Openness to Change). Two other factors were expected to correlated Factor I (Sensitivity) and Factor M (Abstractedness) because these two factors are related to being open to abstract ideas and experiences. The correlational results strongly supported the convergence of IPIP Openness and the 16pf Factor Q1 (Openness to Change) with r=.62. The other hypothesized relationships show some additional support for convergence with Factor I moderately correlated (r=.35) and Factor M significantly related, but with a lower magnitude (r=.20). This IPIP measure showed several moderately significant correlations with 16pf primary factors that were not initially hypothesized. Factors C (Emotional Stability), E (Dominance), F (Liveliness), H (Social Boldness), and Q4 (Tension) showed moderate relationships with the IPIP scale. Looking at the pattern of relationships, it suggests that the IPIP measure of Openness is relating to primary factors in the 16pf that are consistent with Extraversion. Although there were several unexpected relationships, the most direct comparison of IPIP Openness with the 16pf Openness to Change was the strongest and is the main driver of the convergence between the assessments. With regard to divergence, Factor B (Reasoning) was not significantly related to the IPIP factor, showing support. In sum, the pattern of findings for the IPIP Openness factor shows acceptable convergence and divergence with the 16pf.

Agreeableness

Given the definition of Agreeableness as a factor that includes friendliness, compassion, compliance, and group orientation, there are several 16pf factors that were hypothesized to relate: Factor A (Warmth), Factor G (Rule-Consciousness), Factor I (Sensitivity), Factor L (Vigilance), Factor N (Privateness), Factor Q1 (Openness to Change), and Factor Q2 (Self-Reliance). The central factor that embodies the concept of Agreeableness the most is Factor A (Warmth). The pattern of results supports this expectation. Factor A and IPIP Agreeableness correlate strongly (r=.76). The other hypothesized factors are all significantly correlated with IPIP Agreeableness. Three other 16pf factors showed some moderate correlation with IPIP Agreeableness, Factor F (Liveliness), Factor H (Social Boldness), and Factor Q4 (Tension). These results suggest that Agreeable individuals, as measured by the IPIP, are also social, lively, and relaxed. One can see how these traits could lead someone to be considered Agreeable in addition to the other factors described. As such, these unexpected relationships do not detract from the convergence observed. Divergence was observed with the Factor B (Reasoning) scale showing no relationship with the IPIP factor. In sum, the pattern of findings for the IPIP Agreeableness factor shows acceptable convergence and divergence with the 16pf.

Conscientiousness

Descriptions of Conscientiousness describe these individuals as being serious, rule followers, hardworking, perfectionistic and somewhat rigid in their thinking. Given this description, 16pf factor M (Abstractedness) was expected to be negatively related and factors G (Rule-Consciousness) and Q3 (Perfectionism) were expected to be strongly positively related to IPIP Conscientiousness. Perfectionism is the 16pf factor with the closest linkage to the Big Five Conscientiousness and is expected to have the highest relationship of all the 16pf factors. As expected, the strongest observed correlation was between Factor Q3 (Perfectionism) and the IPIP Conscientiousness (r=.67). The results supported the other hypothesized relationships except Factor F (Liveliness), which was not related (r=-.02). Additionally, Factors C (Emotional Stability) and O (Apprehension) were moderately correlated, suggesting the individuals who are deemed Conscientious on the IPIP tend to be more emotionally stable and self-assured. Even though not all hypothesized relationships were supported, the most direct measure (Factor Q3) converged nicely with the IPIP. Divergence was also clearly observed with the lack of relationship seen with the Factor B (Reasoning) scale. In sum, the pattern of findings for the 16pf and IPIP Conscientiousness shows acceptable convergence and divergence.

HPI

Neuroticism (Adjustment)

The same 16pf factors were expected to be related to the HPI's measure of Adjustment as the IPIP's Emotional Stability because both are Big Five measures of Neuroticism. The 16pf's three central factors Factor C (Emotional Stability), Factor O (Apprehension) and Factor Q4 (Tension) were observed as the highest correlates with r=.70, r=-.68, and r=-.69. The additional hypothesized relationship with Factor L (Vigilance) was also strongly supported. The HPI Adjustment measure converges strongly with the expected 16pf factors. Similar to the IPIP measure, Factors H (Social Boldness) and M (Abstractedness) moderately correlated with the Adjustment scale, suggesting that emotionally stable individuals are more likely to be the center of attention and are more grounded in their thinking. Adequate divergence was observed with the lack of relationship between the Factor B (Reasoning) scale and the HPI scale. In sum, there is strong evidence of convergence and divergence between the 16pf and HPI regarding Neuroticism.

Extraversion (Sociability)

HPI's Sociability and IPIP's Extraversion are measures of the same Big Five factor. As such, the same hypothesized relationships were expected, positive relationships with Factor A (Warmth), Factor E (Dominance), Factor F (Liveliness), Factor H (Social Boldness), and negative relationships with Factor N (Privateness), Factor O (Apprehension), and Factor Q2 (Self-Reliance). With respect to convergence, the strongest correlates were Factor F (Liveliness) and Factor H (Social Boldness) with r=.82 and r=.61 respectively, supporting two of the three central factors. Moderate significant correlations were observed with the other hypothesized relationships. Like the IPIP, a strong, positive relationship to Openness to Change was observed, r=.51. The pattern of relationships between the 16pf and the IPIP and HPI with regard to Extraversion are highly consistent and supportive of convergence. Factor B (Reasoning) was not related HPI's Sociability, showing support for divergence. In sum, there is strong evidence for convergence and divergence for the 16pf factors and HPI's Extraversion.

Openness (Inquisitive)

The HPI's measure of Big Five Openness is labeled Inquisitive. Three 16pf factors are expected to be related with the 16pf Factor Q1 (Openness to Change) being the central factor. Factor Q1 clearly emerged as the most highly related factor (r=.71) and established the basis for convergence. The other two 16pf factors hypothesized to relate were significantly correlated but have lower magnitudes. A similar pattern of relationships emerged with unhypothesized factors. 16pf Factors E (Dominance), Factor F (Liveliness), Factor H (Social Boldness), and Factor Q4 (Tension) showed moderately strong relationships with HPI Inquisitiveness. This provides further evidence that some of the central 16pf measures of Extraversion are also related to Openness measures. This relationship is logical as one can see how individuals who are willing to put themselves "out there" to try new things are also comfortable being "out there" socially. Factor B (Reasoning) showed divergence with a nonsignificant relationship with HPI Inquisitiveness. All together these results support acceptable levels of convergence and divergence with this factor.

Agreeableness (Interpersonal Sensitivity)

HPI's Interpersonal Sensitivity measure strikes a close resemblance to the Big Five Agreeableness factor. As with IPIP Agreeableness, there are several 16pf factors expected to relate to HPI Interpersonal Sensitivity: Factor A (Warmth), Factor G (Rule-Consciousness), Factor I (Sensitivity), Factor L (Vigilance), Factor N (Privateness), Factor Q1 (Openness to Change) and Factor Q2 (Self-Reliance). The central 16pf factor that is expected to most highly relate to Agreeableness is Factor A (Warmth). The pattern of results supports this expectation. The highest magnitude relationship observed between the 16pf and HPI Agreeableness was Factor A with r=.64. The other hypothesized factors all significantly correlate with HPI Agreeableness but with Factor G (Rule-Consciousness) showing one of the lowest magnitudes. Unexpectedly, there were five other 16pf primary factors that showed moderate relationships with HPI's Interpersonal Sensitivity, Factors C (Emotional Stability), F (Liveliness), H (Social Boldness), O (Apprehension), and Q4 (Tension). These results are consistent with the IPIP, suggesting that the 16pf factors related to being well-adjusted, relaxed, and socially comfortable are also related to Agreeableness. Even with the unexpected additional relationships, the results support convergence for the 16pf factors and Agreeableness. Divergence is observed with the nonsignificant relationship between HPI Interpersonal Sensitivity and Factor B (Reasoning). In sum, the 16pf shows acceptable levels and convergence and divergence.

Conscientiousness (Prudence)

HPI's Prudence measure is a factor strongly related to Big Five Conscientiousness. The 16pf primary factor expected to be most highly related is Factor Q3 (Perfectionism) with moderate relationships with Factors G (Rule-Consciousness) and M (Abstractedness). The observed correlations partially supported these hypothesized relationships. Perfectionism was not the highest magnitude correlation across the 16pf primary factors, but it was the highest magnitude correlation for Factor Q3 (Perfectionism) across the HPI factors. The other factors did show moderately significant relationships. Two moderate relationships emerged that were not hypothesized, Factors C (Emotional Stability) and Q4 (Apprehension). This pattern was observed with the IPIP as well suggesting that Conscientious individuals also tend to be relaxed and well-adjusted as measured by the 16pf. The pattern still supports convergence and the Factor B (Reasoning) scale is not related showing divergence. In all, there is adequate construct validity support for the 16pf and HPI Prudence.

Ambition and Learning Approach

The HPI also includes two additional personality measures in its assessment, Ambition and Learning Approach. The Ambition factor is related to achievement and goal orientation which is a construct that psychologists have discussed for years. Some believe it is a separate construct, while others suggest it is subfactor of Conscientiousness. The 16pf does not have a clear measure of ambition but the relationships that do exist can help others to better understand the HPI Ambition measure. The pattern of relationships suggests that HPI Ambitious individuals are emotionally stable, socially bold, dominant and confident. These traits seem to logically describe individuals who show ambition.

Learning approach is another trait within the HPI that does not map to the Big Five. This measure asks individuals about the learning style, interest in education, feelings about continuous learning and self-reported math ability. This construct did not correlate highly with any 16pf primary factors; the strongest correlations were with Openness to Change and Reasoning. This was the only measure that correlated with the Factor B (Reasoning) scale. Although interesting, these measures will not be used to draw conclusions about the construct validity of the 16pf.

GPS

Neuroticism (Stability)

The GPS measures Neuroticism with a factor called Stability. The expected relationships are in line with the other two Big Five measures. The results were very consistent with the other measures showing strong positive relationships with the central factors of Factor C (Emotional Stability), Factor O (Apprehension), and Factor Q4 (Tension). Moderately strong and significant relationships were observed with Factors H (Social Boldness), L (Vigilance), and M (Abstractedness). Two moderate relationships observed with the GPS measure of Stability were Factors E (Dominance) and F (Liveliness). Although the other measures did show significant relationships with these factors, there was not a consistent pattern across all Big Five measures. The pattern of relationships overall strongly supports convergence. The nonsignificant relationship with Factor B (Reasoning) shows divergence. In sum, the pattern of findings for the GPS Stability factor show strong convergence and divergence with the 16pf.

Extraversion

The 16pf factors central to Extraversion are Factor E (Dominance), Factor F (Liveliness), and Factor H (Social Boldness). Other factors expected to converge with the GPS

Extraversion factor were Factor A (Warmth), Factor N (Privateness), Factor O (Apprehension), and Factor Q2 (Self-Reliance), where the latter three were expected to be negatively related. The pattern of results observed show strong support for these hypothesized relationships. The high magnitude correlations were seen with the central factors: Factor E (Dominance) r=.60, Factor F (Liveliness) r=.38, and Factor H (Social Boldness) r=.79. Moderate relationships existed for the others. Two unexpected moderate relationships emerged with Factors C (Emotional Stability) and Factor O (Apprehension). These factors emerged with the IPIP Extraversion measure as well. These data suggest that some Extraversion measures relate to Emotional Stability and Apprehension. Overall, the pattern of relationships shows strong convergence between the 16pf and the GPS measure of Stability. Divergence was observed with the nonsignificant relationship between Factor B (Reasoning) and the GPS Stability measure. These results support convergence and divergence for the Extraversion with the 16pf and GPS.

Openness

Three 16pf factors were expected to be related to the GPS Openness measure, Factors I (Sensitivity), M (Abstractedness), and Q1 (Openness to Change). As the central factor, the 16pf Factor Q1 (Openness to Change) was expected to be the strongest correlate. It did clearly emerge as the most highly related factor (r=.66). Sensitivity was significantly correlated but Abstractedness was not related, thus showing partial support for the hypothesized relationships. However, there was additional support for the relationships between Big Five Openness and the 16pf factors E (Dominance), F (Liveliness), H (Social Boldness), and Q4 (Tension). These factors emerged as moderate correlates across all of the measures. Divergence was established with the nonsignificance of the Factor B Reasoning scale with the GPS Openness measure. Overall, there is acceptable convergence and divergence with the 16pf and GPS Openness.

Agreeableness

Several 16pf factors were hypothesized to relate to the GPS Agreeableness scale: Factor A (Warmth), Factor G (Rule-Consciousness), Factor I (Sensitivity), Factor L (Vigilance), Factor N (Privateness), Factor Q1 (Openness to Change) and Factor Q2 (Self-Reliance). The central factor that embodies the concept of Agreeableness the most is Factor A (Warmth). The pattern of results supports this expectation. Factor A and GPS Agreeableness correlate strongly (r=.65). Moderate significant relationships exist with the other hypothesized factors except Openness to Change. Unlike the other two Big Five measures, the relationship was not significant. The GPS measure also strayed from the other measures by showing a significant relationship with Factor G (Rule-Consciousness) and not showing moderate relationships with Factors F (Liveliness), H

(Social Boldness), and I (Sensitivity), suggesting a qualitative difference between the GPS measure and the other two measures. Despite the discrepancies, the overall results show adequate support for convergence given the strong relationship with the central factor. Divergence was established through the lack of relationship between the Factor B (Reasoning) scale and the GPS Agreeableness factor.

Conscientiousness

The 16pf primary factors of Q3 (Perfectionism), G (Rule Consciousness), and M (Abstractedness) were hypothesized to related to GPS Conscientiousness. The central factor of Perfectionism was expected to be the highest correlate because of its focus on planning, organization and attention to detail—all subconstructs of Big Five Conscientiousness. Although all of the hypothesized relationships were supported with moderate significant correlations, the 16pf Factor Q3 did not have the highest correlation. Higher correlations were observed with Factors C (Emotional Stability) r=.53 and Factor E (Dominance) r=.48, and moderate correlations were also observed with Factor H (Social Boldness), Factor O (Apprehension), and Factor Q4 (Tension). The pattern of correlations with the GPS Conscientiousness scale is discrepant from the patterns seen with the other two Big Five measures. It is likely to measure additional aspects of Conscientiousness not included in the other measures. Despite the differences, these results do provide adequate support for the convergence between the 16pf and the GPS Conscientiousness scale with the significant moderate correlations among the hypothesized relationships. Divergence was also clearly observed with the nonsignificant relationship between the Factor B (Reasoning) scale and GPS Conscientiousness.

Overall Construct Validity Conclusions

When comparing the correlations between the 16pf factors and the three Big Five measures used in this study, there are some clear, consistent, and logical patterns of convergence between the 16pf and the Big Five. Table 9.4 shows a summary of correlations between the 16pf and all three Big Five measures. All of the hypothesized central factors were supported across every Big Five measure, showing strong convergence. Discrepancies emerged among the factors that were expected to be correlates but not central to the measurement of the Big Five measure. None of the Big Five factors correlated with the Factor B (Reasoning) scale, showing strong divergence. The results provide strong support for the construct validity of the 16pf as a personality instrument. In general, the primary factors converged where they should have converged and diverged where they should have diverged.

This construct validity study provides consistent results and detailed information about how the 16pf primary factors relate to the Big Five. Table 9.5 below shows a revised version of the construct validity hypothesis table. This is a summary table used to graphically present the expected relationships between the 16pf primary factors and the Big Five measures. Using the results seen in Table 9.4, if moderately strong correlations (>=.30) existed in all three studies, it was included in this table as an expected relationship. For many factors, two of the three studies supported a relationship and future research may support adding it to the model. Some of the initially hypothesized relationships have been removed and others have been added. All of the central factors remain the same. The Big Five factor that is the most different is Openness. Factors I and M have been removed and E, F, H, and Q4 have been added.

Table 9.5: Revised Table of Expected Relationships Between the 16pf Primary Factors and the Big Five

16 pf primary factors	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness
Factor A: Warmth				+*	
Factor B: Reasoning	0	0	0	0	0
Factor C: Emotional Stability	-*				+
Factor E: Dominance		+*	+		
Factor F: Liveliness		+*	+		
Factor G: Rule Consciousness					
Factor H: Social Boldness	-	+*	+		
Factor I: Sensitivity					
Factor L: Vigilance	+			-	
Factor M: Abstractedness	+				-
Factor N: Privateness		-		-	
Factor O: Apprehension	+*				
Factor Q1: Openness to Change		+	+*		
Factor Q2: Self-Reliance		-		-	
Factor Q3: Perfectionism					+*
Factor Q4: Tension	+*		-	-	

* *Indicates a central relationship; expected to be highest magnitude*

Summary

This chapter focused on construct validity evidence to enhance the interpretation of the 16pf scores by weaving them into a nomological network of the Big Five personality factors. The construct validity of the 16pf Sixth Edition primary scales was demonstrated by the relationships resulting from the correlational analyses with three well-established Big Five measures.

The 16pf primary factors have clear connections to the Big Five scales. Overall, the relationships found between the 16pf Sixth Edition Primary Factor scales and the comparison personality inventories are quite consistent with the traditional scale meanings (R. B. Cattell, Eber, & Tatsuoka, 1970; Conn & Rieke, 1994; H. E. P. Cattell & Schuerger, 2003). Strong convergent and divergent evidence was presented.

A reminder that the 16pf was not designed to be a measure of the Big Five. The 16 primary factors were established as narrow traits that can provide specific and detailed information about an individual's personality. The redesign and development of the Sixth Version required a thorough review of its psychometric properties, including construct validity. Given the academic research and acceptance of the Big Five personality model, it was used as a structure for showing the convergence and divergence of the 16pf primary scales. The results of this study provide strong evidence for the content of the scales and what they purport to measure while also establishing a strong model for how the 16pf can be interpreted within the Big Five.

The interpretations presented in this chapter should be considered as representing only part of the whole definition for the primary scales. Other aspects of behavior, such as face-to-face interviews and biographic information, need to be integrated with scale interpretations to gain the richest meanings from 16pf scores. For further discussion of Sixth Edition scale meanings, see Chapter 3.

References

Anastasi, A. (1988). *Psychological Testing*. New York, NY: Macmillan.

Abraham, J. D., & Morrison, J. D. Jr. (2002). *Performance Perspectives Inventory: PPI technical manual, version b1.3.* Tulsa, OK: A & M Psychometrics, LLC.

Abraham, J. D., & Morrison, J. D. Jr. (2003). Relationship between the Performance Perspectives Inventory's Conscientiousness scale and the performance of corporate security guards. *Applied HRM Research, 8*, 45-48.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the 16pf*. Champaign, IL: Institute for Personality and Ability Testing, Inc.

Cattell, H. E. P., & Schuerger, J. M. (2003). *Essentials of 16pf assessment*. New York, NY: John Wiley & Sons.

Conn, S. R., & Rieke, M. L. (1994). *16pf fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.

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

Gough, H. G. (1987). *California Psychological Inventory administrator's guide*. Mountain View, CA: CPP, Inc.

Hogan, R., & Hogan, J. (2007). *Hogan Personality Inventory manual (3rd ed.).* Tulsa, OK: Hogan Assessment Systems.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory--versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Joint Committee on the Standards for Educational and Psychological Testing. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, American Psychological Association, & National Council for Measurement in Education.

Morrison, J. D. Jr., & Abraham, J. D. (2003). Validity study relationship between the Performance Perspectives Inventory's Selling Scale and job performance of Real estate agents. *Applied HRM Research, 8*(2), 85-88.

Morrison, J. D. Jr., Abraham, J. D., & Dennis, G. B. (2004). Validity study relationship between the performance perspectives inventory's leadership scales and job performance of quick service restaurant managers. *Applied HRM Research, 9*, 31-34

Prince, J. P., & Heiser, L. J. (2000). *Essentials of Career Interest Assessment.* New York, NY: Wiley & Sons.

Robson, S. M., Abraham, J. D., & Weiner, J. (2010). Characteristics of successful direct support professionals: an examination of personality and cognitive ability requirements. *International Journal of Selection and Assessment, 18*, 215-219

Skyrme, P., Wilkinson, L., Abraham, J. D., & Morrison, J. D. Jr. (2005) Using personality to predict outbound call center job performance. *Applied HRM Research, 10*, 89-98.

Chapter 10: Validity in Organizations

Personality has a long history of applied use in organizations. Although interest in research, and thus organizational applications, waned in the 1960s, 70s and 80s, personality assessment has once again been recognized as an important component influencing employee behavior at work (Hough, 2001; Seibert & DeGeest, 2017). Two seminal articles demonstrating the relationships between personality traits across a variety of questionnaires and job performance have spurred this renewed interest in personality in organizations (Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991). These articles used a method known as meta-analysis to provide empirical estimates of the relationships based on hundreds of studies and thousands of individuals.

Although these two studies have provided a foundation for the assertion that personality is a valid predictor of behavior at work, they also raise questions, which are still being addressed. Much of the recent work on personality in organizations has been directed toward these additional questions (e.g., Rothstein & Goffin, 2006). One of the questions raised by these early meta-analytic studies stems from the observation that there was substantial variation in the degree to which personality variables were related to performance. Some of this variation is due to the job itself (Dudley, Orvis, Lebiecki, & Cortina, 2006). In other words, not all personality variables are related to performance in all fields. As Tett and his colleagues (1991) observed, careful consideration of the conceptual links between personality traits and job requirements will yield more impressive validity estimates.

Personality has been shown to predict many outcomes relevant to organizations. This includes teamwork, leadership, performance, training outcomes, turnover, work withdrawal (including intentions to quit), goal setting, leader derailment, and mentoring activity (Hough & Oswald, 2008). Despite the impressive body of evidence suggesting that personality does influence important behaviors and outcomes, this influence can be strongly moderated by situational influences. One example of this is the degree of autonomy the job offers (Barrick & Mount, 1993). As a result, at times it can be difficult and misleading to generalize findings from a single study to a local context without considering the similarities and differences of the situations. Although job titles may sound similar, the actual requirements may in fact be quite different.

The 16pf Questionnaire and Organizational Applications

Research using earlier versions of the 16pf Questionnaire has shown the primary factors to be meaningfully related to performance and other significant organizational outcomes. The purpose of the current chapter is to summarize this legacy of validity evidence supporting the use of the 16pf Questionnaire in organizational applications, such as employee selection, and stress and burnout. The research presented and summarized in this chapter spans a variety of diverse occupations, cultures, outcomes, and applications. No single study can be judged as the definitive answer to the question of whether the 16pf Questionnaire is a valid tool for organizational applications. Rather, the pattern of evidence across the body of available research should be judged in its entirety.

Determining the Job Requirements

Most 16pf applications should begin with some form of job analysis (see Brannick, Levine, & Morgeson, 2007) to determine the 16pf primary and global factors most related to the position. Job analysis methods can take several forms, including interviews, surveys, direct observation, and review of job materials or descriptions, and often involve many of these methods.

Most literature on job analysis focuses on abilities where "more" is always "better," but for personality traits, some indication of the direction and level is highly desirable. For example, one position may require higher levels of Factor A+, reflecting greater nurturing interpersonal warmth, whereas another job requires A- scores (e.g., workers who make difficult decisions and need to hold others at arms length).

The optimal level of personality score is often derived from samples of incumbent workers. For example, in a sample of police officers, Factor G was elevated, indicating that police officers tend to be rule abiding. But it was also found that those with more elevated scores were rated as less effective by their supervisors, presumably because higher G+ scores were associated with less successful behaviors (e.g., enforcing minor issues, focusing on rules to the exclusion of building community relationships, generating excessive paperwork, etc.). Of greatest importance is documenting the logical, and when possible, empirical relationship between success on the job and the specified personality dimension including the shape of that relationship (positive, negative, asymptotic).

Using Global Versus Primary Factors in Prediction

The Big Five model is widely known and widely used for many purposes, including criterion-related validity research. The focus of 16pf research, in contrast, has been on the more specific primary factor scores, and there is evidence that prediction is enhanced by using more specific predictors (Mershon & Gorsuch, 1988; Dudley et al., 2006; Judge, Rodell, Klinger, Simon, & Crawford, 2013). This is particularly true when the behavior of interest can be narrowly specified, as opposed to more global, overarching evaluations of behavior.

In theory, prediction using a linear composite of 16pf primary scores should always outperform prediction using a linear composite of global scores. This can be seen by examining the statistical mechanics of a linear composite (which simply means a weighted sum of the scores). Every population has an (unknown) optimal combination of primary scales that best predict performance. Because the global scales are weighted combinations of the primary scores, unless the optimal population equations happen to exactly coincide with the weights used to compute the global scores, a practitioner should generally be better off predicting a criterion using an optimal method like least-squares multiple regression to combine primary scale scores, given a sufficiently large sample.

However, there is a "bandwidth-fidelity dilemma" (Cronbach & Gleser, 1965) because the scores of the shorter primary factors have more measurement error then the longer global factors. In one early study of this effect using the 16pf Questionnaire, Mershon and Gorsuch (1988) examined 16 datasets and found that statistically adjusted R2 values roughly doubled when using the 16 primary scales over models using six factors (the five global scales and Reasoning/B). For example, multiple-R rose from 0.28 to 0.52 in a dataset of apprentice aircraft engineers when moving from using 6 (global scales and Reasoning/B) to the 16 primary scores. Because of results like this, the authors suggested that the primary factor scales were far more effective than the global scales. More recently, Grucza and Goldberg (2007) examined scores at three levels: higher level constructs (e.g., the 16pf global factors); middle-level constructs (e.g., the 16pf primary factors); and lower level constructs (e.g., Hogan HPI HICs). Their results replicated the direction of the earlier study but with far smaller effects. Across six self-reported behavioral criteria (e.g., self-reported undependability) they found that the mean cross-validated multiple-R was 0.40 when 16pf global scores were used and 0.41 when 16pf primary scale scores were used. Furthermore, the Grucza and Goldberg found fairly similar results across several well-known personality instruments. Possibly the differences in effect sizes between these two studies are due to differences in the samples (the earlier study used published data, the more recent study used a single sample, the ESCS dataset; Goldberg & Saucier, 2016) or differences in the criteria (the ESCS dataset has self-reported behaviors; the earlier study used "measurable real-life data such as pay, tenure, supervisor's ratings, or occupation," Mershon & Gorsuch, 1988, p. 678). These two studies suggest that there is at least a small advantage of more specific scores, as exemplified by the 16pf primary factors.

Practitioners should note that this dilemma also affects criterion measurement. Many research studies have tended to use a single overall measure of job performance as the criterion. Yet, there is growing awareness that job performance is a multidimensional construct (Borman & Motowidlo, 1993; Campbell, 1990; Thayer, 2008). As such, research using specific measures of performance is likely to better capture important behavioral differences than research using a single, global measure.

Employee Selection

The 16pf Questionnaire has been found to be predictive of job performance across many studies and occupations (Ones, Viswesveran, & Dilchert, 2005). Recent research using the 16pf Questionnaire to predict performance at work is the focus of the next section, which summarizes the utility of the 16pf Questionnaire in several job families including sales, customer service, manufacturing, and so on. Although the most effective method of summarizing empirical literature on the validity of a construct is meta-analysis, this methodology does require sufficient studies to provide stable estimates of the true relationship. Moreover, as other authors have noted, combining personality validation studies across jobs without regard for the conceptual links between the predictor and criterion can lead to an underestimate of the relationship (Hogan & Holland, 2003). Furthermore, personality variables can be expected to be differentially related to performance across jobs (Tett et al., 1991).

The 16pf literature has a history of presenting job profiles, rather than validity coefficients predicting a criterion of performance (see R. B. Cattell, Eber, & Tatsuoka, 1970) or in focusing on a specific aspect such as entrepreneurship (Fraboni & Saltstone, 1990) or teamwork (Dulewicz, 1995). As a result, most of the research literature is not included in this section, which focuses on "criterion-related validation" studies (see Society for Industrial and Organizational Psychology, 2003). Of the remaining studies, there are several job types for which there was only a single study, and in such cases, no meaningful meta-analysis was possible, so the study was summarized. Where there were at least a few studies of a single job or job family, meta-analytic methods were used. This combination provides the most robust summary of validity evidence for specific occupations.

Sales Jobs

The 16pf Questionnaire has frequently been used in the selection process for sales jobs. This section describes research on salespersons, contact center sales representatives, and business development sales roles. Customer service is covered in the next subsection.

Sales Representatives

Fishbein, Oster, and Bedwell (2007) collected 16pf data and performance ratings on 142 incumbent sales employees in the Midwestern United States. Two positions were evaluated in the organization, Inside sales representatives and territory managers. Primary factors for the 16pf Questionnaire were regressed onto manager ratings of Job Knowledge, Selling Skills, and Attitude Toward Teamwork. In addition, an overall performance rating was also made for each employee by the supervisor. The 16pf primary factors entered into each equation were based on a conceptual mapping of the factor definitions to the descriptions of the performance dimensions.

The multiple regression models for the territory managers were not significant. However, for the Inside sales representatives, the 16pf Questionnaire was a significant predictor of performance. The multiple correlation along with the 16pf primary factors in the model are presented in Table 10.1. The results in Table 10.1 should be interpreted with the understanding that the three performance dimensions are intercorrelated and the multiple coefficients are nonindependent. In spite of the overlap, the results do show substantial relationships between the scales and job performance. Sales representatives rated as being more knowledgeable of products were characterized by the 16pf instrument as being Warm (A+), Practical (M-; low Abstractedness), Worried (O+), and Tolerant of Disorder (Q3-; low Perfectionism). Individuals rated as possessing better sales skills were characterized as being more Warm (A+), Cooperative (E-; low Dominance), Socially Bold (H+), Vigilant (L+), Practical (M-), and Worried (O+). Likewise, individuals rated as having a better attitude toward teamwork were characterized as being Warm (A+), Calm (C+), Rule- Conscious (G+), Vigilant (L+), Practical (M-), and Worried (O+). Although it may seem counterintuitive that individuals who scored higher on Apprehension/O would be rated as more effective in several performance domains, it is important to keep in mind that this was an applicant sample, and the mean for this factor was more than a standard deviation below the population average. As such it may be more reasonable to think of this as indicating that too much self-assurance inhibits sales performance, at least in this study's context.

Table 10.1 16pf Questionnaire Multiple Regression Results for Inside Sales Representatives

Performance dimension	Multiple correlation	Contributing 16pf factors
Product knowledge	.45	A, M-, O, Q3-
Selling skills	.48	A, E-, H, L, M-, O
Attitude	.41	A, C, G, L, M-, O
Overall performance	.43	A, L, M-, O

Note: N = 142; All multiple correlations significant at p<.05

Contact Center Sales Representatives

Examining the performance of a group of inbound sales representatives, Kostman (2003) examined the relative utility of three predictor domains to account for variance in sales ability. The measures included in this study were the Wonderlic Personnel Test (Wonderlic, 1992), the 16pf Questionnaire (R. B. Cattell, A. K. Cattell, & H. E. P. Cattell, 1993), and the Emotional Judgment Inventory (EJI; Bedwell, 2003). All predictor measures significantly predicted sales performance, as measured by sales revenue. Specifically, general mental ability (GMA) as assessed by the Wonderlic demonstrated a correlation of .32 with sales revenue. Kostman created an overall emotional intelligence score by summing across the scales of the EJI and observed a correlation of .28 between emotional intelligence and sales. Three of the five Global Factors from the 16pf Questionnaire were significantly correlated with sales revenue as well, specifically, Self-Control (.30), Independence (−.28), and Anxiety (−.21).

Kostman also examined the incremental validity of the 16pf Questionnaire relative to cognitive ability for sales performance. With only GMA included, the model accounted for about 10% of the variance in performance and resulted in a multiple correlation of .32. Due to Kostman’s hypotheses, only the Self-Control and Independence factors were included in the hierarchical regression model. When the two 16pf Global Factors were added to the model, the multiple correlation increased to about .50 and the model accounted for an additional 15% of the variance in performance. The incremental validity of the 16pf factors (after controlling for cognitive ability) observed in this study indicates that personality factors can help improve hiring decisions for sales positions. As such, HR professionals and hiring managers would do well to consider including a personality assessment to employee selection systems for sales positions.

Business Development Sales

In an exploratory study examining sales personnel whose primary responsibility was developing new clients within a geographic region, the 16pf Questionnaire predicted both sales performance and organizational tenure (Bedwell, 2001). The study employed a concurrent research design using incumbents (N = 64) who had been with the organization for an average of four years. The sales force had not been selected using a personality questionnaire. Indeed, the organization was a franchise and sales staff were hired by the individual franchise owner without a standard selection process across franchises. Sales performance was measured as annual revenues in US dollars.

The multiple correlation was .44 and .47 for sales performance and tenure, respectively. Specifically, lower scores on Rule-Consciousness/G and Openness to Change/Q1 were associated with higher sales, whereas higher scores on Social Boldness/H and Tension/Q4 were related to higher sales performance. For tenure, lower scores on Liveliness/F and Rule-Consciousness/G were associated with longer tenure, whereas higher scores on Social Boldness/H and Tension/Q4 were related to longer tenure. The results for Rule-Consciousness/G and Tension/Q4 for sales are perhaps not surprising, but one might expect the opposite influence on tenure (i.e., that higher Rule-Consciousness/G would lead to longer tenure and higher Tension/Q4 to shorter tenure). Probably this reflects that sales people leave sales roles when they are less successful.

An analysis of the two criterion variables shows that they are highly related (r =.71). Further examination revealed that sales personnel with five or more years of tenure demonstrated significantly higher sales than personnel with fewer than 5 years of tenure. These results suggest that sales volume is likely to be a contaminated performance criterion. That is, factors other than those directly attributable to the employee are influencing sales volume. In this case, the number of years on the job is related to sales volume. One explanation may be that the number of existing respondents is related to both tenure and sales volume. Employees who have been with the organization for fewer years may not have had the same opportunity to make client contacts as those sales employees who have been with the organization longer. Or, possibly those who are more successful in sales were therefore able and willing to stay longer.

Customer Service Jobs

Table 10.2 presents the meta-analytic estimates for the correlation between the 16pf primary factors and overall job performance for customer service representatives. The analyses are based on two independent samples from different organizations. The correlations are not corrected for any artifacts except sampling error. Table 10.2 presents the estimated true correlation, rho, and the upper and lower limits of the estimated value along with the percent of variance accounted for by sampling error.

The last column represents the amount of variation in the observed correlations for the studies included in the analyses that are accounted for by sampling error. Because the number of studies is relatively small—as are the sample sizes for the individual studies—the estimate of the percent of variance accounted for is itself estimated with a significant degree of error, and thus should be treated cautiously.

These positions required taking inbound calls from customers and resolving the customer's request as quickly and courteously as possible. Not surprisingly, being more emotionally resilient was most highly related to overall job performance. Being better able to reason through and troubleshoot problems, being agreeable and practical were also related to more effective performance.

Although being warm (A+) might intuitively seem related to a customer service role at first glance, the 16pf Factor A indicates a willingness to develop an emotional connection to others as well as a desire to take care of their needs. As efficiency in customer response is as important as being helpful, being overly concerned for the customer may lead to too much time interacting with a single customer, resulting in less time to attend to other calls.

Table 10.2 Preliminary Results of Meta-Analytic Research Customer Service

Scale	rho	Lower limit	Upper limit	% Variance accounted for
Warmth/A	.00	−.05	.05	> 100
Reasoning/B	.15*	.10	.20	> 100
Emotional Stability/C	.21*	.11	.31	> 100
Dominance/E	−.12*	−.19	−.04	> 100
Liveliness/F	−.05*	−.07	−.02	> 100
Rule-Consciousness/G	−.03	−.19	.13	92
Social Boldness/H	−.09	−.22	.04	> 100
Sensitivity/I	−.07*	−.13	−.01	> 100
Vigilance/L	−.09	−.26	.08	78
Abstractedness/M	−.11*	−.11	−.10	> 100
Privateness/N	.01	−.08	.10	> 100
Apprehension/O	.08	−.21	.05	> 100
Openness to Change/Q1	.00	−.26	.26	36
Self-Reliance/Q2	.07	−.06	.20	> 100
Perfectionism/Q3	−.02	−.24	.20	49
Tension/Q4	.00	−.09	.10	> 100

Note: *Significantly different from zero (95% confidence interval does not include zero). Sample size is 164. Results based on two correlation coefficients from two studies. Reproduced from IPAT (1999).

Managerial and Executive jobs

Hetland and Sandal (2003) applied Bass and Avolio's (2000) transformational leadership model to mid-level Norwegian managers in five different Norwegian organizations to determine how influential transformational leadership behaviors were in motivating employees beyond transactional leadership behaviors. These authors were also interested in whether personality influenced the use of transformational style of leadership. Based on a content analysis of the transformational leadership behaviors, four 16pf traits were included in the analyses: Warmth/A, Reasoning/B, Openness to Change/Q1, and Tension/Q4. The authors hypothesized that the first three factors listed would be positively related to transformational leadership and the last factor listed would be negatively related to transformational leadership.

The results were mostly consistent with the hypotheses. That is, transformational leadership predicted a substantial amount of variance in three outcome variables (Satisfaction With Leader, Leadership Effectiveness, and Motivation; all rated by one superior and two subordinates) after controlling for transactional leader behaviors. This finding helps to establish the importance of transformational leader behaviors in addition to more common transactional leader behaviors. These authors also found that higher scores on the 16pf traits of Warmth/A, Reasoning/B, Openness to Change/Q1 and lower scores on Tension/Q4 accounted for 10% of the variance in ratings of transformational leadership by subordinates. The multiple correlation after adding these traits to the multiple regression model, controlling for gender and organization type, was .40. However, these 16pf primary factors were less predictive of supervisors' ratings of transformational leadership behaviors. The multiple correlation from the regression model after controlling for manager gender and organization type, was .20. As to the relationship between the 16pf Questionnaire and leadership, the observed differences between managers' and direct reports' ratings are similar to a common finding reported in multirater feedback systems, that rater source groups tend not to agree on ratings of performance for a common target (Bracken, Timmreck, & Church, 2001; Woehr, Sheehan, & Bennett, 2005).

As mentioned earlier, the 16pf primary traits chosen for inclusion in this study were based on hypotheses regarding a conceptual mapping of both personality dimensions and performance dimensions in light of the cultural complexities that distinguish Norway, which is egalitarian with respect to power sharing, as well as valuing cooperation and good working relationships (Hofstede, 1980). As a result, the authors focused on personality traits related to agreeableness and interest in others, resulting in Dominance/E and Social Boldness/H being excluded despite having been linked to leadership (see Chapter 11). Even in cultures that value getting along over getting ahead, leaders still need to be willing to accept responsibility for making decisions and be willing to implement decisions. In addition, putting oneself forward in a leadership position inevitably brings criticism, and leaders need to be able to remain unaffected by the personal sting of criticism or disagreements. Future research on the influence of personality on leader behaviors should bear in mind the requirements of both the larger culture, as in the Hetland and Sandal (2003) study, as well as the requirements of the role itself. These "attribute by situation" interactions are likely to provide a better understanding of the influences on leader behavior than a focus on situational or individual differences alone.

Table 10.3 Meta-Analytic Results for Leadership Dimension and 16pf Primary Factors

16pf Scores	Number of validities	Total sample size	Weighted mean validity r	True population Validity ρ	Observed variance in r_{xy}	Sampling error variance in r_{xy}	Residual variance in r_{xy}	% variance explained	90% credibility Value of ρ
Leadership/LD	6	266	.13	.20	.0116	.0224	.0000	100%	.20
Warmth/A	6	266	.03	.05	.0084	.0232	.0000	100%	.05
Reasoning/B	6	266	.13	.21	.0371	.0224	.0000	60%	-.05
Emotional Stability/C	6	266	.14	.25	.0203	.0223	.0000	100%	.25
Dominance/E	6	266	.09	.15	.0187	.0228	.0000	100%	.15
Liveliness/F	6	266	.04	.07	.0184	.0231	.0000	100%	.07
Rule-Consciousness/G	6	266	-.01	-.02	.0530	.0232	.0000	44%	-.51
Social Boldness/H	6	266	.13	.19	.0094	.0225	.0000	100%	.19
Sensitivity/I	6	266	-.15	-.23	.0173	.0222	.0000	100%	-.23
Vigilance/L	6	266	-.05	-.08	.0246	.0231	.0000	94%	-21
Abstractness/M	6	266	-.01	-.02	.0077	.0232	.0000	100%	-.02
Privateness/N	6	266	-.07	-.11	.0179	.0230	.0000	100%	-.11
Apprehension/O	6	266	-.07	-.11	.0166	.0230	.0000	100%	-.11
Openness to Change/Q1	6	266	.08	.13	.0152	.0229	.0000	100%	.13
Self-Reliance/Q2	6	266	-.01	-.01	.0071	.0232	.0000	100%	-.01
Perfectionism/Q3	6	266	.02	.03	.0141	.0232	.0000	100%	.03
Tension/Q4	6	266	-.05	-.08	.0230	.0231	.0000	100%	-.08

Note: Job performance is measured by either supervisor rating composite scores or objective sales outcome measures. True population validities reflect corrections for unreliability in the criterion measures. When unavailable, supervisor ratings were assumed to have a reliability of .52 (Viswesvaran, Ones, & Schmidt, 1996). Objective outcome criteria were assumed to have a reliability of 1.00.

Meta-Analysis of Managerial Validity Studies

A review of the literature and IPAT's 16pf archives produced six validity studies of managerial roles suitable for meta-analytic combination using a combined sample of 266 managers. Table 10.3 presents the meta-analytic results for the leadership equation (see Chapter 11) and the primary factor scales. The weighted mean validity shows the sample-size weighted mean validity coefficient across all six studies. The true population validity column shows the estimated population validity after correcting for criterion unreliability and range restriction (see Hunter & Schmidt, 1990). We assumed a criterion reliability of 0.52, which is reasonable when the performance criteria consist of ratings, and they did in these six studies (Viswesvaran, Ones, & Schmidt, 1996). Range restriction was calculated using a large sample of managers extracted from operational data (i.e., the standard deviation of each score was calculated from this sample of managers and a ratio of this standard deviation to the general population standard deviation of 2.0 was used to estimate the range restriction correction).

The next four columns document the amount of variation in the six observed validity coefficients for each predictor that can be attributed to sampling error. Values of 100% in the "% Variance Explained" column indicate that all of the variability in the observed validities is attributable to sampling error, that is, not due to any real differences in the validity of these measures across the contexts examined. The 90% credibility interval represents the validity above which 90% of validities should fall in new studies when differences in organizational settings are similar to the differences observed in the meta-analysis studies. It is the lower bound of a confidence band based on variance in observed validities not attributable to sampling error variance. Therefore, if 100% of the variance observed in validities is attributable to sampling error (i.e., differences in study settings did not contribute to any real validity differences), the 90% credibility value is equal to the true population validity because there is no unexplained variance in validities upon which to base a credibility interval.

For Reasoning/B, practitioners are likely to sample population validity coefficients numerically greater than -0.05. Because this interval includes zero, this meta-analysis cannot exclude the possibility that Reasoning/B lacks validity in some settings but note that the majority of the credibility interval includes positive validities. For Rule-Consciousness/G, the credibility interval encompasses mostly negative values. One interpretation of lower percent variance accounted for Reasoning/B (60%) and for Rule-Consciousness/G (44%), and for the small negative population effect size for Rule-Consciousness/G is that unmeasured moderators affect the level of validity. One such unmeasured moderator may be the level of Reasoning/B and Rule-Consciousness/G in the sample, because higher levels of Reasoning are known to be related to performance in many occupations (Neisser et al., 1996), but in samples where all candidates have elevated Reasoning levels, the validity coefficients may be reduced.

Similarly, in samples where all candidates tend to have elevated Rule-Consciousness, even higher levels may not have beneficial effects.

Reviewing the "True Population Validity" column of Table 10.3 demonstrates that the leadership equation, Emotional Stability/C, and Social Boldness/H all have meaningful positive validity (0.20, 0.25, and 0.19, respectively). Sensitivity/I has a similarly sized negative validity (-0.23), and Reasoning/B has positive relationship with success (but as described above, may very across situations). Dominance/E also plays a slightly less influential role, with a 0.15 validity, as does Openness to Change/Q1 (0.13) and lower scores on Privateness/N (-0.11) and Apprehension/O (-0.11). Thus, these results depict an empirical picture of successful managers as tending to have elevated leadership equation scores and to be emotionally stable and socially bold and low on sensitivity. In most situations, they tend to have elevated reasoning ability. To a slightly lesser degree, they are marked by being dominantly forceful, open-minded, forthright, and self-confident.

Public Safety Jobs

The 16pf Questionnaire is used quite frequently in high-risk occupations (e.g., firefighters, sheriff's deputies, correctional officers, security, etc.) and has been shown to predict both on-the-job performance and training outcomes (IPAT, 2003). Validation research with the 16pf Questionnaire in the public safety domain has tended to focus on the four protective services dimensions developed for use with high-risk occupations. However, other studies have used either the Global Factors or the primary factors rather than the composite dimensions.

Love and DeArmond (2007) examined the ability of the 16pf instrument to predict performance, of police sergeants incrementally over assessment center ratings. In their sample of 54 police sergeant candidates, the 16pf Global Factors predicted supervisory ratings of performance after controlling for assessment center ratings during the promotion assessment process. Specifically, the Global Factors of the 16pf Questionnaire accounted for an additional 8% of the variance in performance. The assessment center ratings accounted for 16% of the variance in performance ratings. Supervisors were not aware of either the assessment center results or the 16pf scores when completing the performance ratings.

Given that the 16pf Global Factors provided incremental validity after controlling for the assessment center ratings, it is particularly interesting to note that the assessment center dimensions were identical to the job performance dimensions. The performance dimensions were developed from a job analysis and the assessment center ratings consisted of the same dimensions as the performance rating form. The assessment center ratings were combined into an interpersonal and a problem-solving dimension based on a factor analysis. The performance ratings were summed across all dimensions to arrive at an overall score after observing that the individual dimensions were all highly correlated. Even though assessment center ratings and the performance dimensions were developed from the same data, personality, in the form of the 16pf Global Factors, was still an important predictor of performance.

Meta-analysis of public safety validity studies

A review of the literature and IPAT's 16pf archives produced three validity studies of public safety positions with primary scales as predictors and total samples of 236 participants. In addition, there were seven validity studies with the four PSR dimensions (see IPAT, 2003) and ratings criteria, and five with the PSR dimensions and training criteria. Table 10.4 presents the analysis of the PSR dimensions as predictors and Table 10.5 presents results for the primary scale validities.

As described for the meta-analysis of manager roles, the weighted mean validity shows the sample-size weighted mean validity coefficient across all studies. The true population validity column shows the estimated population validity after correcting for criterion unreliability and range restriction (see Hunter & Schmidt, 1990). We assumed a criterion reliability of 0.52, which is reasonable when the performance criteria consist of ratings, and they did in these studies (Viswesvaran, Ones, & Schmidt, 1996), and 0.80 for the training criteria, which tended to be based on class grades or exam scores. Range restriction corrections were calculated using applicant standard deviations of a sample of selected candidates (IPAT, 2003, Sample B, Hired Deputies) and an assumed population standard deviation of 2.0.

The next four columns document the amount of variation across studies in the observed validity coefficients for each predictor that can be attributed to sampling error. Values of 100% in the "% Variance Explained" column indicate that all of the variability in the observed validities is attributable to sampling error (i.e., not due to any real population differences in the validity of these measures across studies). The 90% credibility interval represents the validity above which 90% of validities should fall in new studies when differences in organizational settings are similar to the differences observed in the meta-analysis studies. It is the lower bound of a confidence band based on variance in observed validities not attributable to sampling error variance. Therefore, if 100% of the variance observed in validities is attributable to sampling error (i.e., differences in study settings did not contribute to any real validity differences), the 90% credibility value is equal to the true population validity because there is no unexplained variance in validities upon which to base a credibility interval.

In Table 10.4 showing the analysis of the PSR dimensions, six of the eight 90% credibility values are positive, indicating that these combinations of dimension and criterion have overwhelmingly positive population values. The Integrity/Control and Interpersonal Relations dimensions have slightly negative 90% credibility values for supervisor ratings. Thus, the vast majority of studies would sample populations with positive validity, but for these two predictor/criteria combinations, this analysis cannot rule out those two dimensions having no validity for some populations. One interpretation of lower percent variance accounted for these two dimensions is that unmeasured moderators affect the level of validity, and one such unmeasured moderator may be the level of these traits in applicant populations, which may already have elevations on Integrity/Control and perhaps on Interpersonal Relations. In such cases, even higher levels may not have beneficial effects. Similarly, in Table 10.5, the primary scales Reasoning/B, Emotional Stability/C, Rule-Consciousness/G, Privateness/N, and Tension/Q4 have some suggestion of the possibility of moderators, although in this analysis with only three studies, these results should be interpreted with caution.

Reviewing the “True Population Validity” column of Table 10.4 demonstrates relatively strong relationships between the PSR dimensions Emotional Adjustment and Intellectual Efficiency, and performance on both subjective and training criteria (true population coefficients ranging from 0.17 to 0.39). Integrity/Control shows modest positive validities (although, as pointed out earlier, one 90% credibility interval includes zero). This probably reflects that most candidates are already elevated on Integrity/Control. Interpersonal Relations also shows modest positive true validities (although the 90% credibility interval includes zero for the subjective criteria studies). Thus, these results depict an empirical picture of successful public safety officers as generally being high on these dimensions, with Emotional Adjustment and Intellectual Efficiency having the most influence on performance.

In practice, the PSR dimensions may be used in combination. A composite of the four dimensions was created by summing (i.e., giving each equal weight) the true population validities. This analysis produced an estimated composite validity of 0.25 for ratings criteria (based on N=1018, K=7) and 0.35 for training criteria (based on N=761, K = 6).

Table 10.4 Law Enforcement/Corrections Job Family 16pf Questionnaire Protective Services Dimensions Meta-Analytic Results

16pf protective service dimensions	**Performance criterion**	**Number of validities**	**Total sample size**	**Weighted mean validity r_{xy}**	**True population validity ρ**	**Observed Variance in r_{xy}**	**Sampling error variance in r_{xy}**	**Residual variance in r_{xy}**	**% variance explained**	**90% credibility value of ρ**
Emotional Adjustment	Ratings	7	1018	.17	.38	.0074	.0067	.0008	90%	.31
	Training	5	761	.21	.39	.0049	.0061	.0000	100%	.38
Integrity/ Control	Ratings	7	1018	.05	.09	.0104	.0070	.0034	67%	-.05
	Training	5	761	.05	.08	.0048	.0067	.0000	100%	.08
Intellectual Efficiency	Ratings	7	1018	.09	.17	.0053	.0069	.0000	100%	.17
	Training	5	761	.29	.39	.0030	.0057	.0000	100%	.39
Interpersonal Relations	Ratings	7	1018	.04	.07	.0085	.0070	.0014	83%	-.01
	Training	5	761	.08	.11	.0042	.0066	.0000	100%	.11

Note: True population validities reflect corrections for unreliability in the criterion measures (but not in predictor measures) and for restriction in the range of test scores in validation samples due to selection. When unavailable, supervisor ratings were assumed to have a reliability of .52 (Viswesvaran, Ones, & Schmidt, 1996), and training criteria were assumed to have a reliability of 0.80.

Table 10.5 Public Safety Job Family Meta-Analytic Results for 16pf Primary Dimensions

16pf Primary Factors	Number of validities	Total sample size	Weighted mean validity r_{xy}	True population validity ρ	Observed variance in r_{xy}	Sampling error variance in r_{xy}	Residual variance in r_{xy}	% variance explained	Lower bound for 80% Credibility interval of ρ	Upper bound for 80% Credibility interval of ρ
Warmth/A	3	236	-.07	-.09	.0021	.0130	.0000	100%	-.09	-.09
Reasoning/B	3	236	.09	.13	.0190	.0129	.0061	68%	-.01	.26
Emotional Stability/C	3	236	.12	.16	.0180	.0127	.0053	71%	.03	.29
Dominance/E	3	236	.00	.00	.0097	.0131	.0000	100%	.00	.00
Liveliness/F	3	236	.13	.17	.0033	.0127	.0000	100%	.17	.17
Rule-Consciousness/G	3	236	-.02	-.03	.0372	.0131	.0241	35%	-.30	.25
Social Boldness/H	3	236	.01	.02	.0045	.0131	.0000	100%	.02	.02
Sensitivity/I	3	236	-.03	-.04	.0133	.0131	.0002	98%	-.07	-.02
Vigilance/L	3	236	-.20	-.27	.0080	.0121	.0000	100%	-.27	-.27
Abstractness/M	3	236	.07	.09	.0084	.0130	.0000	100%	.09	.09
Privateness/N	3	236	.10	.13	.0222	.0129	.0093	58%	-.04	.30
Apprehension/O	3	236	-.13	-.17	.0078	.0127	.0000	100%	-.17	-.17
Openness to Change/Q1	3	236	-.10	-.14	.0168	.0128	.0040	76%	-.25	-.03
Self-Reliance/Q2	3	236	.13	.18	.0023	.0126	.0000	100%	.18	.18
Perfectionism/Q3	3	236	.02	.03	.0137	.0131	.0006	95%	-.02	.07
Tension/Q4	3	236	-.09	-.13	.0149	.0129	.0020	86%	-.21	-.05

Note: Job performance is measured by supervisor rating composite scores. True population validities reflect corrections for unreliability in the criterion measures. When unavailable, supervisor ratings were assumed to have a reliability of .52 (Viswesvaran, Ones, & Schmidt, 1996).

Reviewing the "true population validity" column of Table 10.5 demonstrates relatively strong relationships between higher scores on Reasoning/B (0.13), Emotional Stability/C (0.16), Liveliness/F (0.17), Privateness/N (0.13), and Self-Reliance/Q2 (0.18), as well as lower scores on Vigilance/L (-0.27), Apprehension (-0.17), Openness to Change/Q1 (-0.14), and Tension/Q4 (-0.13). These results are based on only three studies and 236 participants, and should be interpreted with caution, but they depict an empirical picture of successful public safety officers as generally being bright, emotionally stable, with a lively manner, trusting, able to keep confidences, self-confident, traditional, comfortable alone, and relaxed. These traits are generally consistent with the picture that emerged from the PSR dimensions. Two of the most interesting findings were that Rule-Consciousness/G has a modest validity that varied considerably across studies. As previously discussed, this probably arises from mean levels of candidate populations. Although Rule-Consciousness/G certainly must be important for public safety positions, most samples have elevated levels due to self-selection, and even higher levels may have a detrimental effect on performance (e.g., over enforcing on minor issues). It seems imprudent to read these results as Rule-Consciousness/G as being unimportant for public-safety officers. The other surprising finding was the very strong results for better performance with lower Vigilance/L. This finding may also reflect generally elevated levels with a strong deleterious effect of even higher levels

One final note with respect to the use of personality data for predicting success in law enforcement, Cascio, Jacobs, and Silva (2010) present the efficacy of using cognitive ability measures, biodata, and personality scales. In their work they document the positive impact of personality in both increasing validity and simultaneously reducing adverse impact.

Technical and Industrial jobs

Results from unpublished raw data suggest that the 16pf scores can also predict performance in a manufacturing role. As part of a larger project, 38 engineers and production support staff completed the 16pf Questionnaire. Supervisors also rated these individuals on five core competencies: customer satisfaction, communication, team effectiveness, process improvement, and accountability. In addition, an overall rating was also made by the supervisor. The ratings used a scale ranging from 1 (*below average*) to 5 (*above average*). A factor analysis of the five core competency ratings resulted in a single factor accounting for 62% of the variance. In addition, an examination of the plot of eigenvalues indicated a clear one–factor solution. The Overall Performance rating was highly correlated with the resulting factor score (r=.88). Given the high degree of overlap, the overall performance rating was used as the criterion in the study. Means and standard deviations for the 16pf Questionnaire and the overall performance rating are presented in Table 10.6, along with the correlations between the 16 primary factors and the overall performance rating.

As the results in Table 10.6 indicate, Openness to Change/Q1 and Perfectionism/Q3 were significantly related to overall performance. Specifically, individuals who were more comfortable with the status quo, were detail oriented, and who preferred order were rated higher on overall job performance. Due to the small sample size and ad-hoc nature of the study design, these results should be viewed cautiously and verified with future research.

Table 10.6 Means, Standard Deviations, and Correlations with Performance for Engineers and Production Support Staff

Scale	Mean	*SD*	Correlation
Warmth/A	5.25	1.79	0.05
Reasoning/B	6.15	1.97	−0.15
Emotional Stability/C	8.00	1.32	0.13
Dominance/E	5.73	1.89	0.20
Liveliness/F	5.45	1.68	−0.11
Rule-Consciousness/G	6.93	1.49	0.20
Social Boldness/H	6.25	1.82	0.20
Sensitivity/I	3.63	1.41	−0.10
Vigilance/L	4.18	1.63	−0.15
Abstractedness/M	3.93	1.25	−0.16
Privateness/N	4.93	1.69	−0.06
Apprehension/O	3.83	1.66	−0.11
Openness to Change/Q1	5.70	2.10	**−0.32**
Self-Reliance/Q2	3.83	1.45	0.23
Perfectionism/Q3	6.13	1.45	**0.47**
Tension/Q4	3.23	1.46	−0.18
Overall Performance	3.97	0.79	–

Note: *N*=38; Correlations significant at alpha *p*<.05 are listed in bold type.

Other Applications

This subsection describes two organizational applications that outside the realm of selecting employees.

Medical Specialty Choice

The nature of medical training in the US results in a rigorous prescreening process for doctors. In addition, the licensing exams present a final formal hurdle for potential applicants, whereas the required residency programs serve as an apprentice model for admission into the profession. With these stringent requirements in mind for ensuring a minimum level of competency in the medical profession, some researchers have turned to personality, not as a selection tool but as tool support career choices. In this light Borges and Osmon (2001) compared differences in personality traits among 161 physicians who were completing or had finished their medical residency requirements. These authors examined the medical specialties of anesthesiology, surgery, and family practice on the basis of vocational prestige within the medical community as well as differences on Technical versus Person orientation of the specialty. Specifically, surgery is viewed as the most prestigious specialty of the three, followed by family practice and anesthesiology. In addition, family practice is more people oriented whereas the other two specialties are more technique focused.

The authors found some support for these classification schemes. The 16pf global factor Tough-Mindedness discriminated between the surgical specialty and family practice and anesthesiology. The authors noted that the characteristics associated with the Tough-Mindedness trait (less empathetic, reserved, and preferring tradition and established procedures) fit well with another author's description of surgical practice (Coombs, 1978). Doctors choosing family practice were characterized by higher scores on Rule-Consciousness/G and Abstractedness/M, whereas anesthesiologists were characterized by higher scores on Vigilance/L. Although the authors offer some exploratory interpretations of these differences, connections between the results and with the theoretical framework presented were more tenuous than that for the surgical specialty.

In a follow up study, Hartung, Borges, and Jones (2005) explored medical career specialty choice. Their methodology involved matching the 358 participants to 62 people within a reference sample of career specialties. Matching was performed using squared "distance" to find the closest reference profiles. Matching by personality to specific medical specialties provide to be a challenging task with a success rate of 43% to 60%, but considering that there were more than 20 specialties, this represents a practically significant level of prediction as well as a novel method.

Stress and Burnout

The concept of stress and its impact on performance and turnover in organizations is, relatively speaking, a new area of research. There is a growing awareness that stress can lead to burnout, a phenomenon that can eventually lead talented employees to leave organizations or even occupations as a result of too much stress over time.

Rausch and Braverman (2000) note that there is a substantial amount of stress for nurses in reproductive medicine. The nursing role requires both medical care of patients as well as the interpersonal interactions and psychological support for patients who are undergoing an extremely stressful experience. These authors examined the relationship between the 16pf Questionnaire and burnout, assessed using the Maslach Burnout Inventory (MBI), in a sample of nurses (N=110) working in reproductive medicine.

Results indicated that the sample of nurses participating in this research scored high on the MBI and that the number of years practicing in the reproductive and fertility field was substantially correlated with feelings of burnout. In addition, several primary factors from the 16pf Questionnaire were related to feelings of burnout. Specifically, lower scores on Emotional Stability/C, Social Boldness/H, Abstractedness/M, and Privateness/N were associated with increased feelings of burnout. In addition, higher scores on Apprehension/O were also related to burnout. Although the study design was correlational, and causation cannot be inferred, it may be that individuals who are more emotionally stable and thick skinned are buffered from stressful events. In addition, one contributing factor to stress is a perception of not being in control. Lower scores on the Apprehension/O are related to increased self-assurance. These individuals may feel more in control of events and thus not perceive situations to be as stressful as individuals who feel less in control.

Summary

The results of the studies reviewed in this chapter indicate that the 16pf Questionnaire is predictive of a variety of important behaviors at work. Given the high degree of equivalency shown for the Fifth and Sixth Editions (see Chapter 8), these results should generalize to the Sixth Edition.

Although it was noted in the introduction to this chapter that validity generalization techniques were influential in sparking a renewed interest in the influence of personality on behaviors at work, these methods do have their limitations. The expression of personality, and therefore its influence on behavior, is clearly moderated by situational constraints or the lack thereof (Barrick & Mount, 1993). As a result, meta-analytic estimates of the validity of personality in predicting interesting outcomes will be limited, to the extent that they do not take into account aspects of the situation in which personality assessments are being applied. In other words, there is likely to be more situational specificity with regard to personality variables as compared to cognitive ability. For this reason it is always advisable to conduct local validation studies whenever possible. Moreover, as best-practice guidelines and empirical evidence bear out, validation studies are more effective when based on job analysis results and the conceptual links between personality and criterion of interest have been clearly specified in advance (Hogan & Holland, 2003; Joint Committee on the Standards for Educational and Psychological Testing, 2014; Rothstein & Goffin, 2006; Tett et al., 1991).

References

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., (1993). Autonomy as a moderator of relationships between the big five personality dimensions and job performance. *Journal of Applied Psychology, 78*, 111–118.

Bass, B. M., & Avolio, B. J. (2000). *Transformational Leadership: Manual for the Multifactor Leadership Questionnaire*. Mountain View, CA: CPP, Inc.

Bedwell, S. (2001). *Relationship between personality and sales performance in the printing industry*. IPAT Technical Report, Champaign, IL: IPAT.

Bedwell, S. (2003). *Emotional Judgment Inventory manual.*: Champaign, IL: IPAT.

Borges, N.J., & Osmon, W.R. (2001). Personality and medical specialty choice: Technique orientation versus people orientation. *Journal of Vocational Behavior, 58*, 22–35.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), *Personnel selection in organizations* (pp. 71-98). San Francisco, CA: Jossey Bass.

Bracken, D. W., Timmreck, C. W., & Church, A. H. (2001). *The handbook of multisource feedback*. San Francisco, CA: Jossey-Bass.

Brannick, M. T., Levine, E. L., & Morgeson, F. P. (2007). *Job and work analysis: Methods, research, and applications for human resource management (2nd ed.).* Thousand Oaks, CA: Sage Publications, Inc.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.) *Handbook of industrial and organizational psychology (*2nd ed., Vol. 1, pp. 687–732). Mountain View, CA: CPP, Inc.

Cascio, W.F., Jacobs, R.R., & Silva, J. (2010) Validity, utility, and adverse impact: Practical implications from 30 years of data. In J. Outtz (Ed.), *Adverse Impact* (pp. 271-288). New York, NY: Routledge.

Cattell, R. B., Cattell, A. K., & Cattell, H. E. P. (1993) *Sixteen Personality Factor Questionnaire, Fifth Edition*. Champaign, IL: Institute for Personality and Ability Testing, Inc.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the 16pf*. Champaign, IL: Institute for Personality and Ability Testing, Inc.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Oxford, England: U. Illinois Press.

Coombs, R. H. (1978). *Mastering medicine: Professional socialization in medical school*. New York, NY: Free Press.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40–57.

Dulewicz, V. (1995). A validation of Belbin's team roles from 16PF & OPQ using bosses' ratings of competence. *Journal of Occupational & Organizational Psychology, 68*(2), 81-99.

Fishbein, J. A., Oster, C., & Bedwell, S. (2007) *A sales prediction study.* Technical Report, Fishbein Performance Consulting, P.C.

Fraboni, M., & Saltstone, R. (1990). First and second generation entrepreneur typologies: Dimension of personality. *Journal of Social Behavior and Personality, 5*, 105-113.

Goldberg, L. R., & Saucier, G. (2016). The Eugene-Springfield community sample: Information available from the research participants. ORI Technical Report, Vol. 56 No. 1. Retrieved from: https://ipip.ori.org/ESCS_TechnicalReport_January2016.pdf

Grucza, R. A., & Goldberg, L. R. (2007). The comparative validity of 11 modern personality inventories: Predictions of behavioral acts, informant reports, and clinical indicators. *Journal of Personality Assessment, 89*(2), 167-187.

Hartung, P. J., Borges, N. J., & Jones, B. J. (2005). Using person matching to predict career specialty choice. *Journal of Vocational Behavior, 67*, 102–117.

Hetland, H., & Sandal, G. M. (2003). Transformational leadership in Norway: Outcomes and personality correlates. *European Journal of Work and Organizational Psychology, 12*, 147–170.

Hofstede, G. (1980). Culture and organizations. *International Studies of Management and Organization, 4*, 15–41.

Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job performance relations: A socio-analytic perspective. *Journal of Applied Psychology, 88*, 100–112.

Hough, L. M. (2001). I/Owes its advances to personality. In B. W. Roberts & R. Hogan, (Eds.) *Personality Psychology in the Workplace*. Washington, DC: APA.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial-organizational psychology: Reflections, progress and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*(3), 272–290.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage Publications.

IPAT (1999). *16pf® Select Manual*. Champaign, IL: IPAT.

IPAT (2003). *Protective Services Reports manual*. Champaign, IL: IPAT.

Joint Committee on the Standards for Educational and Psychological Testing. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, American Psychological Association, & National Council for Measurement in Education.

Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology, 98*, 875-925.

Kostman, I. (2003). *Multi-dimensional performance requires multi-dimensional predictors: Predicting complex job performance using cognitive ability, personality and emotional intelligence assessment instruments as combinatorial predictors.* Doctoral dissertation, The City University of New York. ProQuest Dissertations Publishing.

Love, K. G., & DeArmond, S. (2007). The validity of assessment center ratings and 16pf personality trait scores in police sergeant promotions: A case of incremental validity. *Public Personnel Management, 36,* 21–32.

Mershon, B., & Gorsuch, R. L. (1988). Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology, 55*, 675–680.

Neisser, U., Boodoo, G., Bouchard, T. J. Jr., Boykin, A. W., Brody, N., Ceci, S. J., …Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101.

Ones, D. S., Viswesveran, C., & Dilchert, S. (2005). Personality at work: Raising awareness and correcting misconceptions. *Human Performance, 18*, 389–404.

Rausch, D. T., & Braverman, A. M. (2000) Burnout rates among reproductive endocrinology nurses: The role of personality and infertility attitudes. *Fertility and Sterility, 74,* S7.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does the current research support? *Human Resource Management Review, 16*, 155–180.

Seibert, S. E., & DeGeest, D. S. (2017). The five factor model of personality in business and industry. In T. A. Wigdor (Ed.), *The Oxford handbook of the five factor model* (pp. 381-401). New York, NY: Oxford University Press.

Society for Industrial and Organizational Psychology, Inc. (2003). Principles for the validation and use of personnel selection procedures (4th ed.). Retrieved from http://www.siop.org/_principles/principles.pdf

Tett, R. P., Jackson, D. N., & Rothstein, M. G. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703–742.

Thayer, P. (2008). That's not the only problem. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*, 372.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557-574.

Wonderlic (1992). *Wonderlic Personnel Test and Scholastic Exam: User's manual*. Libertyville, IL: Wonderlic.

Woehr, D. J., Sheehan, M. K., & Bennett, W. Jr. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology, 90*, 592–600.

Appendices

Appendix A Historical Acknowledgments of Individuals Who Made Key Contributions to the 16pf Questionnaire

Given the historical legacy of the 16pf questionnaire, many individuals have been part of its development, maintenance and enhancement. This manual summarizes the most recent enhancement to the questionnaire itself. While there is a new future for the questionnaire, we do not want to forget those in the past. A large team of individuals have been part of making the 16pf one of the most accurate and popular normal personality questionnaires on the market.

The individuals below should be recognized for members of a historical 16pf project team and/or authors of previous manuals:

Abigail Griffin
Afandi Mohamed
Catherine C. Maraist
Darcie Karol
David G. Watterson, Jr.
Deborah Matthews
Deirdré Gyenes
Heather Cattell
Herb Eber
James M. Schuerger
Julia Aufenast
Mark Rieke
Mary L. (Kelly) Doherty
Mary Russell
Philippa Riley
Robert Bailey
Sarah Hudson
Scott Bedwell
Stephen Guastello
Steve Conn
William Lindemann
Graham Kilian
Callum Welch

Appendix B IRT GRM Item Parameter Estimates for the Items of the Primary Factor Scales

Scale										
A	**Item**	**Slope**	**Slope (se)**	**Location**	**Location (se)**	**t1**	**t2**	**t3**	**t4**	**Chi-square/ df**
	L1	0.54	0.01	-0.36	0.05	2.30	0.76	-0.60	-2.47	3.10
	L127	1.14	0.02	-1.00	0.03	1.73	0.70	-0.38	-2.06	4.42
	L159	0.71	0.01	-0.70	0.04	2.23	0.87	-0.58	-2.52	4.48
	L259	1.01	0.02	-1.32	0.03	1.83	0.87	-0.37	-2.34	2.52
	L311	0.99	0.02	-0.67	0.03	1.94	0.73	-0.39	-2.28	4.04
	L430	0.72	0.01	-0.35	0.04	2.12	0.83	-0.46	-2.49	5.30
	L545	1.23	0.02	-0.88	0.03	1.69	0.70	-0.43	-1.96	3.32
	L65b	0.84	0.02	-0.60	0.03	1.83	0.70	-0.51	-2.02	2.57
	L826	0.97	0.02	-1.04	0.03	1.67	0.88	-0.38	-2.17	5.03
	LE166	0.61	0.01	-0.96	0.04	2.50	0.85	-0.59	-2.77	3.80

C	**Item**	**Slope**	**Slope (se)**	**Location**	**Location (se)**	**t1**	**t2**	**t3**	**t4**	**Chi-square/ df**
	L32	1.07	0.02	-0.27	0.03	1.68	0.45	-0.36	-1.77	4.18
	L131	0.91	0.02	-0.46	0.03	1.56	0.51	-0.33	-1.74	3.54
	L446	0.84	0.01	-0.76	0.03	2.05	0.89	-0.55	-2.39	6.27
	L470	0.83	0.02	-0.70	0.04	1.86	0.80	-0.39	-2.27	2.66
	L503	1.03	0.02	-0.92	0.03	1.72	0.67	-0.25	-2.14	3.94
	L578	1.16	0.02	-0.52	0.03	1.73	0.53	-0.33	-1.92	4.38
	L685	1.00	0.02	0.20	0.03	1.98	0.39	-0.52	-1.85	4.20
	L765	0.99	0.02	-0.40	0.03	1.90	0.49	-0.39	-2.00	3.84
	L788	0.93	0.02	-0.67	0.03	1.56	0.52	-0.26	-1.82	3.51
	L807	0.76	0.01	-0.26	0.04	2.02	0.64	-0.51	-2.16	3.48

E	**Item**	**Slope**	**Slope (se)**	**Location**	**Location (se)**	**t1**	**t2**	**t3**	**t4**	**Chi-square/ df**
	L66	0.83	0.01	-0.52	0.04	2.39	0.82	-0.61	-2.59	2.65
	L245	0.72	0.01	0.06	0.04	2.39	0.86	-0.79	-2.46	2.30
	L275	0.90	0.02	-0.29	0.03	2.27	0.77	-0.61	-2.43	3.88
	L426	0.67	0.01	-0.88	0.04	2.44	0.89	-0.61	-2.72	4.35
	L438	0.95	0.02	-0.28	0.03	1.95	0.66	-0.52	-2.09	6.66
	L511	0.88	0.02	-0.42	0.03	1.84	0.66	-0.46	-2.05	3.90
	L514	0.94	0.02	-0.53	0.03	1.96	0.67	-0.37	-2.26	3.30
	L519	0.66	0.01	-0.55	0.04	2.57	0.78	-0.73	-2.62	2.78
	L554	0.89	0.01	-0.54	0.03	2.07	0.69	-0.48	-2.29	3.56
	L771	0.66	0.01	-0.28	0.04	2.41	0.52	-0.52	-2.42	3.39

F	**Item**	**Slope**	**Slope (se)**	**Location**	**Location (se)**	**t1**	**t2**	**t3**	**t4**	**Chi-square/ df**
	L6	1.09	0.02	-0.24	0.03	1.60	0.53	-0.35	-1.78	3.02

	L68	1.14	0.02	0.10	0.03	1.74	0.52	-0.49	-1.77	3.09
	L164b	1.03	0.02	0.08	0.03	1.59	0.54	-0.40	-1.73	3.68
	L253	0.71	0.01	-0.13	0.04	2.27	0.75	-0.61	-2.41	2.99
	L343	0.81	0.01	-0.02	0.03	2.08	0.70	-0.57	-2.21	3.82
	L691	1.50	0.03	-0.25	0.03	1.32	0.51	-0.32	-1.51	4.86
	L708	0.42	0.01	0.30	0.06	3.67	0.93	-0.98	-3.62	2.91
	L762	0.84	0.01	-0.14	0.03	2.17	0.59	-0.62	-2.15	2.91
	L796	0.54	0.01	0.20	0.05	2.56	0.72	-0.60	-2.67	3.74
	L878	0.59	0.01	1.02	0.04	2.80	0.66	-0.96	-2.50	2.91
	L882	0.56	0.01	0.98	0.05	2.39	0.37	-0.58	-2.18	1.83

G	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L133	0.70	0.01	-0.88	0.04	2.01	0.77	-0.45	-2.33	5.09
	L136	0.70	0.01	-0.29	0.04	2.48	0.70	-0.81	-2.37	4.66
	L166	0.77	0.01	-0.33	0.04	2.18	0.64	-0.64	-2.18	4.62
	L214	1.15	0.02	-1.55	0.03	1.46	0.86	-0.26	-2.05	4.98
	L272	0.89	0.02	-0.38	0.03	1.95	0.63	-0.50	-2.07	4.20
	L613	1.05	0.02	-1.53	0.03	1.71	0.98	-0.36	-2.32	4.64
	L694	0.75	0.01	-1.30	0.04	2.24	0.91	-0.52	-2.62	2.95
	L735	0.82	0.01	0.07	0.04	2.41	0.62	-0.73	-2.30	5.51
	L753	0.74	0.01	-0.01	0.04	2.27	0.61	-0.61	-2.28	3.62
	L794	0.68	0.01	-1.10	0.04	1.99	0.70	-0.30	-2.39	6.21
	L842	0.67	0.01	-0.85	0.04	2.62	1.26	-0.74	-3.15	5.57

H	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L9	0.89	0.01	-0.21	0.03	2.07	0.77	-0.55	-2.29	3.87
	L71	0.92	0.02	0.27	0.03	1.73	0.37	-0.39	-1.71	2.75
	L73	1.16	0.02	-0.11	0.03	1.48	0.42	-0.30	-1.59	6.13
	L135	1.36	0.02	0.16	0.03	1.35	0.36	-0.36	-1.34	3.55
	L137	1.08	0.02	0.08	0.03	1.75	0.55	-0.52	-1.79	3.97
	L169	1.78	0.03	0.09	0.02	1.29	0.29	-0.31	-1.26	5.17
	L543	1.12	0.02	-0.06	0.03	1.61	0.44	-0.42	-1.64	4.28
	L574	1.42	0.02	-0.13	0.03	1.35	0.33	-0.27	-1.41	4.12

I	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L10	0.78	0.02	-0.44	0.04	1.76	0.65	-0.42	-1.99	2.21
	L44	0.73	0.02	-0.60	0.04	1.76	0.58	-0.53	-1.81	3.20
	LE47	0.68	0.01	-0.94	0.04	1.98	0.72	-0.35	-2.35	1.92
	L74	0.61	0.01	-0.04	0.04	2.04	0.59	-0.44	-2.18	2.78
	L77	0.75	0.02	-0.74	0.04	1.93	0.70	-0.55	-2.08	3.72
	L140	0.48	0.01	-0.08	0.05	1.67	0.33	-0.38	-1.62	2.01
	L170	0.48	0.01	0.28	0.05	2.99	0.91	-1.03	-2.87	2.25
	L408	0.79	0.02	-0.94	0.04	2.04	0.75	-0.53	-2.25	2.26

	L532	0.60	0.01	-0.13	0.04	2.05	0.59	-0.55	-2.10	2.32
	L861	0.42	0.01	0.01	0.06	2.92	0.78	-0.81	-2.89	2.32
	L873	0.67	0.01	-1.23	0.04	2.18	1.08	-0.60	-2.66	1.66
	L874	0.64	0.01	0.79	0.04	2.68	0.63	-0.87	-2.44	4.71

L	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L323	1.39	0.02	-0.18	0.03	1.77	0.52	-0.52	-1.78	5.69
	L410	0.93	0.02	0.68	0.03	2.51	0.42	-0.81	-2.12	6.03
	L412	0.89	0.02	-0.82	0.03	2.23	0.76	-0.55	-2.44	3.46
	L427	0.74	0.01	0.25	0.04	2.29	0.35	-0.63	-2.01	4.40
	L563	1.08	0.02	-0.02	0.03	1.98	0.52	-0.58	-1.91	4.96
	L651	1.05	0.02	-0.21	0.03	2.01	0.61	-0.55	-2.07	6.54
	L838	0.51	0.01	0.49	0.05	3.61	0.86	-1.19	-3.29	7.64

M	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L12	0.84	0.02	1.03	0.04	2.07	0.41	-0.58	-1.90	3.91
	L49	0.68	0.01	1.35	0.04	3.10	0.52	-1.01	-2.61	4.15
	L81	0.71	0.01	0.49	0.04	2.82	0.61	-0.87	-2.55	3.57
	L142	0.66	0.01	0.17	0.04	2.81	0.58	-0.73	-2.66	5.10
	L217	0.62	0.01	-0.60	0.04	2.57	0.69	-0.60	-2.66	3.92
	L321	0.48	0.01	0.82	0.05	3.90	1.22	-1.39	-3.72	2.73
	L657	0.75	0.01	-0.03	0.04	2.06	0.44	-0.41	-2.09	3.87
	L662	0.97	0.02	0.45	0.03	2.17	0.43	-0.64	-1.96	3.74
	L760	0.89	0.02	1.36	0.04	2.64	0.47	-0.98	-2.13	2.64
	L847	0.70	0.01	0.71	0.04	2.35	0.42	-0.70	-2.07	4.08

N	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L15	0.95	0.02	-0.13	0.03	1.87	0.41	-0.47	-1.81	5.01
	L47	0.59	0.01	0.01	0.05	2.44	0.61	-0.58	-2.48	2.77
	L50	0.99	0.02	-0.76	0.03	1.74	0.56	-0.32	-1.98	5.99
	L117	0.84	0.01	0.14	0.03	2.18	0.36	-0.61	-1.93	4.45
	L260	1.10	0.02	-1.04	0.03	1.85	0.63	-0.48	-2.00	5.39
	L391	1.05	0.02	-0.40	0.03	1.66	0.42	-0.33	-1.76	3.32
	L631	0.85	0.01	0.01	0.03	2.10	0.51	-0.65	-1.96	6.25
	L660	0.77	0.01	0.12	0.04	2.22	0.52	-0.59	-2.15	3.45
	L792	0.72	0.01	-0.66	0.04	2.27	0.61	-0.56	-2.32	4.04

O	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L19	1.25	0.02	-0.10	0.03	1.41	0.36	-0.27	-1.49	5.06
	L116	1.02	0.02	0.10	0.03	1.59	0.41	-0.29	-1.71	4.20
	L146	0.92	0.02	-0.19	0.03	1.80	0.40	-0.47	-1.74	5.78
	L305	0.91	0.01	0.11	0.03	1.99	0.37	-0.45	-1.91	3.64

	L407	0.91	0.01	-0.20	0.03	1.99	0.49	-0.42	-2.05	4.24
	L520	1.05	0.02	0.20	0.03	1.69	0.42	-0.35	-1.76	3.18
	L718	0.90	0.01	0.09	0.03	2.02	0.44	-0.47	-1.99	6.21
	L773	0.81	0.01	0.29	0.04	1.96	0.40	-0.45	-1.92	4.71

Q1	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L20b	0.51	0.01	-0.51	0.05	2.77	0.95	-0.82	-2.90	3.03
	L83	0.52	0.01	-1.06	0.05	3.42	1.59	-1.06	-3.96	3.23
	L118	0.77	0.01	-1.30	0.04	2.16	0.93	-0.48	-2.61	2.06
	LX166	1.27	0.02	-0.71	0.03	1.70	0.77	-0.51	-1.96	5.86
	L333	0.91	0.02	-0.86	0.03	1.84	0.78	-0.46	-2.16	3.44
	L395	1.36	0.03	-0.87	0.03	1.59	0.76	-0.38	-1.97	5.98
	L501	0.51	0.01	0.76	0.05	3.22	0.79	-0.88	-3.13	4.82
	L605	0.37	0.01	1.51	0.07	4.45	0.97	-1.47	-3.95	4.78
	L673	0.41	0.01	-0.01	0.06	4.15	1.53	-1.32	-4.35	4.31
	L849	0.92	0.02	-1.00	0.03	1.55	0.66	-0.35	-1.86	3.37
	L901	1.06	0.02	-0.91	0.03	1.90	0.72	-0.51	-2.12	3.97

Q2	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L25	0.76	0.01	0.12	0.04	2.51	0.75	-0.90	-2.37	3.83
	L59	0.71	0.01	-0.90	0.04	2.36	0.70	-0.63	-2.43	3.38
	L89	1.59	0.03	-0.68	0.03	1.49	0.51	-0.42	-1.58	7.13
	L92	0.72	0.01	0.53	0.04	2.84	0.69	-1.11	-2.41	5.33
	L121	0.91	0.02	-0.38	0.03	1.86	0.56	-0.53	-1.89	6.24
	L152	0.64	0.01	-1.09	0.04	1.94	0.58	-0.36	-2.16	3.47
	L156	0.88	0.01	-0.08	0.03	2.05	0.58	-0.60	-2.03	11.64
	L476	1.62	0.03	-0.43	0.03	1.44	0.47	-0.43	-1.48	7.77

Q3	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L29	0.55	0.01	0.30	0.05	3.05	0.58	-0.85	-2.79	5.82
	L90	1.16	0.02	-0.31	0.03	1.51	0.23	-0.30	-1.43	3.64
	L122	0.56	0.01	0.03	0.05	2.39	0.42	-0.54	-2.27	3.47
	L449	0.86	0.02	-0.24	0.03	2.12	0.53	-0.50	-2.16	2.75
	L475	0.93	0.02	-1.00	0.03	1.90	0.63	-0.41	-2.11	3.17
	L534	0.49	0.01	-1.87	0.05	2.86	1.26	-0.44	-3.69	2.51
	L625	0.31	0.01	0.75	0.08	4.76	0.81	-1.33	-4.24	4.17
	L683	1.41	0.03	-0.79	0.03	1.42	0.50	-0.35	-1.57	5.60
	L790	0.90	0.02	0.18	0.03	2.08	0.28	-0.56	-1.80	7.68

Q4	Item	Slope	Slope (se)	Location	Location (se)	t1	t2	t3	t4	Chi-square/ df
	L60	0.55	0.01	-0.57	0.04	3.02	0.78	-0.63	-3.17	3.00
	L62	1.14	0.02	-0.16	0.03	1.73	0.40	-0.31	-1.82	5.24

L91	0.92	0.02	0.55	0.03	2.25	0.48	-0.60	-2.13	5.31
L155	0.82	0.02	-0.69	0.04	1.85	0.60	-0.29	-2.16	3.51
L158	0.77	0.01	0.94	0.04	2.47	0.41	-0.82	-2.06	4.17
L320	0.82	0.01	0.48	0.03	2.00	0.42	-0.51	-1.92	2.88
L569	0.65	0.01	-0.27	0.04	2.36	0.67	-0.55	-2.49	3.79
L608	0.89	0.02	0.78	0.03	2.15	0.44	-0.66	-1.92	2.30
L836	1.03	0.02	-0.41	0.03	1.86	0.51	-0.39	-1.98	3.93

Note: Standardization sample, *N*=2,528. A=Warmth, B=Reasoning, C=Emotional Stability, E=Dominance, F=Liveliness, G=Rule-Consciousness, H=Social Boldness, I=Sensitivity, L=Vigilance, M=Abstractedness, N=Privateness, O=Apprehension, Q1=Openness to Change, Q2=Self-Reliance, Q3=Perfectionism, Q4=Tension.

Appendix C.1 Descriptives for the IPIP Construct Validity Sample

Scale	Mean	*SD*
16pf Primary Factor		
Warmth/A	5.7	2.1
Reasoning/B	5.5	1.6
Emotional Stability/C	5.9	2.1
Dominance/E	5.5	2.3
Liveliness/F	5.2	2.1
Rule-Orientation/G	5.5	2.0
Social Boldness/H	5.0	2.2
Sensitivity/I	5.9	1.8
Vigilance/L	5.6	2.5
Abstractedness/M	5.2	2.0
Privateness/N	5.7	2.2
Apprehension/O	5.4	2.3
Openness to Change/Q1	5.8	2.0
Self-Reliance/Q2	5.6	2.1
Perfectionism/Q3	5.8	2.2
Tension/Q4	5.2	2.2
16pf Global Factor		
Extraversion	5.2	2.1
Anxiety	5.3	2.2
Tough-Mindedness	5.1	2.0
Independence	5.5	2.2
Self-Control	5.8	2.1
IPIP Big Five Factor		
Extraversion	28.1	9.2
Agreeableness	39.7	6.2
Conscientiousness	39.0	6.2
Neuroticism	25.3	9.0
Openness	37.9	6.4

Note: N=214 except for Reasoning/B where N=209. 16pf Primary and Global scales are stens.

Appendix C.2 Descriptives for the HPI Construct Validity Sample

Scale	Mean	*SD*
16pf Primary Factor		
Warmth/A	5.7	2.1
Reasoning/B	5.5	1.7
Emotional Stability/C	5.9	2.1
Dominance/E	5.6	2.3
Liveliness/F	5.3	2.2
Rule-Orientation/G	5.5	2.1
Social Boldness/H	5.1	2.2
Sensitivity/I	5.9	1.8
Vigilance/L	5.5	2.5
Abstractedness/M	5.2	2.0
Privateness/N	5.6	2.3
Apprehension/O	5.4	2.3
Openness to Change/Q1	5.8	2.0
Self-Reliance/Q2	5.6	2.2
Perfectionism/Q3	5.9	2.2
Tension/Q4	5.2	2.2
16pf Global Factor		
Extraversion	5.3	2.1
Anxiety	5.3	2.2
Tough-Mindedness	5.1	2.0
Independence	5.6	2.1
Self-Control	5.7	2.1
HPI HIC		
Accomplishment	1.6	1.0
Avoids Trouble	2.3	1.1
Calmness	2.4	1.3
Caring	3.3	1.2
Competitive	2.3	1.3
Culture	3.0	1.3
Curiosity	2.9	1.3
Easy to Live With	2.9	1.3
Education	3.0	1.4
Empathy	2.2	1.4
Entertaining	2.5	1.6
Even Tempered	2.1	1.3
Exhibitionistic	2.1	1.5
Experience Seeking	2.2	1.3
Generates Ideas	2.4	1.3
Good Attachment	1.9	1.4
Good Memory	2.9	1.1
Identity	2.5	1.7
Impulse Control	3.0	1.3

Scale	Mean	*SD*
Intellectual Games	3.4	0.9
Leadership	1.9	1.6
Likes Crowds	1.6	1.5
Likes Parties	1.5	1.5
Likes People	1.6	1.3
Mastery	3.3	0.9
Math Ability	2.1	1.7
Moralistic	2.1	1.3
No Complaints	2.0	1.3
No Guilt	1.8	1.3
No Hostility	2.6	1.3
No Social Anxiety	1.6	1.4
Not Anxious	2.5	1.3
Not Autonomous	2.0	1.6
Not Spontaneous	2.8	1.2
Reading	3.2	1.2
Science Ability	3.3	1.0
Self-Confidence	2.6	1.5
Sensitive	3.1	1.2
Thrill Seeking	2.0	1.5
Trusting	2.0	1.5
Valid	3.8	0.9
Virtuous	2.2	1.1

Note: N=233 except for Reasoning/B where N=223. 16pf Primary and Global scales are stens, HIP HICs are raw scores.

Appendix C.3 Descriptives for the GPS Construct Validity Sample

Scale	Mean	SD
16pf Primary Factor		
Warmth/A	5.6	2.2
Reasoning/B	5.6	1.8
Emotional Stability/C	6.2	2.1
Dominance/E	5.7	2.2
Liveliness/F	5.2	2.1
Rule-Orientation/G	5.6	2.1
Social Boldness/H	5.1	2.2
Sensitivity/I	5.7	1.8
Vigilance/L	5.6	2.6
Abstractedness/M	5.0	2.0
Privateness/N	5.7	2.3
Apprehension/O	5.2	2.4
Openness to Change/Q1	5.9	2.0
Self-Reliance/Q2	5.7	2.2
Perfectionism/Q3	5.9	2.2
Tension/Q4	5.0	2.2
16pf Global Factor		
Extraversion	5.2	2.0
Anxiety	5.1	2.2
Tough-Mindedness	5.3	1.9
Independence	5.6	2.1
Self-Control	6.0	2.2
GPS Component Scale		
Caring	4.2	0.8
Helpful	4.0	0.7
Complying	4.2	0.6
Considerate	4.3	0.6
Trusting	3.3	1.0
Achievement Focus	4.3	0.6
Initiative	3.8	0.7
Organization	4.3	0.6
Thoroughness	4.4	0.6
Diligence	4.6	0.6
Lively	3.7	0.6
Influential	3.5	0.9
Likes Attention	2.4	1.0
Sociable	3.4	1.1

Composure	3.9	0.8
Even Tempered	3.7	0.8
Optimism	3.8	0.9
Self-Confidence	3.7	0.9
Curiosity	4.2	0.6
Flexibility	3.7	0.7
Inventiveness	4.1	0.7
Quick Thinking	4.0	0.7

Note: N=150 except for Reasoning/B where N=144. 16pf Primary and Global scales are stens, GPS Component scales are raw scores.

PSI Services, LLC (PSI)
www.16pf.com

611 N. Brand Blvd, 10th Floor
Glendale, CA 91203

t: 800 367 1565
t: +44 (0) 1763 257100
hello@psionline.com
www.psionline.com