UNIT 2

EVALUATION OF NORMALITY AND INDIVIDUAL STRENGTHS

Assessment does not always entail delving into symptoms, distress level, defense mechanisms, diagnosis, and the like. One example might be a young executive who wants to know about "growth edges" in regard to leadership positions. Another example might be a college student who desires self-knowledge as part of vocational explorations. Even though clinical tests can be employed within the normal spectrum, they do not excel in this application. In fact, the evaluation of normal personality was not the original purpose of tests such as the MMPI or the Rorschach. When a practitioner wants to assess personality within the normal spectrum, tests designed expressly for that purpose typically provide a more helpful perspective than instruments developed from the standpoint of psychopathology. Instead of measuring concepts such as depression, paranoia, anxiety, narcissism, or suicide potential, the focus in these alternative instruments is on qualities pertinent to the normal range of human functioning.

The features like responsibility, social presence, intuition, locus of control, attachment style, or faith maturity are referred here. Normality differs from abnormality by shades of gray rather than revealing a sharp demarcation. In addition to their value in the assessment of client personality, tests also contribute to our understanding of both typical and atypical trajectories of personality across the life span. Other forms of assessment pertinent to the normal spectrum of adult functioning are the evaluation of spiritual, religious, and moral constructs.

BROAD BAND TESTS OF NORMAL PERSONALITY

A broad band test is one that measures the full range of functioning, as opposed to limited aspects. Beginning in the 1940s, researchers sought to capture the nuances of normal personality by developing broad-band self-report instruments. The sheer variety of approaches to this task is a testament to the complexity of human functioning.

MYERS-BRIGGS TYPE INDICATOR (MBTI)

Originally published in 1962, the MBTI is a forced-choice, self-report inventory that attempts to classify persons according to an adaptation of Carl Jung's theory of personality types (Myers & McCaulley, 1985; Tzeng, Ware, & Chen, 1989). Recent adaptations of the test also provide dimensional scores in addition to the well-known four-letter typological codes.

According to the publisher, the MBTI is the most widely used individual test in history, taken by approximately 2 million people a year. Proponents of the instrument deem it valuable in vocational guidance and organizational consulting. It comes in a number of versions, including Form M, a 93-item test which can be purchased by qualified psychologists in a self-scoring paper-and-pencil format, or administered on-line. Other forms such as the 126-item Form G and the 144-item Form Q are available on-line and must be authorized by a psychologist who has agreed to a licensing arrangement with the publisher, Consulting Psychologists Press (www.cpp.com).

Regardless of the version employed, the MBTI is scored on four theoretically independent polarities: Extraversion–Introversion, Sensing–iNtuition, Thinking–Feeling, and Judging–Perceiving. The testtaker is categorized on one side or the other of each polarity, which results in a four-letter code such as ENTJ (Extraversion, iNtuition, Thinking, Judging). Because there are two poles to each of the four dimensions, this allows for 24 or 16 different personality types. Each of the 16 types has been studied extensively over the years. The opposite ends of each polarity are simply different modes of being that may have a variety of

implications for relationships, vocation, leadership, and personal functioning. Possessing the qualities of one polarity or the other may be advantageous (or not) in different situations.

Extraversion–Introversion is probably the easiest to describe. An extravert (E) directs energy outward to people and conversations, whereas an introvert (I) directs energy inward to his or her inner world. A note of clarification: The MBTI retains the original spelling of Extraversion, preferred by Jung, instead of using the synonymous concept of Extroversion, preferred by contemporary psychologists. Sensing–iNtuition involves two opposite ways of perceiving. Those who prefer sensing (S) rely on the immediate senses, whereas those who prefer intuition (N) rely upon "relationships and/or possibilities that have been worked out beyond the reach of the conscious mind" (Myers & McCaulley, 1985). Of course, the letter N is used to designate intuition because the letter I already is taken to label Introversion. Thinking–Feeling refers to basing conclusions on thinking (T), that is, logic and objectivity, as opposed to feeling (F), which involves a reliance on personal values and social harmony. Finally, Judging–Perceiving indicates a preference for decisiveness and closure (J) or an open-ended flexibility and spontaneity (P). Whereas in common parlance the notion of "judging" often has a negative connotation, this is not the case when the term is applied to this polarity of the MBTI.

The 16 possible four-letter types are not equally represented in the general population, and some types are more common in specific occupational groups. Standardization data for the MBTI is extensive and based on large samples collected over many decades. Split-half reliabilities for the four scales are in the .80s for the combined subject pool of nearly 56,000 participants. Test–retest reliabilities for the four scales are somewhat lower and depend on the interval between tests. When the interval is short, on the order of a few weeks, results are strong, with coefficients mainly in the .70s and higher. Yet, when the interval is longer, on the order of several years, the coefficients are predictably lower, in the .40s and .50s. With regard to reliability, an important question with the MBTI is the stability of the four letter code from test to retest. The test manual reports on a dozen studies of code type stability, with retest intervals ranging from 5 weeks to 5 years (most intervals a year or two). On average, about 41 percent of examinees retained their identical code type, that is, all four letters of the code remained the same from test to retest.

In a review of 17 studies reporting reliability coefficients, Capraro and Capraro (2002) found respectably strong reliability coefficients of .84 (E-I), .84 (S-N), .67 (T-F), and .82 (J-P). More than 400 references citing the MBTI were found in PsychINFO from 2000 to 2009, many pertaining to the validity of the instrument.

Another recent study also provides support for the validity of the polarities assessed by the MBTI. Furnham, Moutafi, and Crump (2003) tested 900 adults with two instruments: the MBTI and the Revised NEO-Personality Inventory (NEO-PI-R, Costa & McCrae, 1992). The NEO-PI-R is a well validated measure of personality that evaluates five factors of personality known as the "big five." These factors are Neuroticism, Extraversion, Openness (to experience), Agreeableness, and Conscientiousness. As predicted by the authors, the MBTI dimensions revealed healthy and appropriate correlations with corresponding factors from the NEO-PI-R. Specifically, the following averaged concurrent validity correlations were found between the MBTI dimensions and the NEO-PI-R scales: E-I correlated .71 with Extraversion; S-N correlated –.65 with Openness; T-F correlated –.35 with Agreeableness; and, J-P correlated .46 with Conscientiousness. The negative correlations indicate an inverse relationship, that is, those categorized as S (Sensing) on the MBTI obtained low scores on Openness, whereas those categorized as N (iNtuition) obtained high scores on Openness. In like manner a T or Thinking type tended to obtain low scores on Agreeableness whereas an F or Feeling type tended to obtain high scores.

Recent versions of the MBTI yield additional information beyond the four-letter typological classification. For example, the 144-item form Q, available on-line, provides a highly detailed and sophisticated summary report that partitions each of the four polarities into five facet scores. Hence the report includes a total of 20 facet scores in addition to the four-letter code. For example, the Thinking-Feeling dimension includes bipolar facets such as Logical-Empathetic, Reasonable Compassionate, and Tough-Tender. The dimensions and facets of this version of the MBTI are displayed in Table 9.1. The report includes not only the typological classifications (e.g., T or F) but also a rating for each bipolar facet on an 11-point continuum.

One concern about the MBTI is that the increasing cost of administering the instrument—in the range of $10 to $30 per individual—provides a disincentive for outside researchers who want to conduct reliability or validity studies.

CALIFORNIA PSYCHOLOGICAL INVENTORY (CPI)

Originally published in 1957, the CPI is a true–false test designed expressly to measure the dimensions of normal personality. The instrument is available in two forms, the CPI-434 (Gough, 1995) and the CPI-260 (www.skillsone.com), which is available only online. The component scales and the interpretive strategies are nearly identical for the two versions, which differ mainly in the number of items—434 versus 260. Psychometric properties of both versions are similar and strong. Because of its ease of administration and the immediacy with which the practitioner receives an extensive computer-generated report, the CPI260 rapidly is gaining favor among psychological practitioners.

The CPI-260 is scored for 20 folk measures of personality, 7 work-related scales, and 3 broad vectors. The purpose of the test is to provide a clear picture of the examinee by using descriptors based on the ordinary language of everyday life. Three of the basic personality scales also provide information on test-taking attitudes and therefore function as validity scales. These scales are Good Impression (Gi), which assesses the extent to which the individual presents a favorable image to others; Communality (Cm), which measures unusual responses that might arise from carelessness or faking bad; and Well-being (Wb), which gauges the portrayal of serious emotional problems.

The test developers used an empirical methodology of criterion-keying to develop the majority of the scales. Specifically, extreme groups of participants (mainly college students) were formed on such scale-relevant criteria as school grades, sociability, and participation in curricular activities. Item-endorsement frequencies were then contrasted to ferret out the best statements for each scale. For example, the Sociability (Sy) scale was constructed by contrasting item-endorsement rates for persons reporting a large number of social activities versus those reporting few or no social activities. In constructing four of the folk scales, the authors used a rational basis backed up by indices of internal consistency.

Reliability data for the CPI are respectable. Most alpha coefficients are in the .70s and .80s, with a median value of .76. The test–retest reliability coefficients tend to be somewhat lower, with a median retest correlation of .68. The authors provide a wealth of normative data, including average test scores for 52 samples of males and 42 samples of females, subdivided by education, occupation, college major, gender, and other variables. The basic normative sample consists of 3,000 males and 3,000 females of varying age, social class, and geographic region.

The CPI also is scored on three broad dimensions or vectors derived from decades of factor-analytic studies with the instrument. The three vectors include two basic orientations and a third theme reflecting ego integration. The first basic orientation called vector 1 or v.1 has two polarities: toward people or toward one’s inner life. This vector is similar to the extraversion–introversion dimension found in nearly every personality theory ever proposed.

The second basic orientation or v.2 also has two polarities: rule-favoring or rule-questioning. This vector reflects a conventional–unconventional dimension also found in many studies. These first two bipolar orientations, v.1 and v.2, provide a 2 × 2 typology of four lifestyles termed the Implementer, Supporter, Innovator, and Visualizer lifestyles, described below. The third vector or v.3 assesses a 7-point continuum variously referred to as self-realization, psychological competence, or ego integration.

Results from several correlational studies confirm distinctive psychological portraits for the four lifestyles mentioned above. Briefly, the four life styles are as follows:

- Implementers (extroverted and rule-favoring) tend to do well in managerial and leadership roles.
- Supporters (introverted and rule-favoring) function well in supportive or ancillary positions.
- Innovators (extroverted and rule-questioning) are adept at creating change.
- Visualizers (introverted and rule-questioning) work best alone in fields such as art or literature.

The CPI is useful for helping predict the following: • Psychological and physical health • High school and college achievement • Effectiveness of student-teachers • Effectiveness of police and military personnel • Leadership and management success. The CPI is particularly effective at identifying adolescents or adults who follow a delinquent or criminal lifestyle.

NEO PERSONALITY INVENTORY REVISED (NEO PI-R)

The NEO Personality Inventory-Revised (NEO PIR) embodies decades of factor-analytic research with clinical and normal adult populations (Costa & McCrae, 1992). The test is based upon the five factor model of personality. It is available in two parallel forms consisting of 240 items rated on a five-point dimension. An additional three items are used to check validity. A shorter version, the NEO Five-Factor Inventory (NEO-FFI) is also available (Costa & McCrae, 1989).

Form S is for self-reports whereas Form R is for outside observers (e.g., the spouse of a client). The item format consists of five-point ratings: strongly disagree, disagree, neutral, agree, strongly agree. The items assess emotional, interpersonal, experiential, attitudinal, and motivational variables.

The five domain scales of the NEO PI-R are each based upon six facet (trait) scales. The internal consistency of the scales is superb: .86 to .95 for the domain scales, and .56 to .90 for the facet scales. Stability coefficients range from .51 to .83 in three- to seven-year longitudinal studies. Validity evidence for the NEO PI-R is substantial, based on the correspondence of ratings between self and spouse, correlations with other tests and checklists, and the construct validity of the five-factor model itself.

The NEO PI-R is an excellent measure of personality that is especially useful in research. The test also shows promise as a measure of clinical psychopathology. For example, Clarkin, Hull, Cantor, and Sanderson (1993) found that patients diagnosed with borderline personality disorder scored very high on Neuroticism and very low on Agreeableness, which resonates strongly with every clinician’s response to these challenging patients.

One minor concern about the instrument is that it lacks substantial validity scales—only three items assess validity. The administration of the NEO PI-R assumes that subjects are cooperative and reasonably honest. This is usually a safe assumption in research settings but may not hold true in forensic, personnel, or psychiatric settings. For purposes of education and research, several psychometricians have constructed websites where it is possible to self-administer an equivalent version of the NEO PI-R.

Recently, the developers of the NEO-PI-R produced a new version that is more readable and therefore better suited to students as young as 12 years of age. The NEO-PI-3 is a careful and modest revision of the original instrument that addresses a number of problematic items difficult for adolescents and young adults to comprehend. The NEO-PI-R consists of 240 items rated on a 5-point Likert scale from Strongly Agree to Strongly Disagree. The authors identified 30 items using words on a par with laissez-faire, fastidious, and adhere that even adults might find challenging. The authors rewrote these items for transparency and carefully tested them for equivalence in a new sample of 500 respondents.

An additional 18 items were rewritten because they revealed low item-total correlations with the facet (trait) scale to which they belonged. The resulting instruments, the NEO-PI-3, retained the original five-factor structure and revealed better internal consistency and readability than the previous version.

STABILITY AND CHANGE IN PERSONALITY

Most of us have heard adages like “People don’t change” or “Personality traits become exaggerated with age”. Opinions abound on the stability or malleability of personality. What the lay public seldom recognizes, however, is that issues of stability and change in personality can be approached with empiricism through psychological assessment. A few tests figure prominently in lifespan developmental research, especially instruments that embody the five-factor approach.

One question central to the field of personality psychology is whether personality remains stable throughout life, or reveals predictable shifts in certain qualities as we age. On the surface this question appears amenable to straightforward longitudinal research. Simply administer a suitable instrument to a large sample of the general population, and retest every five years or so. Then, chart the trends in dimensions of personality over the life span. One problem is selective attrition, in which less healthy individuals tend to drop out, disappear, or discontinue the project for reasons known and unknown. Although there are methodological adjustments for minimizing the impact, selective attrition nonetheless may skew results toward an unrealistically optimistic picture of trends in aging. Another problem with longitudinal research is that decades of time are needed to follow individuals over the life span. Long-term developmental research is difficult and expensive.

An alternative strategy is cross-sectional research in which a large sample of individuals of all ages (from teenagers to persons in their 90s) is tested at one point in time, allowing for immediate age comparisons in personality characteristics. This is an appealing technique but also fraught with methodological concerns. In particular, the cross-sectional strategy is vulnerable to a research problem known as cohort effects (Schaie, 2011). A cohort is a group of individuals born at roughly the same time who therefore share particular life experiences and historical influences. A cohort effect is the inference that differences between age groups (cohorts) are due to disparities in the nature and quality of early developmental or historical experiences rather than caused by the impact of aging.

In studying age trends in personality, a certain degree of tentativeness is warranted, because no single study or method is conclusive. Some researchers combine longitudinal and cross-sectional methods in what is known as the cross-sequential approach. This method involves the longitudinal retesting of cross-sectional study participants on at least one additional occasion. The beauty of the cross-sequential method is that cohort effects can be distinguished from genuine longitudinal trends. This allows researchers to identify typical changes resulting from intrinsic maturation.

It is important to mention that core issues of personality change may not be wholly amenable to traditional methods of measurement.

PERSONALITY STABILITY AND CHANGE IN MIDDLE AND LATE LIFE

Do people change in personality traits across the life course? Several researchers have sought to identify mean-level changes or normative changes that are generalizable patterns of development found in most people. Most commonly, investigators use the Big Five model of personality as their measurement perspective. This is the view that personality is best conceived as five factors labelled neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness.

Individual reports of developmental trends in the Big Five factors over the life course often seem inconclusive or contradictory. In a study of 2,274 participants in their forties retested after 6 to 9 years, Costa, Herbst, McCrae, and Siegler (2000) found minimal or no change in mean level of the Big Five factors, even though popular accounts indicate that midlife is a time of crisis and turmoil. In contrast, others report that personality traits continue to transform in middle and old age, with increases in conscientiousness and agreeableness, and decreases in some elements of extraversion.

Perhaps the best approach to this dilemma is a comprehensive synthesis of all relevant studies by means of meta-analysis. Meta-analysis is a sophisticated statistical procedure for combining data from multiple studies. In this method, results from studies using different measurement techniques can be transformed to a common metric, the effect size, and then combined for powerful statistical analyses (Cohen, 1988). One type of effect size is Cohen's d, which is the mean difference on a variable between two comparison groups divided by the standard deviation of the pooled groups on that variable, or $d = (M1 – M2)/sp$. While effect sizes exist theoretically on an infinite range in positive and negative directions, it is rare in everyday research that they exceed the bounds of +3.0 to −3.0, a value of 0 indicating no difference between groups. The beauty of meta analysis is that studies using diverse tests, measuring slightly different constructs, based on varying scales of measurement, nonetheless can be transformed to the common metric of effect size and then combined for comprehensive analysis.

Soto, John, Gosling, and Potter (2011) pursued the question of age differences in personality traits with an intriguing and massive cross-sectional research project. Their sample consisted of an astonishing 1,267,218 individuals (age 10 to 65) who responded to a Web-based questionnaire on Big Five personality traits. Their assessment instrument was the Big Five Inventory (BFI), a simple 44-item measure with excellent psychometric qualities. The BFI is freely available to researchers for non-commercial purposes. The test developers isolated two distinctive subscales, called Facet scales, for each of the Big Five domains. Their assessment tool, the BFI, is appropriate for children and adults of any age with a fifth-grade reading level. However, for participants younger than 10 and older than 65, sample sizes were too small to provide highly reliable estimates. The minimum sample size for each year of age was 922, and at least 422 persons of each gender were included. Their study is vast and comprehensive in its conclusions. The literature on age differences and longitudinal trends in Big Five personality domains is vast.

THE ASSESSMENT OF MORAL JUDGMENT

THE MORAL JUDGMENT SCALE

Kohlberg has proposed one of the few theories of moral development that is both comprehensive and empirically based. Kohlberg generated a method of assessment that is widely used and intensely debated. An adaptation of Kohlberg's approach known as the **Defining Issues Test** is also debated.

Stages of Moral Development: Kohlberg's theory grew out of Piaget's (1932) stage theory of moral development in childhood. Kohlberg extended the stages into adolescence and adulthood. In order to explore reasoning about difficult moral issues, he devised a series of moral dilemmas. Kohlberg concluded that there are three main levels of moral reasoning, with two substages within each level.

Kohlberg's Levels and Stages of Moral Development:

- Level 1: Preconventional
  - Stage 1. Punishment and obedience orientation: The physical consequences determine what is good or bad.
  - Stage 2. Instrumental relativism orientation: What satisfies one's own needs is good.
- Level 2: Conventional
  - Stage 3. Interpersonal concordance orientation: What pleases or helps others is good.
  - Stage 4. "Law-and-order" orientation: Maintaining the social order and doing one's duty is good.
- Level 3: Postconventional or Principled
  - Stage 5. Social contract-legalistic orientation: Values agreed upon by society determine what is good.
  - Stage 6. Universal ethical-principle orientation: What is right is a matter of conscience derived from universal principles.

One use of his measurement instrument, the Moral Judgment Scale, is to determine a respondent's stage of moral reasoning.

The Moral Judgment Scale consists of several hypothetical dilemmas presented one at a time. In its latest revision, the scale comes in three versions called Forms A, B, and C. Scoring is quite complex, based on the examiner's judgment of responses in relation to extensive criteria outlined in a detailed scoring manual. Although there are several different dimensions to scoring, the one element most frequently cited in research studies is the overall stage of moral reasoning that characterizes a respondent.

Critique of the Moral Judgment Scale:

Early versions of the Moral Judgment Scale suffered serious shortcomings of scoring and interpretation. For example, in his doctoral dissertation, Kohlberg proposed two scoring systems: one using the sentence or completed thought as the unit of scoring, the other relying upon a global rating of all the subject's utterances as the unit of analysis. Neither approach was fully satisfactory, and early reviews of the scale were justifiably critical of its reliability and validity.

In response to these criticisms, Kohlberg and his associates developed a scoring system that is unparalleled in its clarity, detail, and sophistication. Fortuitously, since the moral dilemmas of the Moral Judgment Scale have remained constant over the years, it is possible to apply the new scoring system to old data. This investigation reports the results of using the new scoring system in a longitudinal study spanning more than 20 years. The results are impressive and offer strong support for the reliability and validity of the instrument. Test–retest correlations for the three forms were in the high .90s, as were interrater correlations. Longitudinal scores of subjects tested at three- to four-year intervals over 20 years revealed theory-consistent trends. Fifty-six of 58 subjects showed upward change, with no subjects skipping any stages.

The internal consistency of scores was also excellent: about 70 percent of the scores were at one stage, and only 2 percent of the scores were spread further than two adjacent

stages. Cronbach's alpha was in the mid-.90s for the three forms. In sum, the Moral Judgment Scale is reliable, internally consistent, and possesses a theory-confirming developmental coherence.

The Defining Issues Test

The Defining Issues Test (DIT) is similar to the Moral Judgment Scale but incorporates a much simpler and completely objective scoring format (Rest, 1979, 1986). The examinee reads a series of moral dilemmas similar to those designed by Kohlberg and then chooses a proper action for each. For example, one dilemma involves a patient dying a painful death from cancer. In her lucid moments, she requests an overdose of morphine to hasten her death. What should the doctor do? Three options of the following kind are listed:

_____He should give the woman a fatal overdose
_____Should not give the overdose
_____Can't decide

The examinee's choice does not enter directly into the determination of the moral judgment score. The real purpose in forcing a choice is to cause the examinee to think about the importance of various factors in making the decision. Following the choice of proper action, the examinee rates the importance of several factors on a five-point Likert scale: great, much, some, little, or no importance. The factors are distinct for each dilemma. The factors differ in the level of moral judgment they signify, ranging from Kohlberg's stage 1 through stage 6..

The most widely used score is the P score, which is a percentage of principled thinking. Reliability of the P score ranges from .71 to .82 in test–retest studies. Validity has been studied by contrasting groups known to differ on principled thinking. In longitudinal studies, significant upward trends were found over six years and four testings. Recently, Rest has recommended a new measure of moral judgment, the N2 index, calculated on the basis of several complex formulas that use both ranking and rating data. The two indices are highly correlated in the .90s. Nonetheless, in a retrospective analysis of previous studies, the N2 index outperformed the P index by a substantial margin.

Over 600 articles have been published on the Defining Issues Test (McCrae, 1985). In general, the instrument is considered a useful alternative to Kohlberg's Moral Judgment Scale, particularly for research on group differences in moral reasoning.

CAUTIONS ABOUT THE DIT

First, the test uses two moral dilemmas from the Vietnam War and is, therefore, somewhat dated. Many young examinees have little knowledge of (and perhaps no interest in) this topic and may find it difficult to identify with these questions. Another dilemma— the classic case of whether Heinz should steal a drug to save his wife's life—is also of dubious value since it has been widely publicized and reprinted in college textbooks. A significant proportion of prospective examinees are no longer naive about this moral dilemma. The DIT is biased against conservatively religious individuals. Certainly, it is well established that conservative or fundamentalist religious people tend to score lower than average on the P score of the Defining Issues Test.

These researchers demonstrate empirically that certain DIT items measure a different construct for conservative religious persons than for the general population. As a consequence, the validity of the test in these groups is open to question. It is a reasonable assumption that individuals who receive higher P scores on the DIT should also refrain from moral transgressions such as cheating on tests. Another concern about the DIT is the dearth of

norms pertinent to minority groups. In spite of the concerns listed here, the DIT is a widely respected test, particularly for research on moral reasoning.

THE ASSESSMENT OF SPIRITUAL AND RELIGIOUS CONCEPTS

Within the field of psychology, transcendent topics such as spiritual well-being or faith maturity never have received mainstream attention. Many years ago, Gordon Allport (1950) lamented that the subject of religion "seems to have gone into hiding" among intellectuals and academic researchers. The situation is little improved in contemporary times. Yet researchers have no right to retire from the field, given its significance to the average person.

Most people embrace a spiritual perspective in life, and surely this must have some bearing on their adjustment, behavior, and outlook. Unfortunately, the field of psychology, including the specialty area of testing, largely has maintained an indifference to this important aspect of human experience. Worse yet, in many intellectual circles the endorsement of spiritual or religious sentiments is seen as evidence of psychopathology. Among others, Sigmund Freud endorsed a cynical view of religion in his aptly titled essay, The Future of an Illusion (1927/1961). Yet for many persons, a connection with the transcendent is essential to meaning in life. This is especially so in times of extreme duress, as when personal annihilation knocks at the front door.

Spiritual and religious dimensions to life often serve constructive purposes and that assessment within these domains is worthy of additional study.

Challenges and Purposes of Religious and Spiritual Assessment

Other than personal or scholarly curiosity about religious and spiritual matters, what might be the motivation for religious and spiritual assessment? Further, what is spirituality, and how is it distinguished from religiousness? It appears evident that some people can be religious without being spiritual, ghost walking through religious traditions with no involvement of heart. But is it possible to be spiritual without being religious?

According to the Yearbook of American and Canadian Churches (2012), total church membership has declined steadily for many years, even though some denominations have increased in popularity. Alongside this general decline in traditional forms of worship, spiritual practices have expanded in popularity, as witnessed by the proliferation of meditation, 12-step, Eastern, yoga, and other broadly spiritual practices. For example, mindfulness meditation, with roots in Buddhism, is more popular than ever. It is recommended for problems with anxiety, depression, pain, hyperactivity, sleep, parenting, stress, tinnitus, psoriasis, Parkinson's disease—the list goes on and on. Those who practice mindfulness, for whatever initial purpose, often embrace it as a way of being in the world, a spiritual discipline.

But what is spirituality, and how is it distinguished from religiousness? Religion is frequently defined by institutional affiliation, whereas spirituality is not. Religion is also often considered more external or mediated by a group, whereas spirituality is more closely associated with personal experience and is less doctrinaire.

Efforts to develop measures of spirituality and religiousness have flourished in recent years. Hill and Hood (1999) compiled information on 125 measures of spirituality/religiosity. Dozens of new scales have been developed since the release of their compendium. The Search Institute, which serves educators, parents, youth groups, faith communities, and researchers in efforts to create a better world for children, lists 18 easily accessible measures of spirituality, the majority published in recent years (www .search-institute.org). There is an abundance of available instruments.

Richards and Bergin (2005) make a strong case that clinicians need to include spiritual and religious assessment in psychotherapy. They list five reasons for a spiritual-religious assessment of clients, which include: understanding client world view and improving the capacity of the therapist to empathize; establishing the impact of spiritual-religious views on the presenting problem; determining if the spiritual religious views of the client can be used for growth or coping; identifying which spiritual interventions might be useful in therapy; and, recognizing any spiritual doubts that need to be addressed in therapy. These benefits of spiritual-religious assessment can be extended beyond the therapeutic alliance.

Historical Overview on Spiritual and Religious Assessment

Interest in the psychology of religion can be traced to the early 1900s when William James (1902) composed his masterpiece, The Varieties of Religious Experience. In this book, James catalogued the manifold ways in which humans reveal their interest in transcendent matters. His overall conclusion was that religion is "an essential organ of our life, performing a function which no other portion of our nature can so successfully fulfill."

In the late 1960s the assessment of religious variables began to appear. One of the first such measures was the Allport-Ross Religious Orientation scales, which assessed two dimensions of religious expression, the **intrinsic** and the **extrinsic** (Allport & Ross, 1967). Intrinsically religious persons were thought to *live* their religion (e.g., to find meaning, direction, outlook), whereas extrinsically religious persons were believed to *use* their religion (e.g., to seek security, status, sociability). Allport referred to intrinsic religious expression as a genuine or mature religious orientation, whereas extrinsic religious expression was viewed as immature. The impetus for development of these scales was Allport's distressing observation of a positive relationship between religiosity (in certain forms) and authoritarian, bigoted, prejudicial attitudes.

Allport was convinced that intrinsically oriented religious individuals rarely would harbor these attitudes. Yet the evidence was overwhelming to Allport that at least some religious individuals did reveal hatred, bigotry, and prejudice toward their neighbors. The usual targets of these malicious attitudes were racial minorities, Jews, and homosexual persons, among others. He reasoned that religious persons with intolerant attitudes possessed a predominantly extrinsic religious orientation; that is, their faith served external goals such as status in the community, belonging to an in-group, and the like. The investigation of this hypothesis (that extrinsically religious persons would be more authoritarian, bigoted, and prejudiced than intrinsically religious persons) required appropriate tools. For this purpose, Allport and colleagues developed the Religious Orientation scales.

Examples of the kinds of items on the 11-item Extrinsic scale and the 9-item Intrinsic scale are as follows:

- The church is important as a place to develop good social relationships. (Extrinsic)
- Sometimes I find it necessary to compromise my religious beliefs for economic reasons. (Extrinsic)
- I try hard to carry my religion over into other aspects of my life. (Intrinsic)
- My religion is important because it provides meaning to my life. (Intrinsic)

Although originally devised in a yes–no format, modern applications of these scales utilize a nine point continuum from (1) strongly disagree to (9) strongly agree (Batson, Schoenrade, & Ventis, 1993). Research on the Religious Orientation scales has not provided strong support for Allport's original hypothesis (Wulff, 1996). In fact, several studies have

shown that persons scoring high on the Intrinsic scale actually reveal *higher* levels of authoritarianism, close-mindedness, and prejudice toward African Americans, gays, and lesbians.

RELIGION AS QUEST

Increasingly, the conceptual basis for the distinction between intrinsic and extrinsic religious orientation has been questioned. Kirkpatrick and Hood (1990) summarized the major theoretical and methodological criticisms of the scales as follows:

- A lack of conceptual clarity in what the Intrinsic–Extrinsic scales are supposed to be measuring. Are these types of motivation (i.e., the motives associated with religious belief and practice), or personality variables (i.e., pervasive aspects of institutional behavior or involvement), or something else?

• A confusion over the relationship between the Intrinsic–Extrinsic scales. In particular, are these opposite ends of a single bipolar dimension, or do the scales measure separate dimensions (so that conceivably some persons could score high on *both*)?

Other problems cited include weaknesses in the factorial structure, reliability, and construct validity of the scales; excessive reliance on a "good religion" versus "bad religion" dichotomy; and the folly of defining and studying religiousness independent of belief content. In response to the limitations of the Religious Orientation scales, Batson and his associates (1993) developed a measure of a third religious orientation known as Quest. These researchers consider Quest to be a more mature and flexible religious outlook than the intrinsic and extrinsic orientations. Actually, Allport recognized the elements inherent to this orientation but failed to incorporate them in his Intrinsic scale.

Religion as Quest is characterized by complexity, doubt, and tentativeness as ways of being religious. Examples of the kinds of items on the 12-item Quest scale are as follows:

• My life experiences have led me to reconsider my religious convictions.
• I find religious doubts upsetting. (reverse scored)
• As I grow and mature, I expect my religious beliefs to change.
• Questions are more important to my religious faith than answers.

Items are scored on the same nine-point continuum from (1) strongly disagree to (9) strongly agree. Results are reported as an average rating. Research with 424 undergraduates interested in religion indicates that Quest is, indeed, a dimension of religious experience independent from both Intrinsic and Extrinsic orientations. Whereas Intrinsic and Extrinsic scores correlated .72, Quest revealed negligible relationships with both scales (–.05 with Intrinsic and .16 with Extrinsic).

The intention of its authors was that it assess "the degree to which an individual's religion involves an open-ended, responsive dialogue with existential questions raised by the contradictions and tragedies of life". The three components of the Quest orientation are (1) readiness to face existential questions without reducing their complexity, (2) self-criticism and perception of religious doubts as positive, and (3) openness to change. But critics have charged that the scale may not measure anything religious at all, that instead it may assess agnosticism, anti-orthodoxy, religious doubt, or religious conflict.

Quest is its own dimension of religious expression, and substantial research on the meaning and correlates of this faith orientation has been completed. Quest arose as a response to the limitations of the Intrinsic and Extrinsic approach to the measurement of religious orientation. But this brief 12-item scale possesses its own limitations, chief among them its brevity and factorial simplicity.

The Spiritual Well-Being Scale

The concept of spiritual well-being can be traced to a paper by Moberg (1971) that proposed this form of well-being as an essential component of healthy aging. Spiritual well-being was conceptualized as a two-dimensional construct consisting of a vertical dimension and a horizontal dimension. The vertical dimension concerned well-being in relation to God or a higher power, whereas the horizontal dimension involved existential well-being, which is a sense of purpose in life without any specific religious reference.

The challenge of developing a scale to measure these components of well-being was taken up by Ellison (1983) and Paloutzian and Ellison (1982). Their instrument was designated the Spiritual Well-Being Scale (SWB Scale). The SWB Scale consists of two subscales: Religious Well-Being (RWB), which assesses the vertical dimension of well-being in relation to God; and Existential Well-Being (EWB), which measures the horizontal dimension of well-being in relation to life purpose and life satisfaction. Each subscale consists of 10 items that are scored from 1 (strongly disagree) to 6 (strongly agree). The items from the two subscales are combined on the SWB Scale, with odd-numbered items assessing religious well-being and even-numbered items assessing existential well-being. Some items are worded negatively; these are reverse scored so that a higher score always indicates greater wellbeing.

Ellison described the SWB Scale as a measure of psychospiritual personality integration and resultant well-being. According to this view, well-being consists of "the integral experience of a person who is functioning as God intended, in consonant relationship with Him, with others, and within one's self". If this conceptualization is correct, healthy spirituality as measured by the SWB Scale should show positive relationships with independent measures of health and subjective well-being. The one identified shortcoming of the SWB Scale is an apparent low ceiling, especially in religious samples.

The Assessment of Spirituality and Religious Sentiments (ASPIRES) Scale

The Assessment of Spirituality and Religious Sentiments (ASPIRES) scale is a recent and promising measure of spiritual and religious variables (Piedmont, 2010). What makes the test unique is its predictive power above and beyond the Big Five personality factors. In other words, ASPIRES represents an extension of these well-established components into a sixth dimension of personality. The scale also is robust across cultures and useful within nonreligious samples, including agnostics and atheists.

The 35-item ASPIRES scale measures two dimensions, spiritual transcendence and religious sentiments. Spiritual transcendence is further subdivided into three facets: prayer fulfillment, universality, and connectedness. Religious sentiments consists of two facets: religious involvements, and religious crisis.

- Spiritual Transcendence Scale (STS) : The motivational capacity to create a broad sense of personal meaning for one's life
- Prayer Fulfillment (PF) Facet : The ability to create a personal space that enables one to feel a positive connection to some larger reality
- Universality (UN) Facet : The belief in a larger meaning and purpose to life
- Connectedness (CN) Facet : Feelings of belonging and responsibility to a larger human reality that cuts across generations and groups
- Religious Sentiments Scale (RSS) : The extent to which an individual is involved in and committed to the precepts, teachings, and practices of a specific religious tradition
- Religious Involvements (RI) Facet : How actively involved a person is in performing various religious rituals and activities

- Religious Crisis (RC) Facet: Extent to which a person may be experiencing problems, difficulties, or conflicts with the God of their understanding

The ASPIRES scale demonstrates strong psychometric qualities. Alpha reliabilities for the facet scales range from .60 (CN) to .95 (PF) with a mean alpha of .82 (Piedmont, 2010). The normative sample consists of nearly 3,000 individuals, ages 17 to 94, from four geographic areas of the Midwestern and East Coast regions of the United States. The STS portion of the scale correlates with religious and spiritual variables and incrementally predicts (above and beyond the Big Five dimensions) relevant outcomes such as social support and prosocial behavior (Piedmont, 1999, 2001). The test holds up well cross-culturally, revealing a robust factor structure in diverse religious groups and cultures.

The STS component of ASPIRES yields incremental validity in the prediction of treatment outcome in spiritually based programs for alcohol and drug abuse.

The Faith Maturity Scale

In 1987, six major Protestant denominations undertook a national four-year study of personal faith, denominational allegiance, and their determinants (Benson, Donahue, & Erickson, 1993). Funded in part by the Lilly Endowment, this project spawned what is undoubtedly the most sophisticated measure of spiritual maturity ever conceived. The Faith Maturity Scale (FMS) arose as a practical tool to serve three research purposes:

1. Provide baseline data on the vitality of faith in mainstream Protestant congregations

2. Identify the contributions of demographic, personal, and congregational variables to faith development

3. Furnish a criterion variable for evaluating the impact of religious education in mainstream denominations.

The development of the scale was a time-consuming and careful process that began with a working definition:

Faith maturity is the degree to which a person embodies the priorities, commitments, and perspectives characteristic of vibrant and life-transforming faith, as they have been understood in "mainline" Protestant traditions.

Using open-ended questionnaires with a convenience sample of 410 mainline Protestant adults, the test developers next identified eight core dimensions of faith maturity. Three advisory panels provided ongoing counsel during this stage and the next phase of item writing. These interactions assured that the scale possessed face and content validity.

The resulting FMS is a 38-item test that embodies key indicators of faith maturity in eight core areas. Items are answered on a seven point scale from 1 = never true to 7 = always true. Based upon the areas assessed, the reader will notice that right belief is only one aspect of a mature faith. In large measure, faith maturity is defined by value and behavioral consequences. As the authors note, the Faith Maturity Scale "parts company with more traditional ways of defining and measuring personal religion." Yet it does embody the kinds of behaviors and attitudes that derive from a dynamic, life-transforming faith. These behaviors and attitudes are consistent with the theology found in most religious traditions but are especially pertinent for the particular purpose of assessing faith maturity in the Protestant context.

The eight core dimensions of the Faith Maturity Scale are:

- Trusts and believes (5 items): Every day I see evidence that God is at work in the world
- Experiences the fruits of faith (5 items): I feel weighed down by all my responsibilities (reverse scored)
- Integrates faith and life (5 items): My faith influences how I think and act every day

- Seeks spiritual growth (4 items): I take time to meditate or pray
- Experiences and nurtures faith in community (4 items): I talk with others about my faith
- Holds life-affirming values (6 items): I tend to be critical of other persons (reverse scored)
- Advocates social change (4 items): I believe the churches of this nation should get involved in political issues
- Acts and serves (5 items): I offer significant amounts of time to help others

The FMS is scored as the mean of the 38 items, which yields a potential range of 1 to 7. The average score for 3,040 adults in five Protestant denominations was 4.63, which indicates that the instrument avoids the "ceiling effect". The estimated reliability of the scale is very robust across age, gender, occupation, and denomination, with typical coefficient alphas of .88 (Benson et al., 1993). Test–retest reliability was not reported.

The validity of the scale is supported by several lines of evidence, beginning with the careful approach to item selection, by which face validity and content validity were built-in. Construct validity was demonstrated in several ways. First, it was predicted and confirmed that groups presumed to differ in levels of faith maturity would obtain significantly different mean scores on the FMS.

The scale also revealed predictive utility. Specifically, FMS scale scores were strongly related to a variety of prosocial behaviors such as donating time to help those who are poor, hungry, or sick; promoting a greater role for women in the church; and endorsing the use of foreign policy to challenge apartheid.

ASSESSMENT OF CREATIVITY

The topic of creativity has fascinated and yet also vexed psychologists and educators for more than a century. Researchers are beginning to understand fundamental elements common to many forms of creativity, yet, a simple definition of creativity remains elusive, and its assessment continues to be problematic. It is no exaggeration to state that hundreds of tests of creativity have been published. Some of these instruments possess respectable psychometric qualities, but most are of questionable validity. In the field of creativity there are no acknowledged "gold standards" for assessment. In part, this is because of the criterion problem—the difficulty in defining creativity.

Psychologists have sought to understand creativity since at least the early 1900s. For example, John B. Watson, the famous American behaviorist, suggested simplistically that a poem or brilliant essay is the mere product of shifting words around until a new pattern is hit upon (Watson, 1928). Fortunately, Watson's simplistic views were followed by a large number of more thoughtful formulations.

A few perspectives on creativity from eminent researchers:

• If a response is to be called original, it must be to some extent adaptive to reality (Barron, 1955, p. 553).

• We may proceed to define the creative thinking process as the forming of associative elements into new combinations that either meet specified requirements or are in some way useful (Mednick, 1962, p. 221).

• Creativity can be regarded as the quality of products or responses judged to be creative by appropriate observers, and it can also be regarded as the process by which something so judged is produced (Amabile, 1983, p. 31).

• Creativity involves bringing something into being that is original (new, unusual, novel, unexpected) and also valuable (useful, good, adaptive, appropriate) (Ochse, 1990, p. 2).

• Creativity is the ability to produce work that is both novel (i.e., original, unexpected) and appropriate (i.e., useful, adaptive concerning task constraints) (Sternberg & Lubart, 1999, p. 3).
• Creativity is a specific capacity to not only solve problems but to solve them originally and adaptively (Feist & Barron, 2003, p. 63).
• Creativity is the ability to come up with ideas or artifacts that are new, surprising, and valuable (Boden, 2004, p. 1).

These conceptual definitions emphasize novelty and usefulness of the creative product, but also suggest that creativity is a particular kind of process as well. Relevant to assessment, one controversy overshadows the study of creativity. This is the question whether creativity is general or domain-specific in nature. Kaufman and Baer (2004) acknowledge the complexity of the specific versus general debate, noting that the answer hinges on the definition of creativity and the assessment methods employed. But they also render a final conclusion that that the evidence for *c* (general creativity) is weak. We agree with their verdict that creativity appears to be domain-specific.

There are as many domains of creativity as there are fields of inquiry or expression, whether in science, art, economics, service, leadership, entrepreneurship, or whatever. People who are creative in one field typically reveal talent in closely allied fields as well. Gifted writers usually can be good poets, if they choose, and vice versa. A creative scientist might excel at mechanical problem-solving as well. The number of domains must be somewhere between huge (nearly infinite), and small (a handful). But creativity is not a single general factor.

The study by Kaufman (2012) is representative. His investigation was based on the common sense view that layperson perceptions of constructs like intelligence, wisdom, personality, or creativity, when analyzed collectively, embody some degree of practical wisdom.

Participants were 2,318 college students asked to rate an initial collection of 94 items. Students rated themselves on a 5-point Likert scale from 1 (*much less creative*) to 5 (*much more creative*) on each item. The items were gleaned from several prior research projects. The 94 items coalesced into five factors (from factor analysis), which provided a basis for reducing the scale to 50 items organized into 5 domains of about 10 items each. The emergent domains were the following:

• Self/Everyday: Successfully dealing with problems in self and others, teaching creatively.
Items resemble *Helping friends deal with difficult problems*.
• Scholarly: Effectively analyzing problems and coming up with new and creative ideas. Items resemble *Finding a new way to think about old problems*.
• Performance: Successfully composing lyrics and singing a new song in public. Items resemble *Making up lyrics and melody for an amusing song*.
• Mechanical/Scientific: Efficiently solving a scientific or mechanical problem. Items resemble *Designing and conducting a scientific experiment*.
• Artistic: Productively drawing or painting a landscape or still life. Items resemble *Crafting a sculpture or piece of pottery*.

The new instrument, called the **Kaufman Domains of Creativity Scale** (K-DOCS), demonstrated strong psychometric qualities, with internal consistency coefficients of .83 to .86 and test–retest reliabilities (132 participants retested after two weeks) of .78 to .86. In addition to finding a clear-cut five-factor structure for the test, additional evidence of validity was found in the domain scale correlations with Big Five personality dimensions, which were

theoretical sensible, for example, Openness to Experience correlated significantly with all creativity domains except Mechanical/Scientific. Over the years, creativity has been studied in terms of cognitive processes, personal characteristics, and behavioral products.

Creativity as Process

Several theorists and researchers have focused on underlying cognitive processes in their understanding of creativity. Of historical interest is Wertheimer's (1945) suggestion that creativity arises when the thinker grasps the essential features of a problem and their relation to a final solution—the so-called "aha!" phenomenon. Wallas (1926) theorized that such insights often occur after a period of incubation wherein the unconscious mind rearranges the features of the puzzle even while the conscious mind takes "time off" from the problem.

Mednick (1962) proposed that creativity is the capacity to combine remote associations. According to this view, creativity is a matter of novel arrangements of unusual associations to a given stimulus. Based on his process-oriented view of creativity, Mednick (1962) developed **the Remote Associates Test (RAT),** a clever index of the remoteness of verbal associations. The RAT is a timed, 40-minute paper-and-pencil test with inter item reliability consistently above .90. (Mednick & Mednick, 1966).

Some examples of the kinds of items encountered on the RAT:

- rat–blue–cottage _____
- out–dog–cat _____
- wheel–electric–high _____
- surprise–line–birthday _____

For each triplet, the examinee must find a fourth word that "fits" in the sense of having reasonable (but often remote) associations to the other three words. (The correct answers above are *cheese, house, wire,* and *party*.) Competent performance on this test would appear to require a capacity to examine several novel or remote associations at the same time and to search for the one association that is common to all three stimulus words.

Validity studies of the RAT have been mixed in outcome. Early studies were promising and indicated that high RAT-scorers tended to receive higher ratings for the creativity of their products (e.g., architectural designs, research projects, suggestions, and drawings) than low scorers. However, later studies indicated complex patterns between RAT scores and other creativity indices. Ochse (1990) provides a thorough appraisal of RAT validity. He concludes that the test may predict scores on tests of verbal fluency, but fails to predict creativity in general. In other words, the RAT is not so much a general measure of creativity as a specialized measure of verbal intelligence.

Creativity as Personal Characteristics

Guilford (1950) was one of the first researchers to define creativity in terms of the person when he asserted that "creativity refers to the abilities that are most characteristic of creative people." This helped inspire an expansion of research on the personal characteristics of creative persons. Much of this research has relied upon contrasts of peer nominated high- and low-creative persons in various professions. In this methodology, colleagues within a field of study nominate other individuals who are high and low in creativity, and their consensus view is used to identify two select groups of individuals (high-creative, low-creative). These groups are then contrasted on personality measures, including self-checked adjectives and standard personality inventories.

Based on hundreds of studies, a fairly stable set of core characteristics of creative persons has emerged. Interestingly, the distinguishing characteristics of creative individuals appear to be largely temperamental, although a certain minimum level of intelligence also is

required. Harrington (1975) has captured a not altogether flattering portrait of the creative person in his **Composite Creative Personality Scale**, which consists of 42 self-checked adjectives (from a larger list) that empirically distinguish creative from non-creative persons. These adjectives include many positive terms such as *active, curious, imaginative, inventive, original, resourceful,* and *sensitive,* but also embrace negative terms such as *argumentative, cynical, egotistical, impulsive, rebellious,* and *unconventional*.

The particular link between personality characteristics and creative behavior also depends on the specific domain of investigation. For example, compared with their less creative counterparts, creative artists tend to be more spontaneous, creative writers tend to be more nonconforming, creative architects tend to be less flexible, and creative engineers tend to be better adjusted than other groups. Sternberg (2002) has proposed that creative individuals are distinguished not so much by specific traits as by the heartfelt *decision* to be creative. This perspective suggests that creative individuals will be characterized by a stubborn dedication to their creative endeavors, even when rewards for their activities seem to be lacking. The opinion that creativity resides within qualities of the person continues to be popular. From this perspective, self-report measures are the natural and preferred assessment method.

Self-Report Measures of Creativity:

1. Biographical Inventory of Creative Behaviors (BICB) (Batey, 2007) : Based on the implicit assumption that creativity is a general attribute, the BICB consists of 34-items rated yes/no by the respondent. Items consist of behaviourally anchored creative accomplishments "actively involved in" over the last 12 months. Results range from 0 to 34, yielding a single overall score without subscales. Higher scores indicate greater creativity. Domain coverage is broad. Items resemble written a poem, painted a picture, devised a recipe, coached a team, held an office. The scale possesses good internal consistency (a = .74) and correlates appropriately with other measures of creativity (Furnham, Batey, Anand, & Manfield, 2008).
2. Creative Achievement Questionnaire (CAQ) (Carson, Peterson, & Higgins, 2005): Innovative in its measurement approach, the CAQ assesses creativity in 10 domains: Visual Arts, Music, Dance, Architectural Design, Creative Writing, Humor, Inventions, and Scientific Discovery. Although an overall score can be obtained, the implicit assumption of the test is that creativity is domain specific. Hence, a high score in one domain is sufficient to demonstrate creativity. Each domain consists of eight items, numbered 0 through 7, representing increasing levels of creative achievement. Most items are binary, but higher numbered items in each domain require a numerical entry. For example, item 7 in Creative Writing might request the number of stories published in literary sources. The entry for this item (for example, "3") is multiplied by the item number to obtain the score ($7 \times 3 = 21$). This inventive scoring approach allows for the detection of persons with exceptional creativity in one or more domains.
3. Revised Creative Domain Questionnaire (CDQ-R) (Kaufman, Cole, & Baer, 2009): Simple but effective in its format, the CDQ-R consists of 21 items in four domains: Drama (e.g., acting, dancing, writing), Math/science (e.g., chemistry, logic, computers), Arts (e.g., crafts, design, painting), and interaction (e.g., leadership, selling, teaching). Respondents are asked to self-rate their creativity in each activity. Items are completed on a six-point scale (no midpoint) ranging from Not at all creative to Extremely creative. The four domain scores are averaged to obtain an overall creativity score. The scale possesses reasonable reliability, with internal

consistencies of.71 to .76. for the domains and .82 for the overall scale. Regarding validity, the CDQ-R domain scores reveal theoretically appropriate correlations with Big Five personality dimensions (e.g., Openness to Experience correlates with all four domains; Extraversion correlates with Drama but not Math/Science).

Creativity as Product

The most enduring definitions of creativity have used the *product* as the distinguishing sign of this capacity. According to this approach, creative persons create products (ideas, inventions, writings, artistic outputs, etc.) that meet certain criteria. Jackson and Messick (1968) applied four criteria to creativity:

- Novelty: Creative products are new, or at least represent a new application of the familiar.
- Appropriateness: The product must be appropriate to the context, not merely novel.
- Transcendence of constraints: A product transcends constraints when it goes beyond the traditional.
- Coalescence of meaning: The value of creative products may not be apparent at first, the full significance may only be appreciated with time.

The Jackson and Messick (1968) criteria have proved helpful in delineating the special characteristics of a creative outcome, but they do not constitute a psychological *measure* of creativity. For measures of creativity based on the product-oriented approach, we must examine the seminal studies of Joy Paul Guilford and the various tests inspired by his factor-analytic research. Guilford (1959, 1985) formulated a structure of intellect model that parceled intelligence into 150 factors aligned upon three dimensions: operations, constructs, and products. One of the operations that emerged from Guilford's factor analyses was **divergent thinking.** Divergent thinking is defined as the kind that goes off in different directions. It makes possible changes of direction in problem solving and also leads to a diversity of answers, where more than one answer may be acceptable. Divergent thinking is virtually the opposite of convergent thinking. **Convergent thinking** is the production of a single correct answer determined by facts and reason. Unconstrained, freewheeling thought is the hallmark of the creative person. Tests of divergent thinking are therefore considered excellent measures of creativity.

Guilford and his colleagues developed about a dozen experimental measures of divergent thinking, some of which were subsequently standardized and published as the Christensen-Guilford Fluency Tests. Subtests and items similar to his measures include:

- Alternate Uses: List possible but unusual uses for a common object such as a brick (use it as a door stop, hammer, anchor, or wheel stop)
- Consequences: List possible consequences of a specific hypothetical event, for example, "What would happen if clouds had strings hanging down from them?" (macramé would make a comeback, people would be whisked away, air travel would be hazardous, farmers could winch the clouds down for watering, etc.)
- Ideational Fluency: Name things that belong in a given class such as "Long, thin items" (hair, pin, wire, needle, snake, string, spaghetti, pulled taffy)

Guilford's influence is found in the work of E. Paul Torrance (1915–2003), who developed a group of tests still in use today. **The Torrance Tests of Creative Thinking (TTCT)** (Kim, 2006; Torrance, 1966) are based loosely on Guilford's model, although Torrance was more concerned with the interest level of his measures than with their factorial purity. These tests purport to assess a global cognitive construct of creativity— a style of thinking believed to be essential to creative achievements. The TTCT subtests do not assess

motivation, expertise, intelligence, or other capacities that could contribute to creative productivity. The test comes in two parallel forms, A and B, which are highly comparable.

The TTCT consists of two parts: The TTCT Verbal and the TTCT-Figural. Suitable for ages 6 through 18 and beyond, the TTCT-Verbal contains six subtests: Asking Questions, Guessing Causes, Guessing Consequences, Product Improvement, Unusual Uses and Just Suppose.

The first three verbal subtests are based on the same stimulus card which shows a simple pen and ink drawing of one or two human-like figures engaged in ambiguous activity. In the first activity, *Asking Questions,* the child is encouraged to ask questions about the picture. In the second activity, *Guessing Causes,* the child is told to guess the causes of the action in the picture. In the third activity, *Guessing Consequences,* the child is instructed to speculate about the immediate and long-term consequences. The time limit for each activity is five minutes. In the fourth activity of the Verbal subtests, *Product Improvement,* the task is to suggest improvements to a toy that would make it more appealing to children. *Unusual Uses*, the fifth activity, is a familiar standby in creativity assessment, namely, thinking of unusual uses for a common object such as a brick. The final Verbal subtest is *Just Suppose*, which involves asking the examinee to list the problems and benefits that might arise from an improbable situation.

The verbal subtests are scored according to three criteria: **1.** Fluency—the raw number of relevant ideas; **2.** Originality—the inventiveness or creativity of the ideas; and **3.** Flexibility—the flexibility of categories of ideas.

The TTCT-Figural consists of three activities, which are suitable for ages 5 through 18 and beyond: Picture Construction, Picture Completion and Repeated Figures. The time limit for each activity is 10 minutes. In the first activity, *Picture Construction*, the child draws a picture using a simple shape (jelly bean or pear) as a starting point. The stimulus shape must become an integral part of the constructed picture. In the second activity, *Picture Completion,* the examinee encounters 10 incomplete figures and is asked to complete a drawing from each and then to name each drawing. In the last activity, *Repeated Figures,* the child is provided two or three pages of repeated figures (e.g., circles) and asked to use them in constructing pictures that are then named.

Scoring of the TTCT-Figural subtests is based on five norm-referenced measures and 13 criterion-referenced outcomes. The five norm-referenced measures include: **1.** Fluency—the raw number of stimuli provided; **2.** Originality—the number of statistically infrequent drawings; **3.** Abstractness of Titles—the abstraction level of the titles; **4.** Elaboration—the provision of details and elaboration; **5.** Resistance to Premature Closure—the degree of openness for incomplete figures.

The 13 criterion-referenced measures include a variety of creative strengths expressed in the drawings such as emotional fluency, unusual visual perspective, humor, colorful imagery, and fantasy.

Although scoring of the TTCT is tedious and elaborate—especially for the Figural subtests—experienced testers produce interrater reliabilities in the .90s. Test–retest reliability coefficients are lower, in the range of .50 to .93. The validity of the TTCT is a more complicated question, especially in light of the difficulty of defining the criterion—what is creativity? Yet, the instrument is reasonably predictive of later creative accomplishments, even in the long run. Creativity as measured by the TTCT appears to be more predictive of certain forms of achievement than intelligence. The test has been translated into 35 languages and has spawned more research than any other measure in the field. Among its many strong features, age- and grade-norms are available for more than 50,000 participants, kindergarten through high school. Applications of the test are mainly with school-aged children, although norms are provided for adults as well.

MEASURES OF EMOTIONAL INTELLIGENCE

Emotions and intelligence generally have been viewed as distinct capacities of the individual, each capable of influencing the other, but separate nonetheless. American psychologist Henry H. Goddard (1919) proposed a separation of the emotions and intelligence. He argued that intelligence, properly exercised, can modify and influence emotions for the benefit of the individual.

The first person to hint at a possible union of emotional and intellectual factors was the eminent American psychologist E. L. Thorndike (1920). Thorndike spoke of three kinds of intelligence: abstract, mechanical, and social. The first two types are well known in assessment and have been validated repeatedly. However, the third kind of intelligence, social intelligence, has proved more elusive. Thorndike defined social intelligence as "the ability to understand and manage people." An essential part of this ability is the accurate recognition of emotions in others. Unfortunately, early attempts to measure social intelligence proved fruitless.

Recently, the idea that emotions and intellect might constitute a single cluster of intertwined abilities has reemerged in the concept of emotional intelligence, as proposed by Mayer, Salovey, and colleagues. The Mayer-Salovey model boasts the strongest theoretical and empirical underpinnings.

Mayer et al. (2008) define **emotional intelligence** as follows:

- Managing emotions so as to attain specific goals;
- Understanding emotions, emotional language, and the signals conveyed by emotions;
- Using emotions to facilitate thinking; and
- Perceiving emotions accurately in oneself and others.

These theorists propose that emotional intelligence is an instance of traditional intelligence, not something different from it. In other words, emotional intelligence (EI) is an important and overlooked subset of abilities that contribute to human efficiency and adaptation. Mayer et al. (2008) propose that emotional intelligence is a third major subdivision that complements the traditional dichotomy of verbal and perceptual abilities.

Because of the subtlety and complexity of the construct, the assessment of emotional intelligence has proved challenging. However, with innovative forms of testing such as embodied in **the MSCEIT or Mayer-Salovey-Caruso Emotional Intelligence Test** (Mayer, Salovey, & Caruso, 2002), progress is being made. This instrument consists of 141 items that yield a total emotional intelligence score as well as two Area scores, four Branch Scores, and eight Task scores. Table 9.10 provides a brief description of the test, which is designed for adults age 17 and older. Normative data are based on a sample of more than 5,000 individuals.

The overall score on the MSCEIT is called the Emotional Intelligence (EI) score. This score is normed to a mean of 100 and standard deviation of 15. The two Area scores (Experiential and Strategic) and the four Branch scores (Perceiving, Facilitating, Understanding, and Managing) likewise are normed to these traditional benchmarks. While scores are provided for the eight Tasks, he test developers caution against overinterpretation of these elemental scores because of their lower reliability. The overall EI score demonstrates strong internal reliability, in the low .90s, whereas the reliability of the two Area scores is slightly lower and more variable, typically in the high .80s (Mayer, Salovey, & Caruso, 2002). Test–retest reliability of the overall score is respectable at .86.

The authors of the MSCEIT propose two different scoring methods: consensus scoring and expert scoring. In consensus scoring, the majority choices of the normative sample are used to identify the correct options. Respondents would receive lower scores to the extent they deviated from this alternative. This method is also called general scoring because the reference point is the general, normative sample. The second approach, expert

scoring, relies on the judgment of experts in the field of emotion to determine the correct options. In particular, the authors used 21 experts attending a conference of the International Society for Research on Emotion. Scoring for this approach relies on the consensus of these experts. Fortunately, the two scoring approaches (general and expert) reveal a very high agreement, on the order of .96 to .98.

The validity of the MSCEIT has been investigated from numerous perspectives, including factorial, discriminant, and predictive validity. Some results indicate that the instrument measures a unitary skill that can be subdivided into the four branches described above. EI as measured by the MSCEIT reveals generally low correlations with verbal intelligence, general intelligence, and major dimensions of personality, that is, the construct provides something that goes beyond established measures.

In addition to the MSCEIT, a few other measures of emotional intelligence have gained recognition. One of these is **the Emotional Competence Inventory** (Sala, 2002), based on Goleman's (1995) conception of emotional intelligence. The Emotional Competence Inventory (ECI) contains 110 items organized into four clusters: (1) Self-Awareness, (2) Social Awareness, (3) Self-Management, and
(4) Social Skills. One appealing feature of this instrument is the 360-degree feedback that it yields. In this method, self-ratings, peer ratings, and supervisor ratings are reported separately for comparison and contrast. The ECI is used mainly in large corporate settings for formative evaluation of employees. The publishers have maintained tight proprietary control over the test, which has limited independent research on its psychometric qualities.

Another widely used test is **the Bar-On Emotional Quotient Inventory** (Bar-On, 2000), which is traditionally known by the acronym EQ-i. This 133-item self-report instrument yields an overall EQ score as well as five composite scores: (1) intrapersonal, (2) interpersonal, (3) adaptability, (4) general mood, and (5) stress management. The test appears to overlap substantially with major personality constructs. For example, a correlation of $r = -.77$ with the anxiety scale from Cattell's 16PF is reported. The EQ-i appears to demonstrate strong reliability, with test–retest reliability of .85 after one month.

ASSESSMENT OF OPTIMISM

Optimism is another fruitful area for psychometric research and assessment. Typically, this construct is viewed as one end of a bipolar continuum, optimism–pessimism. The difference between the two ends of the spectrum is captured in the familiar adage about the glass of water that is half-full to the optimist and half-empty to the pessimist. Carver and Scheier capture why this area of assessment is important: "Optimists are people who expect good things to happen to them; pessimists are people who expect bad things to happen to them.

Optimists and pessimists differ in several ways that have a big impact on their lives. They differ in how they approach problems and challenges they encounter, and they differ in the manner and the success with which they cope with life's difficulties".

In short, optimism and pessimism have to do with people's expectations for the future. Optimists expect a better future than pessimists and generally have more confidence in their ability to manage challenges when they arise. Generally, optimists fare better than pessimists in terms of personal adjustment and even physical health, although the differences for health are not substantial.

The most widely used instrument is **the revised Life Orientation Test** (LOT-R; Scheier, Carver, & Bridges, 1994). This is an intriguingly simple scale that consists of six scored items and four "filler" items (10 items total). Respondents indicate their extent of

agreement with the items on a five-point Likert scale ranging from 1 or "strongly disagree" to 5 or "strongly agree."

Of course, negatively worded items are reverse scored. Responses on the six scored items are then summed to yield a total from 6 (highly pessimistic) to 30 (highly optimistic). Even though "pessimist" and "optimist" are categories in popular language, the LOT-R instead provides a score on a continuum, without strict cut-offs. In large samples of respondents, the score distribution tends to be skewed toward the optimistic side.

Psychometric analyses by Herzberg, Glaesmer, and Hoyer (2006) with huge samples of adults (N = 46,133) reveal that the optimism and pessimism items on the test measure two independent constructs rather than a single, bipolar trait. Many researchers now report three scores from the LOT-R: an optimism score based on the positively worded items, a pessimism score based on the negatively worded items, and a total score that combines the two. The reliability of the instrument is low (Cronbach alphas of .71 for the Optimism items and .68 for the Pessimism items). Thus, the test is recommended for group research only; it is not suitable for clinical practice with individuals.

Steptoe, Wright, Kunz-Ebrecht, and Iliffe (2006) investigated the relationship between LOT-R scores and numerous health behaviors in 128 community-dwelling seniors 65 to 80 years old. Dispositional optimism as measured by the LOT-R total score was associated with many healthful behaviors, including moderate alcohol consumption, not smoking, brisk walking, and vigorous physical activities (women only). The full scale was more consistently associated with these positive relationships.

ASSESSMENT OF GRATITUDE

Gratitude is difficult to define because the concept can be viewed as an attitude, an emotion, a disposition, or a personality trait. A simple definition is that gratitude is a response of thankfulness and joy when receiving a gift. But delving further, difficulties arise. What constitutes a gift? What are the possible sources of a gift? In other words, does gratitude require a personal benefactor, or can it be
expanded to the countless ways in which life pleasantly surprises the mindful person?

Gratitude is universally recognized as a personal virtue because it promotes social cohesion and provides an inner buffer against the toil and pain of everyday life. In general, people with a grateful disposition experience greater well-being than those without this asset. In general, gratitude has received less attention as a topic of measurement than it deserves. But recent efforts are beginning to redress this deficiency.

One such effort is the Gratitude Questionnaire-Six Item Form (GQ-6) developed by McCullough, Emmons, and Tsang, 2002. The GQ-6 is a simple self-report measure of the disposition to experience gratitude. The test consists of the six best items from a longer list of statements that articulate gratitude and appreciation. GQ-6 is based on a Likert-type format with seven alternatives ranging from 1 (strongly disagree) to 7 (strong agree). Two items are stated in the reverse (and therefore reverse scored) as a way of inhibiting response bias.

The authors determined that gratitude reflects intensity (feeling more intensely grateful), frequency (feeling grateful many times a day), span (grateful for many things), and density (grateful to many individuals). Initially, they proposed 39 items to measure these qualities. The GQ-6 is composed of the six best items, as determined by factor-analytic procedures performed with test results from two samples: 238 undergraduates and 1,228 adult volunteers surveyed via the Internet. Reliability of the instrument is good, with coefficient alphas between .82 and .87. Validity of the GQ-6 is based on numerous theory-confirming relationships with other measures.

Additional studies indicated that the GQ-6 is positively related to optimism, hope, spirituality, religiousness, forgiveness, empathy, and prosocial behavior. The scale is negatively related to depression, anxiety, materialism, and envy.

Many reseachers conceive of gratitude as a multidimensional concept. **The Gratitude, Resentment, and Appreciation Test** (GRAT, Watkins, Woodward, Stone, & Kolts, 2003) proposes three dimensions to gratitude: Appreciation of others, expressed as gratitude toward other people, Simple appreciation, expressed as gratitude toward non-social sources and Sense of abundance, expressed as the absence of general resentment.

The 42 items of the GRAT are rated on a 1 to 5 scale (strongly agree to strongly disagree). The test possesses excellent reliability for the three subscales and the total score. It reveals theory-consistent relationships with external criteria such as spirituality and the absence of materialism. Wood, Maltby, Stewart, and Joseph (2007) conducted a factor analysis of the three GRAT subscales and nine other indices of gratitude (including the GQ-6), and found a clear one-factor solution. The 12 measures were highly intercorrelated, indicating a single latent construct which the researchers called gratitude/appreciation. Gratitude is an essential element of human experience that deserves ongoing psychometric inquiry.

SENSE OF HUMOR: SELF-REPORT MEASURES

Humor is a broad construct that has many meanings. Humor can refer to the characteristics of the material (a funny joke or cartoon) or the responses of the individual (a chuckle or belly laugh). Humor can be constructive when it brings people together, or destructive when it is at someone's expense. It is thought that individuals with a "good" sense of humor will more easily befriend others and also will be able to weather the adversities of life with greater balance.

Martin (2003, p. 315) argues that: "One of the challenges of research on humor in the context of positive psychology is to identify which aspects or components of the humor construct are most relevant to mental health and successful adaptation." With this approach, Martin has developed three instruments used widely in humor research: The Coping Humor Scale, the Situational Humor Response Questionnaire, and the Humor Styles Questionnaire.

The Coping Humor Scale was designed to assess the extent to which individuals report using humor to cope with stress (Martin & Lefcourt, 1983). The CHS consists of 7 items similar to "When things are tense I look for something funny to say" or "I think humor is a useful way of coping with problems." These items are rated on a scale from 1 (strongly disagree) to 4 (strongly agree). There is no neutral point on the scale, which forces the respondent to take a position.

The CHS has good test–retest reliability, with $r = .80$ over a 12-week period, but only fair internal consistency, with coefficient alphas of .60 to .70. Regarding validity, Martin summarizes a number of robust external correlates of the test. CHS total scores correlate strongly with the following constructs: Peer ratings of using humor to cope with stress, Peer ratings of not taking one's self too seriously, Researcher ratings of funniness of monologues produced under stress and Researcher ratings of using laughter and humor before dental surgery.

The Situational Humor Response Questionnaire provides a measure of the degree to which the respondent is easily amused and laughs in a wide range of situations (Martin, 1996; Martin & Lefcourt, 1984). The SHRQ consists of 21 items, the first 18 of which describe ordinary life situations such as "You were at a party and the host accidentally spilled a drink on you." Each item is rated on a scale from 1 ("I would not have been particularly amused") to 5 ("I would have laughed heartily").

The last three items refer to laughing and being amused in general. SHRQ reveals adequate psychometric qualities, including test-retest correlations of around .70 and Cronbach alphas in the vicinity of .70 to .85. An interesting validity criterion used in several studies is the correlation of test scores with observed frequency of laughter, with *r*s ranging from .30 to .60. The validity evidence for this instrument includes a wide base of diverse studies, such as correlations with rated funniness of monologues produced by participants, and correlations with other humor scales. Another concern about the test is that the humor situations were designed with college students in mind and may not generalize to other groups. The humor situations date to the 1980s and earlier; some are no longer funny.

Recently, Martin and colleagues have developed a new humor instrument that represents the culmination of decades of research. **The Humor Styles Questionnaire** (HSQ, Martin, Puhlik-Doris, Larsen, Gray, & Weir, 2003) assesses four dimensions that convey individual differences in uses of humor:

- Affiliative: Use of humor to entertain others and facilitate relationships.
- Self-enhancing: Use of humor to cope with stress and uphold a positive outlook during difficult times.
- Aggressive: Use of mocking, manipulative, put-down, or disparaging humor.
- Self-defeating: Use of humor for undue self-disparagement, ingratiation, or defensive reply.

The HSQ includes 32 self-descriptive statements (8 for each subscale) that depict specific uses of humor. The first two styles, Affiliative and Self-enhancing, embody constructive and healthy uses of humor. The last two styles, Aggressive and Self-defeating, involve unhealthy uses of humor that distance the individual from others. For each item, respondents indicate agreement or disagreement on a 7-point scale ranging from 1 (totally disagree) to 7 (totally agree). The HSQ reveals excellent psychometric properties, with strong internal consistencies of the subscales (around .80), and good test–retest reliabilities (.80 to .85). Validity is based on convergent and discriminant correlations of the subscales with appropriate external criteria including well-being, hostility, intimacy, coping, satisfaction with relationships, and major personality variables.