Self-report inventories (SRIs) have been a mainstay of assessment psychology for over seven decades. Researchers and clinicians use them frequently in a broad range of settings and applications. These assessment devices require that the test taker respond to a series of stimuli, the test items, by indicating whether, or to what extent, they describe some aspect of his or her functioning. The response format varies from a dichotomous “true” or “false” to a Likert scale indication of degree of agreement with the statement as a self-description. Some SRIs focus primarily on abnormal functioning or psychopathology, whereas others concentrate more on the normal personality range. Still others cover both normal and abnormal aspects of personality and psychopathology.

In addition to their relative emphasis on normal versus abnormal personality, self-report inventories differ in several notable ways. One important variable is their conceptual basis. Some SRIs are developed with guidance from a particular personality or psychopathology theory or model, whereas others are based more on the results of empirical analyses. In this context, Ben-Porath (1994) noted that personality test developers have pursued their efforts from two broad, non–mutually exclusive perspectives. One approach, the clinical perspective, is designed to produce clinically useful instruments that help detect psychopathology. Self-report inventory developers who follow this approach typically are clinically trained psychologists who focus on conducting applied research. On the other hand, test developers working from the normal personality perspective typically have backgrounds in personality or developmental psychology and often seek to construct measures of normal-range personality constructs that can serve as tools in basic personality research.

The clinical perspective on self-report instrument development has its origins in the psychiatric medical model, in which psychopathology is viewed generally as typological in nature and measures are designed to identify membership in distinct diagnostic classes. In contrast, the normal personality perspective has its origins in the field of differential psychology. Its focus is on personality traits, and dimensional constructs are used to describe meaningful differences among individuals. As just noted, the two perspectives are not mutually exclusive. Tests developed from the clinical perspective have been found to be quite useful in personality research, and normal-range personality inventories are used in a variety of clinical applications. Self-report inventories can also be distinguished in terms of the approaches used to construct and interpret scores on their scales, the methods used to derive standard scores for these scales, and the availability and types of scales and techniques designed to monitor individuals’ test-taking attitude and its impact on scale scores.

This chapter first describes the history and development of SRIs and summarizes early criticisms of this technique. Next, current issues in SRI interpretation are described...
discussed. Finally, directions for future SRI research are outlined. This chapter’s focus is on general issues related to SRI-based assessment of personality and psychopathology. Literature reviews relating to use of specific instruments have been provided by Craig (1999), Dorfman and Hersen (2001), Groth-Marnat (1997), Maruish, (1999), and Strack and Lorr (1994).

EARLY HISTORY

Ben-Porath and Butcher (1991) identified three primary personality assessment techniques and differentiated between them based on the means and sources used for data collection. Behavioral observations include methods in which personality is assessed by systematically recorded observations of an individual’s behavior. Examples include Cattell’s (1965, 1979) T (systematic experimentation) and L (behavioral observation) data. Somatic examinations consist of techniques that rely on some form of physical measurement as the basis for assessing psychological functioning. Examples include various psychophysiological measures (e.g., Keller, Hicks, & Miller, 2000). Verbal examinations rely on verbalizations (oral, written, or a combination of the two) produced by the individual being assessed or another person who presumably knows the assessment target. Self-report inventories, as defined earlier, are a subclass of the verbal examination techniques. Projective assessment techniques (e.g., the Rorschach and Thematic Apperception Test, or TAT) also fall under this definition and are reviewed by Viglione in his chapter in this volume.

Ben-Porath and Butcher (1991) traced the early origins of verbal examinations to an elaborate system of competitive examinations (described in detail by Dubois, 1970) used for over 3000 years to select personnel for the Chinese civil service. Candidates for government positions were tested (and retested every three years) to determine their suitability for these prestigious appointments. Examinees were required to write essays for hours at a time, over a period of several successive days. The essays were used (among other purposes) to gauge the candidates’ character and fitness for office (DuBois, 1970).

In the modern era, Sir Francis Galton was the first to suggest and try out systematic procedures for measuring psychological variables based on verbalizations (as well as some novel approaches to behavioral observations). Influenced heavily by the writings of his cousin, Charles Darwin, Galton was interested in devising precise methods for measuring individual differences in mental traits he believed were the product of evolution. Laying the foundations for quantitative approaches to personality assessment, Galton wrote:

We want lists of facts, every one of which may be separately verified, valued, and revalued, and the whole accurately summed. It is the statistics of each man’s conduct in small everyday affairs, that will probably be found to give the simplest and most precise measure of his character. (Galton, 1884, p. 185)

Most of Galton’s efforts to elicit such information through verbalizations focused on devising various associative tasks. The Dutch scholars Heymans and Wiersma (1906) were the first to devise a questionnaire for the task of personality assessment. They constructed a 90-item rating scale and asked some 3,000 physicians to use the scale to describe people with whom they were well acquainted. Based upon correlations they found among traits that were rated, Heymans and Wiersma, in essence, developed a crude, hierarchical, factor-analytically generated personality model. They proposed that individuals may be described in terms of their standing on eight lower-order traits: Amorphous, Apathetic, Nervous, Sentimental, Sanguine, Phlegmatic, Choleric, and Impassioned. These traits consisted, in turn, of various combinations of three higher-order traits labeled Activity, Emotionality, and Primary versus Secondary Function.

Hoch and Amsden (1913) and Wells (1914) provided further elaboration on the Heymans and Wiersma (1906) model’s utility for personality description and assessment by adding to it various psychopathology symptoms. Their work, in turn, laid the foundations for the first systematic effort to develop a self-report personality questionnaire, Woodworth’s (1920) Personal Data Sheet. Woodworth developed the Personal Data Sheet to assist in identifying psychoneurotic individuals who were unfit for duty in the U.S. military during World War I. This need arose because of the large number of combat personnel who had developed shell shock during the conflict. The questionnaire was to be used as a screening instrument so that recruits who exceeded a certain threshold would be referred for follow-up examinations.

DuBois (1970) reported that Woodworth initially compiled hundreds of “neurotic” items from various sources as candidates for inclusion on his questionnaire. Candidate items were selected if their content was judged to be potentially relevant to identifying neurosis. Items were phrased in question form, and test takers were instructed to answer “yes” or “no” to indicate whether each item described them accurately. Woodworth conducted a series of empirical investigations
and eliminated items answered “yes” by large numbers of normal individuals. The final questionnaire consisted of 116 items. All were keyed such that a “yes” response was an indication of psychoneurosis. Although the Personal Data Sheet was never used for the purposes for which it was constructed—the war had ended by the time it was completed—both its items and Woodworth’s reliance (in part) on empirical analyses for its construction served as the cornerstones for most subsequent self-report personality inventories.

With the conclusion of World War I Woodworth abandoned his test development efforts and refocused his attention on experimental psychology. However, a number of researchers in the then-novel subdiscipline called personality psychology followed in his footsteps. Downey’s (1923) Will-Temperament tests, Travis’s (1925) Diagnostic Character Test, Heidbreder’s (1926) Extraversion-Introversion test, Thurstone’s (1930) Personality Schedule, and Allport’s (1928) Ascendance-Submission measure were among the more prominent early successors to Woodworth’s efforts. Over the next three decades, a substantial literature evaluating the SRI technique’s merits accumulated. Two comprehensive reviews of this methodology reflected the normal personality and clinical perspectives on assessing personality and psychopathology by self-report. Both Allport (1937), adopting a normal personality perspective, and Ellis (1946), from the clinical perspective, noted SRIs’ rapid proliferation, while expressing concern (for somewhat different reasons) about their scientific foundations.

**Allport’s (1937) Critique**

Allport (1937), among the originators of the field of personality psychology, anticipated (correctly) that SRIs would enjoy widespread use in personality research and compared them (somewhat skeptically) with the then more established use of behavioral ratings as a source for quantitative personality data:

> Though less objective than behavioral scales, standardized questionnaires have the merit of sampling a much wider range of behavior, through the medium of the subject’s report on his customary conduct or attitudes in a wide variety of situations. These paper and pencil tests are popular for a number of reasons. For one thing, they are fun to construct and fun to take. Students find them diverting, and teachers accordingly use them as agreeable classroom demonstrations. Furthermore, the scores on the tests can be manipulated in diverse ways, and when the quantitative yield of the coefficients and group differences is complete, everyone has a comforting assurance concerning the “scientific” status of personality.

In considering self-report personality questionnaires’ merits, Allport (1937) identified several limitations that remain salient in current applications of this methodology. One was that “It is a fallacy to assume that all people have the same psychological reasons for their similar responses [to self-report items]” (p. 449). Allport answered this concern by quoting Binet: “Let the items be crude if only there be enough of them. . . . One hopes through sheer length of a series that the erroneous diagnoses will to a certain extent cancel one another, and that a trustworthy residual score will remain” (p. 449).

In describing a second major limitation of personality tests, Allport stated:

> Another severe criticism lies in the ability of the subject to fake the test if he chooses to do so. . . . Anyone by trying can (on paper) simulate introversion, conservatism, or even happiness. And if he thinks he has something to gain, he is quite likely to do so. . . . Even well intentioned subjects may fail insight or slip into systematic error or bias that vitiates the value of their answers.

Thus, Allport listed their transparent nature and susceptibility to intentional and unintentional manipulation among SRI’s major limitations.

In reviewing the major SRIs of his time, Allport (1937) singled out the Bernreuter Personality Inventory (BPI; Bernreuter, 1933). The BPI consisted of 125 items (originating from several previous SRIs including the Personal Data Sheet) phrased as questions with a “yes” “no” or “?” (i.e., cannot say) response format. The items yielded scores on four common personality traits, labeled Dominance, Self-Sufficiency, Introversion, and Neuroticism. Each of the 125 items was scored on all four scales (although some were scored zero), according to empirically derived criteria. For example, if answered “?,” the item “Do you often feel just miserable” was scored –3 on introversion, –1 on dominance, 0 on neuroticism, and 0 on self-sufficiency. Allport (1937) questioned the logic of this approach and recommended instead that items be scored on single scales only.

Finally, Allport (1937) grappled with the question of whether multiscaled SRIs should be designed to measure independent traits or constructs. Commenting on the then-budding practice of factor analyzing scores on multiscale SRIs to derive “independent factors,” Allport noted:

> Unnecessary trouble springs from assuming, as some testers do, that independent factors are to be preferred to inter-dependent traits. What if certain scales do correlate with each other? . . . Each scale may still represent a well-conceived, measurable
common trait. . . . No harm is done by overlap; indeed, overlap is a reasonable expectation in view of the roughness of approximation which is the very nature of the entire procedure (also in view of the tendency of certain traits to cluster). Well-considered scales with some overlap are preferable to ill-conceived scales without overlap. To seek intelligible units is a better psychological goal than to seek independent units. (p. 329)

In summary, viewing SRIs from the normal personality perspective, Allport (1937) raised several important concerns regarding the early successors to Woodworth’s Personal Data Sheet. Recognizing their simplicity of use and consequent appeal, Allport cautioned that SRIs, by necessity, distill human personality to common traits at the expense of a more complete, individually crafted personality description. He emphasized SRIs’ tremendous vulnerability to intentional and unintentional distortion, viewing it as an inherent feature of this methodology. He criticized the BPI’s method of scoring the same item on multiple scales, as well as early factor analysts’ efforts to reduce multiscale instruments such as the BPI to a small number of independent factors. Allport (1937) offered this rather ambivalent concluding appraisal of the nascent area of personality assessment by self-report: “Historically considered the extension of mental measurements into the field of personality is without doubt one of the outstanding events in American psychology during the twentieth century. The movement is still in its accelerating phase, and the swift output of ingenious tests has quite outstripped progress in criticism and theory” (p. 455).

Ellis’s (1946) Review of Personality Questionnaires
Ellis (1946), writing from the clinical perspective, offered a comprehensive appraisal of personality questionnaires near the midpoint of the twentieth century. He opened his critique with the following generalization:

While the reliabilities of personality questionnaires have been notoriously high, their validities have remained more questionable. Indeed some of the most widely known and used paper and pencil personality tests have been cavalierly marketed without any serious attempts on the part of their authors to validate them objectively . . . no real endeavors have been made to show that, when used according to their standard directions, these instruments will actually do the clinical jobs they are supposed to do: meaning, that they will adequately differentiate neurotics from non-neurotics, introverts from extroverts, dominant from submissive persons, and so on. (p. 385)

Ellis’s (1946) opening comments reflected aptly the clinical perspective’s focus on classifying individuals into dichotomous, typological categories. Ellis noted that several authors had preceded him in criticizing SRIs and outlined the following emerging points of concern:

- Most empirical SRI studies have focused on their reliability (which has been established), while ignoring matters of validity.
- SRIs do not provide a whole, organismic picture of human behavior. Although they may accurately portray a group of individuals, they are not useful in individual diagnosis.
- Some questionnaires (like the BPI) that purport to measure several distinct traits are, at best, measuring the same one under two or more names.
- Different individuals interpret the same SRI questions in different ways.
- Most subjects can easily falsify their answers to SRIs and frequently choose to do so.
- SRIs’ “yes/?/no” response format may compromise the scales’ validity.
- Lack of internal consistency may invalidate a questionnaire, but presence of internal consistency does not necessarily validate it.
- SRIs’ vocabulary range may cause misunderstandings by respondents and thus adversely affect validity.
- Testing is an artificial procedure, which has little to do with real-life situations.
- Some personality questionnaires are validated against other questionnaires from which their items were largely taken, thus rendering their validation spurious.
- Even when a respondent does his best to answer questions truthfully, he may lack insight into his true behavior or may unconsciously be quite a different person from the picture of himself he draws on the test.
- Armchair (rather than empirical) construction and evaluation of test items is frequently used in personality questionnaires.
- Uncritical use of statistical procedures with many personality tests adds a spurious reality to data that were none too accurate in the first place.
- Many personality tests that claim to measure the same traits (e.g., introversion-extroversion) have very low intercorrelations with each other.
- There are no statistical shortcuts to the understanding of human nature; such as the ones many test users try to arrive at through involved factorial analyses.

Although generated from a notably different perspective, Ellis’s (1946) concerns overlap substantially with Allport’s
(1937) reservations. The two authors also shared consternation that, in spite of these glaring deficiencies, SRIs had become quite popular: “In spite of the many assaults that have been made against it, the paper and pencil personality test has got along splendidly as far as usage is concerned. For there can be little doubt that Americans have, to date, taken more of the Woodworth-Thurstone-Bernreuter type of questionnaires than all other kinds of personality tests combined” (Ellis, 1946, p. 388).

To explain their seemingly unfounded popularity, Ellis (1946) identified several advantages that their proponents claimed for SRIs:

- They are relatively easy to administer and score.
- Even if the respondent’s self-description is not taken at face value, it may itself provide some clinically meaningful information.
- Although scale scores may be meaningless, examination of individual responses by experienced clinicians may provide valid clinical material.
- Statistical analyses had shown that the traits posited by questionnaires were not simply the product of chance factors.
- Normal and abnormal test takers tended to give different answers to SRI items.
- It does not matter if respondents answer untruthfully on personality questionnaires, since allowances are made for this in standardization or scoring of the tests.
- Traditional methods of validating questionnaires by outside criteria are themselves faulty and invalid; hence, validation by internal consistency alone is perfectly sound.

Having outlined the prevailing pros and cons for personality questionnaires (from a decidedly con-slanted perspective) Ellis (1946) proceeded to conduct a comprehensive, albeit crude, meta-analysis of the literature on personality tests’ validity, differentiating between two methods for validating personality questionnaires. He dubbed one method subjective and described it rather derogatorily as consisting of “checking the test against itself: that is[,] seeing whether respondents answer its questions in a manner showing it to be internally consistent” (p. 390).

Ellis (1946) described the second method, labeled objective personality test validation, as checking a schedule, preferably item by item, against an outside clinical criterion. Thus, a questionnaire may be given to a group of normal individuals and to another group of subjects who have been diagnosed by competent outside observers as neurotic, or maladjusted, or psychotic, delinquent, or introverted. Often, the clinically diagnosed group makes significantly higher neurotic scores than does the normal group, [so] the test under consideration is said to have been validated. (p. 390)

Ellis (1946) questioned whether the subjective method had any bearing on tests’ validity, stating “Internal consistency of a questionnaire demonstrates, at best, that it is a reliable test of something: but that something may still have little or no relation to the clinical diagnosis for which the test presumably has been designed” (p. 391). He also found very limited utility in the objective methods of test validation, citing their sole reliance on questionable validity criteria. Nonetheless, he proceeded to review over 250 published objective validation studies classified into six types based on the method used to generate criterion validity data. Ellis sought to quantify his findings by keeping count of the number of positive, negative, and questionable findings (based on whether these were statistically significant) in each category of studies. Overall, he found positive results in 31 percent of the studies, questionable ones in 17 percent, and negative findings in 52 percent of the publications included in his survey. Ellis (1946) concluded, “Obviously, this is not a very good record for the validity of paper and pencil personality questionnaires” (p. 422).

In selecting studies for inclusion in his analysis, Ellis (1946) singled one instrument out for separate treatment and analysis, the then relatively unknown Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943). Ellis explained that, unlike the more established instruments included in his review, which were administered anonymously by paper and pencil to groups of subjects, the MMPI was administered individually, simulating more accurately a clinical interview. Of the fifteen MMPI studies he reviewed, Ellis reported positive results in ten studies, questionable ones in three, and negative findings in two investigations.

Ellis’s overall conclusions regarding SRIs’ validity were quite negative:

We may conclude, therefore, that judging from the validity studies on group-administered personality questionnaires thus far reported in the literature, there is at best one chance in two that these tests will validly discriminate between groups of adjusted and maladjusted individuals, and there is very little indication that they can be safely used to diagnose individual cases or to give valid estimations of the personality traits of specific respondents. The older, more conventional, and more widely used forms of these tests seem to be, for practical diagnostic purposes, hardly worth the paper on which they are printed. Among the newer questionnaires, the Minnesota Multiphasic schedule appears to be the most promising one—perhaps
because it gets away from group administration which has hitherto been synonymous with personality test-giving. More research in this direction is well warranted at the present time. (1946, p. 425)

Judged with the hindsight of 55 years, Ellis’s critique appears rather naïve and inherently flawed. As has been shown, Ellis himself questioned the utility of the validation methods used in the studies he included in his analyses, noting (correctly) that many, if not most, relied on questionably valid criteria. Given this limitation, these studies could not adequately demonstrate SRIs’ validity or invalidity. Although the tests he reviewed were indeed psychometrically inadequate, Ellis’s effort to appraise them empirically was hampered significantly by limitations in the literature he reviewed. Moreover, his summary dismissal of internal consistency as having little or no bearing on validity was overstated.

Nonetheless, Ellis’s review, published in the prestigious Psychological Bulletin, had a devastating effect on SRIs’ position within the budding field of clinical psychology. Dahlstrom (1992) described the widespread skepticism with which all SRIs were perceived for the ensuing 10 years following this and several similar analyses. Indeed, use of tests (such as the BPI) singled out for their lack of validity waned dramatically in the years that followed. Ellis (1946) did, however, anticipate correctly that the MMPI might emerge as a viable alternative to the SRIs of the first half of the twentieth century.

Ellis (1953) revisited this issue seven years later, in an updated review of personality tests’ validity. He concluded that there had been limited progress in developing valid personality SRIs and focused his criticism on the instruments’ susceptibility to intentional and unintentional distortion. He was particularly concerned with the effects of unconscious defenses. Ellis (1953) again singled out the MMPI as an instrument whose authors had at least made an attempt to correct for these effects on its scale scores, but he expressed skepticism about such corrections’ success. He also observed that the efforts involved in correcting and properly interpreting MMPI scores might better be otherwise invested, stating, “The clinical psychologist who cannot, in the time it now takes a trained worker to administer, score, and interpret a test like the MMPI according to the best recommendations of its authors, get much more pertinent, incisive, and depth-centered ‘personality’ material from a straightforward interview technique would hardly appear to be worth his salt” (p. 48).

Curiously, Ellis (1953) saw no need to subject the preferred “straightforward interview technique” to the type of scrutiny he applied handily to SRIs. That task would be left to Meehl (1956) in his seminal monograph comparing the validity of clinical and actuarial assessment techniques.

Summary of Early History

Self-report inventories emerged as an attractive but scientifically limited approach to personality assessment during the first half of the twentieth century. Representing the normal personality perspective, Allport (1937) criticized these instruments for being inherently narrow in scope and unnecessarily divorced from any personality theory. Ellis (1946), writing from the clinical perspective, concluded that there was little or no empirical evidence of their validity as diagnostic instruments. Both authors identified their susceptibility to intentional and unintentional distortion and the implicit assumption that test items have the same meaning to different individuals as major and inherent weakness of SRIs as personality and psychopathology measures.

CURRENT ISSUES IN SELF-REPORT INVENTORY INTERPRETATION

In spite of their shaky beginnings, SRIs emerged during the second half of the twentieth century as the most widely used and studied method for assessing personality and psychopathology. Modern SRI developers sought to address the limitations of their predecessors in a variety of ways. Various approaches to SRI scale construction are described next, followed by a review of current issues in SRI scale score interpretation. These include the roles of empirical data and item content in interpreting SRI scale scores, methods used to derive standard scores for SRI interpretation, and threats to the validity of individual SRI protocols.

Throughout this section, examples from the SRI literature are cited, and most of these involve either the MMPI or MMPI-2. Emphasis on the MMPI/MMPI-2 reflects this instrument’s central role in the modern literature as the most widely studied (Butcher & Rouse, 1996) and used (Camara, Nathan, & Puente, 2000) SRI.

Approaches to SRI Scale Construction

Burisch (1984) described three primary, non–mutually exclusive approaches that have been used in SRI scale construction. The external approach involves using collateral (i.e., extratest) data to identify items for an SRI scale. Here, individuals are classified into known groups based on criteria that are independent of scale scores (e.g., psychiatric diagnoses) and items are chosen based on their empirical ability to differentiate among members of different groups. The method is sometimes also called empirical keying. Self-report inventory developers who view personality or psychopathology categorically and seek to develop empirical methods for
classifying individuals into predetermined categories typically use the external scale construction method. Often, these categories correspond to diagnostic classes such as schizophrenia or major depression. As would be expected, scale developers who rely on this approach typically assume a clinical perspective on personality assessment.

Ellis (1946) highlighted a major limitation of the external approach in his critique of SRIs as measures of personality and psychopathology. That is, their validity is constrained by the criteria that are used in their development. Absent consensually agreed-upon criteria for classification (a situation not uncommon in psychological assessment, and what typically motivates efforts to develop a scale to begin with), test developers must rely upon imperfect or controversial external criteria for subject classification, item selection, and subsequent cross-validation. Consequently, scales developed with this method have generally not fared well as predictors of the class membership status that they were designed to predict. However, in some instances (e.g., the MMPI clinical scales), subsequent (to their development) empirical research has guided fruitful application of externally developed scales in ways other than those in which their developers intended originally that they be used, by identifying clinically meaningful correlates of these scales and the patterns of scores among them.

Scale developers who follow the inductive approach, according to Burisch (1984), assume that there exists a basic, probably universal personality structure, which they attempt both to discover and to measure. The approach is considered inductive because its adherents do not set out to measure a preconceived set of traits, but instead leave it up to empirical analyses to reveal important personality dimensions and the relations among them. In the process, an SRI is developed to measure the discovered personality structure. Scale developers who apply the inductive approach often adhere to a normal personality perspective on assessment. They typically rely on various forms of factor analysis, and the constructs they identify characteristically are dimensional. A leading example of an inductively derived SRI is Cattell’s 16 Personality Factor Questionnaire (16PF; Cattell, Cattell, & Cattell, 1993). Inductive scale development often follows an iterative process of item writing, data collection, factor analysis, and item revision, followed by subsequent rounds of data collection, analysis, and item modification (e.g., Tellegen, 1982).

Finally, Burisch (1984) describes the deductive approach to personality scale construction as one in which developers start with a conceptually grounded personality model and rationally write or select items that are consonant with their conceptualization. Most early personality and psychopathology SRI developers followed this approach in developing the MMPI precursors so devastatingly criticized by Allport (1937) and Ellis (1946). Consequently, deductive scale construction was viewed for many years as an inferior, less sophisticated form of SRI development. Burisch argued and demonstrated that these seemingly less sophisticated scale development techniques often yield measures that compare quite favorably with products of external and inductive scale construction.

The three approaches to scale construction are not mutually exclusive. Any combination of the three may be used in constructing an SRI scale, or different sets of scales within the same instrument. For example, the MMPI-2 (Butcher et al., 2001) contains three sets of scales, each initially based on a different one of the three approaches to scale construction—the clinical scales, originally (Hathaway, 1956; Hathaway & McKinley, 1940, 1942; McKinley & Hathaway, 1940, 1942, 1944) based on the external method; the Content Scales (Butcher, Graham, Williams, & Ben-Porath, 1990), constructed with a modified deductive approach; and the Personality Psychopathology Five (PSY-5; Harkness, McNulty, & Ben-Porath, 1995), the end product of an inductive research project (Harkness & McNulty, 1994).

Approaches to SRI Scale Score Interpretation

Two general approaches to SRI scale score interpretation can be identified based on their sources for interpretive conclusions. Empirically grounded interpretations rely on empirical data to form the basis for ascribing meaning to SRI scale scores. Content-based interpretations are guided by SRI scales’ item content. Empirically grounded approaches have played a more central role in personality and psychopathology SRIs; however, more recently, content-based interpretation has gained increasing recognition and use. As will be discussed after the two approaches are described, they are not mutually exclusive.

Empirically Grounded Interpretation

Meehl (1945) outlined the basic logic of empirically grounded SRI scale interpretation in his classic article “The Dynamics of ‘Structured’ Personality Inventories.” Responding to early SRI critics’ contention that the instruments are inherently flawed because their interpretation is predicated on the assumption that test takers are motivated, and able, to respond accurately to their items, he stated:

A “self-rating” constitutes an intrinsically interesting and significant bit of verbal behavior, the non-test correlates of which must be discovered by empirical means. Not only is this approach free from the restriction that the subject must be able to describe his own behavior accurately, but a careful study of structured personality tests built on this basis shows that such a restriction
Thus, according to Meehl, empirical interpretation is neither predicated nor dependent on what the test taker says (or thinks he or she is saying) in responding to SRI items, but rather on the empirical correlates of these statements (as summarized in SRI scale scores).

Two subclasses can be distinguished among the empirical approaches to SRI scale interpretation. Scales constructed with the external approach are expected, based on the method used in their construction, to differentiate empirically between members of the groups used in their development. This form of empirical interpretation, which may be termed empirically keyed interpretation, is predicated on the assumption that if members of different groups (e.g., a target group of depressed patients and a comparison sample of nonpatients) answer a set of items differently, individuals who answer these items similarly to target group members likely belong to that group (i.e., they are depressed). This turns out to be a problematic assumption that requires (and often fails to achieve) empirical verification. Consequently, empirically keyed interpretations, as defined here, are used infrequently in current SRI applications.

The second approach to empirical interpretation is predicated on post hoc statistical identification of variables that are correlated with SRI scale scores (i.e., their empirical correlates). The empirical correlate interpretation approach is independent of the method used to develop a scale and may be applied to measures constructed by any (one or combination) of the three methods just outlined. Unlike the empirically keyed approach, it requires no a priori assumptions regarding the implications of one scale construction technique or another. All that is required are relevant extratest data regarding individuals whose SRI scale scores are available. Statistical analyses are conducted to identify variables that are correlated empirically with SRI scale scores; these are their empirical correlates. For example, if a scale score is empirically correlated with extratest indicators of depressive symptoms, individuals who score higher than others on that scale can be described as more likely than others to display depressive symptomatology.

Empirical correlates can guide SRI scale interpretation at two inference levels. The example just given represents a simple, direct inference level. The empirical fact that a scale score is correlated with an extratest depression indicator is used to gauge the depression of an individual who produces a given score on that scale. The correlation between scale and external indicator represents its criterion validity, which in turn reflects the confidence level we should place in an interpretation based on this correlation.

Although the concept of interpreting scale scores based on their criterion validity represents a simple and direct inference level, the process of establishing and understanding SRIs’ criterion validity is complex and challenging. As already noted, the absence of valid criteria often motivates scale development to begin with. In addition, as with any psychological variable, criterion measures themselves are always, to some extent, unreliable. Consequently, validity coefficients, the observed correlations between SRI scale scores and criteria, always underestimate the scales’ criterion validity. If a criterion’s reliability can be reasonably estimated, correction for attenuation due to unreliability is possible to derive a more accurate estimate of criterion validity. However, this is rarely done, and it does not address limitations in criterion validity coefficients imposed by the criterion measures’ imperfect validity. Self-report inventory critics often point to rather low criterion validity coefficients as indications of these instruments’ psychometric weakness, without giving adequate consideration to the limitations just noted.

A second, more complex, and less direct inference level in empirical interpretation of SRI scale scores involves reliance on their construct validity. Cronbach and Meehl (1955) indicated that construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality, which is not “operationally defined” . . . When an investigator believes that no criterion available to him is fully valid, he perforce becomes interested in construct validity because this is the only way to avoid the “infinite frustration” of relating every criterion to some more ultimate standard . . . Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured. (p. 282)

Cronbach and Meehl (1955) described construct validation as an ongoing process of learning (through empirical research) about the nature of psychological constructs that underlie scale scores and using this knowledge to guide and refine their interpretation. They defined the seemingly paradoxical bootstraps effect, whereby a test may be constructed based on a fallible criterion and, through the process of construct validation, that same test winds up having greater validity than the criterion used in its construction. As an example, they cited the MMPI Pd scale, which was developed using an external scale construction approach with the intent that it be used to identify individuals with a psychopathic personality. Cronbach and Meehl (1955) noted that the scale turned out to have a limited degree of criterion validity for this task. However, as its empirical correlates became elucidated through subsequent research, a construct underlying Pd scores emerged that allowed MMPI interpreters to describe individuals who score high on this scale based on both
a broad range of empirical correlates and a conceptual understanding of the Pd construct. The latter allowed for further predictions about likely Pd correlates to be made and tested empirically. These tests, in turn, broadened or sharpened (depending on the research outcome) the scope of the Pd construct and its empirical correlates.

Knowledge of a scale’s construct validity offers a rich, more comprehensive foundation for empirical interpretation than does criterion validity alone. It links the assessment process to theoretical conceptualizations and formulations in a manner described by Cronbach and Meehl (1955) as involving a construct’s nomological network, “the interlocking system of laws which constitute a theory” (p. 290). Thus, empirical research can enhance our understanding of (and ability to interpret) psychological test results by placing them in the context of well-developed and appropriately tested theories.

Whether it is based on criterion or construct validity, or both, empirically grounded SRI interpretation can occur at two levels, focusing either on individual scale scores or on configurations among them. Configural interpretation involves simultaneous consideration of scores on more than one SRI scale. Linear interpretation involves separate, independent consideration and interpretation of each SRI scale score.

Much of the literature on this topic involves the MMPI. The move toward configural MMPI interpretation came on the heels of the test’s failure to meet its developers’ original goal, differential diagnosis of eight primary forms of psychopathology. Clinical experience, bolstered by findings from a series of studies (e.g., Black, 1953; Guthrie, 1952; Halbower, 1955; Hathaway & Meehl, 1951) led MMPI interpreters to conclude that robust empirical correlates for the test were most likely to be found if individuals were classified into types based on the pattern of scores they generated on the test’s clinical scales. Based partly on this development, Meehl’s (1954) treatise on clinical versus actuarial prediction advocated that researchers pursue a three-pronged task: First, they must identify meaningful classes within which individuals tend to cluster. These would replace the inadequate Kraepelinian nosology that served as the target for the MMPI clinical scales’ original development. Next, investigators would need to devise reliable and valid ways of identifying to which class a given individual belongs. Finally, they would identify the empirical correlates of class membership.

In his subsequent call for a so-called cookbook-based interpretation, Meehl (1956) proposed that MMPI profiles could serve all three purposes. Patterns of scores (i.e., configurations) on MMPI clinical scales could be used to identify clinically meaningful and distinct types of individuals; these scores could be used (based on a series of classification rules) to assign individuals to a specific profile type; and empirical research could be conducted to elucidate the correlates of MMPI profile type group membership. Several investigators (most notably Marks and Seeman, 1963, and Gilberstadt and Duker, 1965) followed Meehl’s call and produced such MMPI-based classification and interpretation systems.

Underlying configural scale score interpretation is the assumption that there is something about a combination of scores on SRI scales that is not captured by consideration of each scale score individually (i.e., a linear interpretation) and that the whole is somehow greater than (or at least different from) the sum of its parts. For example, there is something to be learned about an individual who generates his or her most deviant scores on MMPI scales 1 (Hypochondriasis) and 3 (Hysteria) that is not reflected in the individual’s scores on these scales when they are considered separately. Statistically, this amounts to the expectation of an interaction among scale scores in the prediction of relevant extratest data.

Surprisingly, the assumption that configural interpretation should be more valid than linear approaches has not been tested extensively. Goldberg (1965) conducted the most elaborate examination of this question to date. He found that a linear combination of scores on individual MMPI scale scores was more effective than the configural set of classification rules developed by Meehl and Dahlstrom (1960) to differentiate between neurotic and psychotic test takers. The implicit assumption of an interaction among scales that make up the configuration has yet to be extensively tested.

**Content-Based Interpretation**

Content-based SRI interpretation involves reliance on item content to interpret scale scores. For example, if a scale’s items contain a list of depressive symptoms, scores on that scale are interpreted to reflect the individual’s self-reported depression. It is distinguished from deductive SRI scale construction in that the latter involves using item content for scale development, not necessarily interpretation. Indeed, scales constructed by any of the three primary approaches (external, inductive, or deductive) can be interpreted based on their item content, and SRI measures constructed deductively can be interpreted with an empirically grounded approach.

Content-based SRI interpretation predates empirically grounded approaches and was the focus of many aspects of both Allport’s (1937) and Ellis’s (1946) early SRI critiques. Meehl’s (1945) rationale for empirically grounded SRI scale score interpretation was a reaction to the criticism that content-based interpretation was predicated on the dubious assumptions that test items have the same meaning to test takers that they do to scale developers and that all respondents understand items comparably and approach testing in a motivated and cooperative manner. Meehl (1945) agreed (essentially) that such assumptions were necessary for content-based interpretation.
Empirically Grounded Versus Content-Based SRI Interpretation

Empirically grounded and content-based interpretations are not mutually exclusive. Often, scale scores intended for interpretation based on one approach can also be (and are) interpreted based on the other. For example, although the MMPI-2 clinical scales are interpreted primarily based on their empirical correlates, the Harris-Lingoes subscales augment clinical scale interpretation by identifying content areas that may primarily be responsible for elevation on a given clinical scale. Conversely, although their interpretation is guided primarily by item content, the MMPI-2 Content Scales (Butcher et al., 1990) also are interpreted based on their empirical correlates.

A primary distinction between empirically grounded and content-based interpretation is that the latter (as SRI critics have long argued) is more susceptible to intentional and unintentional distortion. However, as is discussed later in detail, appropriate application of SRIs requires that test-taking attitude be measured and considered as part of the interpretation process. Because of its inherent susceptibility to distortion, content-based interpretation requires that an SRI be particularly effective in measuring and identifying misleading approaches and that its users apply tools designed to do so appropriately.

Generating Scores for SRI Interpretation: Standard Score Derivation

Depending upon their response format, SRI raw scores consist either of a count of the number of items answered in the keyed (true or false) direction or a sum of the respondent’s Likert scale ratings on a scale’s items. These scores have no intrinsic meaning. They are a function of arbitrary factors such as the number of items on a scale and the instrument’s response format. Raw scores typically are transformed to some form of standard score that places an individual’s SRI scale raw score in an interpretable context. Standard scores are typically generated by a comparison of an individual’s raw score on a scale to that of a normative reference group(s) composed of the instrument’s standardization or normative sample(s). Because of their critical role in SRI interpretation, it is important to understand how standard scores are derived as well as the factors that determine their adequacy.

The most common standard score used with SRIs is the T score, which expresses an individual’s standing in reference to the standardization sample on a metric having a mean of 50 and a standard deviation of 10. This is typically accomplished through the following transformation:

$$T = \frac{RS - MRS}{SDRS} * 10 + 50,$$

where T is the individual’s T score, RS is his or her raw score, MRS is the standardization sample’s mean score, and SDRS is the sample’s standard deviation on a given SRI scale. A T score of 50 corresponds to the mean level for the standardization sample. A T score equal to 60 indicates that the person’s score falls one standard deviation above the normative
The accuracy and utility of standard scores rest heavily on the nature and quality of the normative reference sample. To the extent that they aptly represent the target population, standard scores will provide an accurate gauge of the individual’s standing on the construct(s) of interest and allow for comparison of an individual’s standing across constructs and, more generally, facilitate configural SRI interpretation. Conversely, if the normative reference scores somehow misrepresent the target population, the resulting standard scores will hinder all of the tasks just mentioned. Several factors must be considered in determining whether a normative reference sample represents the target population appropriately. These involve various types and effects of normative sampling problems.

**Types and Effects of Normative Sampling Problems**

Identifying potential problems with standard scores can be accomplished by considering the general formula for transforming raw score to standard scores:

\[
SS = \frac{RS - MRS}{SDRS} \times NewSD + NewMean,
\]

where \(SS\) is the individual’s standard score, \(RS\) is his or her raw score, \(MRS\) is the standardization sample’s mean score, \(SDRS\) is the sample’s standard deviation on a given SRI scale, \(NewSD\) is the target standard deviation for the standard scores, and \(NewMean\) is the target mean for these scores. As discussed earlier, the target mean and standard deviations are arbitrary, but common practice in SRI scale development is to use \(T\) scores that have a mean of 50 and a standard deviation of 10. An important consideration in evaluating standard scores’ adequacy is the extent to which the normative reference (or standardization) sample appropriately represents the population’s mean and standard deviation on a given SRI scale.

Examination of the general transformation formula shows that if a normative sample’s mean (\(MRS\)) is higher than the actual population mean, the resulting standard score will underestimate the individual’s standing in reference to the normative population. Consider a hypothetical example in which \(T\) scores are used, the individual’s raw score equals 10, the normative sample’s mean equals 12, the actual population mean equals 8, and both the sample and population standard deviations equal 5. Applying the \(T\) score transformation formula provided earlier,

\[
T = \frac{10 - 12}{5} \times 10 + 50 = 46
\]

we find that this normative sample yields a \(T\) score of 46, suggesting that the individual’s raw score falls nearly half a standard deviation below the normative mean on this construct. However, had the sample mean reflected accurately the population mean, applying the \(T\) score transformation formula

\[
T = \frac{10 - 8}{5} \times 10 + 50 = 54
\]

would have yielded a \(T\) score of 54, indicating that the individual’s raw score falls nearly half a standard deviation above the population mean. Larger discrepancies between sample and population mean would, of course, result in even greater underestimates of the individual’s relative standing on the construct(s) of interest. Conversely, to the extent that the sample mean underestimates the population mean, the resulting standard scores will overestimate the individual’s relative position on a given scale.

A second factor that could result in systematic inaccuracies in standard scores is sampling error in the standard deviation. To the extent that the normative sample’s standard deviation underestimates the population standard deviation, the resulting standard score will overestimate the individual’s relative standing on the scale. As we apply again the \(T\) score transformation formula, consider an example in which the individual’s raw score equals 10, the sample and population means both equal 8, and the sample standard deviation equals 2, but the population standard deviation actually equals 5. Applying the formula based on the sample data

\[
T = \frac{10 - 8}{2} \times 10 + 50 = 60
\]

yields a \(T\) score of 60, indicating that the individual’s score falls one standard deviation above the normative mean.
However, an accurate estimate of the population standard deviation

\[ T = \frac{10 - 8}{5} \times 10 + 50 = 54 \]

would have produced a \( T \) score of 54, reflecting a score that falls just under half a standard deviation above the normative mean. Here, too, a larger discrepancy between the sample and population standard deviation results in an even greater overestimation of the individual’s standing on the construct(s) of interest, and, conversely, an overestimation of the population’s standard deviation would result in an underestimation of the individual’s relative score on a given measure.

**Causes of Normative Sample Inadequacies**

In light of their importance in determining standard scores’ adequacy, it is essential to identify (and thus try to avoid) reasons why standardization samples may inaccurately estimate a target population’s mean or standard deviation on an SRI scale. Three general types of problems may generate inaccurate normative means and standard deviations: sampling problems, population changes, and application changes. Two types of sampling problems error may occur. The simplest among these is random sampling error, in which, as a result of random factors associated with the sampling process, the normative sample mean or standard deviation fails to represent accurately the relevant population statistics. This can be minimized effectively by collecting sufficiently large normative samples.

Systematic sampling errors occur when, due to specific sampling flaws, the sample mean or standard deviation reflects inaccurately the relevant population statistics. In such cases a normative sample fails to represent accurately one or more segments of the target population as a result of sampling bias. This will negatively affect the normative sample’s adequacy if two conditions are met: (1) a sample fails to represent accurately a certain population segment, and (2) the inadequately represented population segment differs systematically (in its mean, its standard deviation, or both) from the remaining population on a particular SRI scale. For example, if as a consequence of the sampling method used younger adults are underrepresented in a normative sample that is designed to represent the entire adult population, and younger adults differ systematically from the remaining adult population on the scale being standardized, this could result in biased estimates of both the population mean and its standard deviation. This might occur with a scale designed to measure depression, a variable that tends to vary as a function of age.

If younger adults are represented inadequately in a normative sample (this could occur if the sampling process failed to incorporate college students and military personnel) used to develop standard scores on a depression scale, the normative sample would overestimate the population mean on the scale and underestimate its standard deviation, resulting in the effects discussed previously.

Note that in order for systematic sampling error to affect scale norms, both conditions just specified must be met. That is, a population segment must be misrepresented and this segment must differ systematically from the remaining population on the scale being standardized. If only the first condition is met, but the misrepresented segment does not differ systematically from the remaining population, this will not result in biased estimates of the population mean and standard deviation. Such a scenario occurred with the updated normative sample used to standardize the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). Data included in the MMPI-2 manual indicated that the new normative sample differed substantially from the general adult population in education. Specifically, the normative sample significantly underrepresented individuals with lower levels of education and overrepresented people with higher levels of education in the general adult population. Some authors (e.g., Duckworth, 1991) expressed concern that this may introduce systematic bias in the updated norms. However, subsequent analyses demonstrated that this sampling bias had no significant impact on resulting test norms because education is not correlated substantially with MMPI scale scores (Schinka & LaLone, 1997).

Population changes are a second reason why normative samples may inadequately represent their target population. These occur when, over the course of time, the target population changes on the construct that a scale measures. For example, Anastasi (1985), in a review of longitudinal research on intelligence, found a trend for population-wide increases in intelligence over the first half of the twentieth century. These were the result primarily of increases in population education levels in general and literacy levels in particular. To account for these changes’ effects on their norms, it has been necessary for intelligence test developers to periodically collect new normative data. To the extent that constructs measured by SRIs are affected similarly by population changes, it becomes necessary to update their normative databases as well. This was one of the considerations that led to the development of new norms for the MMPI (Butcher et al., 1989).

Application changes are a third reason why normative samples may misrepresent target populations. Two types of application changes can be distinguished. Changes in administration practices may affect normative data adequacy. For
example, the original MMPI normative data were collected using the test’s so-called Box Form. Each of the test’s items was typed on a separate card and test takers were instructed to sort the cards (which were presented individually in a random order) into three boxes representing a “true,” “false,” or “cannot say” response. The instructions given to the original normative sample did not discourage the “cannot say” option. However, after the normative data were collected, an administration change was introduced and test takers were instructed to “be sure to put less than 10 cards behind the ‘cannot say’.” (Dahlstrom, Welsh, & Dahlstrom, 1972, p. 32). Later still, the Box Form was largely superseded by the MMPI Group Form, a booklet that presented the test’s items in a fixed order and required that the test taker record his or her responses (true, false, or cannot say) on an answer sheet. Here, too, test takers were admonished to attempt to answer all of the test items.

To the extent that either of these changes in administration practices affected individuals’ responses to the test items, this could have resulted in the original normative sample data’s misrepresenting population statistics under the revised administration procedures. In fact, the updated MMPI-2 norms are significantly different from the original norms in both their means and standard deviations, and to some extent these shifts are a product of the administration changes just described. For example, as a result of the change in instructions regarding the “cannot say” option, the new normative sample members omitted far fewer items than did their original counterparts, which in turn probably contributed to the new sample’s higher mean raw scores on many of the test’s original scales. In other words, the original normative sample underestimated the target population’s mean raw scores on the MMPI scales given the shift in administration procedure, contributing partly to the artificially elevated T scores generated by individuals and groups tested with the original MMPI when they were transformed to standard scores based on the original test norms.

A more recent change in SRI administration practices followed the introduction of computer technology. Although most SRI norms were collected using booklet forms, software is now available to administer most tests by computer. Such a change in administration practice could also, potentially, affect these instruments’ norms’ adequacy if the different administration format resulted in a systematic change in responses to SRI items. Reassuringly, a recent meta-analysis by Finger and Ones (1999) demonstrated that computerized test administration does not affect group means or standard deviations (and thus would have no negative impact on the test’s norms) on MMPI/MMPI-2 scales. Butcher, in his chapter in this volume, provides further discussion of computer applications in psychological assessment.

A second type of application change that could potentially affect norms’ adequacy involves expansion of the target population. When an SRI developed for use with a rather narrowly defined population is considered for application to a broader population, the possibility that its norms will no longer accurately reflect the expanded population’s means and standard deviations on its scales needs to be considered. For example, the MMPI was developed originally for use at the University of Minnesota Hospital, and its normative sample, made up primarily of a group of Caucasian farmers and laborers with an average of eight years of education, represented fairly well this target population. As the test’s use expanded to the broader U.S. population, concerns were raised (e.g., Gynther, 1972) about the MMPI norms’ adequacy for interpreting scores generated by minorities, primarily African Americans, who were not included in the original normative sample.

The effects of expanding an SRI’s population on its normative sample’s adequacy depend upon the new population segment’s performance on its scales. To the extent the new segment differs systematically from the original on a scale’s mean or standard deviation, this would necessitate an expansion of the instrument’s normative sample to reflect more accurately the expanded population’s scale parameters. This was one of the primary considerations that led to the collection of new normative data for the MMPI and publication of the MMPI-2 (Butcher et al., 1989). Similarly, as the test’s use has expanded beyond the United States to other countries, cultures, and languages, researchers throughout the world have collected new normative data for MMPI and later MMPI-2 application in an ever-increasing number of countries (c.f., Butcher, 1996; Butcher & Pancheri, 1976).

The effects of expanding an SRI’s target population on normative data adequacy should not be confused with questions about an instrument’s validity across population segments, although frequently these very separate concerns are confounded in the literature. Ensuring that various population segments are represented adequately in an SRI’s normative sample is not sufficient to guarantee that the test is as valid an indicator of its target psychological constructs in the new segment as it was in the original. To the extent that an instrument’s interpretation is predicated on an SRI’s empirical correlates, its construct validity, or the combination of the two (as discussed earlier), its application to the expanded population is predicated on the assumption that these test attributes apply comparably to the new population segment.

**General Population Versus Population Subsegment Norms**

A final consideration in evaluating normative data adequacy is whether an SRI’s standard scores are derived from general
or more narrowly defined and specific normative samples. When a general population normative sample is used, the same set of standard scores is applied regardless of the assessment setting or the individual’s membership in any specific population subsegments. Thus, for example, the MMPI-2 has just one set of standard scores generated based on a normative sample designed to represent the general U.S. population. The same set of standard scores is used regardless of where the test is applied.

A more recently developed SRI, the Personality Assessment Inventory (PAI; Morey, 1991) provides two sets of norms for its scales, based on a sample of community-dwelling adults and a clinical sample. Morey (1991) explains that clinical norms are designed to assist the interpreter in tasks such as diagnosis:

For example, nearly all patients report depression at their initial evaluation; the question confronting the clinician considering a diagnosis of major depression is one of relative severity of symptomatology. That a patient’s score on the PAI DEP [Depression] scale is elevated in comparison to the standardization sample is of value, but a comparison of the elevation relative to a clinical population may be more critical in formulating diagnostic hypotheses. (p. 11)

The ability to know how an individual compares with others known to have significant psychological problems may indeed contribute useful information to test interpretation. However, this particular approach to generating such information has a significant drawback. If, using Morey’s example, nearly all members of a clinical reference sample report depression when their normative data are collected, then a typical patient experiencing significant problems with depression will produce a nondeviant score on the instrument’s clinically referenced depression measure, thus obscuring depression’s prominence in the presenting clinical picture.

A similar problem results when SRI scales are normed based on other narrowly defined, setting-specific population segments. For example, Roberts, Thompson, and Johnson (1999) developed several additional sets of PAI norms for use in assessing applicants for public safety positions. The additional reference samples were all made up of public safety job applicants. A feature common among individuals undergoing evaluations for possible employment in public safety positions is the tendency to deny or minimize any behavioral and emotional problems that they believe may cast them in a negative light and reduce the likelihood that they will be offered the position they seek. As a result, most individuals tested under these circumstances tend to score higher than the general population on defensive test-taking measures. Deviant scores on defensiveness scales alert the interpreter that the test taker is probably minimizing or denying such problems. However, when compared with other individuals tested under similar circumstances, public safety position applicants produce nondeviant scores on defensiveness measures when they are in fact approaching the assessment with a defensive attitude. Here, too, narrowly defined norms may obscure an important feature (defensiveness) of a test taker.

Threats to SRI Protocol Validity

The impact of test-taking approaches on SRIs has long been the focus of heated debate. As reviewed earlier in this chapter, early SRI critics (e.g., Allport, 1937; Ellis, 1946) cited their vulnerability to intentional and unintentional distortion by the test taker as SRIs’ primary, inherent limitation. The basic concern here is that, even if he or she is responding to a psychometrically sound SRI, an individual test taker may, for a variety of reasons, approach the assessment in a manner that compromises the instrument’s ability to gauge accurately his or her standing on the construct(s) of interest. In such cases, a psychometrically valid test may yield invalid results.

Use of the term validity to refer to both a test’s psychometric properties and an individual’s test scores can be confusing. A distinction should be drawn between instrument validity and protocol validity. Instrument validity refers to a test’s psychometric properties and is typically characterized in terms of content, criterion, and construct validity. Protocol validity refers to the results of an individual test administration. Use of the term validity to refer to these two very different aspects of SRI assessment is unfortunate, but sufficiently well grounded in practice that introduction of new terminology at this point is unlikely to succeed.

A need to distinguish between psychometric and protocol validity has been highlighted in a debate regarding the widely studied NEO Personality Inventory-Revised (NEO-PI-R). Responding to suggestions by Costa and McCrae (1992a; the NEO-PI-R developers) that practitioners use this test in clinical assessment, Ben-Porath and Waller (1992) expressed the concern (among others) that the absence of protocol validity indicators on the NEO-PI-R may limit the instrument’s clinical utility. Costa and McCrae (1992b) responded that validity scales were unnecessary, in part because evidence has shown that test scores may be psychometrically valid even in instances in which validity indicators showed evidence of limited protocol validity.

Most recently, Piedmont, McCrae, Riemann, and Angleitner (2000) sought to demonstrate this point by showing that scores on an SRI’s validity scales (designed to assess protocol validity) were unrelated to the NEO-PI-R’s psychometric validity. However, their analyses were based on data generated by research
volunteers who completed the instruments anonymously. Thus, unlike respondents in most clinical assessment settings, these research volunteers had nothing at stake when responding to the NEO-PI-R. In contrast, as reviewed next, test takers in clinical settings may be motivated by various factors to present themselves in a particular manner. Moreover, psychometric validity in this and similar studies was established based on statistical analyses of group data, whereas protocol validity pertains to individual test results. If, for example, one of the participants in such a study marked his or her answer sheet randomly, without actually reading the SRI items, his or her resulting scale scores are completely invalid and uninterpretable, regardless of how others in the sample responded.

Consideration of protocol validity is one aspect of SRI-based assessment in which users are able to take an individualized perspective on a generally normative enterprise. Allport (1937) distinguished between idiographic (individualized) and nomothetic (generalized) approaches to personality research and assessment. Drawing an analogy to the diagnostic process in medicine, he noted that the two approaches are not mutually exclusive. Rather, a combined idio- nomothetic approach is likely to yield the optimal perspective on diagnosis and assessment. Consideration of protocol validity offers an important window into idiographic aspects of SRI-based assessment.

In sum, instrument validity is necessary but insufficient to guarantee protocol validity. Although it sets the upper limit on protocol validity, information regarding instrument validity does not address a critical question that is at issue in every clinical assessment: Is there anything about an individual’s approach to a particular assessment that might compromise its user’s ability to interpret an SRI’s scores? To answer this question, users must be aware of various threats to protocol validity.

**Types of Threats to Protocol Validity**

Threats to SRI protocol validity need to be considered in each SRI application because of their potential to distort the resulting test scores. This information can be used in two important ways. First, knowledge of threats to protocol validity makes it possible for test users to attempt to prevent or minimize their occurrence. Second, it makes it possible to anticipate invalid responding’s potential impact on the resulting test scores and, on the basis of this information, provide appropriate caveats in test interpretation. Such statements may range from a call for caution in assuming that an interpretation will likely reflect accurately the individual’s standing on the construct(s) of interest to an unambiguous declaration that protocol validity has been compromised to a degree that makes it impossible to draw any valid inferences about the test taker from the resulting SRI scale scores.

Threats to protocol validity fall broadly into two categories that reflect test item content’s role in the invalid responding. Important distinctions can be made within each of these categories as well. Table 24.1 provides a list of the various non-content- and content-based threats to protocol validity identified in this chapter.

**Non-Content-Based Invalid Responding.** Non-content-based invalid responding occurs when the test taker’s answers to an SRI are not based on an accurate reading, processing, and comprehension of the test items. Its deleterious effects on protocol validity are obvious: To the extent that a test taker’s responses do not reflect his or her actual reactions to an SRI’s items, then those responses cannot possibly gauge the individual’s standing on the construct of interest. This invalidating test-taking approach can be divided further into three modes: nonresponding, random responding, and fixed responding.

**Nonresponding** occurs when the test taker fails to provide a usable response to an SRI item. Typically, this takes the form of failing to provide any response to an SRI item, but it may also occur if the test taker provides more than one response to an item. Nonresponding may occur for a variety of reasons. Test takers who are uncooperative or defensive may fail to respond to a large number of an SRI’s items. Less insidious reasons why individuals may fail to respond appropriately to a SRI may include an inability to read or understand its items, cognitive functioning deficits that result in confusion or obsessiveness, or limits in the test taker’s capacity for introspection and insight.

Nonresponding’s effect on protocol validity depends, in part, on the SRI’s response format. In tests that use a “true” “false” response format, a nonresponse is treated typically as a response in the nonkeyed direction. In SRIs with a Likert scale response format, a nonresponse typically receives the

<table>
<thead>
<tr>
<th>TABLE 24.1 Threats to Self-Report Inventory Protocol Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-content-based invalid responding</td>
</tr>
<tr>
<td>Nonresponding</td>
</tr>
<tr>
<td>Random responding</td>
</tr>
<tr>
<td>Intentional random responding</td>
</tr>
<tr>
<td>Unintentional random responding</td>
</tr>
<tr>
<td>Fixed responding</td>
</tr>
<tr>
<td>Content-based invalid responding</td>
</tr>
<tr>
<td>Overreporting</td>
</tr>
<tr>
<td>Intentional overreporting</td>
</tr>
<tr>
<td>Exaggeration versus fabrication</td>
</tr>
<tr>
<td>Unintentional overreporting (negative emotionality)</td>
</tr>
<tr>
<td>Underreporting</td>
</tr>
<tr>
<td>Intentional underreporting</td>
</tr>
<tr>
<td>Minimization versus denial</td>
</tr>
<tr>
<td>Unintentional underreporting (social desirability)</td>
</tr>
</tbody>
</table>
value zero. These ipso facto scores can by no means be assumed to provide a reasonable approximation of how the respondent would have answered had he or she chosen or been able to do so. Therefore, to the extent that nonresponding occurs in a given SRI protocol, this will distort the resulting test scores. For example, in a true/false response format a respondent’s failure to respond appropriately to a large number of items will result in artificial deflation of his or her scores on the instrument’s scales, which, if not identified and considered in scale score interpretation, may result in underestimation of the individual’s standing on the constructs measured by the affected scales.

Random responding is a test-taking approach characterized by an unsystematic response pattern that is not based on an accurate reading, processing, and understanding of an SRI’s items. It is not a dichotomous phenomenon, meaning that random responding may be present to varying degrees in a given test protocol. Two types of random responding can be distinguished. Intentional random responding occurs when the individual has the capacity to respond relevantly to an SRI’s items but chooses instead to respond irrelevantly in an unsystematic manner. An uncooperative test taker who is unwilling to participate meaningfully in an assessment may engage in intentional random responding rather than becoming embroiled in a confrontation with the examiner over his or her refusal to participate. In this example, the test taker provides answers to an SRI’s items without pausing to read and consider them. He or she may do this throughout the test protocol or at various points along the way in responding to an SRI’s items.

Unintentional random responding occurs when the individual lacks the capacity to respond relevantly to an SRI’s items, but, rather than refraining from giving any response to the items, he or she responds without having an accurate understanding of the test items. Often these individuals are not aware that they lack this capacity and have failed to understand and respond relevantly to an SRI’s items.

Several factors may lead to unintentional random responding. Reading difficulties may compromise the test taker’s ability to respond relevantly to an SRI’s items. Most current SRIs require anywhere from a fourth- to a sixth-grade reading level for the test taker to be able to read, comprehend, and respond relevantly to the items. Regrettably, this is not synonymous with having completed four to six years of education. Some high school graduates cannot read at the fourth grade level. If the examiner has doubts about a test taker’s reading ability, a standardized reading test should be administered to determine his or her reading level. For individuals who do not have the requisite reading skills, it may still be possible to administer the test if the problem is strictly one of literacy rather than language comprehension. In such cases, an SRI’s items can be administered orally, preferably using standard stimulus materials such as an audiotaped reading of the test items.

Comprehension deficits can also lead to random responding. In this case the individual may actually be able to read the test items but does not have the necessary language comprehension skills to process and understand them. This could be a product of low verbal abilities. In other instances, comprehension deficits may be found in those lacking familiarity with English language nuances, for example, among individuals for whom English is not their primary language.

Unintentional random responding can also result from confusion and thought disorganization. In some instances, these types of difficulties may have prompted the assessment and SRI administration. Whereas reading and comprehension difficulties tend to be relatively stable test-taker characteristics that will probably compromise protocol validity regardless of when an SRI is administered, confusion and thought disorganization are often (although not always) transitory conditions. If and when the individual’s sensorium clears, she or he may be able to retake an SRI and provide valid responses to its items.

Finally, random responding may result from response recording errors. Many SRIs are administered by having the respondent read a set of items from a booklet and record the responses on a separate answer sheet. If the respondent marks his or her answer to an SRI’s items in the wrong location on an answer sheet, he or she is essentially providing random responses. This could result from the test taker’s missing just one item on the answer sheet or from an overall careless approach to response recording.

Fixed responding is a non-content-based invalidating test-taking approach characterized by a systematic response pattern that is not based on an accurate reading, processing, and understanding of an SRI’s items. In contrast to random responding, here the test taker provides the same non-content-based responses to SRI items. If responding to a true/false format SRI, the test taker indiscriminately marks many of the test items either “true” or “false.” Note that if the test taker provides both “true” and “false” responses indiscriminately, then he or she is engaging in random responding. In fixed responding the indiscriminant responses are predominantly either “true” or “false.” In fixed responding on a Likert scale, the test taker marks items at the same level on the Likert rating scale without properly considering their content. Like nonresponding and random responding, fixed responding is a matter of degree rather than a dichotomous all-or-none phenomenon.

Unlike nonresponding and random responding, fixed responding has received a great deal of attention in the
SRI-based assessment literature. Jackson and Messick (1962) sparked this discussion when they proposed that much (if not all) of the variance in MMPI scale scores was attributable to two response styles, termed acquiescence and social desirability. Acquiescence was defined as a tendency to respond “true” to MMPI items without consideration of their content. This type of non-content-based responding is labeled fixed responding in this chapter.

A detailed examination of Jackson and Messick’s arguments and the data they analyzed in its support is beyond the scope of this chapter. Essentially, Jackson and Messick factor analyzed MMPI scale scores in a broad range of samples and found recurrently that two factors accounted for much of the variance in these scores. They attributed variance on these factors to two response styles, acquiescence and social desirability, and cautioned that MMPI scale scores appear primarily to reflect individual differences on these nonsubstantive dimensions. They suggested that MMPI scales were particularly vulnerable to the effects of acquiescence and its counterpart, counteracquiescence (a tendency to respond “false” to self-report items without consideration of their content), because their scoring keys were unbalanced. That is, for some MMPI scales many, if not most, of the items were keyed “true,” whereas on other scales most of the items were keyed “false.”

In an extensive and sophisticated series of analyses, Block (1965) demonstrated that the two primary MMPI factors reflected substantive personality dimensions rather than stylistic response tendencies. With regard specifically to acquiescence, he showed that completely balanced MMPI scales (i.e., ones with equal numbers of “true” and “false” keyed items) yielded the same factor structure that Jackson and Messick (1962) attributed to the effect of response styles. He showed further that the so-called acquiescence factor was correlated with substantive aspects of personality functioning. Block (1965) labeled this factor ego control and demonstrated that its association with extratest data was unchanged as a function of whether it was measured with balanced or unbalanced scales.

It is important to note that Block’s analyses did not indicate that acquiescence is never a problem in SRI-based assessment. In the relatively rare instances when they occur, acquiescence and counteracquiescence can indeed jeopardize protocol validity. In the most extreme case of acquiescence, if a respondent answers “true” to all of a scale’s items without reference to their content, his or her score on that scale is obviously invalid. In addition, use of a Likert scale format does not obviate the potential effects of this response style, because with this format, as well, it is possible for test takers to provide a fixed response that is independent of item content.

Block’s compelling demonstration notwithstanding, Jackson and Messick and their followers continued to advocate the response style position and argue that acquiescence represented a serious challenge to MMPI use and interpretation. Most recently, Helmes and Reddon (1993) revisited this issue and criticized the MMPI and MMPI-2 (among other things) for their continued susceptibility to the effects of acquiescence. These authors again identified the test’s unbalanced scoring keys as a primary reason for its susceptibility to acquiescence. In constructing his own SRI, the Basic Personality Inventory (BPI), Jackson (1989) indeed adopted the balanced scoring key solution for its scales, each of which is made up of 20 items, half keyed “true” and the others keyed “false.” However, balanced scoring keys actually provide no protection whatsoever against the protocol invalidating effects of fixed responding. Consider the hypothetical example just mentioned, in which a test taker responds “true” to all 20 BPI scale items without actually referring to their content. The only effect a balanced key might have in this instance might be to instill a false sense of security in the test interpreter that the scale is not susceptible to the protocol invalidating effects of acquiescence, when, in fact, it is.

In summary, although fixed responding does not pose as broad a threat to protocol validity as Jackson and Messick would argue, in cases in which a test taker uses this response style extensively, the resulting SRI scale scores will be invalid and uninterpretable. Constructing scales with balanced keys or Likert scale response formats does not make an SRI less susceptible to this threat to protocol validity. Self-report inventory users need to determine in each instance that a test is used whether, to what extent, and with what impact fixed responding may have compromised protocol validity. This requires that the SRI s include measures of fixed responding.

**Content-Based Invalid Responding.** Content-based invalid responding occurs when the test taker skews his or her answers to SRI items and, as a result, creates a misleading impression. This test-taking approach falls broadly into two classes that have been discussed under various labels in the literature. The first of these has been termed alternatively over-reporting, faking bad, and malingering. The second type of content-based invalid responding has been labeled underreporting, faking good, and positive malingering. In this chapter, they will be discussed under the more neutral labels of over- and underreporting.

Overreporting occurs when, in responding to an SRI, a test taker describes him- or herself as having more serious difficulties, a greater number of them, or both than he or she actually has. Underlying this definition is the hypothetical notion that if a completely objective measure of psychological functioning
was available, the overreporter’s subjective self-report would indicate greater dysfunction than does the objective indicator. Two non–mutually exclusive types of overreporting can be distinguished. *Intentional overreporting* occurs when the individual knowingly skews his or her self-report. This test taker is typically motivated by some instrumental gain and thus fits the *DSM-IV* definition of *malinger*ing (APA, 2000). The label *faking bad* also carries with it a connotation of volitional distortion and similarly falls under the category of intentional overreporting.

It is important to note that intentional overreporting is not in itself an indication that psychopathology is absent. That is to say, if an individual intentionally overreports in responding to a SRI, that, in itself, does not indicate that he or she is actually free of bona fide psychological dysfunction. It is, in fact, possible for someone who has genuine psychological difficulties to amplify their extent or significance when responding to SRI items. On the other hand, some people who intentionally overreport in response to an SRI actually have no problems. The distinction here is between *exaggeration* and *fabrication* of difficulties. Both forms of intentional overreporting fall under the *DSM-IV* definition of malingering: “the intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives such as avoiding military duty, avoiding work, obtaining financial compensation, evading criminal prosecution, or obtaining drugs” (APA, 2000, p. 739). In practice, distinguishing between exaggeration and fabrication in SRI protocol validity determination is quite challenging.

In *unintentional overreporting*, the test taker is unaware that she or he is deviating from a hypothetically objective self-description and describing her- or himself in an overly negative manner. Here, it is the test taker’s self-concept that is skewed. Individuals who engage in this test-taking approach believe mistakenly that they are providing an accurate self-description when in fact they are overreporting their difficulties.

Tellegen (1985) has described a primary personality trait, *negative emotionality*, which predisposes individuals to perceive their environment as more threatening than it is in reality, and themselves as having greater subjective difficulty functioning than they actually have. Individuals high in negative emotionality do indeed experience psychological dysfunction; however, they overestimate, and as a result, overreport its extent and significance. As a consequence, they produce deviant scores on SRIs that confound genuine with unintentionally overreported dysfunction.

Demonstrating and evaluating the extent of the confound between genuine and unintentionally overreported psychological dysfunction is quite challenging because of the inherent difficulty in obtaining objective indicators of functioning. Just about any effort to derive an objective measure of psychological functioning relies, at least to some extent, on self-report or self-presentation. Structured diagnostic interviews and even informant reports are influenced by how an individual responds to specific interview questions (asked in person by an interviewer rather than impersonally by a questionnaire) or the impression a person creates on others who are asked to describe her or his psychological functioning.

Watson and Pennebaker (1989) provided a compelling illustration of this phenomenon by focusing on the role negative emotionality plays in assessing physical functioning. Unlike psychological functioning, in assessing physical health it is possible to obtain objective indicators of dysfunction that are independent of self-report. These investigators examined the relation between self-reported negative emotionality, self-reported health complaints, and objectively derived physical functioning indicators (e.g., fitness and lifestyle variables; frequency of illness; health-related visits or absences; objective evidence of risk, dysfunction, or pathology; and overall mortality). They found a consistent correlation between negative emotionality and self-reported health problems, but little or no correlation between self-reported negative emotionality and objective health indicators. The unintentional overreporting associated with negative emotionality accounted almost entirely for its relation with physical health complaints, leading the investigators to conclude that there was little or no association between this construct and actual physical health.

Negative emotionality’s role in assessing mental health and personality functioning is more complex. People high in negative emotionality are genuinely psychologically distressed, and their difficulties are often manifested in multiple areas of psychological dysfunction. In diagnostic terms, this results in substantial levels of psychopathology comorbidity. Mineka, Watson, and Clark (1998) reported, for example, that anxiety and mood disorders have approximately a 50% rate of co-occurrence. Similar levels of comorbidity have been reported among other Axis I diagnostic categories, among Axis II diagnoses, and across Axis I and Axis II. Although diagnostic comorbidity is real, unintentional overreporting associated with negative emotionality probably inflates estimates of its extent. In SRI measures of personality and psychopathology, this inflation has the effect of yielding deviant scores on multiple scales. It also results in phenotypic correlations among SRI scales that overestimate the actual correlations among the latent constructs they are designed to measure. When correlations among SRI scales are factor analyzed, they yield typically one very strong general factor that represents both the genuine psychological sequela of negative
emotionality (i.e., true phenotypic comorbidity) and the confounding effects of unintentional overreporting.

In summary, overreporting in response to SRI items results in scale scores that overestimate the extent or significance of psychological problems the respondent experiences. If overreporting is suspected, the test interpreter is confronted with the challenge of determining whether, and to what extent, it might involve intentional distortion versus manifestations of negative emotionality and, if it is intentional, whether it involves fabrication or exaggeration of problems. Moreover, these threats to protocol validity are not mutually exclusive, and the interpreter needs to consider the possibility that some or all may be manifested in a given protocol.

Underreporting occurs when in responding to an SRI a test taker describes him- or herself as having less serious difficulties, a smaller number of difficulties, or both than he or she actually has. To refer back to the hypothetical objective functioning indicator, in underreporting the individual’s self-report reflects better functioning than would be indicated by an objective assessment. Here, too, a distinction may be drawn between intentional and unintentional underreporting. In intentional underreporting, the individual knowingly denies or minimizes the extent of his or her psychological difficulties or negative characteristics. As a result, the individual’s SRI scale scores underestimate his or her level of dysfunction. Differentiation between denial and minimization is important but complex. The distinction here is between an individual who blatantly denies problems that he or she knows exist and one who may acknowledge some difficulties or negative characteristics but minimizes their impact or extent.

Unintentional underreporting occurs when the individual unknowingly denies or minimizes the extent of his or her psychological difficulties or negative characteristics. Here, too, objective and subjective indicators of psychological functioning would be at odds; however, in unintentional underreporting this discrepancy results from the individual’s self-misperception rather than an intentional effort to produce misleading test results.

Much of the discussion of this topic in the assessment literature has appeared under the label social desirability. Edwards (1957) defined social desirability as “the tendency of subjects to attribute to themselves, in self-description, personality statements with socially desirable scale values and to reject those socially undesirable scale values” (p. vi). As was the case with acquiescence (discussed earlier), social desirability was proposed as a response style, “an organized disposition within individuals to respond in a consistent manner across a variety of substantive domains” (Wiggins, 1973). Edwards (1970) differentiated between social desirability and what he called “impression management,” a deliberate attempt to lie or dissimulate for ulterior motives, that is, intentional underreporting as defined in this chapter. Thus, as conceptualized by Edwards (1957, 1970), social desirability was a form of unintentional underreporting in response to SRI items.

Edwards (1957) argued that much of the variance in MMPI scale scores could be attributed to social desirability. He based this conclusion on research he did with an MMPI scale he constructed and labeled social desirability. The scale was made up of 39 items that 10 judges unanimously deemed to reflect highly desirable self-statements. Edwards (1957, 1970) reported that this scale was correlated highly with most MMPI scales in general, and the strong, omnipotent first factor that emerged from factor analyses of MMPI scale scores. He concluded that MMPI scale scores were thus hopelessly confounded with the social desirability response style and, therefore, could not be used to identify meaningful (rather than stylistic) individual differences.

As was the case with Jackson and Messick’s (1962) argument regarding acquiescence (see the discussion of fixed responding), Block (1965) provided a definitive refutation of Edwards’s (1957) social desirability critique. Block demonstrated that Edwards’s social desirability scale was in fact a marker of a substantive personality dimension he termed ego resiliency. Following the earlier work of Wiggins (1959), Block developed an ego resiliency–free measure of social desirability and found much lower levels of overlap with substantive MMPI scale scores than Edwards reported for his social desirability scale. Moreover, Block demonstrated that both a social-desirability-independent ego resiliency scale he constructed and Edwards’s social desirability scales were correlated with meaningful non-MMPI variables that reflected substantive individual differences.

Commenting on Edwards’s (1957) claim that the MMPI scales were hopelessly confounded with social desirability, Block (1965) observed:

Confounding is a blade that, if held too tightly, will cut its wielder. With the same logic advanced for social desirability as underlying MMPI scales, one can argue that the [first] factor of the MMPI represents a personality dimension that is vital to understanding the SD scale. Many of the MMPI scales have empirical origins and demonstrable validity in separating appropriate criterion groups. The high correlations found between these scales and the SD measure therefore plausibly suggest—not an artifact or naïveté in the construction of the earlier scales—but rather that the SD scale, wittily or not, is an excellent measure of some important variable of personality. (pp. 69–70)

When we reflect on the methods Edwards (1957) used to construct his social desirability scale, the resulting confound
is readily understood. Stated simply, psychopathology is undesirable. Ask a group of persons to identify SRI items that reflect undesirable characteristics, and, if they are included in the pool, participants will undoubtedly generate a list of items describing negative psychological characteristics. Edwards’s assumption that individuals’ responses to such items reflect a substantively meaningless response style proved subsequently to be unwarranted and was refuted by Block’s (1965) analyses. Nonetheless, Edwards and some followers continued to raise these arguments. For example, relying (like Edwards) on scales that reflected desirability judgments, Jackson, Fraboni, and Helmes (1997) criticized the MMPI-2 Content Scales (Butcher et al., 1990) for being highly saturated with social desirability. As Edwards failed to do before them, these authors did not explain how scales that they concluded were highly saturated with irrelevant stylistic variance could account significantly for a wide range of extratest personality and psychopathology variables (Butcher et al., 1990).

Implications of Threats to Protocol Validity

The issues discussed and highlighted in this section illustrate the crucial role played by respondents’ test-taking approaches in determining the interpretability of SRI scale scores. Allport (1937) and Ellis (1946) foresaw accurately that reliance on an individual’s willingness and ability to generate an accurate self-portrayal when responding to test items was one of the greatest challenges facing SRI developers and users. Subsequent decades of research and practice have illuminated a host of threats to protocol validity (just described), all manifestations of the kinds of concerns identified early on by Allport and Ellis. Self-report inventory developers have responded to these threats in various ways, ranging from the development of validity scales, SRI measures designed to assess and, in some instances, correct for the effects of protocol invalidating test-taking approaches (e.g., the MMPI-2 validity scales; Butcher et al., 2001), to declaration and attempts to demonstrate that these threats do not really amount to much (Costa & McCrae, 1992a; Piedmont et al., 2000) and the consequent decision not to include validity scales on some instruments (e.g., the NEO-PI-R; Costa & McCrae, 1992c).

Commenting on the then-prevailing paucity of efforts by SRI developers to address threats to protocol validity, Meehl and Hathaway (1946) observed:

It is almost as though we inventory-makers were afraid to say too much about the problem because we had no effective solution for it, but it was too obvious a fact to be ignored so it was met by a polite nod. Meanwhile the scores obtained are subjected to varied “precise” statistical manipulations which impel the student of behavior to wonder whether it is not the aim of the personality testers to get as far away from any unsanitary contact with the organism as possible. Part of this trend no doubt reflects the lack of clinical experiences of some psychologists who concern themselves with personality testing. . . . (p. 526)

Acting on this concern, Hathaway and McKinley incorporated two validity scales, L and F, in their original MMPI development efforts. The MMPI was not the first SRI to make validity scales available to its users. Cady (1923) modified the Woodworth Psychoneurotic Inventory (derived from of the original Personal Data Sheet) to assess juvenile incorrigibility and incorporated negatively worded repeated items in the revised inventory to examine respondents’ “reliability.” Maller (1932) included items in his Character Sketches measure designed to assess respondents’ “readiness to confide.” Humm and Wadsworth (1935), developers of the Humm-Wadsworth Temperament Scales, incorporated scales designed to identify defensive responding to their SRI. Ruch (1942) developed an “honesty key” for the BPI, the most widely used SRI prior to the MMPI.

Hathaway and McKinley’s inclusion of validity scales on the original MMPI was thus consistent with growing recognition among SRI developers of the need to incorporate formal means for assessing and attempting to correct for threats to protocol validity. In describing their efforts to develop and apply the MMPI K scale and K-correction, Meehl and Hathaway (1946) articulated the conceptual and empirical underpinnings of MMPI approaches to assessing threats to protocol validity. As MMPI use and research proliferated throughout the latter part of the twentieth century, Hathaway, McKinley, and Meehl’s emphasis on assessing threats to protocol validity was continued through efforts to develop a variety of additional MMPI and MMPI-2 validity scales. Following in this tradition, most (but not all) modern SRIs include measures designed to provide information regarding threats to protocol validity.

Future Directions for Self-Report Inventory Research

Self-report measures play a vital role in personality and psychopathology assessment. Self-report inventories are used commonly and routinely in various applied assessment tasks, and they have been the focus of thousands of empirical investigations. Considerable progress was made in developing this technology over the course of the twentieth century, and many of the concerns identified early on by Allport (1937) and Ellis (1946) have been addressed in modern self-report
measures. Three primary aspects of SRI-based assessment were reviewed and analyzed in this chapter: approaches to SRI scale score interpretation, standard score derivation for SRIs, and threats to protocol validity. As discussed earlier, modern SRIs offer a variety of solutions to the challenges posed in each of these areas. However, this review has also pointed out needs for further research-based refinement in each of these aspects of SRI-based assessment. The final part of this chapter highlights needs and directions for further research in SRI-based approaches to assessing personality and psychopathology.

Approaches to SRI Scale Score Interpretation

Two primary approaches to SRI scale score interpretation, empirically grounded and content-based, were identified in this review. Not surprisingly, much of the research in this area has focused on empirically grounded SRI scale score interpretation. This is understandable because, by definition, empirically grounded interpretation is research-dependent. However, content-based interpretation can and should be subjected to rigorous empirical scrutiny. Specifically, research is needed to examine the validity of content-based SRI scale score interpretation. Such investigations should explore the content validity of content-based measures (i.e., the extent to which they adequately canvass the relevant content domain) and the criterion and ultimately construct validity of content-based interpretation. Moreover, as detailed earlier, content-based and empirically grounded approaches are not mutually exclusive, and research is needed to guide SRI users regarding optimal ways to combine them in scale score interpretation.

Several aspects of empirically grounded SRI scale score interpretation also require further elaboration. As reviewed previously, empirically keyed interpretation has garnered limited support in the SRI literature to date. It is unclear whether this is a product of limitations inherent in the external approach to SRI scale construction, in which case further efforts at developing empirically keyed interpretable approaches should be abandoned, or whether the problem rests more in deficiencies of previous efforts at external scale construction that attenuated the validity of their products. There has been no extensive effort at external scale construction since the original MMPI clinical scales were developed. Considerable progress has since been made in other approaches to diagnostic classification (e.g., development of structured diagnostic interviews) and in the methodologies and technology available to test constructors. It is possible (if not likely) that a comprehensive effort to develop SRI scales keyed to differentiate empirically between reliably (with the aid of structured diagnostic interviews) diagnosed classes of individuals will yield diagnostic indicators that are more valid than the original MMPI clinical scales.

As noted previously, most empirically grounded SRI scale score interpretation has followed the empirical correlate approach. Much of the research in this area has focused on the direct, simple inference level afforded by knowledge of a scale score’s criterion validity. Limited attention has been paid in this literature to an issue that receives prominent attention in the industrial/organizational (I/O) assessment literature, the question of validity generalization: Under what circumstances are empirical correlates identified in one setting likely to apply to others? Following the seminal work of I/O researchers Schmidt and Hunter (1977), I/O psychologists have developed various techniques to appraise validity generalization for their assessment instruments. In light of the particularly prominent role of criterion validity in SRI-based assessment of personality and psychopathology, similar research in this area is clearly needed.

Configural interpretation (examination of patterns among SRI scale scores; as distinguished from linear interpretation, which involves independent consideration of SRI scale scores) is another aspect of criterion-validity-based SRI application requiring further examination. As discussed earlier, the primary assumption underlying configural interpretation (that there is something about the pattern of scores on a set of SRI scales that is not captured when they are interpreted linearly) has seldom been tested empirically. Moreover, in the rare cases in which it has been tested, configural interpretation has not demonstrated incremental validity in reference to linear approaches. Configural approaches may improve upon linear interpretation either by enhancing the scales’ convergent validity or by sharpening their discriminant validity. Research is needed to evaluate the extent to which configural interpretation adds (beyond linear interpretation) to either or both.

Finally, with respect to scale score interpretation, research has yet to mine adequately the prospects of construct validity. As a result, SRI users are unable to rely on construct validity adequately as an interpretive source. Most empirically grounded SRI scale score interpretation is guided by the simple, direct inference level afforded by criterion validity data. Concurrent with the move in psychiatry toward a descriptive, atheoretical nosology, research on clinical applications of SRIs has similarly focused narrowly on their scales’ criterion validity. Cronbach and Meehl’s (1955) admonition that psychological tests be used to identify and elucidate the nature of major constructs, and that the resulting enhancement in our understanding of these constructs guide our interpretation of test scores, has not been followed. We remain largely incapable of interpreting SRI scale scores in the context of theoretically grounded nomological networks.
A potential exception to this trend is the five-factor model (FFM) of personality, which focuses on five core personality traits: extraversion, agreeableness, conscientiousness, neuroticism, and openness/intellect. Although not without its critics (e.g., Block, 1995; Loevinger, 1994), this product of the normal personality assessment literature has generated an empirical literature base that can be used to elucidate a rich, theoretically grounded nomological network associated with its five core constructs (e.g., John & Srivastava, 1999). Unfortunately, efforts to date to apply this rich framework to clinical assessment tasks have met with limited success. These difficulties, however, appear largely to be a product of limitations in tests designed to measure the FFM (e.g., questions about the clinical utility of the NEO-PI-R related to its authors’ decision not to measure potential threats to protocol validity; Costa & McCrae, 1992c). Alternative conceptualizations (e.g., Harkness and McNulty’s PSY-5 model; 1994), developed from the clinical rather than normal personality perspective, may ultimately prove more fruitful. In any event, enhancing SRI interpreters’ ability to rely on their construct validity should be a major goal of further research efforts in this area.

**Standard Score Derivation for SRIs**

Two primary needs for further research exist with respect to standard score derivation for SRIs. First, as reviewed earlier, various problems in normative sampling may result in over- or underestimation of an individual’s standing on SRI-measured constructs. Current and future SRIs need to be scrutinized carefully to determine whether, and to what extent, the systematic sampling errors, population changes, and application changes described previously might compromise their normative samples’ adequacy.

A second aspect of standard score derivation for SRIs that should be the focus of further research efforts relates to the advisability and feasibility of using special norms when applying SRIs to specific subpopulations or setting types. Some approaches to incorporating population subsegment information in SRI scale score interpretation involve developing separate norms for use in these applications (e.g., Roberts et al.’s approach to using the PAI in public safety personnel screening; 1999). However, as discussed earlier, use of so-called special norms may obscure features shared commonly by members of a population subsegment or by individuals tested under similar circumstances (e.g., defensiveness among individuals being screened for public safety positions or depression in people tested in clinical settings).

An alternative method for considering how an individual’s SRI scale scores compare with those of population subsegments is to provide interpreters data on group members’ means and standard deviations on the relevant scales. Such data could be provided in professional publications or along with individual test scores generated through automated scoring services. For example, many automated scoring services currently include a graphic printout of the individual’s standard scores on a profile sheet. Group mean profiles, along with their associated standard deviations or errors plotted as confidence intervals, could be added to these printouts. This would allow the test interpreter to learn how the individual’s scores compare with both the general normative standard and with relevant comparison groups without obscuring the effects of group deviations from the mean.

**Assessing Threats to Protocol Validity**

Several types of threats to SRI protocol validity were identified in this chapter. Existing instruments vary in the extent to which they provide interpreters information regarding these threats’ presence in a given protocol. Most SRIs provide means for assessing at least some of the categories of threats outlined in Table 24.1. The recently updated MMPI-2 (Butcher et al., 2001) contains scales designed to tap each of the types and subtypes of threats described earlier. Within the category of Non-Content-Based Invalid Responding, nonresponding is assessed by the Cannot Say scale; random responding by the Variable Response Inconsistency (VRIN) scale; and fixed responding is measured by the True Response Inconsistency (TRIN) scale. In the category of Content-Based Invalid Responding, overreporting is gauged by the infrequency scales F (Infrequency), Fb (Back Infrequency), and Fp (Infrequency psychopathy), and underreporting is assessed by the defensiveness indicators L (Lie), K (Defensiveness), and S (Superlative).

Existing validity scales fall short, however, in their ability to differentiate meaningfully among threats within these subtypes. For example, existing scales do not allow for differentiation among intentional versus unintentional random responding, intentional versus unintentional over- or underreporting, exaggeration versus fabrication, or minimization versus denial. Some of these distinctions may only be possible through consideration of extratextual data; however, further research is needed to explore whether configural interpretation of existing validity scales or development of additional validity scales may allow SRI interpreters to more finely distinguish among the various threats and levels of threats to protocol validity.

**CONCLUSION**

This chapter provided an overview of the historical foundations and early criticisms of self-report measures, current issues and challenges in SRI interpretation, and needs for
future research in this area. A great deal of progress has been made in developing this technology’s conceptual and empirical foundations. Over the past 50 years, the challenges articulated early on by Allport (1937) and Ellis (1946) have been addressed (with varying degrees of success) by subsequent SRI developers and researchers. These efforts have been documented in an elaborate body of scholarly literature that, of course, goes well beyond the scope of this chapter. Other chapters in this volume cover additional aspects of this literature, in particular the chapters by Garb on clinical versus statistical prediction, Bracken and Wasserman on psychometric characteristics of assessment procedures, and Reynolds and Ramsey on cultural test bias. Chapters on assessment in various settings include reviews of more setting-specific aspects of the SRI literature. Overall, these chapters indicate that assessment of personality and psychopathology by self-report rests on solid foundations that leave this technology well positioned for future research and development efforts.

REFERENCES


Hathaway, S. R. (1956). Scales 5 (Masculinity-Femininity), 6 (Paranoia), and 8 (Schizophrenia). In G. S. Welsh & W. G. Dahlstrom (Eds.), Basic readings on the MMPI in psychology and medicine (pp. 104–111). Minneapolis: University of Minnesota Press.


