CHAPTER 3

ISSUES IN PERSONALITY ASSESSMENT

CHAPTER OUTLINE

Sources of Information
- Observer Ratings
- Self-Reports
- Implicit Assessment
- Subjective versus Objective Measures

Reliability of Measurement
- Internal Consistency
- Inter-Rater Reliability
- Stability across Time

Validity of Measurement
- Construct Validity
- Criterion Validity
- Convergent Validity
- Discriminant Validity
- Face Validity
- Culture and Validity
- Response Sets and Loss of Validity

Two Rationales Behind the Development of Assessment Devices
- Rational or Theoretical Approach
- Empirical Approaches

Better Assessment: A Never-Ending Search

Summary

CHAPTER SUMMARY

Assessment (measurement of personality) is something that people constantly do informally. Psychologists formalize this process into several distinct techniques. *Observer ratings* are made by someone other than the person being rated—an interviewer, someone who watches, or someone who knows the people well enough to make ratings of what they are like. Observer ratings often are somewhat subjective, involving interpretations of the person's behavior. *Self-reports* are made by the people being assessed, about themselves. Self-reports can be single scales or multiscale inventories. *Implicit assessment* is measuring patterns of associations within the self that are not open to introspection. Assessment devices can be subjective or objective. Objective techniques require no interpretation as the assessment is made. Subjective techniques involve some sort of interpretation as an intrinsic part of the measure.

One issue for all assessment is *reliability* (the reproducibility of the measurement). Reliability is determined by checking one measurement against another (or several others). Self-report scales usually have many items (each a measurement), leading to indices of *internal reliability*, or *internal consistency*. Observer judgments are checked by inter-rater reliability. Test-retest reliability assesses the reproducibility of the measure over time. In all cases, high correlation among measures means good reliability.

Another important issue is *validity* (whether what you're measuring is what you want to measure). The attempt to determine whether the operational definition (the assessment device) matches the concept you set out to measure is called *construct validation*. Contributors to construct validity are evidence of *criterion*, *convergent*, and *discriminant* validity. *Face validity* is not usually taken as an important element of construct validity. Validity is threatened by the fact that people have *response sets* (*acquiescence* and *social desirability*) that bias their responses.

Development of assessment devices proceeds along one of two paths. The *rational* path uses a theory to decide what should be measured and then figures out the best way to measure it. Most assessment devices developed this way. The *empirical* path involves using data to determine what items should be in a scale. The MMPI was developed this way, using a technique called *criterion keying*, in which the test developers let people's responses tell them which items to use. Test items that members of a diagnostic category answered differently from other people were retained.

KEY TERMS

Acquiesence: The response set of tending to say “yes” (agree) in response to any question.

Assessment: The measuring of personality.

Construct validity: The accuracy with which a measure reflects the underlying concept.

Convergent validity: The degree to which a measure relates to other characteristics that are conceptually similar to what it's supposed to assess.

Criterion keying: The developing of a test by seeing which items distinguish between groups.

Criterion validity: The degree to which the measure correlates with a separate criterion reflecting the same concept.

Discriminant validity: The degree to which a scale does not measure unintended qualities.

Empirical approach (to scale development): The use of data instead of theory to decide what should go into the measure.

Error: Random influences that are incorporated in measurements.

Face validity: The scale “looks” as if it measures what it's supposed to measure.

Implicit assessment: Measuring associations between the sense of self and aspects of personality that are implicit (hard to introspect about).

Internal reliability (internal consistency): Agreement among responses made to the items of a measure.

Inter-rater reliability: The degree of agreement between observers of the same events.

Inventory: A personality test measuring several aspects of personality on distinct subscales.

Objective measure: A measure that incorporates no interpretation.

Observer ratings: An assessment in which someone else produces information about the person being assessed.

Operational definition: The defining of a concept by the concrete events through which it is measured (or manipulated).

Predictive validity: The degree to which the measure predicts other variables it should predict.

Rational approach (to scale development): The use of a theory to decide what you want to measure, then deciding how to measure it.

Reliability: Consistency across repeated measurements.

Response set: A biased orientation to answering.

Self-report: An assessment in which people make ratings pertaining to themselves.

Social desirability: The response set of tending to portray oneself favorably.

Split-half reliability: Assessing internal consistency among responses to items of a measure by splitting the items into halves, then correlating them.

Subjective measure: A measure incorporating personal interpretation.

Test-retest reliability: The stability of measurements across time.

Theoretical approach: See Rational approach.

Validity: The degree to which a measure actually measures what it is intended to measure.

TEST ITEMS

Multiple Choice

(d/34) 1. The process of measuring personality is called:

a. personology.
b. analytic technique.
c. psychometrics.
d. assessment.

(b/34) 2. When a measure comes from someone other than the person being observed, this is called:

a. introspection.
b. observer rating.
c. psychometrics.
d. none of the above

(d/34) 3. Observer ratings can be based on:

a. interviews in which people talk about themselves.
b. direct observations of overt action.
c. interviews in which people talk about something other than themselves.
d. all of the above

(c/35) 4. A measure that assesses several dimensions of personality is called a(n):

a. personology scale.
b. projective test.
c. inventory.
d. none of the above

(d/35-36) 5. Implicit assessment techniques attempt to learn about a person by:

a. directly asking him or her about a characteristic in a survey.
b. directly asking him or her about a characteristic in an interview.
c. hypnotizing him or her and asking about a characteristic.
d. none of the above

(a/36) 6. A scale is "objective" if:

a. a person's responses are recorded directly, with no interpretation until a later time.
b. an interpretation is made before information about behavior is recorded.
c. another researcher has used it.
d. it is included in the *Manual of Personality Inventories*.

(b/37) 7. The reliability of an observation refers to:

a. whether it was collected in a laboratory or field setting.
b. its consistency across repeated observations.
c. its applicability to many persons.
d. its scope covering all possibly related items.

(d/37) 8. Which of the following is a potential source of error in a measure?

a. the way an item is phrased
b. variations in an observer's attention
c. distractors present when observations are made
d. all of the above

(d/37) 9. Which of the following allows one to assess the reliability of a measure?

a. Make the observation more than once.
b. Measure the same quality from a slightly different angle.
c. Measure the same quality with a slightly different measure.
d. all of the above

(c/38) 10. Reliability within a set of observations measuring the same aspect of personality is referred to as:

a. lack of random error.
b. unified reliability.
c. internal consistency.
d. internal clarity.

(b/38) 11. If there is a high degree of correlation among the items on a measure, it is said to have high:

a. validity.
b. reliability.
c. significance.
d. diversity.

(b/38) 12. Split-half reliability refers to the:

a. correlation between two halves of a sample.
b. correlation between the items comprising the first and second halves of a test.
c. correlation between half the items on a test and some other criterion.
d. correlation between half the items on one test and half the items on another test.

(c/39-40) 13. A high correlation between scores on the same test administered at two different points in time demonstrates high:

a. split-half reliability.
b. inter-rater reliability.
c. test-retest reliability.
d. none of the above

(a/40) 14. Which of the following is true about reliability and validity?

a. It is possible for a measure to be reliable but not valid.
b. Once you prove a measure is reliable, you know it is valid.
c. You can assess either the reliability or validity of a measure, but not both.
d. none of the above

(b/41) 15. A measure is high in validity when:

a. the results obtained with the measure match the researcher's predictions.
b. the operational definition closely matches the conceptual definition.
c. others have used the same measure.
d. all of the above

(a/42) 16. The most all-encompassing and, thus, most important kind of validity is:

a. construct.
b. criterion.
c. convergent.
d. discriminant.

(a/42) 17. If an assessment measures the intended conceptual characteristics, it has demonstrated:

a. construct validity.
b. inter-rater reliability.
c. factor strength.
d. conceptual validity.

(d/42) 18. A high correlation between an assessment device and an external standard of comparison is an indication of _____ validity.

a. behavioral
b. inter-rater
c. observer
d. criterion

(a/42) 19. _____ validity is generally seen as the most important means of establishing construct validity.

a. Criterion
b. Discriminant
c. Face
d. Convergent

(c/42-43) 20. If a scale is correlated with other scales that measure similar concepts it is said to have:

a. generalizability.
b. congruent validity.
c. convergent validity.
d. inter-test reliability.

(a/43) 21. If a researcher demonstrates that her measure of loneliness is not correlated with a measure of intelligence, she has provided evidence for the _____ validity of this measure.

a. discriminant
b. criterion
c. convergent
d. true

(d/44) 22. Sometimes researchers actually try to reduce _____ validity by obscuring the true purpose of a measure.

a. construct
b. convergent
c. discriminant
d. face

(a/44) 23. Why is face validity regarded as a convenience by researchers?

a. Some believe it is easier to respond to face-valid instruments.
b. Face-valid instruments make it difficult to guess what the researcher is measuring.
c. Face-valid instruments are often reliable for 40-50 years.
d. none of the above

(d/44) 24. What are some difficulties with using a measure developed in one culture with a sample from a different culture?

a. The personality construct (e.g., self-esteem) might mean different things in different cultures.
b. People from different cultures might interpret the questions differently.
c. Questions might contain phrasing that is difficult to translate into another culture.
d. all of the above

(b/45) 25. Response sets are:

a. answer sheets for a personality test.
b. biased ways in which people respond to personality measures.
c. multiple pieces of information from the same person.
d. the particular response options the researcher gives participants (e.g., 1-7 ratings).

(b/45) 26. A response set in which participants simply tend to answer “yes” to all questions is known as:

a. agreeableness.
b. acquiescence.
c. a positivity bias.
d. social desirability.

(a/45) 27. The attempt to create a good impression on a personality measure is called:

a. social desirability.
b. self-consciousness.
c. assessment error.
d. none of the above

(d/46) 28. Which of the following is NOT suggested as a way to deal with social desirability?

a. Phrase undesirable responses to make them seem acceptable.
b. Find ways for people to admit their undesirable qualities indirectly.
c. Include items that measure one's degree of concern for social desirability, then use it as a correction factor.
d. Tell participants they will not receive credit if they are found to be deceptive.

(b/46) 29. The theoretical approach to assessment often results in measures that have a high degree of _____ validity.

a. construct
b. face
c. discriminant
d. internal

(b/46) 30. Which of the following is NOT required by the rational approach to developing personality measures?

a. demonstrating the measure is reliable
b. demonstrating the measure has never been administered before
c. demonstrating the measure predicts behavioral criteria
d. demonstrating the measure has construct validity

(a/47) 31. Which of the following is NOT true about the empirical approach to developing measures?

a. It relies on theory.
b. It relies on data.
c. It allows psychologists to decide what qualities of personality exist.
d. none of the above

(a/47) 32. The scale construction method that uses predefined groups to select items is called the:

a. criterion keying approach.
b. content keying approach.
c. theoretical keying approach.
d. normative group keying approach.

(a/47) 33. Which of the following statements about the MMPI is NOT true?

a. It uses a 7-point response scale.
b. It is used to diagnose clinical disorders.
c. It was developed using the criterion keying approach.
d. It has a true-false response format.

(c/47) 34. In recent years, the MMPI-2 has become controversial because:

a. it is only valid for inpatient populations.
b. it is only valid for college students.
c. diagnostic categories are not as distinct as they were once thought to be.
d. it is too time consuming to complete.

<u>True and False</u>

(T/34) 1. Observer ratings can involve interviews.

(F/34) 2. If observer ratings involve interviews, participants must talk about themselves to the interviewer.

(T/35 Box 3.1) 3. Researchers have been able to learn about personality by studying people's bedrooms.

(T/35) 4. A measure that assesses several dimensions of personality is called an inventory.

(F/35) 5. A self-report scale can measure only one aspect of personality.

(F/35-36) 6. Implicit assessment involves asking participants directly about themselves.

(F/36) 7. If a trained observer rates how tired a person appears after a test, that observer is making an objective rating.

(T/37) 8. Making multiple observations generally improves reliability.

(F/38) 9. Internal reliability is the extent to which raters agree with one another.

(F/38) 10. Split-half reliability refers to how consistent responses are across time.

(F/38 Box 3.2) 11. Item response theory, because it is such a new technique, has only been applied to a narrow range of assessments thus far.

(T/38-39) 12. If no correlation is found between two halves of a personality measure, then it is said to have very poor split-half reliability.

(T/39) 13. Inter-rater reliability is most applicable to observational measures.

(T/40) 14. When participant responses to a test are consistent across time, the test is said to have strong test-retest reliability.

(F/40) 15. If we know that a scale is reliable, we also know that it is valid.

(T/40) 16. An operational definition is a description of some kind of physical event.

(F/40) 17. The dictionary definition of "envy" is the same as the operational definition of "envy."

(T/40) 18. Even abstract concepts such as love can be operationalized.

(T/42) 19. To say that a measure has construct validity is to say that it measures what the researcher intended.

(T/42) 20. The most important type of validity for personality measurement is construct validity.

(T/42) 21. Establishing construct validity is a long and complicated process.

(T/42) 22. Criterion validity is often regarded as the most important indicator of construct validity.

(F/42-43) 23. Convergent validity and discriminant validity are a “trade-off” such that a measure can be high on only one or the other.

(T/43) 24. If a personality measure has discriminant validity, it does not measure characteristics it was not intended to measure.

(F/43) 25. Most researchers think that face validity is among the most important types of validity.

(F/44) 26. Even if scales must be modified for different cultures, it is safe to assume that personality constructs (e.g., shyness) mean the same thing across cultures.

(T/44) 27. One problem with administering psychological measures across cultures is determining whether the psychological construct has the same meaning.

(T/45) 28. Acquiescence is a response set in which a person is biased toward agreeing with the questions asked.

(F/45) 29. Acquiescence is a response set that can not be resolved.

(T/46) 30. Assessing and correcting for one's concern with social desirability is one means of handling the problem of social desirable responding.

(T/46) 31. The majority of existing personality measures described in the text were developed using the rational approach.

(T/47) 32. According to the empirical approach to test development, it does not matter at all what the items look like.

(T/47) 33. The criterion keying approach emphasizes the importance of a scale's ability to accurately predict whether a person is a member of some group.

(F/47) 34. The MMPI was developed utilizing the content validation approach.

(F/47) 35. The MMPI uses a seven-point rating scale.

(T/47) 36. Interpretation of the MMPI is based on comparing an individual's personality profile with those of patients who have a particular diagnosis.

(F/47) 37. Discriminant validity can be established more quickly than other types of validity.

(F/48) 38. Once a personality measure has been validated, it need never be revised and/or re-validated.

Short Essay

(36) 1. What is the difference between subjective and objective personality measures? Give an example of each.

SUBJECTIVE: An interpretation takes place before the information is recorded. For example, an observer decides a subject is nervous before impression is recorded.

OBJECTIVE: Subject's responses are recorded directly with no interpretation taking place until later. For example, self-report.

(36-39) 2. Identify and briefly explain three different types of reliability.

(1) Internal reliability: Consistency within the test;
(2) Test-retest reliability: Consistency across time;
(3) Inter-rater reliability: Agreement between raters.

(40-44) 3. Describe three different types of validity that are relevant to personality assessment, and briefly discuss the meaning of each.

(1) CONSTRUCT: The measure accurately reflects the construct (conceptual quality) of interest.

(2) CRITERION: The measure relates to other manifestations of the personality quality it is supposed to be measuring; or, uses a behavioral index (or observer judgment) as an external criterion (standard of comparison), and sees how well the measure correlates with it.

(3) CONVERGENT: The measure relates to characteristics that are similar to, but not the same as, what it is supposed to measure.

(4) DISCRIMINANT: The measure does not measure qualities it was not intended to measure, especially qualities that do not fit with what the researcher had in mind as a construct.

(5) FACE: The measure appears, on the "face" of it, to be measuring the construct; it "looks" right.

(43-44) 4. Face validity can sometimes be detrimental to research. Give an instance/example of when this might be the case.

The assessment device is intended to measure something the person being assessed would find undesirable to admit or threatening. For example, if a researcher wants to measure problem drinking, a participant might not want to admit this. In these cases, the test developer usually tries to obscure the measure's face validity.

(44) 5. Identify two ways in which cultural differences might threaten validity.

(1) Psychological constructs might not mean the same thing in different cultures.
(2) People from different cultures might interpret scale items differently.

(44-45) 6. Identify two types of response sets personality psychologists would like to avoid, and give one way to combat each problem.

(1) ACQUIESENCE: Key one-half of the items in a positive direction and the remaining half in a negative direction.

(2) SOCIAL DESIRIBILITY: Try to phrase question in such a way that social desirability is not salient.

(46-47) 7. Two approaches to developing personality measures were described in the text. Identify them and briefly describe or give examples of how personality psychologists might use them.

(1) RATIONAL (Theoretical) Approach: Psychologist first derives a rational basis for believing that a particular dimension of personality is important. Next, she/he creates a test on which this logical dimension is reflected validly and reliably in people's answers.

(2) EMPIRICAL (Data-based) Approach: Relies on data, rather than theory, to decide what items make up the assessment device. The groups into which people are to be sorted represent a set of criteria for the test. Start with a lot of possible test items, and find out which items tend to be answered differently by members of one criterion group than by other people.

TEST YOURSELF 3-1

Source: Crowne, D. P., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*. New York: Wiley.

Relevant Part of Text: Chapter 3, page 46.

Description of Scale: The items on this scale were developed by Crowne and Marlowe (1964) to measure "social desirability." Students respond to each of the items by answering either True or False. The items consist of statements that are unlikely to be true for most people (e.g., "No matter who I'm talking to, I am always a good listener.") or unlikely to be false for most people (e.g., "There have been occasions when I took advantage of someone."). Higher scores reflect greater tendencies to give socially desirable responses. The items were chosen because of their tendency to make one look good. In fact, given the way the items are written, these responses make someone look just a little too good--too good to be true. Nobody finds out everything about politicians before voting; nobody is always a good listener; everyone sometimes tries to get even; and everyone occasionally experiences intense dislike. This, at least, is the reasoning that underlies the development of the scale. An occasional item will happen to be true for everyone. But if too many of the items are answered in the "good" way, either the person who's answering them is a saint or there's a little bit of exaggeration going on.

As is discussed in greater detail in Chapter 3 (Crowne and Marlowe, pp. 47-48), social desirability is often viewed as something that gets in the way of accurate assessment. The reasoning goes like this: If a person's responses are colored too much by trying to make "acceptable" answers, the test won't tell much about the person's true personality. On the other hand, it is sometimes argued that the tendency to make socially appropriate responses is an important personality dimension in its own right. Regardless of which argument you think is more persuasive, this exercise should demonstrate a little about the degree to which social desirability considerations play a role in one's own personality.

PRIMARY SOURCES

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillian.

Campbell, D. T. (1960). Recommendations for the APA test standards regarding construct, trait, and discriminant validity. *American Psychologist*, **15**, 546-553.

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