Reference Support - Psychology Test VIII - IAS (Mains) 2016

1.a. How would you define "minimal risk" involved in an experiment? 10

Reference:

A study is described as involving “minimal risk” when the procedures or activities in the study are similar to those experienced by participants in their everyday life.

A distinction is sometimes made between a participant “at risk” and one who is “at minimal risk.” Minimal risk means that the harm or discomfort participants may experience in the research is not greater than what they might experience in their daily lives or during routine physical or psychological tests. As an example of minimal risk, consider the fact that many psychology laboratory studies involve lengthy paper-and-pencil tests intended to assess various mental abilities. Participants may be asked to complete the tests quickly and may receive specific feedback about their performance. Although there is likely to be stress in this situation, the risk of psychological injury is likely no greater than that typically experienced by students. Therefore, such studies would involve only minimal risk for college students. When the possibility of injury is judged to be more than minimal, individuals are considered to be at risk. When a study places participants at risk, the researcher has more serious obligations to protect their welfare.

1.b. What do you understand by analysis of variance for single factor independent group design? 10

Reference:

- Analysis of variance (ANOVA) is an inferential statistics test used to determine whether an independent variable has had a statistically significant effect on a dependent variable.
- The logic of analysis of variance is based on identifying sources of error variation and systematic variation in the data.
- The F-test is a statistic that represents the ratio of between-group variation to within-group variation in the data.
- The results of the initial overall analysis of an omnibus F-test are presented in an analysis of variance summary table; comparisons of two means can then be used to identify specific sources of systematic variation in an experiment.
- Although analysis of variance can be used to decide whether an independent variable has had a statistically significant effect, researchers examine the descriptive statistics to interpret the meaning of the experiment’s outcome.
- Effect size measures for independent groups designs include eta squared and Cohen’s f.
- A power analysis for independent groups designs should be conducted prior to implementing the study in order to determine the probability of finding a statistically significant effect, and power should be reported whenever non significant results based on NHST are found.
- Comparisons of two means may be carried out to identify specific sources of systematic variation contributing to a statistically significant omnibus F-test.

1.c. What do you understand by incremental validity in Psychological assessment? 10

Reference:

Incremental validity in psychological assessment refers to the extent to which new information increases the accuracy of a classification or prediction above and beyond the accuracy achieved by information already available. Assessors pay adequate attention to incremental validity by collecting the amount and kinds of information they need to answer a referral question, but no more than that. In theory, then, familiarity with the incremental validity of various measures when used for certain purposes, combined with test selection based on this information, minimizes redundancy in psychological assessment and satisfies both professional and scientific requirements for justifiable test selection.

In practice, however, strict adherence to incremental validity guidelines often proves difficult and even disadvantageous to implement. As already noted, it is difficult to anticipate which sources of information will prove to be most useful. Similarly, with respect to which instruments to include in a test battery, there is little way to know whether the tests administered have yielded enough data, and which tests have contributed most to understanding the person being examined, until after the data have been collected and analyzed. In most practice settings, it is reasonable to conduct an interview and review previous records as a basis for deciding whether formal testing would be likely to help answer a referral question—that is, whether it will show enough incremental validity to warrant its cost in time and money. Likewise, reviewing a set of test data can provide a basis for determining what kind of additional testing might be worthwhile. However, it is rarely appropriate to administer only one test at a time, to choose each subsequent test on the basis of the preceding one, and to schedule a further testing session for each additional test administration. For this reason, responsible psychological assessment usually consists of one or two testing sessions comprising a battery of tests selected to serve specific additive, confirmatory, and complementary functions.

1.d. What are the advantages and disadvantages of computer aided assessment? 10

Reference:

Computer-assisted assessment has a number of advantages. Computers can save valuable professional time, potentially improve test-retest reliability, reduce possible tester bias, and reduce the cost to the consumer by improving efficiency (Butcher, Perry, & Atlis, 2000; Groth-Marnat, 1999). Even greater benefits may someday be realized by incorporating more complicated decision rules in interpretation, collecting data on response latency and key pressure, incorporating computer-based models of personality, tailoring future questions to a client based on past responses, and estimating the degree of certainty of various interpretations (Groth-Marnat, 2000a, 2000b).

In the past, computer-assisted assessment has resulted in considerable controversy within mental health publications (Faust & Ziskin, 1989; Groth-Marnat & Schumaker, 1989), the popular media (C. Hall, 1983), and professional publications outside the mental health area (Groth-Marnat, 1985). A primary issue has been untested reliability and validity. Research on reliability, however, has typically indicated that computerized administrations have generally excellent reliability that is at least equivalent to the paper-pencil versions (Campbell et al., 1999). In addition, computer-administered versus paper-pencil results for traditional tests have generally been found to result in negligible differences in scores (Finger

& Ones, 1999). This supports the view that if a paper and pencil version of the test is valid, a computerized version will also have equal validity resulting from the comparability in scores.

A further issue is the validity of computer-based test interpretation. Butcher et al. (2000) concluded a narrative review on the validity of computer-based interpretations by stating that in the vast majority of computer-based interpretations, 60% of the interpretations were appropriate. Shorter to mid-length narratives were generally considered to have a higher proportion of valid interpretations when compared with longer ones. In addition, the narrative statements contained in the computer-based reports were comparable to the types of statements made by clinicians. While this generally supports the use of computer-based interpretations, the finding that 40% or more of interpretations were not considered accurate means that the computer-based reports should be carefully evaluated. Thus, cutting and pasting computerized narratives into reports, results in unacceptably high error rates. Indeed, 42% of psychologists surveyed felt this procedure raised ethical concerns (McMinn, Ellens, et al., 1999). The previous summary clearly emphasizes that computer-based reports should not be used to replace clinical judgment but should instead be used as an adjunct to provide possible interpretations that the clinician needs to verify.

One concern is that many software packages are available to persons who do not possess appropriate professional qualifications. Ideally, qualified persons should be those who meet the requirements for using psychological tests in general (Turner et al., 2001). The American Psychological Association (1986) has also attempted to clarify these standards in their Guidelines for Computer-Based Test Interpretation. However, Krug's (1993) Psychware Sourcebook indicated that approximately a fifth of the programs could be sold to the general public. The American Psychological Association guidelines specify that users "have an understanding of psychological or educational measurement, validation problems, and test research" and that practitioners "will limit their use of computerized testing to techniques which they are familiar and competent to use" (American Psychological Association, 1986, p. 8). Users should also "be aware of the method used in generating the scores and interpretation and be able to evaluate its applicability to the purpose for which it will be used" (American Psychological Association, 1986, pp. 8–9).

The preceding difficulties associated with computer-assisted instruction suggest a number of guidelines for users (Groth-Marnat & Schumaker, 1989). First, practitioners should not blindly accept computer-based narrative statements, but rather should ensure, to the best of their ability, that the statements are both linked to empirically based research and placed in the context of the unique history and unique situation of the client. Computers have, among other benefits, the strong advantage of offering a wide variety of possible interpretations to the clinician, but these interpretations still need to be critically evaluated. Far greater research needs to be performed on both the meaning of computer-administered test scores and on the narrative interpretations based on these scores. The developers of software should also be encouraged to provide enough information in the manual to allow proper evaluation of the programs and should develop mechanisms to ensure the updating of obsolete programs.

1.e. List the diagnostic criteria for Gender Identity Disorder. 10

Reference:

Gender identity disorder

Gender identity disorder (GID) is the formal diagnosis used by psychologists and physicians to describe persons who experience significant gender dysphoria (discontent with their biological sex and/or the gender they were assigned at birth). It describes the symptoms related to transsexualism, as well as less severe manifestations of gender dysphoria. Although it is a psychiatric classification, GID is not medically classified as a mental illness.

Gender identity disorder in children is usually reported as "having always been there" since childhood, and is considered clinically distinct from GID that appears in adolescence or adulthood, which has been reported by some as intensifying over time. As gender identity develops in children, so do sex-role stereotypes. Sex-role stereotypes are the beliefs, characteristics and behaviors of individual cultures that are deemed normal and appropriate for boys and girls to possess. These "norms" are influenced by family and friends, the mass-media, community and other socializing agents. Since many cultures strongly disapprove of cross-gender behavior, it often results in significant problems for affected persons and those in close relationships with them. In many cases, transgendered individuals report discomfort stemming from the feeling that their bodies are "wrong" or meant to be different.

Many transgendered people and researchers support the declassification of GID as a mental disorder for several reasons. Recent medical research on the brain structures of transgendered individuals have shown that some transgendered individuals have the physical brain structures that resemble their desired sex even before hormone treatment. In addition, recent studies are indicating more possible causes for gender dysphoria, stemming from genetic reasons and prenatal exposure to hormones, as well as other psychological and behavioral reasons. (See Causes of transsexualism).

One contemporary treatment for this disorder consists primarily of physical modifications to bring the body into harmony with one's perception of mental (psychological, emotional) gender identity, rather than vice versa.

Diagnostic criteria

In the United States, the American Psychiatric Association permits a diagnosis of gender identity disorder if the four diagnostic criteria in the Diagnostic and Statistical Manual of Mental Disorders, 4thEdition, Text-Revised (DSM-IV-TR) are met. The criteria are:

- Long-standing and strong identification with another gender

- Long-standing disquiet about the sex assigned or a sense of incongruity in the gender-assigned role of that sex
- The diagnosis is not made if the individual also has physical intersex characteristics.
- Significant clinical discomfort or impairment at work, social situations, or other important life areas.

If the four criteria are met under the DSM-IV-TR, a diagnosis is made under ICD-9 code 302.85. See the classification and external resources sidebar at right for other diagnostic codes for gender identity disorder.

The International Classification of Diseases (ICD-10) list three diagnostic criteria:

Transsexualism (F64.0) has three criteria:

- The desire to live and be accepted as a member of the opposite sex, usually accompanied by the wish to make his or her body as congruent as possible with the preferred sex through surgery and hormone treatment
- The transsexual identity has been present persistently for at least two years
- The disorder is not a symptom of another mental disorder or a chromosomal abnormality

Uncertainty about gender identity, which causes anxiety or stress, is diagnosed as sexual maturation disorder.

2.a. "The use of deception reveals psychologists are willing to lie, which seemingly contradicts their supposed dedication to seeking truth". Critically evaluate whether these kinds of "technical illusion" should be permitted in the interests of scientific inquiry. Give examples.

15

Reference:

DECEPTION IN PSYCHOLOGICAL RESEARCH

- Deception in psychological research occurs when researchers withhold information or intentionally misinform participants about the research. By its nature, deception violates the ethical principle of informed consent.
- Deception is considered a necessary research strategy in some psychological research.
- Deceiving individuals in order to get them to participate in the research is always unethical.
- Researchers must carefully weigh the costs of deception against the potential benefits of the research when considering the use of deception.

One of the most controversial ethical issues related to research is the use of deception. Deception can occur either through omission, the withholding of information, or commission, intentionally misinforming participants about an aspect of the research. Some people argue that research participants should never be deceived because ethical practice requires that the relationship between experimenter and participant be open and honest (e.g., Baumrind, 1995).

To some, deception is morally repugnant; it is no different from lying. Deception contradicts the principle of informed consent. Despite the increased attention given to deception in research over the last several decades, the use of deception in psychological research has not declined and remains a popular research strategy (Sharpe, Adair, & Roese, 1992). For example, Skitka and Sargis (2005) surveyed social psychologists who used the Internet as a data collection tool and found that 27 percent of the reported studies involved deception of Internet participants.

How is it that deception is still widely used, despite ethical controversies? One reason is that it is impossible to carry out certain kinds of research without withholding information from participants about some aspects of the research (see Figure 3.6). In other situations, it is necessary to misinform participants in order to have them adopt certain attitudes or behaviors. For example, Kassin and Kiechel (1996) investigated factors affecting whether people will falsely confess to having done something they

did not do. Their goal was to understand factors that lead criminal suspects to falsely confess to a crime. In their experiment, the participants' task was to type letters that were being read aloud. They were told not to hit the Alt key while typing because this would crash the computer. The computer was rigged to crash after a brief time and the experimenter accused the participant of hitting the Alt key. Even though none of the participants had hit the Alt key, nearly 70% of the participants signed a written confession that they had done so. If the participants had known in advance that the procedures were designed to elicit their false confessions, they probably would not have confessed. The disclosure required for informed consent would have made it impossible to study the likelihood that people would make a false confession.

Although deception is sometimes justified to make it possible to investigate important research questions, deceiving participants for the purpose of getting them to participate in research that involves more than minimal risk is always unethical. As stated in the Ethics Code, "Psychologists do not deceive prospective participants about research that is reasonably expected to cause physical pain or severe emotional distress" (Standard 8.07b).

A goal of research is to observe individuals' normal behavior. A basic assumption underlying the use of deception is that sometimes it's necessary to conceal the true nature of an experiment so that participants will behave as they normally would, or act according to the instructions provided by the experimenter. Problems may arise, however, with frequent and casual use of deception (Kelman, 1967). If people believe that researchers often mislead participants, they may expect to be deceived when participating in a psychology experiment. Participants' suspicions about the research may prevent them from behaving as they normally would (see Box 3.2). This is exactly the opposite of what the researchers hope to achieve. Interestingly, Epley and Huff (1998) directly compared reactions of participants who were told or not told in a debriefing following the experiment that they had been deceived. Those who were told of the deception were subsequently more suspicious about future psychological research than were participants who were unaware of the deception. As the frequency of online research increases, it is important that researchers give particular attention to the use of deception, not only because of the potential for increasing the distrust of researchers by society's members, but also because deception has the potential to "poison" a system (i.e., the Internet) that people use for social support and connecting with others (Skitka & Sargis, 2005).

Kelman (1972) suggests that, before using deception, a researcher must give very serious consideration to (1) the importance of the study to our scientific knowledge, (2) the availability of alternative, deception-free methods, and (3) the "noxiousness" of the deception. This last consideration refers to the degree of deception involved and to the possibility of injury to the participants. In Kelman's view: "Only if a study is very important and no alternative methods are available can anything more than the mildest form of deception be justified" (p. 997).

TO DECEIVE OR NOT TO DECEIVE: THAT'S A TOUGH QUESTION

Researchers continue to use deceptive practices in psychological research (e.g., Sieber, Iannuzzo, & Rodriguez, 1995). The debate in the scientific community concerning the use of deception also has not abated (see, for example, Bröder, 1998; Fisher & Fryberg, 1994; Ortmann & Hertwig, 1997). It is a complex issue, with those taking part in the debate sometimes at odds over the definition of deception (see Ortmann & Hertwig, 1998). Fisher and Fryberg (1994) summarized the debate as follows: "Ethical arguments have focused on whether deceptive research practices are justified on the basis of their potential societal benefit or violate moral principles of beneficence and respect for individuals and the

fiduciary obligations of psychologists to research participants" (p. 417). This is quite a mouthful; so let's break it down.

A moral principle of "beneficence" refers to the idea that research activities should be beneficent (bring benefits) for individuals and society. If deception is shown to harm individuals or society, then the beneficence of the research can be questioned. The moral principle of "respect for individuals" is just that: People should be treated as persons and not "objects" for study, for example.

This principle would suggest that people have a right to make their own judgments about the procedures and purpose of the research in which they are participating (Fisher & Fryberg, 1994). "Fiduciary obligations of psychologists" refer to the responsibilities of individuals who are given trust over others, even if only temporarily. In the case of psychological research, the researcher is considered to have responsibility for the welfare of participants during the study and for the consequences of their participation.

These ideas and principles can perhaps be illustrated through the arguments of Baumrind (1985), who argues persuasively that "the use of intentional deception in the research setting is unethical, imprudent, and unwarranted scientifically" (p. 165). Specifically, she argues that the costs to the participants, to the profession, and to society of the use of deception are too great to warrant its continued use. Although these arguments are lengthy and complex, let us attempt a brief summary. First, according to Baumrind, deception exacts a cost to participants because it undermines the participants' trust in their own judgment and in a "fiduciary" (someone who is holding something in trust for another person). When research participants find they have been duped or tricked, Baumrind believes this may lead the participants to question what they have learned about themselves and to lead them to distrust individuals (e.g., social scientists) whom they might have previously trusted to provide valid information and advice. A cost to the profession is exacted because participants (and society at large) soon come to realize that psychologists are "tricksters" and not to be believed when giving instructions about research participation. If participants tend to suspect psychologists of lying, then one may question whether deception will work as it is intended by the researcher, a point raised earlier by Kelman (1972). Baumrind also argues that the use of deception reveals psychologists are willing to lie, which seemingly contradicts their supposed dedication to seeking truth. Finally, there is harm done to society because deception undermines people's trust in experts and makes them suspicious in general about all contrived events.

Of course, these are not the views of all psychologists (see Christensen, 1988; Kimmel, 1998). Milgram (1977), for instance, suggested that deceptive practices of psychologists are really a kind of "technical illusion" and should be permitted in the interests of scientific inquiry. After all, illusions are sometimes created in real-life situations in order to make people believe something. When listening to a radio program, people are not generally bothered by the fact that the thunder they hear or the sound of a horse galloping is merely a technical illusion created by a sound effects specialist. Milgram argues that technical illusions should be permitted in the case of scientific inquiry. We deceive children into believing in Santa Claus. Why cannot scientists create illusions in order to help them understand human behavior? Just as illusions are often created in real-life situations, in other situations, Milgram points out, there can be a suspension of a general moral principle. If we learn of a crime, we are ethically bound to report it to the authorities. On the other hand, a lawyer who is given information by a client must consider this information privileged even if it reveals that the client is guilty. Physicians perform very personal examinations of our bodies.

Although it is morally permissible in a physician's office, the same type of behavior would not be condoned outside the office. Milgram argues that, in the interest of science, psychologists should occasionally be allowed to suspend the moral principle of truthfulness and honesty.

Those who defend deception point to studies showing that participants typically do not appear to react negatively to being deceived (e.g., Christensen, 1988; Epley & Huff, 1998; Kimmel, 1996). Although people's "suspiciousness" about psychological research may increase, the overall effects seem to be small (see Kimmel, 1998). Nevertheless, the bottom line according to those who argue for the continued use of deception is well summarized by Kimmel (1998): "An absolute rule prohibiting the use of deception in all psychological research would have the egregious consequence of preventing researchers from carrying out a wide range of important studies" (p. 805). No one in the scientific community suggests that deceptive practices be taken lightly; however, for many scientists the use of deception is less noxious (to use Kelman's term) than doing without the knowledge gained by such studies.

2.b. Critically evaluate the criticism levelled at intelligence tests that almost all have an inherent bias toward emphasizing convergent, analytical, and scientific modes of thought. Give examples in support of your answer. 20

Reference:

A criticism levelled at intelligence tests is that almost all have an inherent bias toward emphasizing convergent, analytical, and scientific modes of thought. Thus, a person who emphasizes divergent, artistic, and imaginative modes of thought may be at a distinct disadvantage. Some critics have even stressed that the current approach to intelligence testing has become a social mechanism used by people with similar values to pass on educational advantages to children who resemble themselves. Not only might IQ tests tend to place creative individuals at a disadvantage but also they are limited in assessing nonacademically oriented intellectual abilities (Gardner, 1999; Snyderman & Rothman, 1987). Thus, social acumen, success in dealing with people, the ability to handle the concrete realities of the individual's daily world, social fluency, and specific tasks, such as purchasing merchandise, are not measured by any intelligence test (Greenspan & Driscoll, 1997; Sternberg, 1999). More succinctly, people are capable of many more cognitive abilities than can possibly be measured on an intelligence test.

Misunderstanding and potential misuse of intelligence tests frequently occur when scores are treated as measures of innate capacity. The IQ is not a measure of an innate fixed ability, nor is it representative of all problem-solving situations. It is a specific and limited sample, made at a certain point in time, of abilities that are susceptible to change because of a variety of circumstances. It reflects, to a large extent, the richness of an individual's past experiences. Although interpretation guidelines are quite clear in pointing out the limited nature of a test score, there is a tendency to look at test results as absolute facts reflecting permanent characteristics in an individual. People often want a quick, easy, and reductionist method to quantify, understand, and assess cognitive abilities, and the IQ score has become the most widely misused test score to fill this need.

An important limitation of intelligence tests is that, for the most part, they are not concerned with the underlying processes involved in problem solving. They focus on the final product or outcome rather than on the steps involved in reaching the outcome. They look at the "what " rather than the "how" (Embretson, 1986; E. Kaplan et al., 1999; Milberg et al., 1996). Thus, a low score on Arithmetic might

result from poor attention, difficulty understanding the examiner because of disturbances in comprehension, or low educational attainment. The extreme example of this "end product" emphasis is the global IQ score. When the examiner looks at the myriad assortment of intellectual abilities as a global ability, the complexity of cognitive functioning may be simplified to the point of being almost useless. The practitioner can apply labels quickly and easily, without attempting to examine the specific strengths and weaknesses that might make precise therapeutic interventions or knowledgeable recommendations possible. Such thinking detracts significantly from the search for a wider, more precise, and more process-oriented understanding of mental abilities.

A further concern about intelligence tests involves their limited usefulness in assessing minority groups with divergent cultural backgrounds. It has been stated that intelligence-test content is biased in favor of European American, middle-class values. Critics stress that minorities tend to be at a disadvantage when taking the tests because of deficiencies in motivation, lack of practice, lack of familiarity with culturally loaded items, and difficulties in establishing rapport. Numerous arguments against using intelligence tests for the assessment and placement of minorities have culminated in legal restrictions on the use of IQ scores. However, traditional defenses of IQ scores suggest that they are less biased than has been accused. For example, the removal of biased items has done little to alter overall test scores, and IQs still provide mostly accurate predictions for many minorities (see Chapter 2 for a further discussion). The issue has certainly not been resolved, but clinicians should continue to be aware of this dilemma, pay attention to subgroup norms, and interpret minority group IQ scores cautiously (see Lopez, 1997). Finally, many people feel that their IQs are deeply personal pieces of information. They would prefer that others, even a psychologist who is expected to observe confidentiality, not be allowed access to this information. This problem is further compounded when IQ scores might be given to several different persons, such as during legal proceedings or personnel selection.

Intelligence tests provide a number of useful and well-respected functions. They can adequately predict short-term scholastic performance; assess an individual's relative strengths and weaknesses; predict occupational achievement; reveal important personality variables; and permit the researcher, educator, or clinician to trace possible changes in an individual or population. However, these assets are helpful only if the limitations of intelligence tests are adequately understood and appropriately taken into consideration. They are limited in predicting certain aspects of occupational success and nonacademic skills, such as creativity, motivational level, social acumen, and success in dealing with people. Furthermore, IQ scores are not measures of an innate, fixed ability, and their use in classifying minority groups has been questioned. Finally, there has been an overemphasis on understanding the end product of cognitive functioning and a relative neglect in appreciating underlying cognitive processes.

2.c. What do you understand by reliability and validity of self-report measures? Discuss in detail the correspondence between reported and actual behavior. Give suitable examples.

15

Reference:

- Reliability refers to the consistency of measurement and is frequently assessed using the test–retest reliability method.
- Reliability is increased by including many similar items on a measure, by testing a diverse sample of individuals, and by using uniform testing procedures.
- Validity refers to the truthfulness of a measure: Does it measure what it intends to measure?

- Construct validity represents the extent to which a measure assesses the theoretical construct it is designed to assess; construct validity is determined by assessing convergent validity and discriminant validity.

Reliable self-report measures, like reliable observers or any other reliable measurements, are characterized by consistency. A reliable self-report measure is one that yields similar (consistent) results each time it is administered. Self-report measures must be reliable when making predictions about behavior. For example, in order to predict stress-related health problems, measures of individuals' life stress must be reliable. There are several ways to determine a test's reliability. One common method is to compute a test–retest reliability. Usually, test–retest reliability involves administering the same questionnaire to a large sample of people at two different times (hence, test and retest). For a questionnaire to yield reliable measurements, people need not obtain identical scores on the two administrations of the questionnaire, but a person's relative position in the distribution of scores should be similar at the two test times. The consistency of this relative positioning is determined by computing a correlation coefficient using the two scores on the questionnaire for each person in the sample. A desirable value for test–retest reliability coefficients is .80 or above, but the size of the coefficient will depend on factors such as the number and types of items.

A self-report measure with many items to measure a construct will be more reliable than a measure with few items. For example, we are likely to have unreliable measures if we try to measure a baseball player's hitting ability based on a single time at bat or a person's attitude toward the death penalty based on a single question on a survey. The reliability of our measures will increase greatly if we average the behavior in question across a large number of observations—many at-bats and many survey questions (Epstein, 1979). Of course, researchers must walk a fine line between too few items and too many items. Too many items on a survey can cause respondents to become tired or careless about their responses.

In general, measurements will also be more reliable when there is greater variability on the factor being measured among the individuals being tested. Often the goal of measurement is to determine the extent to which individuals differ. A sample of individuals who vary a great deal from one another is easier to differentiate reliably than are individuals who differ by only a small amount. Consider this example. Suppose we wish to assess soccer players' ability to pass the ball effectively to other players. We will be able to differentiate more reliably good players from poor players if we include in our sample a wider range of players—for example, professionals, high school players, and peewee players. It would be much harder to differentiate players reliably if we tested only professional players—they'd all be good! Thus, a test is often more reliable when administered to a diverse sample than when given to a restricted sample of individuals.

A third and final factor affecting reliability is related to the conditions under which the questionnaire is administered. Questionnaires will yield more reliable measurements when the testing situation is free of distractions and when clear instructions are provided for completing the questionnaire. You may remember times when your own test performance was hindered by noise or when you weren't sure what a question was asking. The reliability of a survey measure is easier to determine and to achieve than the validity of a measure. The definition of validity is deceptively straightforward—a valid questionnaire measures what it is intended to measure. Have you ever heard students complain that questions on a test didn't seem to address the material covered in class? This is an issue of validity.

At this point, we will focus on construct validity, which is just one of the many ways in which the validity of a measurement is assessed. The construct validity of a measure represents the extent to which it measures the theoretical construct it is designed to measure. One approach to determining the construct validity of a test relies on two other kinds of validity: convergent validity and discriminant validity. These concepts can best be understood by considering an example.

Lucas, Diener, and Suh (1996) note that psychologists are increasingly examining factors such as happiness, life satisfaction, self-esteem, optimism, and other indicators of well-being. However, it's not clear whether these different indicators all measure the same construct (e.g., well-being) or whether each is a distinguishable construct. Lucas and his colleagues conducted several studies in which they asked individuals to complete questionnaire measures of these different indicators of well-being. For our purposes we will focus on a portion of their data from their third study, in which they asked participants to complete three scales: two life satisfaction measures, the Satisfaction with Life Scale (SWLS) and a 5-item Life Satisfaction measure (LS-5); and a measure of Positive Affect (PA). At issue in this example is whether the construct of life satisfaction—the quality of being happy with one's life—can be distinguished from being happy more generally (positive affect).

The data in Table 5.1 are presented in the form of a correlation matrix. A correlation matrix is an easy way to present a number of correlations. Look first at the values in parentheses that appear on the diagonal. These parenthesized correlation coefficients represent the values for the reliability of each of the three measures. As you can see, the three measures show good reliability (each is above .80). Our focus, however, is on measuring the construct validity of "life satisfaction," so let's look at what else is in Table 5.1.

It is reasonable to expect that scores on the Satisfaction with Life Scale (SWLS) should correlate with scores on the 5-item Life Satisfaction measure; after all, both measures were designed to assess the life satisfaction construct. In fact, Lucas et al. observed a correlation between these two measures of .77, which indicates that they correlate as expected. This finding provides evidence for convergent validity of the measures; the two measures converge (or "go together") as measures of life satisfaction.

The case for the construct validity of life satisfaction can be made even more strongly when the measures are shown to have discriminant validity. As can be seen in Table 5.1, the correlations between the Satisfaction with Life Scale (SWLS) and Positive Affect (.42) and between the 5-item Life Satisfaction measure (LS-5) and Positive Affect (.47) are lower. These findings show that life satisfaction measures do not correlate as well with a measure of another theoretical construct—namely, positive affect. The lower correlations between the life satisfaction tests and the positive affect test indicate that different constructs are being measured. Thus, there is evidence for discriminant validity of the life satisfaction measures because they seem to "discriminate" life satisfaction from positive affect—being satisfied with one's life is not the same as general happiness. The construct validity of life satisfaction gains support in our example because there is evidence for both convergent validity and discriminant validity.

Correspondence Between Reported and Actual Behavior

- Survey research involves reactive measurement because individuals are aware that their responses are being recorded.
- Social desirability refers to pressure that respondents sometimes feel to respond as they "should" believe rather than how they actually believe.

- Researchers can assess the accuracy of survey responses by comparing these results with archival data or behavioral observations.

Regardless of how carefully survey data are collected and analyzed, the value of these data depends on the truthfulness of respondents' answers to the survey questions. Should we believe that people's responses on surveys reflect their true thoughts, opinions, feelings, and behavior? The question of the truthfulness of verbal reports has been debated extensively, and no clearcut conclusion has emerged. In everyday life, however, we regularly accept the verbal reports of others as valid. If a friend tells you she enjoyed reading a certain novel, you may ask why, but you do not usually question whether the statement accurately reflects your friend's feelings. There are some situations in everyday life, however, when we do have reason to suspect the truthfulness of someone's statements. When looking for a used car, for instance, we might not always want to trust the "sales pitch" we receive. Generally, however, we accept people's remarks at their face value unless we have reason to suspect otherwise. We apply the same standards to the information we obtain from survey respondents.

By its very nature, survey research involves reactive measurement. Respondents know their responses are being recorded, and they may also suspect their responses may prompt some social, political, or commercial action. Hence, pressures are strong for people to respond as they "should" believe, not as they actually believe. The term often used to describe these pressures is social desirability (the term "politically correct" refers to similar pressures). For example, if respondents are asked whether they favor giving help to the needy, they may say "yes" because they believe this is the most socially acceptable attitude to have. In survey research, as was true with observational research, the best protection against reactive measurement is to be aware of its existence.

Sometimes researchers can examine the accuracy of verbal reports directly. For example, Judd, Smith, and Kidder (1991) describe research by Parry and Crossley (1950) wherein responses obtained by experienced interviewers were subsequently compared with archival records for the same respondents kept by various agencies. Their comparisons revealed that 40% of respondents gave inaccurate reports to a question concerning contributions to United Fund (a charitable organization), 25% reported they had registered and voted in a recent election (but they did not), and 17% misrepresented their age. A pessimist might find these figures disturbingly high, but an optimist would note that a majority of respondents' reports were accurate even when social desirability pressures were high, as in the question pertaining to charitable contributions.

Another way researchers can assess the accuracy of verbal reports is by directly observing respondents' behavior. An experiment done by Latané and Darley (1970) illustrates this approach. They found that bystanders are more likely to help a victim when the bystander is alone than when other witnesses are present. Subsequently, a second group of participants was asked whether the presence of others would influence the likelihood they would help a victim. They uniformly said that it would not. Thus, individuals' verbal reports may not correspond well to behavior (see Figure 5.6). Research findings such as these should make us extremely cautious of reaching conclusions about people's behavior solely on the basis of verbal reports. Of course, we should be equally cautious of reaching conclusions about what people think solely on the basis of direct observation of their behavior. The potential discrepancy between observed behavior and verbal reports illustrates again the wisdom of a multi-method approach in helping us identify and address potential problems in understanding behavior and mental processes.

3.a. What are the advantages and disadvantages of psychological assessment done via web based applications? 15

Reference:

Scalability and convenient availability to a broad audience of consumers make Internet testing attractive to practicing psychologists. Several new companies have been formed because Internet testing creates business opportunities; traditional test publishers have also entered this playing field. Individual practicing psychologists can also utilize the Internet. For example, routine assessment activities can be conducted via the Internet, leaving psychologists to devote their time to interpretation and feedback. Wilson Learning Corporation explicitly considers what psychologists do best and what computers do best as they design an assessment system for a customer (Burroughs et al., 1999). Psychologists prepare for face-to-face meetings, conduct feedback sessions, and write final interpretive reports. Computers store scores, mechanically combine scores, and generate feedback report support information.

Psychometric Advantages

Computerized tests provide some psychometric advantages in comparison to paper-and-pencil assessments. In fact, considerable research has been conducted to document and demonstrate these advantages. A brief summary is provided here; more detail can be found in Sands, Waters, and McBride (1997) and Drasgow and Olson-Buchanan (1999).

An Internet test and assessment provides more accurate scoring compared with a traditional paper-and-pencil test. Optical scanning of paper test forms encounters difficulties with stray pencil marks, incomplete erasures, and insufficiently darkened answers. In computerized testing, an examinee enters a response, the response is displayed on-screen, and the examinee is provided an opportunity to change the answer. Suppose an examinee has selected “B” as his or her response. The computer monitor will then display a darkened circle next to option “B” and will allow the examinee to change the response or proceed to the next item. If the examinee goes to the next item, well-designed software will correctly record and score the “B” response. By eliminating optical scanning, a significant source of errors is removed.

Internet testing and assessment is especially well suited for the use of item response theory (IRT; Hambleton & Swaminathan, 1985; Hulin, Drasgow, & Parsons, 1983; Lord, 1980) . For example, computerized adaptive tests (CATs) that tailor difficulty to the ability level of each examinee can be efficiently delivered through this medium. In this process, IRT technology would be used to select which items are given so that they are of appropriate difficulty for each examinee. Internet assessment can also offer the potential for assessing abilities and skills not easily assessed by paper-and-pencil methods. For example, Vispoel (1999) developed a computerized assessment of musical aptitude; Ackerman, Evans, Park, Tamassia, and Turner (1999) created a dermatological test that allows examinees to pan in and out as they examine color images of skin disorders; and Drasgow, Olson-Buchanan and Moberg (1999) developed an assessment that uses video clips to assess respondents’ conflict resolution skills. These and many other needs can be effectively met using computerized technology that can be delivered via the Internet.

New Problems yet Old Issues

Test/Client Integrity

In the same way that we do not allow clients to take tests at home, given that they might not take them privately, Internet testing encounters this old problem with a new twist. When the goals of the test taker differ from the goals of the test user, it is important to confirm the identity of the person answering items. The simplest and most effective method is to require test takers to go to a secure test site and show a government issued photo ID such as a driver's license or passport. Of course, test administration at such sites is inconvenient and expensive. When a test or assessment is not administered at a secure test center, there are a number of ways to check a test taker's identity (e.g., "What is your mother's maiden name?"). Unfortunately, such methods can easily be circumvented; a more talented accomplice can sit with the supposed test taker and provide answers to items.

Segall (2001) suggested a clever means of confirming the validity of test takers' administered tests remotely via the Internet. He has proposed Internet administration of the lengthy Armed Services Vocational Aptitude Battery (ASVAB) enlistment test used by the U.S. military. Segall's idea is that individuals could take the ASVAB at their convenience in nonsecure locations. Individuals who obtain scores qualifying for enlistment would then travel to secure test centers, where they would be administered a much shorter confirmation test composed of highly discriminating items. A statistical procedure developed by Segall would then be used to check whether the test taker's original responses are consistent with the responses from the confirmation test. This method was found to be very effective at detecting cheating in a simulation study. A combination of informing examinees that a confirmation test will be administered as well as applying Segall's (2001) statistical analysis may also prove to be effective in discouraging cheating.

Technical Issues

Host/Server Hardware Considerations

Internet testing poses a number of technical considerations that are not at issue in conventional testing. Prominent among these considerations are characteristics of the hardware involved, which can be generally classified into Host and Client issues. On the Host side, Internet test developers should have many of the same concerns that any Internet delivered application might have. Primary among these concerns should be high availability of the host network and hardware, a high degree of fault tolerance through all components of the host configuration, data protection through frequent backup and secure storage procedures, and sufficient network bandwidth and hardware capacity for the testing requirements. Many useful checklists are available for evaluating the quality of services offered by prospective host providers. Hardware should be Enterprise class servers with built in redundancy in order to minimize potential impacts of hardware failures. Most providers will recommend and lease the appropriate hardware as part of the service agreement. Most host provider plans limit the amount of data that will be transferred to and from a site within a monthly period. Clients are charged for data exchanges that exceed this maximum data transfer rate. Therefore, it is important to consider the amount of testing (inbound) and, if appropriate, reporting (outbound) traffic that will be conducted in a monthly period.

The amount of data storage capacity that will be required will depend on testing volume, the amount of data collected per test and the duration that the data need to remain accessible. While storage costs are relatively inexpensive, these costs can be reduced by periodically archiving aged data. Virtually all reputable providers offer some type of network monitoring; however, many providers offer active monitoring services for additional fees. These services are intended to identify and correct network or site problems immediately when and, to the extent possible, before they occur. Reputable providers

should be able to provide a guarantee of reliability for their site and should be able to provide historical reliability data.

A major consideration for any Internet-delivered application is available network bandwidth. Host providers will guarantee a minimum amount of network bandwidth available to handle all traffic coming to a site for all applications operating on the site. Bandwidth limitations are more typically observed on the client side. While client side broadband is becoming more widely available due to decreased costs of high-speed services, conventional dial-up 56K connections still represent the vast majority of Internet traffic, particularly in the consumer and home markets. While Internet connectivity has significantly increased within the schools, much of that connectivity is dial-up. Items that include high quality color graphics will be constrained by bandwidth available through conventional dial-up 56K connections. Bandwidth constraints and bottlenecks can be a significant threat to the standardized delivery of an assessment; the extent to which variability in network performance can be tolerated depends on the nature of the assessment and the degree to which variation in time between stimulus or item presentation affects examinee performance. Where strict control is necessary, a locally installed stimulus control applet should be used as discussed below.

Client Hardware Considerations

The required client hardware configuration will largely depend on the type of test and items being administered. In general, tests that display items with graphics, present timed stimuli, or collect response time data require greater consideration and specificity of the client hardware configuration. Graphics are memory intensive and often require workstations with additional Random Access Memory (RAM) and Video RAM (VRAM) to operate efficiently. The graphics and stimulus presentation characteristics of the test will constrain the minimum hardware necessary to present the items within the parameters defined for the test. Due to inherent constraints in an application's ability to control the timing of the delivery of Internet packets to a client, accurate presentation of timed stimuli and recording of response time data require a stimulus control application running locally on the client's workstation. Timing applications control item presentation, save examinees' responses at the client's workstation, and send response data back to the host application. Although such applications are generally small, the resource requirements of the application should be considered in the overall hardware requirements for the client. Peripherals will generally include a mouse and possibly a microphone if speech input is required or is an option.

The hardware characteristics of the client's display require special consideration where timed stimuli are presented. The first consideration is the display refresh rate, which indicates how many times per second the screen gets redrawn from top to bottom. When presentation of stimuli must be controlled to the millisecond, applications must take into consideration or attempt to control the refresh rate of the display. Furthermore, while most individuals cannot perceive any flicker on displays operating at refresh rates in excess of 72 Hz, a small percentage of individuals can detect flicker at rates as high as 85 Hz on conventional (non-LCD) displays. Because flicker is known to cause discomfort, it is likely to have some effect on performance on items that require visual decoding, particularly if performance is timed.

The standardized presentation of stimuli presents a particular challenge in the development of computer-administered assessments. While on the surface it may seem relatively easy to present exactly the same picture or graphic on any computer display, it is, in fact, almost impossible to do so without a recalibration of the display environment. All displays present the same colors at slightly different hues and, based on user preferences, have different contrast and brightness settings.

Furthermore, the actual size of an object will vary depending on the resolution settings of the display. Text may be displayed differently depending on the fonts installed on the user's workstation. While such variations may not be of concern for some assessments, they can raise serious questions regarding the standardized administration of items with graphic stimuli. To address the issue of stimulus color, test developers should include a methodology and template that would allow the user to calibrate the display to a template standard. To control the standard size of a stimulus on different displays operating at different resolutions, the developer must either a) require the examinee to operate the test at a particular resolution or b) dynamically adjust the size of the stimulus based on the resolution setting. These technical issues will require careful consideration in the Internet testing environment.

Test Security

Levels of security can range from highly secure and restrictive (e.g., high-stakes testing programs) to unsecured and permissive (low stakes testing). As might be expected, the greater the level of security, the higher the cost for implementing and maintaining an application. The level of security implemented for a given test or test site should be appropriately matched to the usage of the test. Secure test environments should use a 3-tier server model. Within this model, the test system is actually made up of 3 independent servers: an Internet server, a test application server, and a database server. It is imperative that the application server is solely dedicated to the test application. In order to maximize the security of client data, a separate data server should be maintained behind a secure firewall. This configuration reduces the possibility of unauthorized intrusions into client test data. If scoring and reporting services are required, it is recommended that these applications be placed on yet a fourth server in the middle tier with the application server in order to minimize processing bottlenecks that may affect the test application or data access. Regular and frequent backups of all collected data should be conducted and the provider should be able to give prospective customers a detailed disaster recovery plan. Redundancy allows a site to continue to operate even if one of its components completely fails. A reputable provider will have redundancy on all systems throughout its site including incoming and outgoing communications lines. As with any secure application, client and administrator password formats need to be robust (non-trivial) and actively maintained. Finally, server traffic should be actively and continuously monitored for intrusions.

On the client side, one of the most important security considerations is the prevention of unauthorized copying by the examinee or an observer and printing of test items. This can partially be achieved within a browser by disabling access to menu selections such as cut, copy, paste, export, save, save as, print, print screen, etc. Hot keys and right mouse context menu selections should also be disabled. However, it is only possible to partially secure items by controlling browser functions. Even with such controls in place, it is still possible for more technically knowledgeable examinees to make use of operating system features and other applications to capture items from the screen. Therefore, where full client-side security is required, it is necessary to install a test security agent on the client's desktop, which completely prohibits an examinee from dropping out of the test application while it is in operation. Such an application prevents users from launching screen recorders, word processors, e-mail applications, and any other unrelated application that may be used to compromise the security of test items.

Issues for Special Populations

The delivery of psychological tests through the Internet provides the opportunity to meet the needs of a wide variety of individuals, in particular, important special populations including people with disabling conditions and culturally and linguistically diverse persons.

People with Disabling Conditions

A critical issue in determining appropriate accommodations for a person with a disability is demonstrating the clear relationship between the individual's deficit and the nature of the accommodation. The challenge of determining the type of accommodations required for Internet-based assessment arises in part because little is known about the unique aspects of testing in this format. Although many of the accommodations developed for paper-and-pencil testing can be used for Internet assessments, new issues will likely arise. As psychologists begin to make recommendations to institutions on behalf of individuals with disabilities or on behalf of institutions attempting to design fair testing practices for groups of individuals with disabilities, it is important to consider new types of accommodations to address the unique problems inherent in Internet assessment.

Accommodations may be considered in terms of operating at the level of the individual or at the level of the group. At the individual level, adequate accommodations include alterations in the testing environment. Although this is standard practice with paper-and-pencil testing, unique challenges may be encountered with Internet assessments because the computer may be permanently affixed in one position. Adjusting the height and placement of the table on which the computer sits is critical for an individual in a wheelchair. For some disabilities, accommodations require alterations to test administration itself rather than alterations to the environment. For example, a reader is often recommended for individuals who are sight-impaired or who have a specific reading disability (e.g., dyslexia). During Internet-based assessment, there is likely to be a new vocabulary to describe the spatial layout of the material and the actions taken by the reader (e.g., rather than stating, "I am filling in answer a on the scantron sheet," the reader might say, "I am clicking choice a on the answer screen"). These accommodations are not unique to testing over the Internet, but are unique to testing on a computer platform, the frequency of which will likely increase as Internet technology advances.
At the group level, a lack of equal access to technology may result in poorer test performance for some groups. Households with lower incomes have reduced access to computers and therefore the Internet. Assessment over the Internet, therefore, may be confounded by the novelty of the format. For example, during cognitive testing, individuals who are less familiar with computers will have a greater cognitive load due to divided attention than individuals who are familiar with computers. Further, a lack of familiarity with the security and privacy features of the Internet may influence performance. In this sense, low access to computers may be viewed as a disability that requires accommodations to ensure fair testing.

Culturally and Linguistically Diverse Groups

Many culturally and linguistically diverse groups, including Latinos and African Americans, have been among the last to connect to the Internet because of economic and/or access issues. Yet, the number of people from these and other minority groups that have access to the Internet is increasing dramatically. For these groups, the Internet is proving to be a tool that connects them to their country of origin, resources in a particular language or dialect, and so forth.

Like the majority Euroamerican population, members of these groups have also begun to access the Internet for information related to mental health and psychology. For example, it is not unusual for a Spanish-speaking Latino to seek out information about a particular mental condition or even a psychological test instrument through the Internet. The person is now likely to find information in Spanish, usually from an Internet site in Latin America. Similarly, the person may seek information about

a particular test that he or she is about to take that will be administered by a psychologist (e.g., for employment screening or child custody purposes).

There remain many unanswered questions regarding the psychological testing and assessment of these groups via the Internet. In many ways, these issues are similar to concerns related to test use with culturally diverse or minority groups (e.g., fair assessment). For example, it is unclear if it is necessary to have separate norms, including norms for minorities, for an instrument that is administered via the Internet versus administered in the traditional manner. Also, a review of various webpages indicated that many instruments are poorly translated, or have been modified for use with Latinos in the United States or Latin America, by Spanish-speaking professionals usually outside of the United States. Additionally, older measures, such as the MMPI, can still be found in Spanish despite the appearance of recent translations of the MMPI-2 that are superior to the translations of the older MMPI. People may use old and outdated instruments in a manner that is inappropriate or problematic, which may then result in negative consequences for clients and the public.

3.b. "It has been extremely difficult to find independent operationalizations of personality traits and personality disorder constructs that are consistent across assessment devices." Discuss and suggest measures to improve the assessment of personality disorders with clinical interview.

20

Reference:

MISUSE OF THE INTERVIEW

Many clinicians have such faith in the clinical interview (and in their own skills) that interviews can be misused. One such current venue is that occasioned by managed care constraints that often preclude the use of other methods (e.g., psychological tests, collateral interviews) that would either add incremental validity in clinical practice or possibly confirm hypotheses gleaned from the interview itself. Psychologists need to guard against such practices and to advocate for the best possible psychological practice for a given problem.

WHAT NEEDS TO BE DONE?

Most problems in any classification system of personality disorders are endemically and systematically related to the issue of construct validity. One continuing problem in assessment is that it has been extremely difficult to find independent operationalizations of personality traits and personality disorder constructs that are consistent across assessment devices. Convergent validity between self-report measures and interview-based assessments range from poor to modest.

Median correlations between structured psychiatric interviews range from .30 to .50; median correlations between self-report measures range from .39 to .68; and median correlations between questionnaires and structured interviews range from .08 to .42. Consistently moderate correlations between questionnaires have been reported for the diagnoses of borderline, dependent, passive-aggressive, and schizotypal personality disorders. Better convergent validity between questionnaires and clinical interviews has been found with diagnoses of borderline and avoidant personality disorders. For clinical interviews, consistently good convergence has been found for only avoidant personality disorder (Clark, Livesley, & Morey, 1997).

Although method variance and general measurement error may account for some of the findings, the real problem is a lack of clear and explicit definitions of the diagnostic constructs and behavioral anchors that explicate examples of specific items that define the disorder and aid the diagnostician (and researcher) to diagnose the disorder. For example, with a criteria set of eight items, of which five are need to make a diagnosis of borderline personality disorder, there are 95 different possible sets of symptoms that would qualify for this diagnosis (Widiger, Frances, Spitzer, & Williams, 1988). Are there really 95 different types of borderline personality disorders? Obviously not! The criteria merely reflect our confusion on the diagnosis itself. This situation is clearly absurd and serves to illustrate the problems that accrue when the construct and defining criteria are obfuscating. Similarly, the problem of a patient's meeting two or more of the personality disorder diagnoses will continue to exist, due largely to definitional problems. A clinician can reduce this bias somewhat by carefully assessing all criterion symptoms and traits, but the problem in the criteria themselves remains. Associated with the need for more conceptual clarity is the need to reduce terminological confusion inherent in the criteria set. For example, when does spontaneity become impulsivity?

There is also a need for improved accuracy in clinician diagnosis. Evidence exists that trained interviewers are able to maintain high levels of interrater reliability, diagnostic accuracy, and interviewing skills, such that quality assurance procedures should be systematically presented in both research and clinical settings (Ventura, Liberman, Green, Shaner, & Mintz, 1998). For example, 18 clinical vignettes were sent to 15 therapists, along with DSM personality disorder criteria sets. Fourteen of the vignettes were based on DSM criteria and 14 were made up and suggested diagnoses of no personality disorder. Results showed and 82% rate of agreement in diagnosis. This type of procedure can be cost-effective to establish and to assess continuing competency in diagnosing personality disorders (Gude, Dammen, & Frilis, 1997).

Some have called for the explicit recognition of dimensional structures in official classification systems because such structures recognize the continuous nature of personality functioning (Widiger, 2000). Millon (2000) called for adoption of a coherent classification-guiding theory. However, it is unlikely that theorists would ever agree as to the parsimonious system to be adopted. Others suggested the use of prototype criteria sets to define pure cases (Oldham & Skodol, 2000; Westen & Shedler, 2000), but such prototypes might only rarely be observed in clinical practice, and hence such a system would live little practical utility, although Millon (2000) has persuasively argued otherwise. He has also called for the inclusion of personality disorder subtypes hierarchically subsumed under the major prototypes. Still others call for the inclusion of level of functioning (e.g., mild, moderate, severe), within the personality diagnostic system.

Hunter (1998) suggested that personality disorder criteria be rewritten from the patient's perspective. This would have the effect of removing negative language and provide a simplified and more straightforward and objective means of assessment. Cloninger (2000) suggested that personality disorders be diagnosed in terms of four core features: (a) low affective stability, (b) low self-directedness, (c) low cooperativeness, and (d) low self-transcendence. Perhaps a blend of both the categorical and dimensional systems is preferable. The clinician could diagnose a personality disorder in a categorical system, reference personality (disorder) traits that are specific to the individual, and include a specifier that depicts level of functioning.

The aforementioned suggestions apply more to assessing personality disorders with interview. Karg and Wiens (1998) have recommended the following activities to improve clinical interviewing in general:

- Prepare for the initial interview. Get as much information beforehand as possible; be well-informed about the patient's problem area. This preparation will allow you to ask more meaningful questions. There may be important information learned from records or from other sources that warrant more detailed inquiry within the assessment interview. If this information is not available to you at the time of the interview, the opportunity for further inquiry may be lost.
- Determine the purpose of the interview. Have a clear understanding of what you want to accomplish. Have an interview structure in mind and follow it.
- Clarify the purpose and parameters of the interview to the client. If the client has a good understanding of what is trying to be accomplished, his or her willingness to provide you with meaningful information should increase.
- Conceptualize the interview as a collaborative process. Explain how the information will be used to help the client with his or her situation.
- Truly hear what the interviewee has to say. This may be accomplished by using active listening and by clarifying the major points of understanding with the interviewee during the interview.
- Use structured interviews. These interviews promote a systematic review of content areas and are more reliable.
- Encourage the client to describe complaints in concrete behavioral terms. This will help the psychologist to understand the client better and will provide examples of the potential problematic behavior in relevant context.
- Complement the interview with other assessment methods, particularly psychological testing. This may provide both convergent and incremental validity.
- Identify the antecedents and consequences of problem behaviors. This will provide more targeted interventions.
- Differentiate between skill and motivation. Some patients may have the desire to accomplish goals that are beyond their capacities, and vice versa.
- Obtain base rates of behaviors. This will provide a benchmark for later assessment of progress.
- Avoid expectations and biases. Self-monitor your own feelings, attitudes, beliefs, and counter-transference to determine whether you are remaining objective.
- Use a disconfirmation strategy. Look for information that might disprove your hypothesis.
- Counter the fundamental attribution error. This occurs when the clinician attributes the cause of a problem to one set of factors, when it may be due to other sets of factors.
- Combine testing with interviewing mechanistically. This is because combining data from interview with data from other sources will be more accurate and valid than data from one source alone.
- Delay reaching decisions while the interview is being conducted. Don't rush to judgments or to conclusions.
- Consider the alternatives. Offering a menu of choices and possibilities should engender greater client acceptance of goals and interventions.
- Provide a proper termination. Suggest a course of action, a plan of intervention, recommended behavioral changes, and so on, that the person can take with them from the interview. It is pointless for the psychologist to conduct thorough assessments and evaluations without providing some feedback to the client.

The future will no doubt actively address, research, refine, and even eliminate some of these problems discussed in this chapter. We can look forward to improvements in diagnostic criteria, improved clarity in criteria sets, increased training so that clinicians can self-monitor and reduce any potential biases in diagnostic decision-making, and take the role of culture more into account in the evaluation of clients. I

hope that these advances will lead to improvements in therapeutic interventions designed to ameliorate pathological conditions.

3.c. "A statistically significant outcome is one that has only a small likelihood of occurring if the null hypothesis were true." Discuss in detail with the help of suitable examples. 15

Reference:

NULL HYPOTHESIS SIGNIFICANCE TESTING (NHST)

- Null hypothesis testing is used to determine whether mean differences among groups in an experiment are greater than the differences that are expected simply because of error variation.
- The first step in null hypothesis testing is to assume that the groups do not differ—that is, that the independent variable did not have an effect (the null hypothesis).
- Probability theory is used to estimate the likelihood of the experiment's ob served outcome, assuming the null hypothesis is true.
- A statistically significant outcome is one that has a small likelihood of occurring if the null hypothesis were true.
- Because decisions about the outcome of an experiment are based on probabilities, Type I (rejecting a true null hypothesis) or Type II (failing to reject a false null hypothesis) errors may occur.

Statistical inference is both inductive and indirect. It is inductive because we draw general conclusions about populations on the basis of the specific samples we test in our experiments, as we do when constructing confidence intervals. However, unlike the approach using confidence intervals, this form of statistical inference is also indirect because it begins by assuming the null hypothesis. The null hypothesis (H0) is the assumption that the independent variable has had no effect. Once we make this assumption, we can use probability theory to determine the likelihood of obtaining this difference (or a larger difference) observed in our experiment IF the null hypothesis were true. If this likelihood is small, we reject the null hypothesis and conclude that the independent variable did have an effect on the dependent variable. Outcomes that lead us to reject the null hypothesis are said to be statistically significant. A statistically significant outcome means only that the difference we obtained in our experiment is larger than would be expected if error variation alone (i.e., chance) were responsible for the outcome (see Box 12.1).

A statistically significant outcome is one that has only a small likelihood of occurring if the null hypothesis were true. But just how small is small enough? Although there is no definitive answer to this important question, the consensus among members of the scientific community is that outcomes associated with probabilities of less than 5 times out of 100 (or .05) if the null hypothesis were true are judged to be statistically significant. The probability we elect to use to indicate an outcome is statistically significant is called the level of significance. The level of significance is indicated by the Greek letter alpha (_). Thus, we speak of the .05 level of significance, which we report as _ _ .05. Just what do our results tell us when they are statistically significant? The most useful information we gain is that we know that something interesting has happened. More specifically, we know that the smaller the exact probability of the observed outcome, the greater is the probability that an exact replication will produce a statistically significant finding. But we must be careful what we mean by this statement. Researchers sometimes mistakenly say that when a result occurs with p _ .05, "This outcome will be obtained 95/100

times if the study is repeated." This is simply not true. Simply achieving statistical significance (i.e., p _ .05) does not tell us about the probability of replicating the results. For example, a result just below .05 probability (and thus statistically significant) has only about a 50:50 chance of being statistically significant (i.e., p _ .05) if replicated exactly (Greenwald et al., 1996). On the other hand, knowing the exact probability of the results does convey information about what will happen if a replication were done. The smaller the exact probability of an initial finding, the greater the probability that an exact replication will produce a statistically significant (p _ .05) finding (e.g., Posavac, 2002). Consequently, and as recommended by the American Psychological Association (APA), always report the exact probability of results when carrying out NHST.

Strictly speaking, there are only two conclusions possible when you do an inferential statistics test: Either you reject the null hypothesis or you fail to reject the null hypothesis. Note that we did not say that one alternative is to accept the null hypothesis. Let us explain. When we conduct an experiment and observe the effect of the independent variable is not statistically significant, we do not reject the null hypothesis. However, neither do we necessarily accept the null hypothesis of no difference. There may have been some factor in our experiment that prevented us from observing an effect of the independent variable (e.g., ambiguous instructions to subjects, poor operationalizations of the independent variable). As we will show later, too small a sample often is a major reason why a null hypothesis is not rejected. Although we recognize the logical impossibility of proving that a null hypothesis is true, we also must have some method of deciding which independent variables are not worth pursuing. NHST can help with that decision. A result that is not statistically significant suggests we should be cautious about concluding that the independent variable influenced behavior in more than a trivial way. At this point you will want to seek more information, for example, by noting the size of the sample and the effect size (see the next section, "Experimental Sensitivity and Statistical Power").

There is a troublesome aspect to the process of statistical inference and our reliance on probabilities for making decisions. No matter what decision you reach, and no matter how carefully you reach it, there is always some chance you are making an error. The two possible "states of the world" and the two possible decisions an experimenter can reach are listed in Table 12.1. The two "states of the world" are that the independent variable either does or does not have an effect on behavior. The two possible correct decisions the researcher can make are represented by the upper-left and lower-right cells of the table. If the independent variable does have an effect, the researcher should reject the null hypothesis; if it does not, the researcher should fail to reject the null hypothesis.

The two potential errors (Type I error and Type II error) are represented by the other two cells of Table 12.1. These errors arise because of the probabilistic nature of statistical inference. When we decide an outcome is statistically significant because the outcome's probability of occurring under the null hypothesis is less than .05, we acknowledge that in 5 out of every 100 tests, the outcome could occur even if the null hypothesis were true. The level of significance, therefore, represents the probability of making a Type I error: rejecting the null hypothesis when it is true. The probability of making a Type I error can be reduced simply by making the level of significance more stringent, perhaps .01. The problem with this approach is that it increases the likelihood of making a Type II error: failing to reject the null hypothesis when it is false. The problem of Type I errors and Type II errors should not immobilize us, but it should help us understand why researchers rarely use the word "prove" when they describe the results of an experiment that involved tests of statistical significance. Instead, they describe the results as "consistent with the hypothesis," or "confirming the hypothesis," or "supporting the hypothesis." These tentative statements are a way of indirectly acknowledging that the possibility of making a Type I error or a Type II error always exists. The .05 level of significance represents a

compromise position that allows us to strike a balance and avoid making too many of either type of error. The problem of Type I errors and Type II errors also reminds us that statistical inference can never replace replication as the best test of the reliability of an experimental outcome.

4.a. Elaborate the characteristics of true experiments. Discuss the obstacles in conducting true experiments in natural settings. 15

Reference:

In true experiments, researchers manipulate an independent variable with treatment and comparison condition(s) and exercise a high degree of control (especially through random assignment to conditions). As we have noted, although many everyday activities (such as altering the ingredients of a recipe) might be called experiments, we would not consider them "true" experiments in the sense in which experimentation has been discussed in this textbook. Analogously, many "social experiments" carried out by the government and those that are conducted by company officials or educational administrators are also not true experiments. A true experiment is one that
leads to an unambiguous outcome regarding what caused an event.

True experiments exhibit three important characteristics:

1. In a true experiment some type of intervention or treatment is implemented.
2. True experiments are marked by the high degree of control that an experimenter has over the arrangement of experimental conditions, assignment of participants, systematic manipulation of independent variables, and choice of dependent variables. The ability to assign participants randomly to experimental conditions is often seen as the most critical defining characteristic of the true experiment (Judd, Smith, & Kidder, 1991).
3. Finally, true experiments are characterized by an appropriate comparison. Indeed, the experimenter exerts control over a situation to establish a proper comparison to evaluate the effectiveness of a treatment. In the simplest of experimental situations, this comparison is one between two comparable groups that are treated exactly alike except for the variable of interest. When the conditions of a true experiment are met, any differences in a dependent variable that arise can logically be attributed to the differences between levels of the independent variable. There are differences, however, between true experiments done in natural settings and experiments done in a laboratory.

Obstacles in Conducting True Experiments in Natural Settings

- Researchers may experience difficulty obtaining permission to conduct true experiments in natural settings and gaining access to participants.
- Although random assignment is perceived by some as unfair because it may deprive individuals of a new treatment, it is still the best way and fairest way to determine if a new treatment is effective.

Experimental research is an effective tool for solving problems and answering practical questions. Nevertheless, two major obstacles often arise when we try to carry out experiments in natural settings. The first problem is obtaining permission to do the research from individuals in positions of authority. Unless they believe that the research will be useful, school board presidents and government and business leaders are unlikely to support research financially or otherwise. The second, and often more

pressing, obstacle to doing experiments in natural settings is the problem of access to participants. This problem can prove especially troublesome if participants are to be randomly assigned to either a treatment group or a comparison group.

Random assignment to conditions appears unfair at first—after all, random assignment requires that a potentially beneficial treatment be withheld from some participants. Suppose that a new approach to the teaching of foreign languages was to be tested at your college or university. Suppose further that, when you went to register for your next semester's classes, you were told that you would be randomly assigned to one of two sections taught at the time you selected—one section involving the old method and one involving the new method. How would you react? Your knowledge of research methods tells you that the two methods must be administered to comparable groups of students and that random assignment is the best way to ensure such comparability. Nonetheless, you might be tempted to feel that random assignment is not fair, especially if you are assigned to the section using the old (old-fashioned?) method. Let's take a closer look at the fairness of random assignment.

If those responsible for selecting the method of foreign language instruction already knew that the new method was more effective than the old method at schools such as yours, there would be little justification for testing the method again. Under such circumstances we would agree that withholding the new method from students in the control group would be unjust. If we do not know whether the new method is better, however, any approach other than conducting a true experiment will leave us in doubt about the new method's effectiveness. Random assignment to treatments—call it a "lottery" if you prefer—may be the fairest procedure for assigning students to sections. The old method of instruction, after all, was considered effective before the development of the new method. If the new method proves less effective, random assignment will have actually "protected" the control participants from receiving an ineffective treatment. There are ways to offer a potentially effective treatment to all participants while still maintaining comparable groups. One way is to alternate treatments. For example, Atkinson (1968) randomly assigned students to receive computer-assisted instruction (the treatment) in either English or math and then tested both groups in English and math. Each group served as a control for the other on the test for which its members had not received computer-assisted instruction. After completing the experiment, both groups could then be given computer-assisted instruction in the subject matter to which they had not been previously exposed. Thus, all participants received all potentially beneficial treatments. Establishing a proper control group is also possible if there is more demand for a service than an agency can meet. People who are waiting to receive the service can become a *waiting-list control group.* It is essential, however, that people be assigned to the waiting list randomly. People who are first in line are no doubt different on important dimensions from those who arrive last (e.g., more eager for treatment). Random assignment is necessary to distribute these characteristics in an unbiased way between treatment and comparison groups.

There will always be circumstances in which random assignment simply cannot be used. For example, in clinical trials involving tests of new medical treatments, it may be extremely difficult to get patients to agree to be randomly assigned to either the treatment group or the control (no treatment) group. As you will see, *quasi- experimental designs* can be used in these situations. The logic and procedures for these quasi-experimental designs will be described later in this chapter.

4.b. Discuss in brief the Standardization Features and Psychometric Adequacy of Wechsler Intelligence Scales. 15

Reference:

The Wechsler scales are renowned for their rigorous standardizations, and their revisions with normative updates are now occurring about every 15 years, apparently in response to the Flynn effect (see the chapter byWasserman & Bracken in this volume). The Wechsler scales tend to utilize a demographically stratified (and quasi-random) sampling approach, collecting a sample at most age levels of about n = 200 divided equally by sex. Larger sample sizes are most important during ages undergoing changes such as the rapid cognitive development in young school-aged children and the deterioration in older individuals. Unfortunately, the WAIS-III sample reduces its sample size requirements (to n = 150 and n = 100) at the two age levels between 80 and 90, although these individuals by virtue of their deterioration actually merit an increased sample size. Stratification targets are based on the most contemporary census figures for race-ethnicity, educational level (or parent educational level for children), and geographic region. The manuals for the Wechsler scales typically report demographic characteristics of the standardization sample across stratification variables, so it is possible to ascertain that characteristics were accurately and proportionally distributed across groups rather than concentrated in a single group. Individuals with sensory deficits or known or suspected neurological or psychiatric disorders were excluded from the WAIS-III sample in an effort to enhance the clinical sensitivity of the measure.

Internal consistency tends to be adequate for the Wechsler scales, although there are some isolated subtests with problems. Composite scores (FSIQ; Verbal IQ, VIQ; Performance IQ, PIQ; Verbal Comprehension Index, VCI; Perceptual Organization Index, POI; and Freedom From Distractibility Index and Working Memory Index, FDI-WMI) tend to yield average rs >.90 for the WISC-III and WAIS-III, although the FDI tends to be slightly lower. Test-retest stability coefficients are reported instead of internal consistency for the PSI. At the WISC-III subtest level, Arithmetic, Comprehension, and all performance subtests (with the exception of Block Design) have average reliabilities below .80. At the WAIS-III subtest level, only Picture Arrangement, Symbol Search, and Object Assembly have average reliability coefficients below .80, and Object Assembly in particular appears to decline in measurement precision after about age 70. Accordingly, the Wechsler scales show measurement precision slightly less than considered optimal for their intended decision-making applications.

Test-retest reliability tends to be adequate for WISC-III and the WAIS-III composite indexes and verbal scale subtests, although some performance subtests have less-than-optimal stability. For six age groups undergoing serial testing with test-retest intervals ranging from12 to 63 days (Mdn = 23 days), the WISC-III yielded a mean corrected stability coefficient of .94 for FSIQ and in the .80s and .90s for composite scores, with the exception of a low FDI corrected stability coefficient of .74 for 6- to 7-year-old children. Corrected reliability coefficients for individual subtests ranged from a low of .54–.62 for Mazes to a high of .82–.93 for Vocabulary. Four subtests (Vocabulary, Information, Similarities, and Picture Completion) have an average corrected stability coefficient above .80 (Wechsler, 1991). Over an interval ranging from 2 to 12 weeks (M = 34.6 days) across four age groups, the WAIS-III FSIQ has a mean stability coefficient of .96 corrected for the variability of scores in the standardization sample. Mean corrected stability coefficients for the WAIS-III subtests range from the .90s for Vocabulary and Information to the .60s and .70s for Picture Arrangement and Picture Completion. Composite indexes all have corrected stability coefficients in the .80s and .90s (The Psychological Corporation, 1997).

The four-factor structure of the WISC-III and the WAIS-III, corresponding to the four interpretive indexes, have been found to be largely resilient across a variety of samples. The WISC-III has been reported to be factorially invariant across age (Keith & Witta, 1997), racial groups (Kush et al., 2001),

deaf and hearing samples (Maller & Ferron, 1997), and Canadian and British samples (Cooper, 1995; Roid & Worrall, 1997). Among clinical and exceptional groups, the factor structure is consistent across samples of children in special education (Grice, Krohn, & Logerquist, 1999; Konold, Kush, & Canivez, 1997), children with psychiatric diagnoses (Tupa, Wright, & Fristad, 1997), and children with traumatic brain injury (Donders & Warschausky, 1997). The WAIS-III has been found to be factorially stable across the United States and Canada (Saklofske, Hildebrand, & Gorsuch, 2000) and across mixed psychiatric and neurologically impaired samples (Ryan & Paolo, 2001).

WISC-III and WAIS-III subtest floors and ceilings tend to be good, spanning at least ± 2 SDs at every age and usually larger. The lowest possible FSIQ yielded by the WISC-III is 40, and the highest possible FSIQ is 160. The WAIS-III has slightly less range, with FSIQs from 45 to 155. Ceilings on several of the performance subtests are obtained through the use of bonus points for speed. Perhaps one of the central weaknesses of the Wechsler scales is that most performance tests are timed. Although measuring speed of performance on subtests such as Block Design, Picture Arrangement, and Object Assembly allows for heightened ceilings and increased reliabilities, it may detract from the construct validity of the tests. The Wechsler scales now include a processing speed index, so the inclusion of speed dependency in other subtests is unnecessary and redundant.

4.c. Elucidate the grounded theory approach to qualitative analysis with the help of suitable examples. 20

Reference:

Grounded theory is probably the most common form of qualitative analysis used today. It was developed by two North American sociologists, Glaser and Strauss, in their 1967 book, The discovery of grounded theory: Strategies for qualitative research. As the title suggests, Glaser and Strauss were attempting to articulate how qualitative data could be used not just to provide rich descriptions, but also to generate theory. Grounded theory can be used with a range of qualitative material, such as semi-structured interviews, focus groups, participant observation, and diaries.

The term ‘‘grounded theory’’ is potentially confusing, as it refers both to a method—a set of systematic procedures for analyzing data—and also to the outcome or product of the analysis, which is theory ‘‘grounded’’ in the data. The basic process involves identifying categories at a low level of abstraction and then building up to more abstract theoretical concepts. The end point is often one or more core categories, which capture the essence of the phenomenon (see Chapter 12). This process of analysis occurs concurrently with the process of data collection, and the developing theory guides the sampling strategy (‘‘theoretical sampling’’: see Chapter 10).

The original Glaser and Strauss (1967) volume was more theoretical and polemical rather than practical; it was aimed at challenging the prevailing quantitative paradigm in American sociology. The practical implications for researchers, i.e., the steps in actually carrying out a grounded theory study, are developed in Glaser (1978) and Strauss and Corbin (1998). The grounded theory method was taken up by psychologists in the 1980s and 1990s. Articles by Rennie et al. (1988) and by Henwood and Pidgeon (1992) were aimed at introducing the grounded theory approach to an audience of psychologists. Rennie and Brewer’s (1987) study entitled ‘‘A grounded theory of thesis blocking’’ (i.e., writer’s block among research students) may well be of personal interest to some readers of this text! As more psychologists have taken up the invitation of Rennie et al., and of Henwood and Pidgeon, grounded theory has become a popular approach to qualitative research. One example of the method in clinical

psychology is Bolger's (1999) study of the phenomenon of emotional pain. The participants were women in a therapy group for adult children of alcoholics; they were interviewed on several occasions following group therapy sessions in which they had explored painful life experiences. The interviews focused on how pain was experienced and what was significant in that experience for them. The core category that emerged from the analysis was labeled the ''broken self,'' characterized by four sub-categories of woundedness, disconnection, loss of self, and awareness of self. Another, well-known, example of a grounded theory study, in a more popularized book format, is Charmaz's (1991) analysis of the experience of living with chronic illness (see box).

5.a. Examine ethical issues involved in the disclosure of assessment reports after clinical evaluation and suggest guidelines for minimizing harm and misuse of test data. 20

Reference:

Following completion of their evaluations and reports, psychologists often receive requests for additional clarification, feedback, release of data, or other information and materials related to the evaluation. Release of confidential client information is addressed in the ethics code and highly regulated under many state and federal laws, but many other issues arise when psychological testing is involved.

Feedback Requests

Psychologists are expected to provide explanatory feedback to the people they assess unless the nature of the client relationship precludes provision of an explanation of results. Examples of relationships in which feedback might not be owed to the person tested would include some organizational consulting, pre-employment or security screening, and some forensic evaluations. In every case the nature of feedback to be provided and any limitations must be clearly explained to the person being assessed in advance of the evaluation. Ideally, any such limitations are provided in both written and oral form at the outset of the professional relationship. In normal circumstances, people who are tested can reasonably expect an interpretation of the test results and answers to questions they may have in a timely manner. Copies of actual test reports may also be provided as permitted under applicable law.

Requests for Modification of Reports

On some occasions, people who have been evaluated or their legal guardians may request modification of a psychologist's assessment report. One valid reason for altering or revising a report would be to allow for the correction of factual errors. Another appropriate reason might involve release of information on a need-to-know basis for the protection of the client. For example, suppose that in the course of conducting a psychological evaluation of a child who has experienced sexual abuse, a significant verbal learning disability is uncovered. This disability is fully described in the psychologist's report. In an effort to secure special education services for the learning problem, the parents of the child ask the psychologist to tailor a report for the school focusing only on matters relevant to the child's educational needs—that is to say, the parents would prefer that information on the child's sexual abuse is not included in the report sent to the school's learning disability assessment team. Such requests to tailor or omit certain information gleaned during an evaluation may be appropriately honored as long as doing so does not tend to mislead or misrepresent the relevant findings.

Psychologists must also be mindful of their professional integrity and obligation to fairly and accurately represent relevant findings. A psychologist may be approached by a case management firm with a request to perform an independent examination and to send a draft of the report so that editorial changes can be made. This request presents serious ethical considerations, particularly in forensic settings. Psychologists are ethically responsible for the content of all reports issued over their signature. One can always listen to requests or suggestions, but professional integrity and oversight of one's work cannot be delegated to another. Reports should not be altered to conceal crucial information, mislead recipients, commit fraud, or otherwise falsely represent findings of a psychological evaluation. The psychologist has no obligation to modify a valid report at the insistence of a client if the ultimate result would misinform the intended recipient.

Release of Data

Who should have access to the data on which psychologists predicate their assessments? This issue comes into focus most dramatically when the conclusions or recommendations resulting from an assessment are challenged. In such disputes, the opposing parties often seek review of the raw data by experts not involved in the original collection and analyses. The purpose of the review might include actual rescoring raw data or reviewing interpretations of scored data. In this context, test data may refer to any test protocols, transcripts of responses, record forms, scores, and notes regarding an individual's responses to test items in any medium (ECTF, 2001). Under long-standing accepted ethical practices, psychologists may release test data to a psychologist or another qualified professional after being authorized by a valid release or court order. Psychologists are exhorted to generally refrain from releasing test data to persons who are not qualified to use such information, except (a) as required by law or court order, (b) to an attorney or court based on a client's valid release, or (c) to the client as appropriate (ECTF, 2001). Psychologists may also refrain from releasing test data to protect a client from harm or to protect test security (ECTF, 2001).

In recent years, psychologists have worried about exactly how far their responsibility goes in upholding such standards. It is one thing to express reservations about a release, but it is quite another matter to contend within the legal system. For example, if a psychologist receives a valid release from the client to provide the data to another professional, is the sending psychologist obligated to determine the specific competence of the intended recipient? Is it reasonable to assume that any other psychologist is qualified to evaluate all psychological test data? If psychologists asked to release data are worried about possible harm or test security, must they retain legal counsel at their own expense to vigorously resist releasing the data?

The intent of the APA ethical standards is to minimize harm and misuse of test data. The standards were never intended to require psychologists to screen the credentials of intended recipients, become litigants, or incur significant legal expenses in defense of the ethics code. In addition, many attorneys do not want the names of their potential experts released to the other side until required to do so under discovery rules. Some attorneys may wish to show test data to a number of potential experts and choose to use only the expert(s) most supportive of their case. In such situations, the attorney seeing the file may prefer not to provide the transmitting psychologist with the name of the intended recipient. Although such strategies are alien to the training of many psychologists trained to think as scientific investigators, they are quite common and ethical in the practice of law. It is ethically sufficient for transmitting psychologists to express their concerns and rely on the assurance of receiving clinicians or attorneys that the recipients are competent to interpret those data. Ethical responsibility in such circumstances shifts to receiving experts insofar as justifying their own competence and the foundation

of their own expert opinions is concerned, if a question is subsequently raised in that regard. The bottom line is that although psychologists should seek appropriate confidentiality and competence assurances, they cannot use the ethics code as a shield to bar the release of their complete testing file.

5.b. Discuss the effects of language and culture in adapting measures of personality traits and the experience and expression of different emotions. 15

Reference:

According to Hall and Lindzey (1970), "no substantive definition of personality can be applied with any generality" (p. 9). Definitions of personality vary from comprehensive accounts of behavior in all of its complex details to specific descriptions of individual personality traits (Anastasi, 1988; Guthrie & Lonner, 1986). Anastasi emphasized the importance of defining personality in terms of meaningful trait concepts that describe categories into which behavior must be classified if it is to be accurately measured. Consistent with Anastasi's emphasis on fundamental traits, Cohen et al. (1992) defined personality as "an individual's unique constellation of psychological states and traits" (p.401). Anxiety, anger, and curiosity are examples of meaningful states and traits that are uniquely related to personality (Spielberger, Reheiser, & Sydeman, 1995). The cross-cultural equivalence of anxiety and anger as emotional states and personality traits is facilitated by the fact that these fundamental emotions appear to be universal products of evolution. In his classic book, Expressions of Emotions in Man and Animals, Darwin (1872/1965) concluded, and others have confirmed (Ekman, 1973; Izard, 1977; Tomkins, 1962), that fear and rage are intense emotions that can be identified by facial expressions, not only in humans, but also in many animal species. Consistent with these research findings, Dimberg (1994, 1998) observed that distinctive facial reactions are manifested after very brief exposure to fear and anger-related relevant stimuli, such as snakes and angry faces, indicating that the perception of threatening stimuli can instantaneously evoke specific emotions.

Plutchik (1984) has proposed a "psycho-evolutionary" theory that defines emotions as complex states that can be inferred from subjective reports, physiological changes, and various forms of behavior, which can best be understood in an evolutionary context. In endorsing a Darwinian ethological perspective, Plutchik (1984) pointed out the adaptive role of emotions in motivating what Cannon (1963) described as behavioral fight-or-flight reactions to environmental emergencies that increased the organism's chances for survival.

However, as was noted by Plutchik, description of the feelings associated with these behavioral reactions will depend on a person's experience with a particular language. The words used in different languages to describe emotional states and personality traits generally have a wide range of connotations (Rogler, 1999; Wierzbicka, 1994). Even within a particular language, the same word may have a variety of meanings in different subcultures (Anastasi, 1988). Therefore, differences between and within cultures, in the meaning of the words used to describe emotional states and personality traits, are especially problematic in the cross-cultural adaptation of measures of these constructs (Rogler, 1999). The following are examples of subcultural differences in the meaning of Spanish words (Cabrera, 1998):

- In Carribbean countries guagua means bus, but this same word refers to a baby or child in Chile, Colombia, and Peru.
- Verraco is a pig in Cuba, but has the connotation in Colombia of a person who is tough.

- In Cuba, bicho refers to an insect, but describes a penis in Puerto Rico.
- In Spain, the verb coger has the innocuous meaning to take or to seize, but means having sex in Mexico and Venezuela.

These examples clearly indicate that the successful adaptation of self-report measures of emotional states and personality traits requires the careful selection of key words (or idioms) that have essentially the same meaning in both the original (source) and second (target) languages. However, ensuring accurate representation of the psychological concepts that are assessed is often difficult because languages differ in the connotations of words used to describe the feelings and cognitions associated with different emotional states and personality traits. Moreover, as noted by Wierzbicka (1994) "the set of emotion terms available in any given language is unique and reflects a culture's unique perspective on people's ways of feeling" (p. 135).

Self-report measures of anxiety and other emotions cannot be simply translated and back-translated, but must be adapted for cross-cultural research. The process of 'back-translation' is traditionally used to facilitate adapting educational and psychological tests from one language into another language (Brislin, 1970, 1986). In the back-translation of test items, from the target language into the original language, the literal translation of words is emphasized. However, the back-translation of an original scale item is often less adequate than constructing a new item based on an equivalent cross-cultural conceptual definition of the emotional state or personality dimension that is being measured (Spielberger & Diaz-Guerrero, 1983). This is especially true in adapting idiomatic expressions.

LeCompte and Oner (1976) maintained that translating of key words and idiomatic expressions is especially difficult, and may require frequent consultations with language experts. From the standpoint of the literalness or exactness of the translation, they recommended that items be grouped into three categories: (a) items with key words whose translations closely fit the meaning of the word in the source language, (b) items with keywords for which it is difficult to find corresponding items in the target language, and (c) items with a linguistic form that cannot be translated from the source language to the target language without changing the grammatical construction. A number of cycles of translation and back-translation maybe required before an adequate adaptation can be developed for the latter type of item (Spielberger & Sharma, 1976). In adaptating measures of emotional states and personality traits, the keyword for an item in the source language may have several different translations that are equally acceptable in the target language. Different key words in two or more items in the source language may also be represented by a single word in the target language. Where the literal translation of a test item is not possible, it is important to retain the essential meaning of the original item by selecting a synonym of the keyword that reflects its basic meaning in the target language.

When adapting idiomatic expressions, special care must be taken to translate the feeling connotation of the idiom, rather than translating the literal meaning of the individual words (Guthrie & Lonner, 1986). Identifying comparable idiomatic expressions in the language into which a scale is being translated is preferable to the literal translation of the original idiom. Consequently, in translating and adapting idioms, the cross-cultural equivalence of the theoretical concepts that are being measured is essential. Given the difficulties that are likely to be encountered in translating key words and idiomatic expressions, a substantially larger pool of items than will be eventually needed should be constructed in order to capture the full meaning of the construct that is being measured. Statistical procedures can then be used to determine which items have the best internal consistency as measures of the specified construct. Measuring State and Trait Anxiety Though contemporary interest in anxiety phenomena has

historical roots in the philosophical and theological views of Pascal and Kierkegaard (May, 1977), it was Freud (1924,1936) who first attempted to explicate the meaning of anxiety within the context of psychological theory. He regarded anxiety as "something felt," an unpleasant affective state or condition. According to Freud (1924), this state, as observed in patients with anxiety-neurosis, was characterized by all that is covered by the word nervousness, which includes apprehension or anxious expectation, and efferent discharge phenomena.

Anxiety is distinguishable from other unpleasant affective (emotional) states such as anger, grief, or sorrow, by its unique combination of phenomenological and physiological qualities. These give to anxiety a special "character of unpleasure" that, although difficult to describe, seems "to possess a particular note of its own" (Freud, 1936, p. 69). The subjective, phenomenological qualities of anxiety—the feelings of apprehensive expectation or dread—were emphasized by Freud, especially in his later formulations, whereas the physiological-behavioral (efferent) discharge phenomena, although considered an essential part of an anxiety state and an important contributor to its unpleasantness, was of relatively little theoretical interest to him. Freud was mainly concerned with identifying the sources of stimulation that evoked anxiety reactions, rather than analyzing the properties of such states. He hoped to discover, in the prior experience of his patients, "the historical element... which binds the afferent and efferent elements of anxiety firmly together" (1936, p. 70).

Anxiety has been investigated in numerous studies in which participants who were presumed to differ in motivation or drive level (Spence, 1958) were selected on the basis of their extreme scores on questionnaires such as the Taylor (1953) Manifest Anxiety Scale (MAS), a self-report measure consisting of 50 MMPI items. The performance of high- and low-anxious subjects was then compared on a variety of tasks to test hypotheses derived from Hullian Learning Theory (Spence, 1958). The findings in these studies suggested that high MAS scores predicted performance on learning tasks, but only in situations involving some degree of stress (Spielberger, 1966a). Research on anxiety and learning has also shown that task difficulty, individual differences in intelligence, and factors that influence the relative strengths of correct and competing responses in a particular learning situation must be taken into account.

Cattell and Scheier (1958, 1961) pioneered the application of multivariate techniques to measuring the intensity of anxiety as an emotional state, and individual differences in anxiety proneness as a personality trait (Cattell, 1961, 1963). In investigations of the covariation, over time, of a number of different anxiety measures, relatively independent state and trait anxiety factors consistently emerged (Cattell, 1966). Physiological variables associated with activation (arousal) of the autonomic nervous system, which fluctuated over time and covaried over occasions of measurement (e.g., respiration rate and blood pressure), had strong loadings on the state anxiety factor, but only slight loadings on the trait anxiety factor. Measures with strong loadings on Cattell's (1961) trait anxiety factor included self-reports of anxiety that were relatively stable over time. Scores on Cattell and Scheier's (1963) IPAT Anxiety Scale, a measure of trait anxiety, correlated .85 with the Taylor (1953) MAS. This finding provides strong evidence that the MAS measures anxiety proneness, or trait anxiety, rather than drive level, which is conceptually more closely related to the level of intensity of state anxiety at a particular time. The IPAT and the MAS appear to measure individual differences in anxiety as a personality trait, that is, the disposition to respond to situations perceived as stressful with more intense elevations in state anxiety, which contribute to higher drive level.

The concepts of state anxiety (S-Anxiety) and trait anxiety (T-Anxiety) refer to two related, yet logically quite different constructs (Spielberger & Krasner, 1988). S-Anxiety may be defined as a psycho-physiological emotional state that consists of subjective feelings of tension, apprehension, nervousness

and worry, and activation (arousal) of the autonomic nervous system (Spielberger, 1966b, 1972). Valid measures of S-Anxiety vary in intensity and fluctuate over time as a function of perceived threat. T-Anxiety has the characteristics of a class of constructs that Atkinson (1964) described as motives, and that Campbell (1963) called acquired behavioral dispositions (Spielberger & Diaz-Guerrero, 1983). The attributes of T-Anxiety include relatively stable differences between people in the tendency to perceive stressful situations as more or less dangerous or threatening, and in the disposition to respond to such situations with corresponding elevations in S-Anxiety. Measures of T-Anxiety assess the frequency that anxiety states have been experienced in the past, and the probability that S-Anxiety will be manifested in the future as a reaction to threatening stimuli (Spielberger, 1983; Spielberger & Krasner, 1988).

State-Trait Anxiety Theory posits that people who are high in T-Anxiety perceive social-evaluative situations as more threatening than do persons who are low in T-Anxiety (Spielberger, 1972,1979). Consequently, persons high in T-Anxiety are more likely to experience intense elevations in S-Anxiety in such situations. The State-Trait Anxiety Inventory (STAI) was developed to provide reliable, relatively brief self-report scales for assessing state and trait anxiety in research and clinical practice (Spielberger et al., 1970). Freud's (1936) danger signal theory and Cattell's (1963, 1966) concepts of state and trait anxiety (Cattell & Scheier, 1958, 1961), as refined and elaborated by Spielberger (1966b, 1972, 1979), provided the conceptual framework that guided the test construction process.

5.c. Discuss the sources of invalidity in the nonequivalent control group design of Quasi-Experiments.
15

Reference:

Sources of Invalidity in the Nonequivalent Control Group Design

- To interpret the findings in quasi-experimental designs, researchers examine the study to determine if any threats to internal validity are present.
- The threats to internal validity that must be considered when using the nonequivalent control group design include additive effects with selection, differential regression, observer bias, contamination, and novelty effects.
- Although groups may be comparable on a pretest measure, this does not ensure that the groups are comparable in all possible ways that are relevant to the outcome of the study.

According to Cook and Campbell (1979), the nonequivalent control group design generally controls for all major classes of potential threats to internal validity except those due to additive effects of (1) selection and maturation, (2) selection and history, (3) selection and instrumentation, and (4) those due to differential statistical regression. We will explore how each of these potential sources of invalidity might pose problems for Langer and Rodin's interpretation of their findings. We will then explain how Langer and Rodin offered both logical argument and empirical evidence to refute the possible threats to the internal validity of their study. We will also examine how experimenter bias and problems of contamination were controlled. Finally, we will comment briefly on challenges of establishing external validity that are inherent in the nonequivalent control group design.

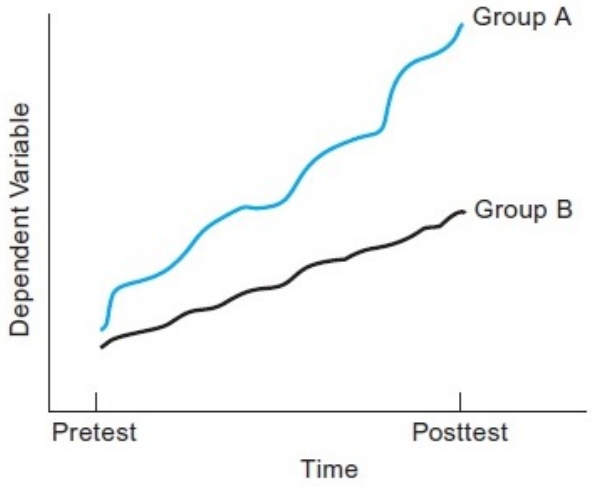
An important initial finding in Langer and Rodin's study was that the residents in the two groups did not differ significantly on the pretest measures. It would not have been surprising to find a difference between the two groups before the treatment was introduced because the residents were not randomly assigned to conditions. Even when pretest scores show no difference between groups, however, we

cannot assume that the groups are "equivalent" (Campbell & Stanley, 1966). We will explain why we cannot conclude that the groups are equivalent in the discussion that follows.

Selection-Maturation Effect An additive effect of selection and maturation occurs when individuals in one group grow more experienced, more tired, or more bored at a faster rate than individuals in another group (Shadish et al., 2002). A selection-maturation effect is more likely to be a threat to internal validity when the treatment group is self-selected (the members deliberately sought out exposure to the treatment) and when the comparison group is from a different population from the treatment group (Campbell & Stanley, 1966). Langer and Rodin selected their groups (but not individuals) randomly from the same population of individuals. Consequently, their design more closely approaches a true experiment than it would if individuals in the two groups had come from different populations (Campbell & Stanley, 1966). A selection-maturation effect would have been more likely, for example, if residents in a nursing home were compared with those attending a sheltered workshop program for the elderly, or if residents on different floors of a nursing facility require different levels of care.

The possibility of a selection-maturation effect is one reason we cannot conclude the groups are equivalent (comparable) even when pretest scores are the same on average for the treatment and control groups. The natural growth rate of two groups from different populations might be different, but the pretest may have been taken at a time when both groups happened to be about the same. This problem is illustrated in Figure 10.4. The normal rate of change is greater in Group A than in Group B, but the pretest is likely to show that the groups do not differ. Because of the differential growth rate, however, the groups would probably show a difference at the posttest that could be mistaken for a treatment effect. There is a second, and more general, reason why we cannot conclude that groups are comparable based only on the absence of a difference between the groups on the pretest. The pretest is likely to measure respondents on only one measure, or at best on a few measures. The mere fact that individuals do not differ on one measure does not mean they don't differ on other measures that are relevant to their behavior in this situation.

FIGURE 10.4 Possible differential growth rates for two groups (A and B) in the absence of treatment.



Is there any reason to suspect a selection-maturation effect in the Langer and Rodin study? That is, would it be reasonable to expect that residents on the treatment floor would change naturally at a faster rate than would patients on the no-treatment floor? Several kinds of evidence suggest that this would not be the case. First, the procedure the nursing home used to assign residents to the two floors was basically random, and the floors were assigned randomly to the treatment and no-treatment conditions. Langer and Rodin also reported that the residents of the two floors were, on the average, equivalent on measures such as socioeconomic status and length of time at the nursing home. Finally, although it is not sufficient evidence in itself, residents on the two floors did not differ on the pretest measures. Thus, the evidence strongly indicates that there was not a threat to the internal validity of the Langer and Rodin study due to the additive effects of selection and maturation.

Selection-History Effect Another threat to internal validity that is not controlled in the nonequivalent control group design is the additive effect of selection and history. Cook and Campbell (1979) refer to this problem as *local history effects*. This problem arises when an event other than the treatment affects one group and not the other. Local history, for example, could be a problem in the Langer and Rodin study if an event affecting the residents' happiness and alertness occurred on one floor of the nursing home but not on the other. You can probably imagine a number of possibilities. A change in nursing staff on one floor, for instance, might bring about either an increase or a decrease in residents' morale, depending on the nature of the change and any differences between the behavior of a new nurse and that of the previous one. Problems of local history become more problematic the more the settings of the individuals in the treatment and comparison groups differ. Langer and Rodin do not specifically address the problem of local history effects.

Selection-Instrumentation Effect A threat due to the combination of selection and instrumentation occurs when changes in a measuring instrument are more likely to be detected in one group than they are in another. Floor or ceiling effects, for instance, could make it difficult to detect changes in behavior from pretest to posttest. If this is more of a problem in one group than in another, a selection-instrumentation effect is present. Shadish et al. (2002) point out that this threat to internal validity is more likely to be a problem the greater the nonequivalence of the groups and the closer the group scores are to the end of the scale. Because Langer and Rodin's groups did not differ on the pretest, and because performance of the groups did not suggest floor or ceiling effects on the measurement scales that were used, this threat to internal validity seems implausible in their study.

Differential Statistical Regression The final threat to internal validity that is not controlled in the nonequivalent control group design is differential statistical regression (Shadish et al., 2002). As we described earlier, regression toward the mean is to be expected when individuals are selected on the basis of extreme scores (e.g., the poorest readers, the workers with the lowest productivity, the patients with the most severe problems). *Differential regression* can occur when regression is more likely in one group than in another. For example, consider a nonequivalent control group design in which the participants with the most serious problems are placed in the treatment group. It is possible, even likely, that regression would occur for this group. The changes from pretest to posttest may be mistakenly interpreted as a treatment effect if regression is more likely in the treatment group than in the control group. Because the groups in the Langer and Rodin study came from the same population and there is no evidence that one group's pretest scores were more extreme than another's, a threat to internal validity due to differential statistical regression is not plausible in their study.

Expectancy Effects, Contamination, and Novelty Effects Langer and Rodin's study could also have been influenced by three additional threats to internal validity that can even affect true experiments—

expectancy effects, contamination, and novelty effects. If observers in their study had been aware of the research hypothesis, it is possible that they inadvertently might have rated residents as being better after the responsibility instructions than before. This observer bias, or expectancy effect, appears to have been controlled, however, because all the observers were kept unaware of the research hypothesis. Langer and Rodin were also aware of possible contamination effects. Residents in the control group might have become demoralized if they learned that residents on another floor were given more opportunity to make decisions. In this case, the use of different floors of the nursing home was advantageous; Langer and Rodin (1976) indicate that "there was not a great deal of communication between floors" (p. 193). Thus, contamination effects do not seem to be present, at least on a scale that would destroy the internal validity of the study.

Novelty effects would be present in the Langer and Rodin study if residents on the treatment floor gained enthusiasm and energy as a result of the innovative responsibility-inducing treatment. Thus, this new enthusiasm, rather than treatment residents' increased responsibility, may explain any treatment effects. In addition, the special attention given the treatment group may have produced a Hawthorne effect in which residents on the treated floor felt better about themselves. It is difficult to rule out completely novelty effects or a Hawthorne effect in this study. According to the authors, however, "There was no difference in the amount of attention paid to the two groups" (p. 194). In fact, communications to both groups stressed that the staff cared for them and wanted them "to be happy." Thus, without additional evidence to the contrary, we can conclude that the changes in behavior Langer and Rodin observed were due to the effect of the independent variable, not to the effect of an extraneous variable that the investigators failed to control.

For investigators to decide whether an independent variable "worked" in the context of a particular experiment, they must systematically collect and carefully weigh evidence for and against the interpretation that the treatment caused behavior to change. As Cook and Campbell (1979) explain: Estimating the internal validity of a relationship is a deductive process in which the investigator has to systematically think through how each of the internal validity threats may have influenced the data. Then, the investigator has to examine the data to test which relevant threats can be ruled out. In all of this process, the researcher has to be his or her own best critic, trenchantly examining all of the threats he or she can imagine. When all of the threats can plausibly be eliminated, it is possible to make confident conclusions about whether a relationship is probably causal. When all of them cannot, perhaps because the appropriate data are not available or because the data indicate that a particular threat may indeed have operated, then the investigator has to conclude that a demonstrated relationship between two variables may or may not be causal. (pp. 55–56)