EVALUATING AND SELECTING
PSYCHOLOGICAL MEASURES
FOR RESEARCH PURPOSES

MADONNA G. CONSTANTINE
JOSEPH G. PONTEROTTO

Even from the beginning of the research process, some individuals become overwhelmed or intimidated by the complex and intricate processes associated with conducting sound psychological inquiry. In particular, evaluating and selecting ideal psychological measures may be quite challenging in light of the broad array of tests currently available (e.g., Robinson, Shafer, & Wrightsman, 1991, 1999). Although some of these instruments are carefully designed and validated, many psychological measures are not carefully thought out and crafted. Unfortunately, we have found that a good number of researchers select measures out of convenience (e.g., no cost and easily available) rather than out of a systematic evaluation process (Ponterotto, 1996).

The most widely used psychological tests are self-report instruments. These measures typically are presented in the format of questionnaires or surveys wherein respondents share their feelings, attitudes, behaviors, or abilities regarding specific issues or constructs. Self-report scales may be valuable tools in assessing variables of interest for quantitative research purposes if proper analysis is applied to the process of selection, utilization, and evaluation of these instruments. We hope this chapter serves as a practical guide to researchers regarding some basic considerations in identifying, selecting, and using appropriate psychological self-report measures.

LOCATING PSYCHOLOGICAL
INSTRUMENTS FOR RESEARCH

Once researchers specify their constructs of interest, they must choose instruments that properly assess these attributes. Using published journal articles, reference books, and computer database searches might be helpful to them in locating possible scales. Subsequently, examining critical reviews of scales helps many investigators to narrow down from a larger group of instruments those most plausible for use.

Given the myriad journal articles and books in the field of psychology, as discussed by Reed and Baxter in Chapter 3, literature databases such as PsycINFO, PsycLIT, and ERIC can be used to locate appropriate references that discuss relevant research variables. For example,
simply using keywords in the bibliographic PsycINFO database, which encompasses most of the published psychological research since 1872, researchers can view a collection of citations for journal articles, books, and doctoral dissertations in their areas of interest. An investigator interested in children’s motivation regarding school performance, for example, might use the keywords motivation, children, and school in a PsycINFO search. After the results of this search present a potential list of citations related to various forms of children’s school motivation, the researcher could retrieve and review articles, books, and dissertations that may be of interest (see Oleson & Arkin in Chapter 4 for more information). If such a search yielded an unwieldy number of entries, the investigator could further specify search criteria, such as focusing on elementary or secondary school children. If any of the aforementioned keywords in isolation were searched in the PsycINFO database, then a very large number of references would be listed for each term. This phenomenon could create difficulties for the researcher in terms of narrowing down his or her area of focus.

The American Psychological Association (APA) suggests using test reference books to locate psychological instruments. Four of the most well-known of these books are Tests in Print (TIP), the Buros Mental Measurements Yearbook (MMY), Tests and Test Critiques. TIP is perhaps the most inclusive reference book because it contains the most entries per volume; the latest edition has 3,009 entries (APA, 2004). It describes scales’ purposes, intended populations, publication dates, prices, authors, time needed to complete instruments, publishers, and test references. TIP VI (Murphy, Plake, Impara, & Spies, 2002) serves as the most recent index to MMY.

Additional reference books for locating instruments include the Directory of Unpublished Experimental Measures; Dictionary of Behavioral Assessment Techniques; Measures of Psychological Assessment: A Guide to 3,000 Original Sources and Their Application; Measures of Personality and Social Psychological Attitudes; Measures of Political Attitudes; Measures for Clinical Practice: A Source Book; ETS Test Collection Catalogue; Handbook of Research Design and Social Measurement; A Sourcebook for Mental Health Measures; and Tests: A Comprehensive Reference for Assessments in Psychology, Education, and Business (Ponterotto, 1996). Reading the reviews of a test also can provide useful information pertaining to the specific population for which the tool was created, as well as its intended purpose. Moreover, test reviews may enlighten readers about issues such as biases and flaws within measures.

Another APA (2004) recommendation for finding appropriate psychological instruments is the use of the Internet. For example, the MMY’s Web site (www.unl.edu/buros) sells over 4,000 commercial reviews that appear in the print version of that book. Each review can be ordered through the Web site for $15. In addition, the Educational Testing Service’s Web site (http://ets.org) contains a database of abstracts of research reports and research memorandums that are available to view online free of charge. More specific information about these test review resources can be found in Table 7.1.

**PRACTICAL CONSIDERATIONS IN SELECTING PSYCHOLOGICAL RESEARCH INSTRUMENTS**

**Cost**

Researchers must take into account the cost to obtain tests from authors or publishers to ensure that it is not precluded by their budget. Some instruments may be free of charge, whereas others may be quite expensive. For-profit test publishers charge for the use of their copyrighted instruments. Prices for these measurements typically are listed in the publishers’ catalogues. Many test authors, however, do not sell their copyright to for-profit test publishing companies, and these authors often allow free use of their measures to both graduate students and experienced researchers.

**Permission and Responsibility for Using Instruments**

It often is mandatory to receive permission from either the authors or publishers of psychological instruments before incorporating them into research studies. Often, permission is
Table 7.1 Test Review Resources

<table>
<thead>
<tr>
<th>Test Review Source</th>
<th>Information Provided</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental Measurements Yearbook</td>
<td>Names of tests, summary of test purpose, intended population, test acronym, price, authors, publishers, and test reviews.</td>
</tr>
<tr>
<td>The Buros Institute of Mental Measurement</td>
<td></td>
</tr>
<tr>
<td><em>Tests in Print</em></td>
<td>Names of over 3,000 published tests by subject index, summary of test purpose, intended population, test acronym, price, authors, publishers, and references of articles related to test.</td>
</tr>
<tr>
<td>The Buros Institute of Mental Measurement</td>
<td></td>
</tr>
<tr>
<td><em>Tests</em></td>
<td>Test titles, descriptions of tests, intended population, authors, tests’ purpose, major test features, administration time, scoring methods, cost, availability, publishers, and information about self- or examiner-administered tests.</td>
</tr>
<tr>
<td>Pro-Ed, Inc.</td>
<td></td>
</tr>
<tr>
<td><em>Test Critiques</em></td>
<td>Companion to <em>Tests</em>; includes information about test purposes, test reliability and validity, test norms, practical applications and uses, technical aspects, test critiques, authors, publishers, intended population, administration procedures, expert opinions regarding test adequacy, and user-friendly explanations of technical terms.</td>
</tr>
<tr>
<td>Pro-Ed, Inc.</td>
<td></td>
</tr>
<tr>
<td><em>ETS Test Collection Catalogue</em></td>
<td>Describes over 10,000 instruments by subject index, including test titles, appropriate uses, authors, and publishers’ addresses.</td>
</tr>
<tr>
<td>Educational Testing Service</td>
<td></td>
</tr>
</tbody>
</table>

requested by sending a formal or e-mail letter to the authors or publishers of the instruments. It is common for publishing companies to list specific requirements of the purchasers (e.g., must possess a master’s degree in psychology or a related field, or a supervisor’s signature must be obtained, as in the case of a student or uncertified professional). Although requesting permission for the use of a scale might seem tedious to some researchers, the test authors’ or publishers’ selective distribution of these instruments can provide substantial security of test materials and prevent possible misuse. Nonetheless, it is virtually impossible to be certain that all tests will remain in qualified hands.

Because different tests require different levels of competence in their use and interpretation, researchers selecting and using psychological tests should have the requisite training and experience to use these scales (Ponterotto, 1996). In addition to minimum qualifications required for the use of certain measures, investigators have a responsibility to contribute to the profession by reporting relevant validity, reliability, and uses of the instruments they select. At times, some researchers are asked to provide raw data from their studies to the instruments’ developers in order to conduct large-scale validity studies.

**Time and Length**

The amount of time required to complete instruments and the length of these instruments are additional considerations in the selection of appropriate tests for research investigations. For example, if one of the scales in a survey packet includes an instrument containing 250 items, the study may attract fewer respondents because of the potential amount of time it would take to complete a scale or packet of scales perceived as lengthy. It may appear to be more intimidating than a shorter measure. Further, excessively long instruments may limit the number of constructs researchers may be interested in studying simultaneously. For example, in the same 40-minute period, research participants can complete either two lengthy scales (measuring two constructs) or five brief scales (measuring five constructs). Moreover, the more time-efficient a measurement, the more potential there is for respondents to complete all the items once they have begun the testing process. We have found,
however, that if participants are compensated with money, food, course credit, or other types of incentives, they may not be as concerned with completing more lengthy instruments.

In addition to the aforementioned practical considerations in selecting psychological research instruments, it is important to ask yourself the following questions with regard to your potential scales:

1. To what degree do these scales address my primary constructs or variables of interest?
2. Based on previous researchers' assertions, what have been the chief strengths and limitations regarding how my constructs of interest have been assessed through these self-report measures?
3. What do I anticipate will be the strengths and limitations of my research study in light of my use of these self-report instruments?

**Psychometric Issues**

**Reliability**

*Reliability* is concerned with the degree to which test scores are free from measurement error. Although in most situations the results of tests are rarely exactly identical when repeated, the results generally should reflect some level of consistency from one trial to another (Carmines & Zeller, 1979; Thorndike, 1997). The lower the degree of measurement error in tests, the higher their reliability. Reliability estimates range from 0 to 1, and the appropriateness of reliability for a test depends on the purpose of the test (Thorndike, 1997). A reliability coefficient of at least .70 is typically considered to be acceptable (Loewenthal, 2001; Ponterotto, 1996). According to Carmines and Zeller (1979), a reliability coefficient of .80 or higher is viewed as satisfactory because it indicates that 80% or more of the variance relies on true test scores and substantially excludes the possibility of it being due to random measurement error.

Because there is more than one technique to calculate measurement error, it is essential for researchers to include details of the method employed. Test-retest, equivalent form, split-half, and internal consistency are the most widely used approaches to calculating reliability.

**Test-Retest**

In test-retest methods of calculating reliability, participants are given the same instrument to complete several times (Carmines & Zeller, 1979; Thorndike, 1997). Researchers then calculate the correlation coefficient by comparing scores from a test session with those of the next test session. When reporting test-retest reliability, noting the time interval between each test administration is recommended (Anastasi & Urbina, 1997). The time interval between measures could range from a few hours to a few months but preferably from 1 to 2 weeks (Pedhazur & Schmelkin, 1991). Intervals of less than 2 weeks might assess substantial alterations within a person rather than random slight differences in test responding, thus altering more accurate representations of instruments' reliability (Anastasi & Urbina, 1997).

Even when the retest is given within the appropriate time interval, however, there are still other difficulties associated with the test-retest method. Participants' previous exposure to the test may result in improved score because of a practice effect. In addition, if the time between test and retest is especially short, there is an increased likelihood of individuals remembering their previous responses and repeating them instead of rereading the same questions (Carmines & Zeller, 1979; Thorndike, 1997). Hence, the carryover effects of practice and memory may falsely increase consistency between tests scores (Pedhazur & Schmelkin, 1991). On the other hand, with an extended time interval, the difficulties associated with the longer time intervals noted above might occur. Furthermore, random variability within participants, such as illness, fatigue, stress, and worry, can account for differences between scores. Although appearing straightforward, the test-retest method for assessing reliability has several limitations and may not be the most appropriate technique to achieve this goal.

**Equivalent or Alternate Form**

Because of possible practice and memory biases associated with the test-retest reliability method, distributing an equivalent reliability form (also referred to as an alternate form) of a test might reduce these problems. Using the same
participants from the original instruments, equivalent forms intend to measure the identical attributes as the original test (Carmines & Zeller, 1979; Thorndike, 1997). The correlation between both measures represents an estimation of the instruments’ reliability, which is expressed in the form of a reliability coefficient (Anastasi & Urbina, 1997; Pedhazur & Schmelkin, 1991).

Similar to the test-retest method, however, equivalent form reliability will reflect random differences within individuals from the original test to the alternate form (Pedhazur & Schmelkin, 1991). Subsequently, researchers should delineate the length of intervals between test administrations (Anastasi & Urbina, 1997). Unlike test-retest reliability, an alternate form may be administered immediately after the completion of an original test because researchers do not need to take into consideration memory and practice effects. Nonetheless, despite the optimistic prognosis of eliminating certain errors associated with the test-retest method, researchers must take into account the complexity of creating a satisfactory alternate form. There is limited availability of acceptable equivalent forms that truly measure the identical construct as the original form of an instrument (Thorndike, 1997).

**Split-Half**

In contrast to the two methods discussed above, the split-half reliability technique yields reliability estimations based on only one test administration. The split-half reliability process entails dividing one test into two equal parts. The correlation between the two distinct halves represents the reliability estimate. The Spearman-Brown correction formula is used to calculate split-half reliability, and this formula assumes that the instruments’ two sections being compared are basically parallel (Pedhazur & Schmelkin, 1991). Although fairly convenient because it only requires one administration, the Spearman-Brown formula of computing split-half reliability has its limitations. For example, this method exhibits some difficulty in proving two test halves as equivalent. Furthermore, erratic scores could result for each of the halves, depending on the manner in which the instruments were divided (Thorndike, 1997). Therefore, the split-half procedure and Spearman-Brown formula could lead to biased estimations of reliability, as well as neglecting reliability associated with stability over time (Anastasi & Urbina, 1997).

**Internal Consistency**

Similar to the split-half method, the internal consistency method of reliability is calculated following a single test administration (Ponterotto, 1996; Thorndike, 1997). The most commonly used coefficient in which internal consistency is expressed is the Cronbach’s alpha (Carmines & Zeller, 1979). Unlike other methods of calculating reliability, Cronbach’s alpha considers the average intercorrelation between every item of a given measure. Thus, the magnitude of alpha depends on the average inter-item correlation and the total number of items in a scale (Anastasi & Urbina, 1997; Ponterotto, 1996). For example, a driver’s test that measures only specific skills on an enclosed course will have higher inter-item consistency in comparison with a driver’s test that includes a written section, parallel parking skills, on-road driving, and an enclosed course section of the test. Although more homogeneous instruments may be considered preferable for determining reliability because the outcomes frequently yield more clarity, more heterogeneous tests may be more precise predictors of issues or behaviors involving a multifaceted construct (Anastasi & Urbina, 1997).

In addition, the more items on tests, the higher the reliability of those instruments (assuming the items do not reduce the inter-item coefficients); however, if there are excessive numbers of items on these measures, the overall reliability may be reduced substantially (Carmines & Zeller, 1979). Although alpha is sometimes more complicated to use compared with other reliability techniques because it compares all items with one another, it is a practical and efficient technique to calculate reliability, and it is considered to be the preferred measure of internal consistency reliability (Carmines & Zeller, 1979; Ponterotto, 1996).

**Other Reliability Considerations**

Determining the reliability of some instruments might be an easy task when the researchers
are concerned with measuring something concrete and stable. But what happens when the subject of interest is constantly changing in moods, thoughts, knowledge, and attitudes? Working with human participants in research can be quite complicated, not only because they are always changing but also because the actual measurement process might affect their behavior as well. For example, in the case of within-group reliability (consistency within one group from one time to the next), individuals completing a test at another time might be concerned with how they answered the questions previously rather than with their experience at that moment. Therefore, it is common to discuss salient characteristics of individuals within a given group, setting, or testing situation when estimating reliability. Pedhazur and Schmelkin (1991) have stressed the importance of researchers’ exploring instruments’ reliability and not overgeneralizing reliability from one study to another. Noting reliability from a previous investigation may be a beneficial comparison tool, as long as investigators realize that samples may vary from study to study. Finally, it is pertinent for researchers to present their own evaluation of measures’ reliability by indicating this information in their findings.

**Validity**

Validity is one of the most important factors to consider when evaluating instruments’ efficacy in measuring a specific construct (Ponterotto, 1996; Thorndike, 1997). An instrument is considered valid when it properly evaluates what it is intended to measure. The correlation between a test score and a criterion measure is known as its validity coefficient. Validity estimates also range from 0 to 1. In general, validity is classified into three basic types: content validity, criterion-related validity, and construct validity (Anastasi & Urbina, 1997).

**Content Validity**

Content validity examines the extent to which instruments properly measure the defined domain of interest. For example, scales designed to measure social anxiety would not have an acceptable content validity if they only examined situations pertaining to anxiety in the work environment and neglected inquiries reflecting anxiety in nonwork situations. Content validity requires a thorough exploration of the constructs of interest to assure that no vital aspects are excluded in the content of the measures.

**Criterion-Related Validity**

Criterion-related validity occurs when the instruments’ results can be connected to one or more outcomes. More specifically, it relates to the extent to which an outcome may be predicted based on the scores obtained on a measure (Thorndike, 1997). Criterion-related validity is not concerned with why something occurs but rather if it predicts it will occur.

When evaluating criterion-related validity, both concurrent and predictive methods may be applied depending on the construct of interest (Ponterotto, 1996). When a criterion exists in the present, concurrent validity is assessed by correlating an instrument with the criterion at an identical point in time (Anastasi & Urbina, 1997); however, when the instruments regard a future criterion, predictive validity correlates this future criterion with the current scales (Carmines & Zeller, 1979). For example, concurrent validity may be applicable to instruments concerned with reactions after an earthquake, administered directly following the earthquake. In contrast, predictive validity would apply when a work performance measure completed before work begins is correlated with a work evaluation measure completed after a certain amount of work was completed.

**Construct Validity**

Construct validity is concerned with the extent to which an instrument is appropriately measuring the variable of interest. There are two primary types of construct validity: convergent and discriminant validity. To determine convergent validity, a scale must correlate significantly and positively with other instruments designed to measure the same construct. To show discriminant validity, also known as divergent validity, a scale should not correlate significantly with
other measures with which theory suggests it should not correlate.

COMMON FLAWS OF SELF-REPORT MEASURES

Self-report psychological instruments are not error free and have several potential shortcomings. According to Laing (1988), these measures may fail to address adequately researchers’ areas of interest, or they may be too costly or inefficient for widespread use. Moreover, some self-report questionnaires are unclear as to their intent. For example, a questionnaire’s item might read, “How often do you communicate with your mother?” The author of this measure could expect respondents to answer the question according to how often they have a verbal conversation with their mother; however, respondents also could answer this question in reference to the degree to which there was e-mail communication, verbal communication, and even in-person interactions with their mother. Because of the possible variety of ways in which such a question could be interpreted by respondents, the responses to some psychological instruments might lead to self-report inconsistencies that ultimately are reflected in the conclusions of the study.

Another problem with self-report measures may be related to participants’ not knowing information required to respond to a question. For example, a question asking about the earned income of their household may inquire about information not accessible to dependent participants still living with their parents. Therefore, respondents could choose to leave this item unanswered or even wager a guess to this question, although they could be providing inaccurate information in the latter case.

Besides the possibility of incorrect information being reported, it is plausible to consider that some individuals might answer self-report questions inaccurately in order to look more socially acceptable (Holtgraves, 2004). This is referred to as social desirability (Paulhus, 1984). Impression management, a form of social desirability, can occur when individuals respond to questions in the hopes of having others view them as competent or healthy in some fashion. Social desirability also may emerge unintentionally when some individuals are unaware they are reporting fictitious information but believe their answers are true. In order to prevent both intentional and unintentional social desirability, some researchers include a social desirability scale, in addition to their primary instruments of interest, to assess the extent of participants’ socially desirable responses (Anastasi & Urbina, 1997; Loewenthal, 2001).

Tests or scales that address potentially uncomfortable or anxiety-producing topics, such as sexual activity, drug use, weight, family conflict, or history of mental health, also may be more likely to be left unanswered. Participants may feel embarrassed or concerned about who might have access to their personal information or whether it can be traced back to them. With increased assurance of anonymity, however, it may be possible to increase the number of answered questions related to unpleasant subjects (Loewenthal, 2001).

Researchers’ inability to understand the context in which the participants are reporting can represent another unfortunate limitation of self-report psychological instruments (Laing, 1988). For instance, when researchers assess participants’ feelings of depression through self-report instruments, some respondents, who may recently have experienced the death of a loved one, might respond affirmatively to items on these instruments, as compared with other respondents who might not have experienced similar life circumstances. Thus, participants’ life circumstances often are not taken into account when interpreting the results of some self-report psychological tests, which might affect the conclusions that some researchers make about data they obtain from respondents.

TECHNOLOGY AND RESEARCH

In many ways, today’s investigators are fortunate because of numerous advances in technology that have occurred over the past 30 years. In particular, computers have contributed immensely to the ease of conducting sound psychological research (see Vaux & Briggs in Chapter 13 for more information). Computers can serve as vital tools to help investigators achieve their research goals and objectives from
the preliminary to the final phases of their research studies.

**Collecting Primary Information**

In the initial stages of research, collecting background information about the constructs of interest and finding ideal measures to assess these constructs has never been easier than with the use of computers. Before the Internet was widely accessible, individuals primarily used library card catalogues to locate specific references. In many ways, the inconvenience and inefficiency of catalogues inhibited some researchers’ access to newly published reference sources because it took many months to update these card catalogues. Today, we are fortunate that many psychology literature bases such as PsycINFO are updated nearly every week or two. With up-to-the-minute information at our fingertips, computer technology has improved and simplified the process of collecting a vast array of psychological knowledge, particularly regarding psychological tests.

**Distributing Measures**

Computers are not only valuable when retrieving background information pertinent to conducting research but also for dispensing information to potential participants. For instance, e-mail can be an excellent device to recruit possible participants for a study or to distribute actual research measures. Sending online surveys to potential respondents also can lower many of the costs associated with conducting research, and data can be collected from literally thousands of participants with minimal intervention on the part of researchers. Electronic surveys are becoming increasingly common, and some research comparing electronic versus postal surveys suggests that electronic survey results may not be significantly different from those obtained from postal surveys (Kraut et al., 2004).

**Test Administration and Scoring**

In addition to computers’ usefulness in increasing the number of participants for psychological research studies, computers frequently can facilitate the testing or evaluation process as well. For instance, some psychological measures feature a *narrative computer interpretation*, which generates an explanation detailing response patterns of the participants (Anastasi & Urbina, 1997). Measures such as the Minnesota Multiphasic Personality Inventory (MMPI), which has a narrative computer interpretation, provide important information about respondents’ personality tendencies and emotional conditions (Anastasi & Urbina, 1997). Another computer application is the *interactive computer system*, which not only can assist researchers in the scoring process but also can allow respondents to complete research measures directly on the computer (as opposed to traditional paper-and-pencil methods).

**Analyzing Data**

Once psychological test scores are acquired, data analysis can begin (see Chapters 18–20 for more information). Although there are a variety of computer programs that can help in the data analysis process, there are at least three common components of most statistical software packages (Pedhazur & Schmelkin, 1991). The first component is *data description*, which includes techniques for defining and naming variables, identifying the format of data, assigning an external data file when needed, supplying extended labels for variables and/or values, and identifying absent values. The second common component of many statistical software packages includes *data transformation techniques*, which can categorize data by certain specified criteria, recode or transform data, generate new variables, and integrate two or more data files. The third common component includes *data analysis techniques*, which can provide means by which researchers can calculate descriptive and inferential statistics.

Although there are variations from one statistical software program to another in terms of the types of analyses they can perform, three of the most commonly used programs among research psychologists are SPSS, SAS, and MINITAB (Pedhazur & Schmelkin, 1991).

**Technical Difficulties**

Researchers must consider that human error can occur when psychological data are being
entered into computers. Another potential difficulty with computer programs includes “bugs,” or problems with hardware or software of these programs. Such problems often go undetected, and the results could be detrimental to the accuracy of researchers’ findings. Fortunately, because most statistical software programs are thoroughly critiqued by experts and because these programs are sometimes upgraded or modified by their developers, researchers are advised to remain updated about potential modifications or changes to these programs (Pedhazur & Schmelkin, 1991).

CONCLUSION

This chapter presents an overview of important considerations related to choosing appropriate self-report instruments for conducting psychological research. By thoroughly identifying literature related to topics of interest, reviewing possible self-report measures with regard to their reliability and validity, and integrating computer technology in the research process, investigators can maximize their likelihood of obtaining successful outcomes related to their psychological research.

EXERCISES

1. Read through a recent issue of one of the following three journals: *Educational and Psychological Measurement*, *Psychological Assessment*, or *Measurement and Evaluation in Counseling and Development*. Select an article that reports on the development and validation of a new psychological instrument. Write a one-page summary of the article and address the following questions:

   A. What construct is the research team trying to measure, and do you believe the Introduction to the article presents a strong rationale for the need for this new instrument?
   
   B. Describe how the researchers went about developing items for the instrument. Was this item development logical to you?
   
   C. Describe briefly the sample(s) and procedures used in the study. Do you think the sample and procedures were appropriate given the purpose of the instrument?
   
   D. Describe how the authors established levels of score validity and reliability for the instrument. Were you impressed by this aspect of the article? Why or why not?
   
   E. Do you think score reliability and validity would hold up with another sample? Why or why not?
   
   F. If you were conducting a study on this psychological construct, would you use this instrument? Why or why not?

2. Now that you have read this chapter, as well as a recent journal article on instrument development, consider a construct in psychology that you would like to measure. In one page or so, address the following questions:

   A. What construct have you selected and why?
   
   B. Using the Internet and various texts cited in this chapter, find out if there are already instruments measuring the construct you have selected. If so, write down their names.
   
   C. Write five items that could be used to measure your construct. If instruments already exist that measure your construct of interest, then come up with a unique aspect of your instrument (for example, if there are already instruments for adolescents and adults that measure Attitudes Toward the Elderly, then come up with an instrument that measures Children’s Attitudes Toward the Elderly). Your five items should be placed on a Likert-type scale (e.g., a *Strongly Agree to Strongly Disagree* continuum with a number of choice points).
   
   D. Now that you have five new instrument items, discuss how you would go about establishing evidence for score validity and reliability for this new brief measure.
RECOMMENDED READINGS

Some recommended resources pertaining to topics highlighted in this chapter include Anastasi and Urbina’s (1997) *Psychological Testing*, which is a comprehensive guide to understanding the use of psychological instruments. Pedhazur and Schmelkin’s (1991) *Measurement, Design, and Analysis: An Integrated Approach* also is a thorough and well-written book that presents pertinent background and technical information related to psychological measures. For a thorough overview of reliability, we recommend Thompson’s (2003) *Score Reliability: Contemporary Thinking on Reliability Issues*. Finally, Loewenthal’s (2001) *An Introduction to Psychological Tests and Scales* is an outstanding resource that explains the basics of conducting psychological research using self-report measures.

REFERENCES


