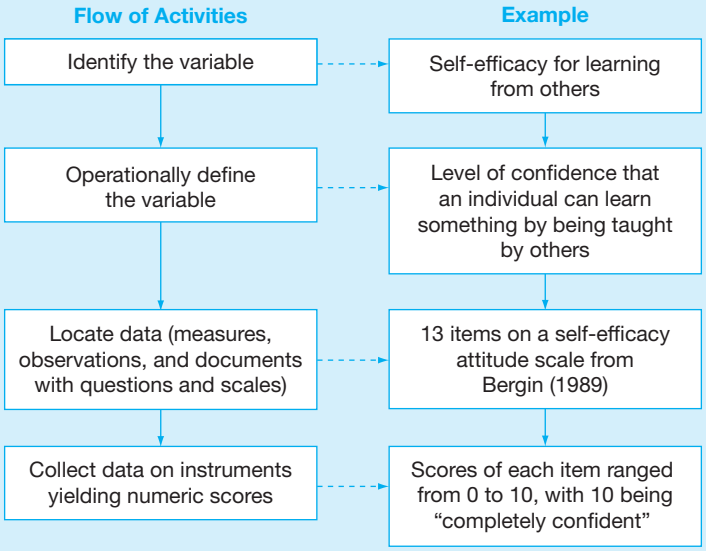
WHAT ARE YOUR OPTIONS FOR COLLECTING INFORMATION?

With the identification of participants and a procedure for gaining permission, you next turn to the specific forms of data that will help you answer your research questions or address your research hypotheses. This step involves identifying the variables in your

FIGURE 5.5

The Flow of Activities in Collecting Data



questions and hypotheses, finding definitions for these variables, and considering types of information that will help you assess these variables, a process outlined in Figure 5.5 using the variable *self-efficacy*.

Specify Variables from Research Questions and Hypotheses

Research questions and hypotheses contain variables. To determine what data need to be collected, you must identify clearly the variables in your study. This will include independent, dependent, and control variables. A useful strategy is to make a list of the variables so that you can determine what variables are operating in a study.

Operationally Define Each Variable

Once you have identified the variables to study, you next need to come up with some way to measure the variables on question items. The first step in this process is to develop a definition for your variable, called an operational definition. An **operational definition** is the specification of how you will define and measure the variable in your study. Unquestionably, a single variable, like *parents' role construction*, as used in the parent involvement study by Deslandes and Bertrand (2005) in Chapter 1, could be defined in many different ways. However, if you look closely at their study, you see that they defined this variable (they called it a construct) as the extent to which parents believed that it was their responsibility to help the school educate their adolescents. The authors did not look up this variable in a dictionary or come up with their own definition of the variable. Instead, they cited other authors who had used the definition and had identified how to measure it. Thus, they came up with questions about parent-focused involvement on six items, school-focused involvement on five items, and partnership-focused

involvement on six items to measure parents' role construction. The authors' logic was to first identify the variable they wanted to study, to provide a definition for it, and to rely on a definition published in the literature by other authors and the question items used by these authors. This is certainly one way to find an operational definition for your variable. Other approaches may be to look in published studies for sections titled "Definition of Terms" or to examine definitions in research summaries such as handbooks or encyclopedias. In some situations, a clear, applied definition suitable for finding a measure is not available, and you will need to construct your own definition. If this is the case, you should test it with other students or individuals knowledgeable about your topic and variable before you use it in your research.

Consider the variable *weapon possession* that Maria needs to define operationally. Write down two or three possible definitions for this variable, such as "a student who has been discovered carrying a knife to school." What other definitions might you use that will help Maria measure the extent to which high school students possess weapons at school? (*Hint:* Think about what happens when a teacher or administrator finds students with weapons in their possession.)

Choose Types of Data and Measures

With operational definitions for your variables, you next need to identify types of data that will measure your variables. Researchers collect data on instruments. An **instrument** is a tool for measuring, observing, or documenting quantitative data. Identified before the researchers collect data, the instrument may be a test, a questionnaire, a tally sheet, a log, an observational checklist, an inventory, or an assessment instrument. Researchers use instruments to measure achievement, assess individual ability, observe behavior, develop a psychological profile of an individual, or interview a person. In quantitative research, four major types of information are gathered, as shown in Table 5.1. Definitions and examples in this table should help you apply your understanding of different forms of quantitative measures.

Performance Measures

You collect **performance measures** to assess an individual's ability to perform on an achievement test, intelligence test, aptitude test, interest inventory, or personality assessment inventory. Participants take tests that measure their achievement (e.g., the Iowa Test of Basic Skills), their intelligence (e.g., Wechsler, Test of Nonverbal Intelligence), or their aptitude (e.g., Stanford–Binet). In addition, you could gather data that measure an individual's career interests or assess personality traits. These measures are all available through instruments reported in the literature. Through past research, researchers have developed "norms" for these tests (conducted the tests with a number of individuals, averaged their scores, and looked at the differences in their scores) so that they can compare individual scores with typical scores for people who have taken the test. However, one drawback of performance data is that they do not measure individual attitudes, and performance data may be costly, time-consuming to gather, and potentially biased toward specific cultural groups.

In Figure 5.6, we examine an example of an instrument used to assess aptitude, abstract reasoning, and problem solving. The Test of Nonverbal Intelligence, fourth edition (TONI-4), may be administered to individuals age 6 through 89 and 11 months to assess general intellectual functioning and identify impairments (Brown, Sherbenou, & Johnsen, 2010). The 60 items on the two forms of this test are actually a sequence of abstract figures with missing figures in the sequence. The images are language-free, allowing the test to be used with individuals with limited language ability. The examiner giving the test requires training in administration and scoring. The examiner scores each item as correct or incorrect along

TABLE 5.1

Types of Quantitative Data and Measures

Types of Data	Types of Tests, Instruments, or Documents to Collect Data	Definition of the Type of Test, Instruments, or Document	Example of the Specific Tests, Instrument, or Source of Information
Measures of individual performance	Achievement test: norm-referenced tests	A test where the individual's grade is a measure of how well he or she did in comparison with a large group of test takers (Vogt & Johnson, 2011)	Iowa Test of Basic Skills
	Criterion-referenced tests	A test where the individual's grade is a measure of how well he or she did in comparison to a criterion or score	General Educational Development or GED Test Metropolitan Achievement Test Series on Reading
	Intelligence test	A test that measures an individual's intellectual ability	Wechsler Intelligence Scale for Children
	Aptitude test	A test to measure a person's ability to estimate how he or she will perform at some time in the future or in a different situation	Cognitive ability: Binet–Simon Scale to identify a child's mental level General ability: Stanford–Binet IQ Scale
	Interest inventory	A test that provides information about an individual's interests and helps them make career choices	Strong Interest Inventory
	Personality assessment	A test that helps a person identify and measure human characteristics that help predict or explain behavior over time and across situations (Thorndike & Thorndike-Christ, 2010)	Minnesota Multiphasic Personality Inventory
Measures of individual attitude	Affective scale	An instrument that measures positive or negative effect for or against a topic	Attitudes toward Self-Esteem Scale Adaptive Behavior Scales
Observation of individual behavior	Behavioral checklist	An instrument used to record observations about individual behavior	Flanders' Interaction Analysis Behavioral Checklist in Reading Vineland Adaptive Behavior Scale
Factual information	Public documents or school records	Information from public sources that provides data about a sample or population	Census data School grade reports School attendance reports

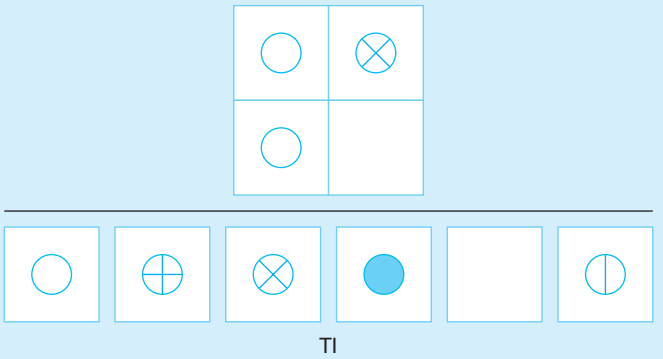
with scoring rules to calculate a raw score. The examiner then uses a table, based on norming values, to convert the raw score to an intelligence score.

Attitudinal Measures

Alternatively, you can measure attitudes of individuals, a popular form of quantitative data for surveys, correlational studies, and experiments. Researchers use **attitudinal measures** when they measure feelings toward educational topics (e.g., assessing positive or negative attitudes toward giving students a choice of school to attend). To develop attitudinal measures, researchers often write their own questions or find an instrument to use that measures the attitudes. Regardless of the approach, these measures need to

FIGURE 5.6

Example of an Instrument That Measures Performance



Source: Sample content from the Test of Nonverbal Intelligence, fourth edition (TONI-4)

contain unbiased questions (e.g., rather than "Should students carry weapons to schools?," ask, "How do you feel about students carrying weapons to school?") and encourage participants to answer questions honestly. One drawback of attitudinal measures is that they do not provide direct evidence of specific behaviors (e.g., whether students actually carry a weapon to school).

Let's examine a research instrument used to gather attitudinal information. Examine the first few questions (out of 74 on the instrument) on the "Student Adaptation to College Questionnaire" available commercially from Western Psychological Services (Baker & Siryk, 1989) in Figure 5.7. This questionnaire begins with personal information questions (e.g., sex, date of birth, current academic standing, and ethnic background) and then asks students to indicate their attitude toward adapting to college on questions using a 9-point response scale from "applies very closely to me" to "doesn't apply to me at all." Overall, the questions focus on the quality of the student's adjustment to the college environment (e.g., whether the student fits in well, feels tense, keeps up to date on academic work, makes friends, attends class, and is satisfied with social life). To analyze these questions, the researcher groups these questions into four subscales: Academic Adjustment (24 questions), Social Adjustment (20 questions), Emotional Adjustment (15 questions), and Goal Commitment–Institutional Attachment (15 questions). Then the analyst sums the scores to the questions on each subscale to identify an individual's score on each subscale. Dahmus, Bernardin, and Bernardin (1992) provided a review of these procedures and background about this questionnaire.

Behavioral Observations

To collect data on specific behaviors, you can observe behavior and record scores on a checklist or scoring sheet. **Behavioral observations** are made by selecting an instrument (or using a behavioral protocol) on which to record a behavior, observing individuals for that behavior, and checking points on a scale that reflect the behavior (behavioral checklists). The advantage of this form of data is that you can identify an individual's actual behavior, rather than simply record his or her views or perceptions. However, behaviors may be difficult to score, and gathering them is a time-consuming form of data collection. Furthermore, if more than one observer gathers data for a study, you need to

FIGURE 5.7

Example of an Instrument That Measures Attitudes

Student Adaptation to College Questionnaire (SACQ)

Robert W. Baker, Ph.D. and Bohdan Siryk, M.A.

Published by

WPS WESTERN PSYCHOLOGICAL SERVICES
Publishers and Distributors
12031 Maple Boulevard
Los Angeles, California 50023

Name:_____ Date:_____

ID Number:_____ Sex: ☐F ☐M Date of Birth:_____

Current Academic Standing: ☐Freshman ☐Sophomore ☐Junior ☐Senior

Semester: ☐1 ☐2 ☐Summer *or* Quarter: ☐1 ☐2 ☐3 ☐Summer

Ethnic Background (optional): ☐Asian ☐Black ☐Hispanic ☐Native American ☐White ☐Other

Directions

Please provide the identifying information requested on the right.

The 67 items on the front and back of this form are statements that describe college experiences. Read each one and decide how well it applies to you at the present time (within the past few days). For each item, circle the asterisk at the point in the continuum that best represents how closely the statement applies to you. Circle only one asterisk for each item. To change an answer, draw an *X* through the incorrect response and circle the desired response. Be sure to use a hard-tipped pen or pencil and press very firmly. Do not erase.

In the example on the right, Item A applied very closely, and Item B was changed from "doesn't apply at all" to "applies somewhat."

Example

A. ⊛ * * * * * * * *

B. * * * ⊛ * * * ⊠

	Applies Very Closely to Me ←	→ Doesn't Apply to Me at All
1. I feel that I fit in well as part of the college environment	*	* * * * * * *
2. I have been feeling tense or nervous lately	*	* * * * * * *
3. I have been keeping up to date on my academic work	*	* * * * * * *
4. I am meeting as many people, and making as many friends as I would like in college	*	* * * * * * *
5. I know why I'm in college and what I want out of it	*	* * * * * * *
6. I am finding academic work at college difficult	*	* * * * * * *
7. Lately I have been feeling blue and moody a lot	*	* * * * * * *
8. I am very involved with social activities in college	*	* * * * * * *
9. I am adjusting well to college	*	* * * * * * *
10. I have not been functioning well during examinations	*	* * * * * * *
11. I have felt tired much of the time lately	*	* * * * * * *
12. Being on my own, taking responsibility for myself, has not been easy	*	* * * * * * *
13. I am satisfied with the level at which I am performing academically	*	* * * * * * *
14. I have had informal, personal contacts with college professors	*	* * * * * * *
15. I am pleased now about my decision to go to college	*	* * * * * * *

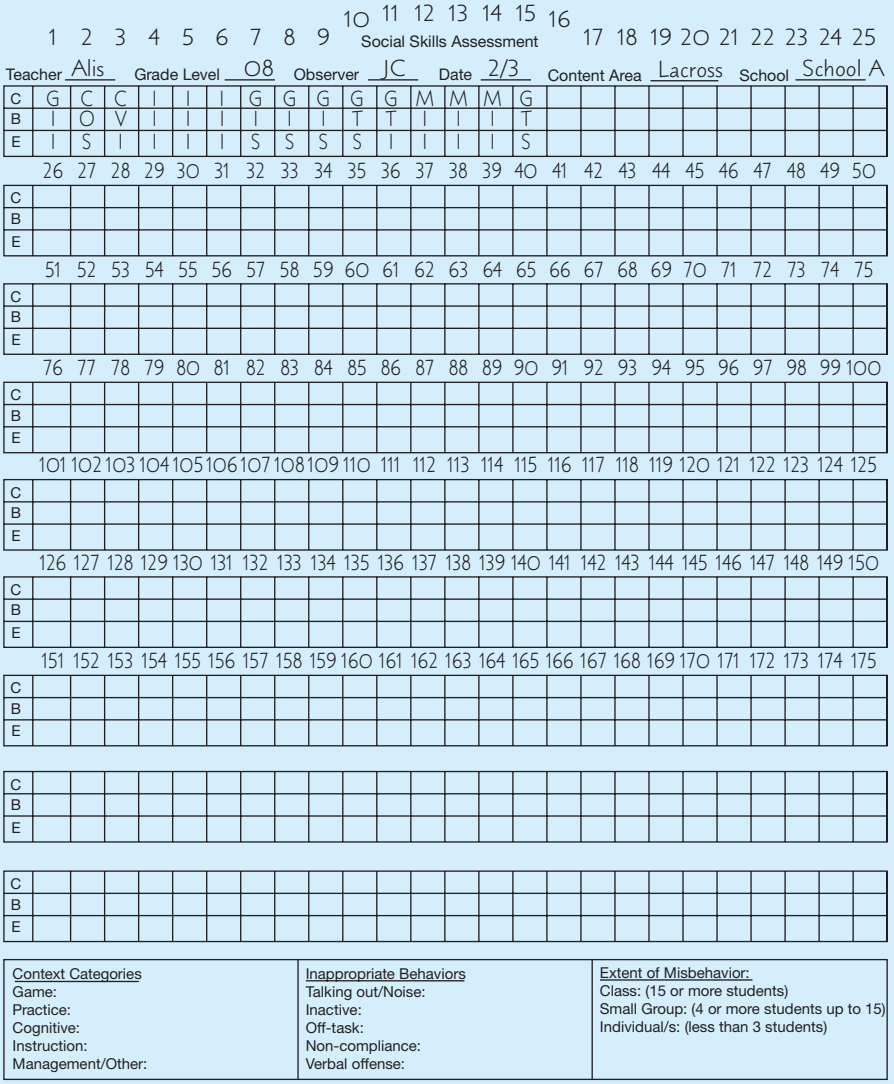
train observers to provide consistent procedures and periodically check that the observers apply consistent scoring.

An example of a behavioral checklist is the Measurement of Inappropriate and Disruptive Interactions (MIDI) developed and used in the Saber-Tooth Project, which studied physical education curriculum changes in one middle school and two comparison schools (Ward, 1999), as shown in Figure 5.8. The investigators used this checklist in a study of four classes in which the teachers provided an instructional unit on lacrosse to eighth-grade students (Ward et al., 1999). During this classroom unit, the researchers observed the students and scored student behaviors using the MIDI scoring sheet in each class portrayed in Figure 5.8.

The legend for this scoring sheet, located at the bottom, lists the codes that observers recorded in each cell. These codes were the first letter of the appropriate word used to describe the context or focus of the lesson in which the behavior occurred (i.e., game, practice, cognitive, instruction, or management/other). The observers also recorded the type of inappropriate behavior during the primary event that involved the most students during the interval (i.e., talking out/noise, inactive, off task, noncompliance, verbal

FIGURE 5.8

Example of an Observational Scoring Sheet with Fictitious Data



Source: Ward, P., Barrett, T. M., Evans, S. A., Doutis, P., Nguyen, P. T., & Johnson, M. K. (1999). Chapter 5: Curriculum effects in eighth grade lacrosse. *Journal of Teaching in Physical Education, 18,* 428–443. Reprinted with permission of the authors.

offense). Finally, observers indicated who was engaged in the disruption (i.e., class, small group, or individual) to assess the extent of misbehavior in the class. Numbers at the top of each column on the scoring sheet represent students (e.g., 1, 2, 3, through 115). Data were collected on-site using this scoring sheet as the instrument, and three observers in the classes recorded their observations (identified by the column numbers) at an interval of 6-second observations. The investigators trained the observers in their scoring

procedures so that they would score the behavior consistently. An audio recording cued the observers as to when they would mark their observations on the checklist sheet. For example, in the fictitious data shown in Figure 5.8 for student 1, the observer recorded the following on the rows:

Context = Game (G)

Inappropriate Behavior = Inactive (I)

Extent of Misbehavior = Individual/s (less than 3 students) (I)

After recording scores for all students, the observers analyzed the differences among students in their disruptive behaviors. Their analysis might have included the number of inappropriate behaviors, a simple count of the number of times a behavior occurs, a percent of the number of inappropriate behaviors divided by the number of opportunities multiplied by 100 (or rate), or the number of occurrences divided by the number of time units (Ayres & Ledford, 2014). Other factors analyzed might be duration, the amount of time the behavior occurs during an observation period, latency, magnitude, or the amount of time to learn a new behavior.

Factual Information

Quantitative, numeric data are also available in public educational records. **Factual information** or **personal documents** consist of numeric, individual data available in public records. Examples of these types of data include grade reports, school attendance records, student demographic data, and census information. As long as these documents are available in the public domain, researchers can access and use them. Investigators cannot easily access some documents, such as health information about students, because federal regulations protect the privacy of individuals. In addition, researchers need to scrutinize the public documents carefully to determine if they represent accurate data. The availability of public documents does not infer that researchers have collected the data carefully with an eye toward accuracy.

Digital Methods of Data Collection

At this time, digital methods of data collection consist of the use of websites and the Internet for administering surveys (Solomon, 2001); gathering interview data (Persichitte, Young, & Tharp, 1997); mining social media data, or using existing databases for analysis (e.g., Texas Lotto, U.S. Census Bureau, or Louis Harris Poll; Pachnowski, Newman, & Jurczyk, 1997). Surveys are typically developed in online survey systems, which store questions, send surveys to participants' e-mail, and store the survey data. An advantage of online surveys is that they can validate responses, such as not allowing a text response for a question that requires a number. In theory, the data will be cleaner than what is generated from a paper form, provided the researcher has applied rigorous survey design principles (Mills & Gay, 2016). The online survey tools will generate reports of descriptive statistics and graphs as well as allow the researcher to download the data set for more sophisticated analyses. Other options include sending text messages with survey questions that the participant completes via a text response. Digital tools may also be used for interviews, as in the case of robopolls, which are automated telephone-based surveys that speak prerecorded questions to participants and collect response through detecting the participants' speech or numbers entered on the telephone keypad (Babbie, 2017). Finally, researchers using digital tools find data sources, as in the case of social media mining (e.g., Facebook, Twitter, blogs) and secondary data analysis of public data sets. Researchers can more easily access large databases for analysis as long as they have obtained necessary permissions and considered the ethical

implications of using data. Digital tools provide an easy, quick form of data collection. However, use of the digital tools may be limited because of (a) limitations involving the use of listservs and obtaining e-mail addresses, (b) limitations of the technology itself, (c) lack of a population list, and (d) the questionable representativeness of the sample data (Mertler, 2001).

How to Decide What Types to Choose

Confronted by these many options for collecting quantitative data, which one or ones will you use? To select your data sources, ask yourself the following questions:

- *What am I trying to learn about participants from my research questions and hypotheses?* If you are trying to learn about individual behaviors of parents at a student–parent conference meeting, you could use a behavioral checklist and record observations. If you are trying to measure the attitudes of teachers toward a bond issue, attitudinal questions or an attitudinal instrument will be required.
- *What information can you realistically collect?* Some types of data may not be collectible in a study because individuals are unwilling to supply them. For example, precise data on the frequency of substance abuse in middle schools may be difficult to collect; identifying the number of student suspensions for substance abuse is much more realistic.
- *How do the advantages of the data collection compare with its disadvantages?* In our discussion of each data source, we have talked about the ideal situations for data collection. Given the ease or difficulty of collecting data, each type needs to be assessed.

How would you now advise that Maria collect her data? Assume that she now seeks to answer the general quantitative research question "Why do students carry weapons in high school?" and the following subquestions:

a. "How frequently do students feel weapons are carried into high school?"
b. "What general attitudes do high school students hold toward the possession of weapons in the schools?"
c. "Does participation in extracurricular activities at school influence attitudes of students toward possession of weapons?"
d. "Are student suspensions for possession of weapons on the increase in high schools?"

Before looking at the answers provided, list the type of information that Maria might collect for subquestions a through d.

To answer these subquestions, Maria first needs to locate or develop a questionnaire to send out to a sample of high school students in the school district. Her data collection will consist mainly of attitudinal data. This questionnaire will measure student *attitudes* toward frequency of weapon possession (question a), assess student *attitudes* toward possession of weapons (question b), and gather *factual data* about the students (question c), such as age, level of education, race, gender, and extent of participation in extracurricular activities. To answer question d, she will contact the school officials of several high schools and ask if she can obtain reports on student suspensions—school documents that report quantitative data. In summary, she will collect both attitudinal and factual information.

WHAT INSTRUMENT WILL YOU USE TO COLLECT DATA?

Let's assume that you will collect performance, attitudinal, or observational data. These forms of data collection all involve using an instrument. What instrument will you use to collect your data? Do you find one to use or develop one yourself? If you search for one to use, how will you locate this instrument? Once you find the instrument, what criteria will you use to determine if it is a good instrument to use?

Locate or Develop an Instrument

Three options exist for obtaining an instrument to use: You can develop one yourself, locate one and modify it, or locate one and use it in its entirety. Of these choices, locating one to use (either modifying it or using it in its original form) represents the easiest approach. It is more difficult to develop an instrument than to locate one and modify it for use in a study. **Modifying an instrument** means locating an existing instrument, obtaining permission to change it, and making changes in it to fit your requirements. Typically, authors of the original instrument will ask for a copy of your modified version and the results from your study in exchange for your use of their instrument.

An instrument to measure the variables in your study may not be available in the literature or commercially. If this is the case, you will have to develop your own instrument, which is a long and arduous process. Developing an instrument consists of several steps, such as identifying the purpose of the instrument, reviewing the literature, writing the questions, and testing the questions with individuals similar to those you plan to study. The four phases of development, recommended by Benson and Clark (1983) and shown in Figure 5.9, illustrate the rigorous steps of planning, constructing, evaluating, and checking to see if the questions work (i.e., validating an instrument). In this process, the basic steps consist of reviewing the literature, presenting general questions to a target group,

FIGURE 5.9

Steps in Developing or Constructing an Instrument

Phase I: Planning
- State purpose of test and target groups
- Identify and define domain of test
- Review literature on construct or variable of interest
- Give open-ended questions to target group
- Interpret open-ended comments
- Write objectives
- Select item format

Phase II: Construction
- Develop table of specifications
- Hire and train item writers
- Write pool items
- Validate content
- Have judges complete qualitative evaluation
- Develop new or revise items

Phase III: Quantitative Evaluation
- Prepare instrument for first pilot test
- Administer first pilot test
- Debrief subjects
- Calculate reliability
- Run item analysis
- Revise instrument
- Prepare for second pilot test

Phase IV: Validation
- Administer second pilot test
- Run item analysis
- Repeat steps of revision, pilot administration, and item analysis
- Begin validation
- Administer for validation data
- Continue validation

Source: Adapted from a flowchart provided by Benson and Clark (1983). Copyright © 1978, CCC Republication.

constructing questions for the item pool, and pilot testing the items. The statistical procedures of calculating reliability and item analysis are available in software programs.

Search for an Instrument

If you decide to use an existing instrument, the publisher or author will typically charge you a fee for its use. Finding a good instrument that measures your independent, dependent, and control variables is not easy. In fact, you may need to assemble a new instrument that consists of parts of existing instruments. Although the publisher will likely share research about the instrument, we encourage you to also review the literature and test reviews. Whether you search for one instrument or several to use, several strategies can aid in your search:

- *Look in published journal articles.* Often authors of journal articles will report instruments and provide a few sample items so that you can see the basic content included in the instrument. Examine references in published journal articles that cite specific instruments and contact the authors for inspection copies. Before you use the instrument, seek permission from the author. With limited space in journals, authors are including fewer examples of their items or copies of their instruments.
- *Run an ERIC search.* Use the term *instruments* and the topic of the study to search the ERIC system for instruments. Use the online search process of the ERIC database. Use the same search procedure to locate abstracts to articles where the authors mention instruments that they have used in their studies.
- *Examine guides to tests and instruments that are available commercially.* Examine the *Mental Measurements Yearbook* (*MMY*; Carlson, Geisinger, & Jonson, 2017) or the *Tests in Print* (*TIP*; Anderson, Schlueter, Carlson, & Geisinger, 2016), both of which are available from the Buros Center for Testing (buros.org). More than 400 commercial firms develop instruments that are available for sale to individuals and institutions. Published since 1938, these guides contain extensive information about tests and measures available for educational research use. You can locate reviews and descriptions of English-language commercially published tests in the *MMY*, which is available online from many academic libraries.

Criteria for Choosing a Good Instrument

Once you find an instrument, several criteria can be used to assess whether it is a good instrument to use. Ask yourself the following:

- Have authors developed the instrument recently, and can you obtain the most recent version? With knowledge expanding in educational research, instruments over 5 years old might be outdated. To stay current, authors update their instruments periodically, and you need to find the most recent copy of an instrument.
- Is the instrument widely cited by other authors? Frequent use by other researchers will provide some indication of its endorsement by others. Use by other researchers may provide some evidence about whether the items on the instrument provide good and consistent measures.
- Are reviews available for the instrument? Look for published reviews about the instrument in the *MMY* or in journals such as *Measurement and Evaluation in Counseling and Development*. If reviews exist, it means that other researchers have taken the instrument seriously and seek to document its worth.
- Is there information about the reliability and validity of scores from past uses of the instrument? When using a performance measure, has it been normed?

- Does the procedure for recording data fit the research questions/hypotheses in your study?
- Does the instrument contain accepted scales of measurement?

Because of the importance of the last three criteria—reliability and validity, recording information, and scales of measurement—a discussion will explore these ideas in more depth.

Are Scores on Past Use of the Instrument Reliable and Valid?

You want to select an instrument that reports individual scores that are reliable and valid. **Reliability** means that scores from an instrument are stable and consistent. Scores should be nearly the same when researchers administer the instrument multiple times at different times. In addition, scores need to be consistent. When an individual responds to certain items one way, the individual should consistently answer closely related items in the same way. **Validity** is the development of sound evidence to demonstrate that the test interpretation (of scores about the concept or construct that the test is assumed to measure) matches its proposed use (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). This definition, in place since 1985, changes the traditional focus on the threefold types of validity—construct, criterion referenced, and content—and shifts the emphasis from "types" of validity to the "evidence" and "use" of the test or instrument (Thorndike & Thorndike-Christ, 2010). Thus, validity is seen now as a single unitary concept, rather than three types. Validity is the degree to which all the evidence points to the intended interpretation of test scores for the proposed purpose. Thus, a focus is on the consequences of using the scores from an instrument (Hubley & Zumbo, 1996; Messick, 1980).

Reliability and validity are bound together in complex ways. These two terms sometimes overlap and at other times are mutually exclusive. Validity can be thought of as the larger, more encompassing term when you assess the choice of an instrument. Reliability is generally easier to understand, as it is a measure of consistency. If scores are not reliable, they are not valid; scores need to be stable and consistent before they can be meaningful. Additionally, the more reliable the scores from an instrument are, the more valid the scores may be (however, scores may still not measure the particular construct and may remain invalid). The ideal situation exists when scores are both reliable and valid. In addition, the more reliable the scores from an instrument are, the more valid the scores will be. Scores need to be stable and consistent before they can be meaningful.

Reliability A goal of good research is to have measures or observations that are reliable. Several factors can result in unreliable data, including when the following occur:

- Items on instruments are ambiguous and unclear.
- Procedures of test administration vary and are not standardized.
- Participants are fatigued, are nervous, misinterpret questions, or guess on tests (Rudner, 1993).

Researchers can use any one or more of five available procedures to examine an instrument's reliability, as shown in Table 5.2. You can distinguish these procedures by the number of times the instrument is administered, the number of versions of the instrument administered by researchers, and the number of individuals who make an assessment of information.

The **test–retest reliability** procedure examines the extent to which scores from one sample are stable over time from one test administration to another. To determine this form of reliability, the researcher administers the test at two different times to the same participants at a sufficient time interval. If the scores are reliable, then they will relate (or will correlate) at a positive, reasonably high level, such as .6. This approach has the

TABLE 5.2

Types of Reliability

Form of Reliability	Number of Times Instrument Administered	Number of Different Versions of the Instrument	Number of Individuals Who Provide Information
Test–retest reliability	Twice at different time intervals	One version of the instrument	Each participant in the study completes the instrument twice.
Alternate forms reliability	Each instrument administered once	Two different versions of the same concept or variable	Each participant in the study completes each instrument.
Alternate forms and test–retest reliability	Twice at different time intervals	Two different versions of the same concept or variable	Each participant in the study completes each instrument.
Interrater reliability	Instrument administered once	One version of the instrument	More than one individual observes behavior of the participants.
Internal consistency reliability	Instrument administered once	One version of the instrument	Each participant in the study completes the instrument.

advantage of requiring only one form of the instrument; however, an individual's scores on the first administration of the instrument may influence the scores on the second administration. Consider this example:

> A researcher measures a stable characteristic, such as creativity, for sixth graders at the beginning of the year. Measured again at the end of the year, the researcher assumes that the scores will be stable during the sixth-grade experience. If scores at the beginning and the end of the year relate, there is evidence for test–retest reliability.

Another approach is **alternative forms reliability**. This involves using two instruments, both measuring the same variables and relating (or correlating) the scores for the same group of individuals to the two instruments. In practice, both instruments need to be similar, such as the same content, same level of difficulty, and same types of scales. Thus, the items for both instruments represent the same universe or population of items. The advantage of this approach is that it allows you to see if the scores from one instrument are equivalent to scores from another instrument for two instruments intended to measure the same variables. The difficulty is whether the two instruments are equivalent in the first place. Assuming that they are, the researchers relate or correlate the items from the one instrument with its equivalent instrument. Consider this example:

> An instrument with 45 vocabulary items yields scores from first graders. The researcher compares these scores with those from another instrument that also measures a similar set of 45 vocabulary items. Both instruments contain items of approximately equal difficulty. When the researcher finds the items to relate positively, we have confidence in the accuracy or reliability of the scores from the first instrument.

The **alternate forms and test–retest reliability** approach is simply a variety of the two previous types of reliability. In this approach, the researcher administers the test

twice *and* uses an alternate form of the test from the first administration to the second. This type of reliability offers the advantages of both examining the stability of scores over time and having the equivalence of items from the potential universe of items. It also has all the disadvantages of both test–retest and alternate forms of reliability. Scores may reflect differences in content or difficulty or in changes over time. An example follows:

> The researcher administers the 45 vocabulary items to first graders twice at two different times, and the actual tests are equivalent in content and level of difficulty. The researcher correlates or relates scores of both tests and finds that they correlate positively and highly. The scores to the initial instrument are reliable.

Interrater reliability is a procedure used when making observations of behavior. It involves observations made by two or more individuals of an individual's or several individuals' behavior. The observers record their scores of the behavior and then compare scores to see if their scores are similar or different. Because this method obtains observational scores from two or more individuals, it has the advantage of negating any bias that any one individual might bring to scoring. It has the disadvantages of requiring the researcher to train the observers and requiring the observers to negotiate outcomes and reconcile differences in their observations, something that may not be easy to do. Here is an example:

> Two observers view preschool children at play in their activity center. They observe the spatial skills of the children and record on a checklist the number of times each child builds something in the activity center. After the observations, the observers compare their checklists to determine how close their scores were during the observation. Assuming that their scores were close, they can average their scores and conclude that their assessment demonstrates interrater reliability.

Scores from an instrument are reliable and accurate if an individual's scores are **internally consistent** across the items on the instrument. If someone completes items at the beginning of the instrument one way (e.g., positive about health effects of tobacco), then they should answer the questions later in the instrument in a similar way (e.g., positive about the health effects of tobacco).

The consistency of responses can be examined in several ways. One way is to split the test in half and relate or correlate the items. This test is called the **Kuder–Richardson split half test (KR-20, KR-21)**, and it is used when (a) the items on an instrument are scored right or wrong as categorical scores, (b) the responses are not influenced by speed, and (c) the items measure a common factor. Since the split half test relies on information from only half of the instrument, a modification in this procedure is to use the **Spearman–Brown formula**, which estimates full-length test reliability using all questions on an instrument. This is important because the reliability of an instrument increases as researchers add more items to the instrument. Finally, the **coefficient alpha** is used to test for internal consistency (Cronbach, 1984). If the items are scored as continuous variables (e.g., *strongly agree* to *strongly disagree*), the alpha provides a coefficient to estimate consistency of scores on an instrument. Calculations for the Kuder–Richardson split half, Spearman–Brown prophecy formula, and coefficient alpha are available in Thorndike & Thorndike-Christ (2010).

Validity In addition to reliability, it is critical to examine whether the scores from the instrument (not the instrument itself) are valid. As a researcher, here are the steps you will likely employ:

- Identify an instrument (or test) that you would like to use
- Look for evidence of validity by examining prior studies that have reported scores and use of the instrument
- Look closely at the purpose for which the instrument was used in these studies

- Look as well at how the researchers have interpreted (discussed if the instrument measured what it is intended to measure) the scores in light of their intended use
- Evaluate whether the authors provide good evidence that links their interpretation to their use

What types of evidence should researchers seek to establish validity? Impara (2010) provided a useful summary of AERA et al.'s *Standards* (1999). He directed readers to examine closely Chapter 1 from the *Standards* on "validity," and then presented an extended list of examples of evidence to document validity. Only a few of the examples are mentioned here.

The *Standards* mention five categories of evidence as shown in Table 5.3: evidence based on test content, response processes, internal structure, relations to other variables, and the consequences of testing. In the discussion to follow, the word "testing" will be equivalent to "instrument."

TABLE 5.3

Sources of Validity Evidence and Examples

Validity Evidence	Types of Tests or Instruments to Which Validity Evidence Is Applicable	Type of Evidence Sought	Examples of Evidence
Evidence based on test content	Achievement tests, credentialing tests, and employment tests	Evidence of an analysis of the test's content (e.g., themes, wording, and format) and the construct it is intended to measure	• Examine logical or empirical evidence (e.g., syllabi, textbooks, and teachers' lesson plans) • Have experts in the area judge
Evidence based on response processes	Tests that assess cognitive processes, rate behaviors, and require observations	Evidence of the fit between the construct and how individuals taking the test actually performed	• Interviews with individuals taking tests to report what they experienced/were thinking • Interviews or other data with observers to determine if all are responding to the same stimulus in the same way
Evidence based on internal structure	Applicable to all tests	Evidence of the relationship among test items, test parts, and the dimensions of the test	• Statistical analysis to determine if factor structure (scales) relates to theory, correlation of items
Evidence based on relations to other variables	Applicable to all tests	Evidence of the relationship of test scores to variables external to the test	• Correlations of scores with tests measuring the same or different constructs (convergent/discriminant validity) • Correlations with scores and some external criterion (e.g., performance assessment—test-criterion validity) • Correlations of tests scores and their prediction of a criterion based on cumulative databases (called meta-analysis—validity generalization)
Evidence based on the consequences of testing	Applicable to all tests	Evidence of the intended and unintended consequences of the test	• Benefits of the test for positive treatments for therapy, for placement of workers in suitable jobs, for prevention of unqualified individuals from entering a profession, for improvement of classroom instructional practices, and so forth

Source: Adapted from Impara (2010) and AERA et al. (1999).

Evidence Based on Test Content Often, instruments will be used that measure achievement, assess applicants for credentials, or are used for employment in jobs. The question is whether the scores from the instrument show that the test's content relates to what the test is intended to measure. This idea relates to the traditional idea of content validity. Typically, researchers go to a panel of judges or experts and have them identify whether the questions are valid. This form of validity is useful when the possibilities of questions (e.g., achievement tests in science education) are well known and easily identifiable. It is less useful in assessing personality or aptitude scores (e.g., on the Stanford–Binet IQ test), when the universe of questions is less certain.

Evidence Based on Response Processes Instruments can be evaluated for the fit between the construct being measured and nature of the responses of the individuals completing the instrument or the individuals conducting an observation using the instrument. Do the scores reflect accurate responses of the participants to the actual instrument? Validity evidence can be assembled through interviews of the participants to report what they experienced or were thinking when they completed the instrument. The responses of observers can be compared to determine whether they are responding in a similar way when they observe. The more the response processes fit what the instrument is intended to measure, the better the evidence for validity.

Evidence Based on Internal Structure Are the test score interpretations consistent with a conceptual framework for the instrument? This form of validity evidence is gathered by conducting statistical procedures to determine the relationship among test item and test parts. It relates to the traditional notion of construct validity. Through statistical procedures, you can do the following:

- See if scores to items are related in a way that is expected (e.g., examine the relationship of a question on a "student depression instrument" to see if it relates to the overall scale measuring depression)
- Test a theory and see if the scores, as expected, support the theory (e.g., test a theory of depression and see if the evidence or data supports the relationships in the theory)

Evidence Based on Relations to Other Variables This is a large category of evidence that relates to the traditional idea of criterion-related validity (predictive and concurrent). Basically, the researcher looks for evidence of the validity of scores by examining other measures outside of the test. The researcher can look at similar or dissimilar tests to see if the scores can be related positively or negatively. The researcher can see if the scores predict an outside criterion (or test scores) based on many different studies. For example, when the results of a current study showed that boys in middle schools have lower self-esteem than girls, can this prediction hold true when many studies have been assessed? Collecting validity evidence from these many studies provides support for the validation of scores on an instrument.

Evidence Based on the Consequences of Testing This form of evidence is the factor that has been introduced into the quantitative validity discussion. Validity evidence can be organized to support both the intended and the unintended consequences of using an instrument. What benefits (or liabilities) have resulted from using the instrument? Researchers can assemble evidence to demonstrate the consequences of testing, such as the enhanced classroom instruction that results as a consequence of testing. Not all the consequences may be intended; for example, an educational test may be supported on the grounds that it improves student attendance or motivation in classes.

After reviewing the forms of reliability and validity, we can now step back and review what questions you should ask when selecting or evaluating an instrument. A short list of questions, provided in Figure 5.10, should aid in this process.

FIGURE 5.10

Reliability and Validity Questions for Selecting/Evaluating a Test or Instrument

When selecting or evaluating an instrument, look for:

Reliability	Validity
1. Did the author check for reliability?	1. Did the author check for validity?
2. If so, what form of reliability was reported?	2. If so, what type of validity was reported?
3. Was an appropriate type used?	3. Was more than one type reported?
4. Were the reliability values (coefficients) reported?	4. Was the validity evidence reported with appropriate statistics?
5. Were they positive, high coefficients?	5. Was the evidence strong?

To practice applying these questions, consider the choice of an instrument by Maria. She finds an instrument titled "Attitudes toward Possession of Weapons in Schools." An author reports this instrument in a published journal article. What might be two forms of evidence about reliability and validity she could look for in the author's discussion about this instrument? Write down these forms of reliability and validity.

Do the Instrument's Data-Recording Procedures Fit the Research Questions/Hypotheses?

Returning to our question of criteria for assessing a good instrument to use, another criterion was that instruments contain recording procedures that fit the data you need to answer the research questions or hypotheses. Who records the data on the instruments or checklists? Data may be self-reported; that is, the participants provide the information, such as on achievement tests or on attitudinal questionnaires. Alternatively, the researcher may record the data on forms by observing, interviewing, or collecting documents. Having participants supply the data is less time-consuming for the researcher. However, when the researcher records the data, he or she becomes familiar with how the participants respond and hence can control for a higher level of quality of the data.

Are Adequate Scales of Measurement Used?

Another criterion is that the instrument should contain good response options to the questions. Variables can be measured as categories or on a continuous range of scores. It is helpful to assess instruments that you might use in research in terms of the adequacy of their scales of measurement. For example, for a study of student attitudes toward the use of tablets in a college classroom, a researcher might ask the question "To what extent does the tablet help you learn in the classroom?" The student might answer this question using a categorical scale such as the following:

_____ To a great extent
_____ Somewhat
_____ To a less extent

The easiest way to think about the types of scales of measurement is to remember that there are two basic types: categorical and continuous scales. Categorical scales have two types: nominal and ordinal scales. Continuous scales (often called *scale scores*

TABLE 5.4

Types of Scales Used in Quantitative Research

Type of Scale	Examples of Questions Using the Scale
Nominal scale (uses categories)	How much education have you completed? _____ No college _____ Some college _____ Bachelor's degree _____ Graduate or professional work What is your class rank? _____ Freshman _____ Sophomore _____ Junior _____ Senior
Ordinal scale (uses categories that imply or express rank order)	Has your adviser helped you select courses? _____ Not at all _____ To a small extent _____ To some extent _____ To a great extent _____ To a very great extent Rank your preference for type of graduate-level instruction from 1 to 4. _____ Activity-based learning _____ Lecture _____ Small-group learning _____ Discussion
Quasi-interval or interval/ratio scale (uses continuous equal intervals)	School is a place where I am thought of as a person who matters. _____ Strongly agree _____ Agree _____ Undecided _____ Disagree _____ Strongly disagree Colleges and universities should conduct research to solve economic problems of cities. _____ Strongly agree _____ Agree _____ Undecided _____ Disagree _____ Strongly disagree

in data analysis programs) also have two types: interval/quasi-interval and ratio scales. These types of scales are shown in Table 5.4.

Scales of measurement are response options to questions that measure (or observe) variables in categorical or continuous units. It is important to understand scales of measurement to assess the quality of an instrument and to determine the appropriate statistics to use in data analysis.

Nominal Scales Researchers use **nominal (**or **categorical) scales** to provide response options where participants check one or more categories that describe their traits, attributes, or characteristics. These scales do not have any order. An example of a nominal scale would be gender, divided into the two categories of male and female (either one could be listed first as a response option). Another form of a nominal scale would be a checklist of "yes" or "no" responses. A semantic differential scale, popular in psychological research, is another type of nominal scale. This scale consists of bipolar adjectives that the participant uses to check his or her position. For example, in a psychological study of talented teenagers, researchers were interested in studying the teenagers' emotional responses to their everyday activities (Csikszentmihalyi, Rathunde, Whalen, & Wong, 1993). The researchers used a semantic differential scale for teenagers

to record their mood on several adjectives at certain times of the day. The researchers used a beeping device (p. 52), and the participants were asked to describe their mood as they were beeped, using the following scale:

	Very	Quite	Some	Neither	Some	Quite	Very	
Alert	0	o	·	–	·	o	0	Drowsy

Although the researchers summed scores of teenagers across several questions such as this one, the response scale to each question was nominal or categorical.

Ordinal Scales Researchers use **ordinal (**or **ranking** or **categorical) scales** to provide response options where participants rank from best, or most important, to worst, or least important, some trait, attribute, or characteristic. These scales have an implied intrinsic order. For example, a researcher might record individual performance in a race for each runner from first to last place. Many attitudinal measures imply an ordinal scale because they ask participants to rank order the importance (*highly important* to *of no importance*) or the extent (*to a great extent* to *a little extent*) of topics. As this example illustrates, the information is categorical in a ranked order. A semantic differential scale that orders responses might also fit into the ordinal scale category. In this scale, the respondent is asked to choose where his or her position lies, on a scale between two polar adjectives (for example, "Good-Evil," "Happy-Sad"). Semantic differentials can be used to measure opinions, values, and attitudes.

Interval/Ratio Scales Another popular scale researchers use is an interval or rating scale. **Interval (**or **rating** or **continuous) scales** provide "continuous" response options to questions with assumed equal distances between options. These scales may have three, four, or more response options. Although an ordinal scale, such as *highly important* to *of no importance*, may seem like an interval scale, we have no guarantee that the intervals are equal. An achievement test such as the Iowa Test of Basic Skills is assumed to be an interval scale because researchers have substantiated that the response choices are of equal distance from each other.

The popular Likert scale (*strongly agree* to *strongly disagree*) illustrates a scale with theoretically equal intervals among responses. It has become common practice to treat this scale as a rating scale and assume that the equal intervals hold between the response categories (Blaikie, 2003). However, we have no guarantee that we have equal intervals unless the researcher establishes it. Hence, often the Likert scale (*strongly agree* to *strongly disagree*) is treated as both ordinal and interval data in educational research (hence the term *quasi-interval* in Table 5.4). How researchers consider this scale (or a similar scale, such as *highly important* to *of no importance*) is critical in the choice of statistic to use to analyze the data. Ordinal scales require nonparametric statistical tests, whereas interval scales require parametric tests. Parametric tests require the underlying data to meet strict assumptions (e.g., normal distributions), while nonparametric statistics do not require those assumptions. Some researchers stress the importance of viewing Likert scales as ordinal data (Jamieson, 2004). Others indicate that the errors for treating the Likert scale results as interval data are minimal (Jaccard & Wan, 1996). In order to consider treating Likert data on an interval scale, researchers should develop multiple response options, determine whether their data are normally distributed, and establish whether the distance between each value on the scale is equal. If this cannot be done, then you should treat the Likert scale and scales like "extent of importance" or "degree of agreement" as ordinal scales for purposes of data analysis.

Finally, a **ratio (**or **true zero) scale** is a response scale in which participants check a response option with a true zero and equal distances between units. Although educational researchers seldom use this type of scale, examples of it are the height of individuals (e.g., 50 inches or 60 inches) and income levels (from zero dollars to $50,000 in increments of $10,000).

Combined Scales In educational research, quantitative investigators often use a combination of categorical and continuous scales. Of these, interval scales provide the most variation of responses and lend themselves to stronger statistical analysis. The best rule of thumb is that if you do not know in advance what statistical analysis you will use, create an interval or continuous scale. Continuous scales can always be converted into ordinal or nominal scales (Tuckman & Harper, 2012) but not vice versa.

Think-Aloud about Finding and Selecting an Instrument

Often, we find beginning researchers developing their own instruments, rather than taking the time to locate an existing instrument suitable for their study. Unquestionably, developing your own instrument requires knowledge about item construction, scale development, format, and length. Although some campuses may have courses that teach this information, most students develop instruments with little feedback from advisers or consultants about how to design the instrument.

Instead of developing your own instrument, we would encourage you to locate or modify an existing instrument. An example can illustrate how you might find this instrument. Figure 5.7 showed you an instrument on students' attitudes toward adaptation to college. How did we find this instrument?

We knew that we wanted to measure the variable "student adjustment to college" because we had formed a general quantitative research question: "What factors influence how freshman students adjust to college?" We began by searching the ERIC database for an instrument using the descriptors of "students" and "college" and "adjustment" in my online key word search. Although we found several good journal articles on the topic, none included a useful instrument. Examining the references in these articles still did not net an instrument that would work.

Two references in our academic library index available instruments: the Buros Center for Testing's *Tests in Print (TIP)* and *Mental Measurements Yearbook (MMY)*. You may recall from earlier in this chapter that these publications contain information about commercially available tests and instruments, including attitudinal instruments. Although our library contains book copies of the *TIP* and *MMY*, we typically use the online version of these books available in our library or through Buros website at buros.org.

We accessed the *MMY* electronically and searched for any instruments that related to students, especially college students. After trying out several words, we found the Student Adaptation to College Questionnaire (SACQ). The brief description of the SACQ gives basic information about the instrument, such as its purpose, the population for its use (i.e., college freshman), publication date (1989), and scales. This review also contained price information at the time of the review, the time required to administer it (20 minutes), authors, publishers, and a cross-reference to a review of the instrument to be found in the *MMY*, 11th edition (Kramer & Conoley, 1992).

Next, we were curious about whether scores reported on this instrument had evidence of validity and reliability, so we read the linked review from the 11th edition of the *MMY* (Kramer & Conoley, 1992) written by E. Jack Asher Jr., professor emeritus of

psychology at Western Michigan University in Kalamazoo. We also searched the ERIC database and located a meta-analytic review article based on the questionnaire by Credé and Niehorster (2012) in the journal *Educational Psychology Review.*

Focusing mainly on the review by Asher, we found that it addressed the following:

- The purpose of the questionnaire
- The subscales on the questionnaire
- Norms on the questionnaire obtained by the authors by administering it from 1980 through 1984
- Evidence for validity of the scores from the instrument (i.e., criterion-related and construct validity)
- Evidence for reliability of the scores based on coefficients of internal consistency
- The value of the manual, especially the inclusion of potential ethical issues in using the instrument
- The overall value of the instrument for college counselors and research applications
- The limitations of the instrument

After reviewing all these topics about the questionnaire, Asher concluded by summarizing an overall positive reaction to the instrument. Although somewhat dated (1989), the instrument has been widely and recently used and positively reviewed. We decided it would be a good instrument to survey college students.

Next, we visited the website of the publisher, Western Psychological Services, for permission to use the instrument and to obtain copies for my study.

Using an instrument already developed by someone else, finding one with good validity and reliability scores, and locating a manual for using the instrument led to the early identification of means for collecting my data. You may not be as fortunate in locating an instrument as quickly, but certainly this process is better than developing numerous versions of your own instrument that might have questionable validity and reliability.