41. DESCRIPTIVE RESEARCH METHODOLOGIES

Nancy Nelson Knupfer
KANSAS STATE UNIVERSITY

Hilary McLellan
MCLELLAN WYATT DIGITAL

Descriptive statistics play an important role in educational research, and thus it is essential to understand the nature and function of such research. As researchers consider possible study designs, statistical analysis, and final reports about any particular topic, it is critical that they maintain focus on the questions to be answered by the research. Those questions will determine the appropriate approach to the investigation and its resulting methodology. The research questions will position the analysis into one of two areas: that which describes data according to a particular organization, and that which draws inferences about cause and effect.

The former, descriptive research, holds a valuable place within education, because in contrast to laboratory experiments, the human nature of educational research is critical to the result. Educational environments and experiences inherently contain many extraneous variables that cannot be controlled in a realistic situation, often call for careful observation of specific life situations, and can require the collection of data from a large number of people spread throughout a wide geographic region.

While writing this chapter, we discussed the topic of descriptive research with several people and discovered that some people were confused about the definition and purpose of descriptive research. As we discussed the possibilities of descriptive research and compared it to other types of research methodologies, we had to conclude that all research contains some degree of description, and thus the term itself could be confusing to the reader. The research discussions that one typically encounters are "quantitative" versus "qualitative" methodology, designs labeled "experimental," "quasi-experimental," "case study," and so on. It is rare to find a research methods class or even a chapter that focuses strictly on descriptive research. Indeed, the term is often given a paragraph or two of importance or ignored altogether. Yet a review of the leading journals related to the field of educational technology shows that descriptive research holds an important place in the study of human interaction and learning (see 39.1, 39.5.2.2, 40.1, 40.1.1, 42.1). Indeed, the descriptive component is critical to educational research, because educational events cannot be reduced to a controlled laboratory environment. The types of questions generated in educational research, particularly with respect to the constructivist paradigm and social implications, require descriptions that help to explain the data and direct emergent prescriptions for educational events.

41.1 WHAT IS DESCRIPTIVE RESEARCH?

Descriptive research does not fit neatly into the definition of either quantitative or qualitative research methodologies, but instead it can utilize elements of both, often within the same study. The term descriptive research refers to the type of research question, design, and data analysis that will be applied to a given topic. Descriptive statistics tell what is, while inferential statistics try to determine cause and effect.

The type of question asked by the researcher will ultimately determine the type of approach necessary to complete an accurate assessment of the topic at hand. Descriptive studies, primarily concerned with finding out "what is," might be applied to investigate the following questions: Do teachers hold favorable attitudes toward using computers in schools? What kinds of activities that involve technology occur in sixth-grade classrooms and how frequently do they occur? What have been the reactions of school administrators to technological innovations in teaching the social sciences? How have high school computing courses changed over the last 10 years? How do the new multimedia textbooks compare to the print-based textbooks? How are decisions being made about using Channel One in schools, and for those schools that choose to use it, how is Channel One being implemented? What is the best way to provide access to computer equipment in schools? How should instructional designers improve software design to make the software more appealing to students? To what degree are special-education teachers well versed concerning assistive technology? Is there a relationship between experience with multimedia computers and problem-solving
41. DESCRIPTIVE RESEARCH METHODOLOGIES

skills? How successful is a certain satellite-delivered Spanish course in terms of motivational value and academic achievement? Do teachers actually implement technology in the way they perceive? How many people use the AECT gopher server, and what do they use if for?

Descriptive research can be either quantitative or qualitative. It can involve collections of quantitative information that can be tabulated along a continuum in numerical form, such as scores on a test or the number of times a person chooses to use a certain feature of a multimedia program, or it can describe categories of information such as gender or patterns of interaction when using technology in a group situation. Descriptive research involves gathering data that describe events and then organizes, tabulates, depicts, and describes the data collection (Glass & Hopkins, 1984). It often uses visual aids such as graphs and charts to aid the reader in understanding the data distribution. Because the human mind cannot extract the full import of a large mass of raw data, descriptive statistics are very important in reducing the data to manageable form. When in-depth, narrative descriptions of small numbers of cases are involved, the research uses description as a tool to organize data into patterns that emerge during analysis. Those patterns aid the mind in comprehending a qualitative study and its implications.

Most quantitative research falls into two areas: studies that describe events and studies aimed at discovering inferences or causal relationships. Descriptive studies are aimed at finding out “what is,” so observational and survey methods are frequently used to collect descriptive data (Borg & Gall, 1989). Studies of this type might describe the current state of multimedia usage in schools or patterns of activity resulting from group work at the computer. An example of this is Cochenour, Hakes, and Neal’s (1994) study of trends in compressed video applications with education and the private sector.

Descriptive studies report summary data such as measures of central tendency including the mean, median, mode, deviance from the mean, variation, percentage, and correlation between variables. Survey research commonly includes that type of measurement, but often goes beyond the descriptive statistics in order to draw inferences. See, for example, Signer’s (1991) survey of computer-assisted instruction and at-risk students, or Nolan, McKinnon, and Soler’s (1992) research on achieving equitable access to school computers. Thick, rich descriptions of phenomena can also emerge from qualitative studies, case studies, observational studies, interviews, and portfolio assessments. Robinson’s (1994) case study of a televised news program in classrooms and Lee’s (1994) case study about identifying values concerning school restructuring are excellent examples of case studies.

Descriptive research is unique in the number of variables employed. Like other types of research, descriptive research can include multiple variables for analysis, yet unlike other methods, it requires only one variable (Borg & Gall, 1989). For example, a descriptive study might employ methods of analyzing correlations between multiple variables by using tests such as Pearson’s Product Moment correlation, regression, or multiple regression analysis. Good examples of this are the Knupfer and Hayes (1994) study about the effects of the Channel One broadcast on knowledge of current events, Manaev’s (1991) study about mass media effectiveness, McKenna’s (1993) study of the relationship between attributes of a radio program and it’s appeal to listeners, Orey and Nelson’s (1994) examination of learner interactions with hypermedia environments, and Shapiro’s (1991) study of memory and decision processes.

On the other hand, descriptive research might simply report the percentage summary on a single variable. Examples of this are the tally of reference citations in selected instructional design and technology journals by Anglin and Towers (1992); Barry’s (1994) investigation of the controversy surrounding advertising and Channel One; Lu, Moran, Letchlorlam, Lee, and Dike’s (1993) investigation of the international utilization of media in education (1993); and Pettersson, Metallinos, Muffoletto, Shaw, and Takakuwa’s (1993) analysis of the use of verbo-visual information in teaching geography in various countries.

Descriptive statistics utilize data collection and analysis techniques that yield reports concerning the measures of central tendency, variation, and correlation. The combination of its characteristic summary and correlational statistics, along with its focus on specific types of research questions, methods, and outcomes is what distinguishes descriptive research from other research types.

Three main purposes of research are to describe, explain, and validate findings. Description emerges following creative exploration, and serves to organize the findings in order to fit them with explanations, and then test or validate those explanations (Krathwohl, 1993). Many research studies call for the description of natural or man-made phenomena such as their form, structure, activity, change over time, relation to other phenomena, and so on. The description often illuminates knowledge that we might not otherwise notice or even encounter. Several important scientific discoveries as well as anthropological information about events outside of our common experiences have resulted from making such descriptions. For example, astronomers use their telescopes to develop descriptions of different parts of the universe, anthropologists describe life events of socially atypical situations or cultures uniquely different from our own, and educational researchers describe activities within classrooms concerning the implementation of technology. This process sometimes results in the discovery of stars and stellar events, new knowledge about value systems or practices of other cultures, or even the reality of classroom life as new technologies are implemented within schools.

Educational researchers might use observational, survey, and interview techniques to collect data about group dynamics during computer-based activities. These data could then be used to recommend specific strategies for implementing computers or improving teaching strategies. Two excellent studies concerning the role of collaborative groups were conducted by Webb (1982), and Rysavy and
Sales (1991). Noreen Webb’s landmark study used descriptive research techniques to investigate collaborative groups as they worked within classrooms. Rysavy and Sales also apply a descriptive approach to study the role of group collaboration for working at computers. The Rysavy and Sales approach did not observe students in classrooms, but reported certain common findings that emerged through a literature search.

Descriptive studies have an important role in educational research. They have greatly increased our knowledge about what happens in schools. Some of the important books in education have reported studies of this type: Life in Classrooms, by Philip Jackson; The Good High School, by Sara Lawrence Lightfoot; Teachers and Machines: The Classroom Use of Technology Since 1920, by Larry Cuban; A Place Called School, by John Goodlad; Visual Literacy: A Spectrum of Learning, by D. M. Moore and Dwyer; Computers in Education: Social, Political, and Historical Perspectives, by Muffoletto and Knupfer; and Contemporary Issues in American Distance Education, by M. G. Moore.

Henry J. Becker’s (1986) series of survey reports concerning the implementation of computers into schools across the United States as well as Nancy Nelson Knupfer’s (1988) reports about teacher’s opinions and patterns of computer usage also fit partially within the realm of descriptive research. Both studies describe categories of data and use statistical analysis to examine correlations between specific variables. Both also go beyond the bounds of descriptive research and conduct further statistical procedures appropriate to their research questions, thus enabling them to make further recommendations about implementing computing technology in ways to support grassroots change and equitable practices within the schools. Finally, Knupfer’s study extended the analysis and conclusions in order to yield suggestions for instructional designs involved with educational computing.

### 41.1.1 The Nature of Descriptive Research

The descriptive function of research is heavily dependent on instrumentation for measurement and observation (Borg & Gall, 1989). Researchers may work for many years to perfect such instrumentation so that the resulting measurement will be accurate, reliable, and generalizable. Instruments such as the electron microscope, standardized tests for various purposes, the United States census, Michael Simonson’s questionnaires about computer usage, and scores of thoroughly validated questionnaires are examples of some instruments that yield valuable descriptive data. Once the instruments are developed, they can be used to describe phenomena of interest to the researchers.

The intent of some descriptive research is to produce tabistical information about aspects of education that interests policymakers and educators. The National Center for Education Statistics specializes in this kind of research. Many of its findings are published in an annual volume called Digest of Educational Statistics. The center also administers the National Assessment of Educational Progress (NAEP), which collects descriptive information about how well the nation’s youth are doing in various subject areas. A typical NAEP publication is The Reading Report Card, which provides descriptive information about the reading achievement of junior high and high school students during the past 2 decades.

On a larger scale, the International Association for the Evaluation of Education Achievement (IEA) has done major descriptive studies comparing the academic achievement levels of students in many different nations, including the United States (Borg & Gall, 1989). Within the United States, huge amounts of information are being gathered continuously by the Office of Technology Assessment, which influences policy concerning technology in education. As a way of offering guidance about the potential of technologies for distance education, that office has published a book called Linking for Learning: A New Course for Education, which offers descriptions of distance education and its potential.

There has been an ongoing debate among researchers about the value of quantitative (see 40.1.2) versus qualitative research, and certain remarks have targeted descriptive research as being less pure than traditional experimental, quantitative designs. Rumors abound that young researchers must conduct quantitative research in order to get published in Educational Technology Research and Development and other prestigious journals in the field. One camp argues the benefits of a scientific approach to educational research, thus preferring the experimental, quantitative approach, while the other camp posits the need to recognize the unique human side of educational research questions and thus prefers to use qualitative research methodology. Because descriptive research spans both quantitative and qualitative methodologies, it brings the ability to describe events in greater or less depth as needed, to focus on various elements of different research techniques, and to engage quantitative statistics to organize information in meaningful ways. The citations within this chapter provide ample evidence that descriptive research can indeed be published in prestigious journals.

Descriptive studies can yield rich data that lead to important recommendations. For example, Galloway (1992) bases recommendations for teaching with computer analogies on descriptive data, and Wehrs (1992) draws reasonable conclusions about using expert systems to support academic advising. On the other hand, descriptive research can be misused by those who do not understand its purpose and limitations. For example, one cannot try to draw conclusions that show cause and effect, because that is beyond the bounds of the statistics employed.

Borg and Gall (1989) classify the outcomes of educational research into the four categories of description, prediction, improvement, and explanation. They say that descriptive research describes natural or man-made educational phenomena that is of interest to policymakers and
educators. Predictions of educational phenomenon seek to determine whether certain students are at risk and if teachers should use different techniques to instruct them. Research about improvement asks whether a certain technique does something to help students learn better and whether certain interventions can improve student learning by applying causal-comparative, correlational, and experimental methods. The final category of explanation posits that research is able to explain a set of phenomena that leads to our ability to describe, predict, and control the phenomena with a high level of certainty and accuracy. This usually takes the form of theories.

The methods of collecting data for descriptive research can be employed singly or in various combinations, depending on the research questions at hand. Descriptive research often calls upon quasi-experimental research design (Campbell & Stanley, 1963). Some of the common data collection methods applied to questions within the realm of descriptive research include surveys, interviews, observations, and portfolios.

41.2 DESCRIPTIVE RESEARCH METHODOLOGY

41.2.1 Survey Methods

Although survey research can yield data that are compared and analyzed at a more complicated level, the simplest use to which survey data can be put is a description of how the total sample has distributed itself on the response alternatives for a single questionnaire item (see 6.6.1, 37.4). These are sometimes called marginal tabulations. For example, news media often report the result of public-opinion polls in terms of marginal tabulations: 50% of the sample were in favor of a particular governmental policy, 30% disagreed with it, and 20% were unsure or had no opinion. Survey research in education often yields this type of normative description.

As an illustration, we may consider a study by N. N. Knupfer (1994) that was designed to investigate the interest of junior high and high school students in viewing the Channel One news program. A Likert-type scale of 34 items plus open-ended questions, each describing an element of the televised news program, was administered to a sample of approximately 2,000 students randomly drawn from an eligible population of approximately 16,000 high school students in a Southwest school district. Students were asked to describe their level of agreement with the statements about the value of the Channel One program based on a five-point scale. In the data analysis, the mean score of the entire sample on each opinion item was determined. This form of data analysis provides an interesting description of students' interest in this form of educational programming and their opinion about its quality. We find, for instance, that the junior high students liked the program more than the high school students, even though Channel One was originally designed for the high school audience. This, of course, has implications for the future design of the program.

Descriptions of this type often provide important leads in identifying needed tactics and changes within the instructional design of such material. Also, we should note that since proper sampling procedures were employed, Knupfer was able to generalize these descriptive findings from the sample to the population from which they were drawn.

In addition to summary statistics, survey data may be used to explore relationships between two or more variables. Researchers who are aware of the possibilities of investigating relationships in the survey data will make a more substantial research contribution than researchers who limit their data analysis to single variable descriptions. Questionnaires may refer to past, present, or future phenomena, so they have a great deal of flexibility.

Because of the various data analysis techniques that can be applied to survey research, surveys cannot automatically be assumed to be strictly descriptive. Consider a study by Davidson and Ritchie (1994) as an example. Their survey collected information from students, teachers, and parents within a school, analyzed the data with descriptive methods, conducted $t$ tests to look for significant differences between groups, and was labeled as a "case study" within the title of their publication.

Two critical components of survey research are sound methodology and well-designed data collection instruments. It is important that data collection instruments used for surveys ensure the ability to collect standardized information and do so in a way that will yield quantifiable results. The same instruments should be distributed to all subjects so that data can be summarized and compared. Although there are a variety of potential methods and instruments, questionnaires and individual interviews are the most common collection techniques used in survey research (Borg & Gall, 1989).

One also can conduct telephone interviews or examine records as part of the main data collection or as follow-up with nonrespondents to assess any differences between the respondent and nonrespondent groups. A common mistake that researchers make is to ignore the nonrespondents. Nonrespondents can indicate bias in the study results and need some follow-up analysis.

It is also important to determine if the variables are related to each other in terms of time-bound or time-ordered association. Time-bound association refers to the same point in time for all questions worded a certain way. Time-ordered association means that the items can be temporarily ordered relative to each other. For example, now ask respondents to tell how they feel about $X$ and to recall how they felt about $Y$.

Surveys can use cross-sectional, longitudinal, or Delphi techniques (Borg & Gall, 1989). Cross-sectional surveys use simple random, stratified, or cluster sampling techniques. Longitudinal surveys are administered at different points of time in order to study changes or explore time-ordered associations. The Delphi technique employs questionnaires but is quite different from the typical questionnaire survey.
Common types of longitudinal studies include trend studies, cohort studies, and panel studies. Trend studies sample a given general population at each data collection point, and although the same individuals are not surveyed, the sample is drawn from the same population. Because the population is general, the specific individuals in it are constantly changing.

Cohort studies follow a specific population over a period of time. Like trend studies, different individuals are drawn from the same population at each data collection point. The difference is that the cohort group does not shift in the makeup of the population, so the sample is always drawn from the same pool of people. To illustrate this concept, envision the population as being fixed, such as the group of people who received the doctor of philosophy degree in educational technology in a specific year. The sample will always be drawn from that group of people.

Finally, panel studies involve the repeated surveying of the same individual subjects at the outset of the study and then again at each subsequent data collection point. Since the study follows the same individuals, the researcher can note data about individual changes over time and explore possible reasons why the individuals have changed. Loss of participants in this type of study would be a serious problem (Borg & Gall, 1989).

The Delphi technique was developed by the RAND Corporation as a method of predicting future defense needs, but it can be modified to various situations where a consensus is needed. It can be used to identify problems, define needs, establish priorities, and identify and evaluate solutions (Borg & Gall, 1989). Basically, the Delphi technique asks a set of questions, then revises that set of questions repeatedly, based on the responses to each step before it. Constant refining eventually whittles the information down to a key set of refined data.

Surveys can take several forms and data can be collected in many ways. Surveys can be in the form of written questionnaires, personal interviews, or telephone interviews. They can be distributed by mail (paper or electronic), in person, or electronically. Commercial sources often use electronic devices to help collect the data, but educational surveys tend to be the low-budget variety that lean toward inexpensive data collection techniques.

No matter what method is eventually used, there are certain major issues involved in choosing a data collection strategy. Some of these issues include sampling, type of population, question form, question content, response rates, costs, available facilities, length of data collection, and computer-assisted techniques for data collection. The type of data collection needs to vary with each situation, and the selection of a particular method is a complex process involving consideration of many aspects of the survey research process (Fowler, 1993).

Subjects are more likely to answer a survey if they can relate to it as being purposeful, if it is not time consuming, if it appears aesthetically pleasing and legible, and if it presents itself as an ethical piece of research. But even the most well-designed study and survey instrument will do no good unless the researcher has determined a way to obtain the best possible return rate. Although specific respondents should not be matched with their answers, the researcher will need a way to keep track of respondents and nonrespondents. Mailed surveys always require a second or third mailing to obtain a higher return rate. Keeping track of the respondents will save the trouble, expense, and waste associated with resending the survey to the entire sample group, not to mention the irritation it will cause and the possibility of tainted data if people were to send in two responses. The goal is to collect one response from each subject, not multiple responses from some of the subjects. Color coding survey instruments by groups can greatly aid organization and follow-up data collection, as can a good recording system for tracking survey returns.

41.2.2 Interviews

Face-to-face interviews and telephone interviews provide an opportunity to follow a line of questioning to obtain more in-depth information. Telephone interviews can be more time efficient, but face-to-face interviews would be appropriate for studies that need the feeling of personal contact, because they allow the researcher to establish rapport with the respondent. In general, interviews produce a better return rate than mailed questionnaires (Fowler, 1993).

To increase the reliability of data collected by interviews, it is important to follow certain guidelines and to train all interviewers to use the same techniques. The most effective interviews require that the researcher develop a guide to use during the interview process. The guide will help organize the information and is usually followed in order.

Personal-interview procedures can be advantageous because they promote the highest response rate and allow the researcher to respond to questions from the respondent, probe for adequate answers, follow complex instructions that might otherwise confuse the respondent, and provide longer interviews than those conducted by telephone. In addition, personal interviews offer the opportunity to use multiple methods of data collection such as observations and visual cues (Fowler, 1993). Disadvantages of personal interviews are that they require more staff time and travel, leading to more expense than telephone interviews and mailed questionnaires. In addition, certain populations are more difficult to reach due to physical factors.

Telephone interviews are potentially advantageous because they are less time consuming and less expensive. They can also employ random dialing to access a truly randomized sample. Disadvantages of telephone interviews include sampling limitations due to restrictions imposed by access to telephones, lack of ability to provide visual information, and lack of ability for the researcher to observe the participant.

41.2.3 Mailed Questionnaires

The major advantages of mailed questionnaires are that one can reach large numbers of people from wide geographic areas, the respondents have time to reflect on their answers
or even check information prior to responding, and the relatively low cost of administration that results from the distribution method, as well as the need for only a small staff of people and minimal facilities (Fowler, 1993). The major disadvantages include the lower response rate and the need to take special care with designing questions that will be self-administered.

Although all surveys should contain some open-ended questions to allow for participant ideas to emerge, a significant number of respondents will not take the time to answer those questions with in-depth answers, and some respondents will not understand the instructions for completing the open-ended questions. Therefore, simple formats that allow users simply to check responses on the questionnaire work the best (Belson, 1981; J. Converse, 1987; Fowler, 1993). Brevity will be helpful there.

A letter of transmittal on letterhead stationery will be critical for mailed questionnaires. The content of the letter as well as its presentation will have an influence on the response rate. To provide a good visual presentation, make the letter look professional by its style and print quality. Be sure to include a postage-paid return envelope.

Within the body of the letter, it is important to state the purpose of the research, the importance of study, the importance of respondents, a reasonable but specific time limit, an assurance of confidentiality, and an offer to share results, as well as to thank the respondent. Concise wording should make it possible to keep this letter to one page. Special delivery options increase the look of importance but also can be detrimental in terms of attitude. Some people get annoyed with the packaging, so you need to make decisions about how it would likely be perceived by the respondents.

41.2.3.1. Designing Good Questions. Good surveys use questions that are interesting because of their relationship to what they are supposed to be measuring. Good questionnaires provide a script for the entire process that the researchers are to follow, word questions so that they mean the same thing to everyone, and inform participants about the kinds of answers that are appropriate. Questionnaires should provide definitions for any terminology that might be confusing; use complete sentences, and offer a limited set of answers. The expression “I don’t know” should be used very carefully because it can encourage participants to mark it rather than think more deeply about the question (Fowler, 1993).

41.2.4 Observational Research Methods

Observational research has a long tradition across a large number of disciplines. Observational methods are concerned with naturally occurring behaviors observed in natural contexts or in contexts that are contrived to be realistic. Suen and Ary (1989) delineate two approaches to the direct observation of behavior: a quantitative approach that employs structured observation methods, and a qualitative approach that features unstructured observation. Unstructured observation-al methods are often used for pilot studies; the data collected yield categories shaped into instruments that are then employed for structured observation on a larger scale.

Observational research requires the systematic, direct observation of behavior. Many analysts consider it superior to other methods of data collection because the data are gathered directly without intermediary instruments such as tests or questionnaires. Since observation provides a direct approach, it requires very little conceptual inference to connect the data to the phenomenon under scrutiny (Suen & Ary, 1989). Further, if used properly, the observational method can overcome the limitations of the self-report method, for example, biased or selective reporting (Borg & Gall, 1989).

The systematic observation of behavior in educational and psychological research can yield a wealth of valuable information, but it is a complicated and labor-intensive process. The process of defining and developing coding schemes, then training observers to acceptable levels of agreement, can be time consuming (Bakeman & Gottman, 1986; Borg & Gall, 1989). The complexity of both the behavior and observation methods can lead to technical errors, such as miscoding, as well as problems in achieving and maintaining satisfactory interrater agreement (Borg & Gall, 1989). But structured observation offers a degree of certainty and replicability that unstructured narrative reports cannot achieve (Bakeman & Gottman, 1986). Both structured and unstructured observational methods can yield thick, rich descriptions that achieve the depth that other research methodologies lack.

This section deals primarily with structured observation research. However, unstructured observation, naturalistic contrived situations, situational testing, and portfolios will also be considered as topics in this section on observational research. 41.2.4.1. Purposes of Observational Research in Educational Technology. Observation methods can be employed productively to support many purposes in the area of educational communication and technology. Observation may be used to understand how people interact with technology in various stages of design and implementation. Observational data can help to improve the materials design as well as their utilization within particular settings. For example, observation can be used in both formative evaluation and summative evaluation of educational software. Questions answered might be: How do learners respond to and interact with a specific program? How do learners interact with a new hardware system? Observational research can help to determine if one has a technology that works. Observation makes it possible both to explore the implementation of a particular technological innovation and assess the instructional outcomes.

Flagg (1990) reports that observation is especially useful with children, because they frequently produce verbal and nonverbal behaviors from which one infers their comprehension of media program segments. Many analysts (Amarel, 1982; Chen, 1985; Dickson, 1985; Fraser & Burkhardt, 1984; McLellan, 1991; McLellan, 1994; Shein-
gold, Kane & Endreweit, 1983; Webb, 1982; Webb, Ender & Lewis, 1986) have suggested that the social interactions that emerge in a computer-learning environment are a fundamental aspect of the computer-mediated experience for students and teachers. Thus an understanding of the characteristics of social interaction is vital to understanding the educational significance of any implementation of computers in the schools.

41.2.4.2. Types of Observational Research

41.2.4.2.1. Structured Observations. Structured observation can be defined as a particular approach to quantifying behavior. Two fundamental components of observational research include the use of predefined, behavior-code catalogs and reliable observers. Because human observers are so important to the instrumentation, reliability issues loom especially large in observational research (Bakeman & Gottman, 1986).

In contrast to informal observations, structured observations are systematic and controlled. Many parameters must be rigidly prescribed and followed, including what, when, where, and how to observe, as well as the methods for recording and analyzing the data. Investigators must articulate beforehand what they hope to find out, what seems important conceptually, and specify accurate, reliable ways to measure those concepts (Suen & Ary, 1989). A hypothesis that tentatively describes the relationships among the variables is also established before the actual observation is carried out. The process of observation is rigidly controlled, and the nature of the data gathered is well defined so that data quality can be assessed by interobserver comparisons, and quantitative data analysis can be employed. The analysis of structured, observational data may examine frequency, duration, and interrelationships between events in an attempt to identify meaningful patterns that are reported with descriptive statistics.

Thus, complex phenomenon, such as human interaction with technology, is first conceptually reduced to a number of measurable and observable behavioral variables. These variables are fine tuned and defined, often through unstructured observation and review of videotaped recordings. A system for measuring these variables is determined before the actual observation is implemented. Structured observation codes define specifically, before observation begins, the behaviors and situations thought to be relevant to the focus and purpose of the study (Flagg, 1990). Other behaviors are inferential, requiring some conclusions on the part of the observer. For example, from verbal or nonverbal behavior the observer infers why someone initiates the use of a technology system.

The structured observation record usually includes frequency counts of how often specific behaviors occur and the duration or interval involved (Flagg, 1990). The data can provide a quantitative description or explanation of the phenomenon (Suen & Ary, 1989; Flagg, 1990). It may be possible to gather observational data "live," during ongoing events as they unfold, or it may be possible to videotape events for later observation and analysis. In some instances, observations are made through a one-way mirror so that the subjects of the study are not aware of the observation. Statistical analyses are performed on these quantitative observational data to determine whether evidence of the hypothesized relationships indeed exist.

Observational research is susceptible to several common mistakes that the researcher must take into consideration and work to avoid. These mistakes include insufficient training of observers, use of a complex observation form that requires too much from the observer or requires observers to make excessively precise discriminations among behaviors, failure to take adequate precautions to avoid having observers disturb or change the situation they are to observe, failure to use at least two observers in order to determine interrater reliability, failure to ensure that observers work independently of each other, contamination of data collection, failure to use random sampling techniques when appropriate, and lack of tight controls that help prevent observer drift and reliability decay (Bakeman & Gottman, 1986).

41.2.4.2.2. Unstructured Observations. The purpose of unstructured observation is to furnish an unselective, detailed, continuous description of behavior (Flagg, 1990). In unstructured observation, the evaluator records all of the user's activities, and if the user is interacting with a media program, the program status is continually recorded. Thus, the observation is continuous and unselective while it is ongoing. This technique may be valuable in a preliminary study to determine observational categories for subsequent structured observation. Or it may be useful for the formative evaluation of media programs, studying the actual interaction between the user and the program content and the machine (interfaces, screen, etc.).

Flagg (1990) emphasizes that unintended effects of a media system are best detected through unstructured observation. Related to this, Norman (1991) claims the best kind of task analysis is in the field. You have to learn how to watch people. It can help to bring in a trained observer, an anthropologist, psychologist, or sociologist who knows about the people who are under study, can make videotapes of typical interactions of people with technology, and show you how computers are used in the environment for which your design is destined. Studying people as they perform familiar, well-understood tasks can provide valuable insights about possible ways to represent tasks on the computer, providing a valuable resource for the design of future media as well as formative evaluation of new designs that are in development (Mountfort, 1991).

Unstructured observation offers several advantages in evaluating user friendliness issues. For example, the frequency of use of functions and the length of time the system is used can be calculated from an unstructured observation record, just as from structured observation. Furthermore, the unstructured record adds a richness to the evaluation that is missing from the structured approach. Unstructured observation can be continuous, or it can be structured around time sampling for example, recording
observations once every 5 minutes for 1 minute, or by sampling only when certain events occur, for example, whenever the help mode in a program is accessed (Flagg, 1990).

Flagg (1990) reports that eventually, for analysis and interpretation, the unstructured observation record is coded in categories and tallied in a manner similar to the structured observation. At the same time, the rich qualitative observations are still available to whatever happens, including unexpected interactions or sources of confusion that are extremely valuable to document. Consequently, the categories resulting from this method might be different from those defined a priori in the structured observation schedule. It is possible to go back into an unstructured narrative to study questions that might not have occurred to the evaluators earlier.

According to Flagg (1990), the main disadvantage of unstructured observation for formative evaluation is the time and labor required to collect and analyze sets of extensive observations. And training is needed so that observers are recording similar data in an unbiased manner. Since it is time intensive, unstructured observations on a small sample are sometimes used to define the behaviors for a larger-scale, structured-observation study.

41.2.4.3. Considerations within Observation

41.2.4.3.1. Types of Observational Variables. Three types of observational variables may be distinguished: descriptive, inferential, and evaluative. Borg and Gall (1989) report that descriptive variables have the advantage that they require little inference on the part of the observer, so that they generally yield reliable data. Descriptive variables are considered to be "low inference" variables.

Other observational variables call upon the observer to make an inference before a variable can be scored. As Borg and Gall (1989) explain:

Observers may be asked to record the self-confidence with which a teacher explains a mathematical concept. Some teachers may speak with a good deal of confidence, whereas others may appear uncertain, confused, or anxious because their understanding of the topic is weak. Confidence, uncertainty, confusion, and anxiety are not behaviors but rather are inferences made from behavior. These are often referred to as "high inference variables." It is much more difficult to collect reliable data when observers are asked to make inferences from behavior (p. 493).

Finally, with the third type of observational variable, the observer must make an evaluative judgment. The researcher may be interested in evaluating the quality of the teacher's explanation of a mathematical concept. Quality ratings are not behavior but rather are inferences made from behavior. Because it is difficult to make reliable observations of evaluative variables, we need to collect examples of behavior that define points along the continuum of excellent-to-poor explanations and use these in training the observers (Borg & Gall, 1989).

In the progression from low inference to high inference to evaluative variables, the observers' task becomes more complex (Borg & Gall, 1989). Any changes in the observational task that make it more complex tends to lower observer reliability or agreement (Reid, 1987). Reid has devised some methods for estimating the complexity of the variables to be observed.

To ensure accurate recording, require observers to score only one behavior at a given point in time. For example, most observers would find it quite difficult to record certain aspects of the teacher's talk and at the same time record the percentage of children who appear to be paying attention to the teacher. Concomitantly, the reliability of both sets of observations would probably be low, although in order to understand complex interactions, including the duration of events that start and stop in overlap, it may be desirable to monitor more than one event simultaneously. Thus mediated observations such as video records or a computer-based coding system such as PLEXYN, which automates the observations as much as possible through a carefully designed coding system, may be extremely valuable.

41.2.4.3.2. The Observation Form. Once the observational variables to be used in the research study are identified, you need to develop a form on which they can be recorded. A paper-and-pencil observational form is one option that can accommodate a variety of scoring procedures. This form of scoring can be designed so as to require a minimum of effort on the part of the observer and can usually be developed so as to require the observer to make a few inferences. However, it is also possible to develop media-based observation forms where data are entered with a hand-held device or a computer keyboard.

With some more complex observation forms, the observer must record not only the behavior as it occurs but also evaluate the behavior on a rating scale. This scoring procedure requires a higher level of inference on the part of the observer (Borg & Gall, 1989). The observer must not only record the behavior but must also evaluate it, and this is much more difficult to do objectively. If you use a rating scale as part of your scoring procedure, you should avoid the common mistake of attempting to obtain excessively precise discrimination from the observer. Most human behavior studied in educational research cannot be reliably rated on more than five levels. The 3-point rating scale, breaking the behavior observed into such categories as "above-average," "average," and "below-average," is often as fine a discrimination as can be made with satisfactory reliability. Five-point rating scales, however, are often used in educational research and can be employed effectively in observing well-defined behavior. It is almost never advisable to attempt to obtain ratings for finer than the 5-point scale. Furthermore, the more inference the observer must use in making the rating, the fewer ratings levels should be employed (Borg & Gall, 1989).

41.2.4.3.3. Mediated Observation Techniques. Traditionally, paper and pencil have been used to record observational data. However, newer mediated observation techniques are available to support data recording and analysis. Behaviors can be recorded on audio and videotape so that observations can be made from these records rather
than, or in addition to, the "live" event. Computers and related hand-held devices can be used as tools to facilitate and improve data gathering by human observers either during the "live" event or during a review of an audio- or videotape recording (Borg & Gall, 1989; Stephenson, 1979, 1983; Fraser & Burkhart, 1984; McLellan, 1991; McLellan, 1994). Observational codes can be entered directly via keyboard or other input, making later analysis easier and more efficient.

Computers offer the advantage of on-line monitoring. On-line monitoring is the process of capturing characteristics of the human-computer interaction automatically, in real time, from an operating system (Borgman, 1986). Keystroke records, audit trails, and logging data are other terms used to describe on-line monitoring. Computers can directly record and store information about a user's interactions with a program, including both transactional (keystroke, screen viewed, etc.) and temporal (duration) features of the interactions (Neal & Simons, 1984; Chen, Liberman & Paisley, 1985; Hawkins, Bosworth, Chewning, Day & Gustafson, 1985; Hasselbring, 1992).

With virtual-reality technology, position-tracking devices make it possible to record human performance precisely and in great detail. These data can be used to provide performance replays as well as performance reports and data graphing. For example, a baseball pitch or a golf swing can be recorded electronically with position-monitoring devices linked to a computer (McLellan, 1994). Sophisticated simulators can electronically "play" a learner's performance based on the learner's precise movements and interactions with simulator controls.

Video- or audiotaping the target behavior has several advantages. These technologies are relatively unobtrusive, and their use can decrease the number of observers necessary, thereby increasing observational reliability (Flagg, 1990). Furthermore, when recorders are used, observers no longer need to make ratings at the time particular events are occurring. These events may be recorded on audiotape or videotape so that they can be replayed several times for careful study or for several observers to rate at their convenience.

Another advantage of recording observations is that the recordings permit you to obtain data on behavior that you did not anticipate you would need at the outset of your study (Borg & Gall, 1989; Flagg, 1990). Recording devices can provide a permanent record for alternate analyses. And recording devices can be used to collect samples of behavior to facilitate the development of an observation form and to provide the basis for training observers. The recordings can be replayed repeatedly, thus making it easier to develop observational categories and to test the reliability with which observers can use these categories in rating behavior (Borg & Gall, 1989). In formative evaluation studies, the recordings can play an additional role by showing user reactions directly to the production staff that must revise the product (Flagg, 1990).

The cost and availability of audio, video, and computer recording equipment is a vital consideration. Each type of media is limited in terms of what it captures. The audiotape records only verbal comments, with no contextual information. The videotape has a restricted field of view. And online monitoring by the computer results in a record of user activities only as they impact on the machine itself. To increase the completeness of the observational data, these methods can be used together or in conjunction with a human observer (Flagg, 1990). In certain situations, technical problems or complexities may pose serious problems. For example, it may be necessary to have several microphones and to adjust the camera frequently so that a reasonably complete record of classroom behavior can be obtained (Borg & Gall, 1989). Technical competence may be required in order to obtain satisfactory audio and video recording.

In some situations it is humanly impractical to collect all of the desired observational data at the same time, so that supplemental data gathering via technology may be all the more helpful in meeting research goals. For example, if several of the behaviors to be rated occur at the same time or closely together, or if more than one person's behavior is monitored, media records may help to capture the data more fully. Media-based recordings offer another advantage in assessing the intrater reliability of observational data; observers can rate the same sample of behavior recorded on videotape (Borg & Gall, 1989).

For interactive media, keystroke records furnish an objective way to observe the learning process. One can pinpoint comprehension difficulties by analyzing what segments of the program were accessed in what sequence, by reviewing right and wrong answers in task and test situations, and by examining error messages and help requests. Keystroke records identify what pathways to revise or eliminate and what new branches to add (Flagg, 1990).

In addition to collecting and examining data recorded directly by the computer, computer tools have been devised to alleviate some of the technical problems of research observation. Borg and Gall (1989) and other analysts report that microcomputers, handheld recording devices, and even handheld calculators have all been used as tools in observational data collection. Possible-event recorders appropriate for behavioral observation have been identified by, among others, Celhoffer, Boukydis, Minde, and Muir (1977); Crossman, Williams, and Chambers (1978); Fernald and Heinecke (1974); Gass (1977); Repp et al. (1985); Sackett, Stephenson, and Ruppenthal (1973); Sanson-Fisher, Poole, Small, and Fleming (1979); Stephenson and Roberts (1977); Stephenson (1979); and Torgerson (1977). With real-time observations of more complicated designs (e.g., multiple subject or multiple behavior), some type of electronic aid such as an electronic real-time event recorder may have to be used.

These tools make it possible to increase both the amount and the complexity of data that a human observer can collect. One example is PLEXYN (Stephenson, 1979, 1983), an event-recording system that encodes the incidence, duration, coincidence, and sequence of human observation entries in real time for subsequent high-speed transcription by computer. The PLEXYN system includes statistical
analysis programs for analyzing the data once the observations have been recorded and reviewed for errors. And it includes programs that make it possible to edit and correct the data transcript and combine files (for example, where video recordings make it possible to record data for the same event more than once). PLEXYN has been applied to a wide array of research problems, including the interactions of student partners in a computer lab (McLellan, 1991; McLellan, 1994).

Berger, Shermis, and Stemmer (1992) list several ways that microcomputers can be used to support observational research. These include recording and timing the events being observed, transcribing the data onto coding sheets, transferring the data from the coding sheets into computer storage ready for data analysis, cleaning up the data by locating coding errors, aggregating and analyzing the data, and interpreting the results of data analysis. Microcomputers have the capacity to produce a variety of graphic data representations, which can greatly help the researcher in understanding his results. Often the cost of doing this activity on a large computer is prohibitive.

41.2.4.3.4. Coding Schemes. A critical component of structured observational research is the catalog of behavior codes identifying the behaviors that will be the basis for data gathering. In fact, Bakeman and Gottman (1986) suggest that the coding scheme is the most important component of observational research. The coding catalog is the measuring instrument of observational research, specifying which behavior is to be selected from the passing stream of activity and recorded for subsequent study.

Developing a coding scheme can be time consuming, involving extensive preparation and fine tuning. Many analysts (Bakeman & Gottman, 1986; Flagg, 1990; Suen & Ary, 1989) recommend that a good starting point for developing a coding scheme is to conduct extensive preparatory unstructured observation, either “live” or using videotapes that can be reviewed many times. As successively refined versions of the coding scheme are developed, they should be extensively reviewed, critiqued, and finally validated by other appropriate experts.

In addition to selecting a coding scheme, the researcher must select a recording procedure. Most recording procedures can be classified into four major categories, including frequency-count recording, duration recording, interval recording, and continuous recording (Bakeman & Gottman, 1986). Before selecting a particular recording strategy, the researcher must determine the units and types of measures that should be used.

Frequency count refers to the number of times a behavioral event occurs. Here, observers wait for an event of interest to occur, and when such an event occurs, it is recorded. Duration events can be conceptualized as “behavioral states” or the specific length of time during which a behavioral state continues to take place. The onset and offset times for the event are recorded, as well as the type of event under observation. Interval recording is similar to this, but instead of recording specific time duration, codes are assigned for successive time intervals during which behavioral events transpire. Continuous recording can refer to the use of some type of automated event recorder or to human recorders who are continuously alert, paying attention, ready to record whenever a behavioral state changes, or whenever a specific time interval elapses (Bakeman & Gottman, 1986, p. 50).

Is it better to code events, intervals, or durations? That depends on several factors, including the kind and complexity of the coding scheme, the desired accuracy, and the kind of recording equipment available (Bakeman & Gottman, 1986). Sometimes investigators are concerned only with how often certain events occur, or in what order they occur, and are not much concerned with how long they last. At other times, there is a fundamental concern with duration or the mean amount of time a particular kind of event lasts or the proportion of time devoted to a particular kind of event. A common procedure used in observations of group interaction is to record the frequency of a certain behavior (e.g., helping) in a group without tying this information to specific group members.

Information at the group level has limited utility, however, for predicting and understanding the impact of the group experience on the achievement of individual members. For example, a high frequency of helping in a group may not be beneficial for the achievement of all group members if explanations are not directed to those who need it most. Furthermore, even a high correlation between the frequency of helping behavior in a group and achievement sheds no light on the effects of giving help separate from those of receiving help (Suen & Ary, 1989).

41.2.4.3.5. Observational Schedules. Many standard observational schedules that have been developed by researchers in education and other fields are available. Instead of developing a new observational schedule for a research project, it may be possible that one of these existing schedules will fit the project at hand (Bakeman & Gottman, 1986; Borg & Gall, 1989; Suen & Ary, 1989).

The standard observational schedules that have been developed vary in complexity, the type of behavior they record, and the setting in which they can be used. They offer several advantages. First, as is true of standardized personality and aptitude tests, standard observational schedules have usually reached a stage of development where they are valid and reliable measuring instruments. Second, use of a standard schedule saves the considerable amount of time that it takes to develop and validate a new schedule. Third, because most of these standard schedules have been used in a number of research studies, you can compare your findings with those obtained by other researchers using the same instrument.

But the standard schedules may not include all the variables that you are interested in measuring. In this case, it is possible to use only the part of the schedule that is needed, keeping in mind that previously reported reliability and validity data will not apply if only part of the instrument is used.

The book Mirrors for Behavior and Evaluating Classroom Instruction by Simon and Boyer (1974) provides an extensive inventory of older classroom observation forms.
The references are now 20 years old, but many of the instruments described are still in use. The Educational Testing Service (ETS) offers a test collection bibliography entitled Systematic Observation Techniques. This is a valuable resource for locating observation schedules, since it is updated frequently. Many of the observation forms listed in the ETS bibliography are experimental forms developed for use in a specific research project; however, many of the instruments listed have been carefully developed and are comparable to standard observational schedules (Borg & Gall, 1989).

41.2.4.3.6. Selecting and Training Observers. Selecting and training observers is critically important since careless or unmotivated observers can destroy the most carefully planned study. The observers who produce the most reliable data tend to be persons of above-average intelligence and verbal fluency who are highly motivated to do a good job (Borg & Gall, 1989; Hartman, 1982).

Observer training should begin with a thorough understanding of what is to be observed and how it is to be recorded, including the observation categories and their definitions, as well as the form that data gathering will take (paper and pencil, recording device, etc.). The observer trainees should become very familiar with the observation coding form and behavior definitions before moving to the next level of training. After the initial training, observers can practice data recording with videotape recordings of situations similar to those to be observed in the study. Practice observations on site, in which all observer trainees participate, are also highly recommended (Borg & Gall, 1989).

During the practice sessions, watch the trainees to determine if there are still deficiencies in the observation form or the instructions for implementing the observations, in case further fine tuning is warranted. Borg and Gall (1989) report that common problems at this stage include: (1) requiring the observer to record too many behaviors, (2) including behaviors on the form that cannot be reliably identified by the observers, even after extensive training, and (3) poor format that slows down recording or causes observer errors or omissions. Sometimes the observation form will include two behaviors that are quite similar. Thus, when borderline cases occur, the observers have difficulty deciding which behavior to record. This problem can be resolved in several ways. The two behaviors can be combined into one redefined behavior, or, if it is essential to get data separately on the two behaviors, a rule can be established to resolve borderline cases. Such cases can be placed in a "can't decide" category, or can be assigned to the two categories alternately, or can be assigned to one of the two categories randomly by flipping a coin (Borg & Gall, 1989, p. 488).

Different videotape segments should be used for practice observations until the observer trainees reach the desired level of agreement. Borg and Gall (1989, p. 489) recommend that:

- For tallying highly specific descriptive behavior, such as counting the number of times the teacher smiles, the percentage of agreement between observers should be above 90%. When the observer must make inferences or evaluations about the behavior being observed, however, 70 to 90% agreement is usually considered satisfactory.

If observational data gathering is to extend for more than 1 week, observers should be checked frequently, and a weekly refresher training session should be held for the observers to maintain the reliability of the observations (Borg & Gall, 1989; Taplin & Reid, 1973). Otherwise, observers will gradually lose the common frame of reference they developed during training. This "observer drift" can be a major source of error in observational studies. Research has determined that observers sometimes lose their objectivity, the proper frame of reference, in their use of observational schedules. This must be guarded against since the resulting research data will reflect the biases and characteristics of the observer rather than the observational variables that the research seeks to measure (Borg & Gall, 1989).

41.2.4.3.7. Reliability and Validity. In earlier years, observational researchers assumed that direct observation of behavior was by definition bias-free and valid (Cone, 1978). At that time, conventional psychometric issues such as reliability and validity were regarded as irrelevant. However, it is now understood that reliability and validity are as significant for observational recording as for any other data-gathering measure (Suen & Ary, 1989). Reliability refers to the dependability, consistency, predictability, and stability of the data. Validity refers to the ability of the data to reflect the underlying attribute of interest. The fact that the observer is more often a human being rather than a mechanical device highlights the importance of recording reliability (Flagg, 1990).

Suen and Ary (1989) report that in literature concerning psychometrics of behavioral observation, much time and space has been devoted to the issue of reliability, whereas validity has not received much attention. Part of this neglect can be attributed to the widespread belief that observational data are inherently valid because of the minimal amount of inferences required on the part of the observers. This belief is justified for some behaviors but not for others, depending on the degree of complexity (Suen & Ary, 1989).

41.2.4.3.8. Validation of the Coding Scheme. The use of predefined catalogs of behavior is a central feature of structured observational research. Coding systems reduce the task of an observer either to assigning the behavioral occurrences to categories or to assigning a numerical value to each behavioral occurrence (Suen & Ary, 1989). The behavioral category codes are measuring instruments that must be validated to ensure that these categories do indeed represent the phenomenon of interest.

Validity can be improved by fine tuning behavioral categories based on unstructured (and increasingly structured) observation within the study setting and by review of videotaped recordings of sample behavior streams. Validity is further enhanced if this process of identifying and fine tuning behavioral categories is carried out by more than one researcher trained in observational methods.

As the categories are fine tuned and ultimately selected, they should be reviewed and approved by outside experts.
Each category should be defined briefly and precisely so that there is no possible confusion. The ultimate approach to validating an observational catalog is repeated testing. Standard observational schedules have usually reached a stage of development where they have been extensively validated.

In educational communication and technology, as well as in other fields, observational methods may be applied to unique or innovative settings and purposes that offer less opportunity for extensive validation. For example, classroom research focusing on student interactions with computers is less than 15 years old; few catalogs of behavior have as yet been devised specifically for this area (Amarel, 1982; Chen, 1985; Dickson, 1985; Fraser & Burkhardt, 1984; McLellan, 1991; Sheingold, Kane & Endreweit, 1983; Webb, 1982; Webb, Ender & Lewis, 1986).

Another specialized application is the formative evaluation of a specific media program. A catalog of behaviors must be developed and validated specifically for that purpose, where users are observed interacting with the media program or a prototype. In sum, adequate validation can be obtained with careful fine tuning and repeated review of categories by qualified experts. The ideal would be to validate thoroughly a catalog of behaviors through rigorous testing, but this is extremely time consuming, often prohibitively so. Suen and Ary (1989) advise that validity can never be proved, and it cannot be described quantitatively; it is only possible to accumulate evidence in support of the validity of data obtained through a specific observation system.

Traditionally, observation codes were fairly simple and were based on readily observable and identifiable behaviors so that validation was fairly easy. But in recent years, the goals of observational research have often become much more complex, in tandem with sophisticated recording devices that make it possible to gather more observational data. A trend has emerged that threatens effective validation of observation codes. Increasingly, more than one behavior is observed in a study within a sophisticated scheme. In other situations, certain behaviors are observed to reflect an abstract variable or construct. For example, the occurrences and nonoccurrences of eye contact, appropriate speech, and so on may be observed to assess a subject’s social skills. The earlier direct connection between the observed behavior and the object of inquiry no longer exists.

This trend toward greater complexity suggests that research should be directed toward improving observational techniques to meet emerging research needs. Research methods will need to be adapted to meet the growing awareness of the complexity of human behavior and interactions. Already there has been a shift in focus from observing individuals to observing groups. For example, instead of observing only teacher behavior, researchers attempt to observe both teacher and student behavior in coordination (Amarel, 1982; Peterson & Janicki, 1979; Peterson, Wilkinson & Hallinan, 1984; Peterson, Wilkinson, Spinelli & Swing, 1984; Webb, 1982; Webb, 1983; Webb, Ender & Lewis, 1986). Mediated observation techniques can help in this effort.

41.2.4.3.9. Interobserver and Intraobserver Agreement and Reliability. Human observers require a much greater concern for recording reliability than do mechanical recording devices. Human observers may have biasing expectations, and their recording methods may change over time due to fatigue or practice. Observer training is critically important; it may be necessary for ongoing training to maintain consistency over time on the part of individual observers.

Observational research calls for an assessment of both interobserver and intraobserver reliability. Interobserver reliability refers to data gathered by two or more observers, while intraobserver reliability refers to data gathered by the same observer on different occasions. The simplistic definition of reliability is the classical theory, as data consistency is inherently incapable of accommodating the idea that there can be more than one reliability estimate. The classical theory is incapable of accommodating the idea that for a given set of data there can be more than one reliability coefficient, since it assumes that a score contains only a true score component and a random error component and that reliability is the proportion of true variance relative to the sum of true and random error variances. Since two types of reliability are of concern in observational research, the generalizability approach to reliability is utilized in place of classical theory (Suen & Ary, 1989).

The generalizability approach, introduced by Cronbach, Gleser, Nanda, and Rajaratnam (1972), is designed to accommodate the coexistence of more than one reliability estimate, making it possible to come to terms with the multifaceted nature of reliability. The generalizability theory has been praised as a powerful new psychometric tool by many analysts (Berk, 1979; Brennan, 1983; Brennan, 1984; Bakeman & Gottman, 1986; Hartmann, 1982; Mitchell, 1979; Suen & Ary, 1989). One advantage of this approach is that it makes it possible to identify sources of error so that an investigator can take steps to improve data (Hartmann, 1982).

The generalizability theory departs from conventions in two respects: (1) Statistically, it approaches the task of estimating reliabilities through an alternative technique known as the intraclass correlation; and (2) conceptually, it systematically defines for each reliability assessment situation its unique frame of reference. It is unique in its fundamental idea of reliability and error. Within the generalizability framework, reliability is not an absolute concept. Rather, depending on the question asked, it is relative to a certain context with a specific set of dimensions (e.g., time, observer, environment, instrumentation, etc.). When a set of observational data is said to be reliable, it implies that the results can be expected to be consistent over a variety of conditions (Suen & Ary, 1989).

In the classical theory, the context for reliability is undefined. But by contrast, within the generalizability theory, reliability is not an absolute concept without a context. Thus, whether or not a set of scores are reliable depends on what question is being asked.

A third group of techniques has been suggested as a superior approach to reliability (Suen & Ary, 1989). This
set of techniques is known alternatively as observer accuracy, criterion-referenced agreement, and transduction accuracy. At this time, however, there is no clear consensus as to exactly what is being determined by these techniques.

41.2.4.3.10. Reducing Observer Effect. Another problem involved in conducting observational research is the degree to which the presence of the observer changes the situation being observed. In some cases, it may be possible to disguise the observer's role in order to direct less attention to the presence of a researcher. Any changes in the observational situation that will make the observation less obtrusive so that it appears to be more a part of the regular classroom situation will reduce reactivity (Borg & Gall, 1989). Observer effects can include the following:

1. Effect of observer on the observed: Person(s) observed change their behavior because they are aware of the observation.
2. Effect of the observer on the setting: Presence of the observer may lead to anxieties or expectations that change the climate of the observed situation.
3. Observer personal bias: Systematic errors are traceable to characteristics of the observer or the observational situation.
4. Error of leniency: When using a rating scale, observer tends to make most ratings at the favorable end of the scale.
5. Error of central tendency: When using a rating scale, observer tends to make most ratings around the midpoint.
6. Halo effect: Observer's initial impression distorts later evaluations or judgments of the subject.
7. Observer omissions: Because the observation system includes variables that occur very rapidly or simultaneously, the observer overlooks some behavior that should be recorded.
8. Observer drift: The tendency for observers to redefine gradually the observational variables, so that the data collected do not reflect the original categories.
9. Reliability decay: Toward the end of training, observer reliability is high, but in the field, as monitoring and motivation decrease, observers become less reliable.
10. Contamination: The observer's knowledge of one aspect of a study influences his or her perception of events observed in another part of the study. Observer expectations are a common form of contamination (Everton & Green, 1986; Borg & Gall, 1989).

Some analysts recommend that the following precautions should be taken as appropriate to minimize these observer effects and errors (Borg & Gall, 1989; Kazdin, 1977; Rosenthal, 1978; Suen & Ary, 1989):

1. Structure the observational situation so that the observer is as unobtrusive as possible.
2. Explain the common rating errors to the observers, and structure the observation schedule to minimize these errors.
3. Be sure the observation schedule does not require the observer to record more data or record at a higher rate than can be done accurately.
4. Make the observational task as objective as possible. Avoid requiring the observer to make evaluations, interpretations, or high-level inferences.
5. Give the observer as little information as possible about your hypotheses, research design, and expectations.
6. Do not reveal to the observer information about the characteristics of your subjects—such as social class, IQ, or composition of experimental and control groups—that the observer need not know.
7. Train observers to a high level of reliability and objectivity, and retrain as necessary to avoid observer drift.
8. Monitor observers on a random basis to minimize reliability decay.
9. Construct your observation form to minimize recording errors.
10. Check for bias when training observers, and eliminate those who submit biased observations.

41.2.4.4. Naturalistic Contrived Situations. In the naturalistic contrived approach, the researcher intervenes in a natural situation in a manner that cannot be detected by the subject; thus, the naturalness of the situation is preserved (Flagg, 1990). Manipulating the situation helps to ensure that the events of interest will occur. For example, by using pupil confederates, contrived situations can be set up to collect the necessary data in a reasonably short time. Naturalistic observations can require many hours of observation to record a short, 2-minute event. For example, suppose you want to observe teachers' responses to deviant student behavior such as cheating, fighting, or abusing computer equipment. Since these behaviors occur at a very low frequency in most classrooms, a great deal of observer time would be needed to gather data on a reasonable sample of such behavior.

There is another advantage of establishing contrived situations to be observed: The level of intensity of the situation can be manipulated. Thus, behavior can be observed at several specific and clearly defined levels of intensity. These variations in intensity are fairly easy to achieve in contrived situations, but in many cases they are virtually impossible to observe in natural situations (Flagg, 1990).

Borg and Gall (1989, p. 507) report that there are two serious limitations to observing in contrived situations:

First, in many cases believable situations cannot be contrived without arousing the suspicion of the subjects. Second, difficult ethical problems may arise because of the deception involved. When one or both of these problems rule out the use of naturalistic contrived observation, a similar technique, situational testing, can be employed.

41.2.4.5. Situational Testing. Situational testing is another form of contrived observation. In this case, the subjects are aware that they are playing a role (Flagg, 1990). The researcher "devises a situation and assigns appropriate roles to the subjects, who are asked to play the roles to the
best of their ability.” Typically, all participants except the person being tested have been trained to play their roles. The roles create a situation toward which the person being tested must respond. For example, the person might make a decision or try to resolve a conflict. The situations are aimed at teasing out the specific types of behavior that the researcher is interested in observing (Flagg, 1990).

Like other forms of contrived observation, situational testing has advantages over the observation of behavior in natural settings (Flagg, 1990). By setting up the situation, you can control the behavior that is likely to occur to a greater degree than is usually possible in the naturalistic, contrived situation. This permits you to focus the observation on behavior that appears to be critical in the area being studied. In order to observe such critical behavior in a natural situation, the observer may need to be present for weeks or even months (Flagg, 1990).

41.2.4.6. Portfolios. Portfolios provide a descriptive measure of student work based on actual performance. Portfolios consist of learner-created products that reflect the processes of learning and development over time. Many analysts have recommended portfolios as a mechanism for integrating learning and assessment (Belland & Best, 1991; Collins, 1990; Paulson, Paulson & Meyer, 1991; Pearlman, 1990; Schon, 1987; Wolf, 1989). There is currently a trend toward using portfolios of student work as a basis for assessing learner performance and mastery. For example, in Vermont, 135 schools are piloting the use of student portfolios to measure student abilities in math and writing (Pearlman, 1990).

41.2.4.6.1. Types of Portfolios. The portfolios can take many forms. For example, portfolios can record the student’s best compositions, game performances, multimedia projects, artwork, or problem solutions. Portfolios serve as relatively complete process records of a student’s learning, providing a cognitive map of his involvement in projects and his exploration of a particular domain or craft. One advantage of portfolios is that they enlarge the view of what is learned. Portfolios provide a developmental point of view so they can provide a mechanism for assessing progress. Furthermore, portfolios offer students an opportunity to learn about their own learning while they require active student responsibility and self-assessment (Paulson, Paulson & Meyer, 1991).

41.2.4.6.2. Constructing Portfolios. There is a strong tradition of using portfolios in the creative arts, and this is now being extended to other fields of study, including various subjects within education. Portfolios may center around such mediated projects produced by students as print-based materials produced with desktop publishing, videotaped recordings of student presentations, audiotaped recordings of student performances, student-produced videos, as well as computer-based and multimedia projects. The process of assessing portfolios is of central concern, and calls for systematic and refined analysis.

In the context of the Key School, a private school in Indianapolis that is piloting innovative educational prac-
tices based on his model of seven intelligences, Howard Gardner (1993) recommends that student portfolios be assessed in terms of five separate dimensions, including individual profile, quality of work, communication, reflection, and mastery of facts, skills, and concepts. The individual profile includes consideration of the specific cognitive strengths, weaknesses, and propensities of the student. The profile includes the student’s disposition toward work, such as taking risks and persevering, as well as the student’s particular intellectual propensities, such as linguistic, logical, spatial, and interpersonal skills.

Usually a bargain is struck between the student and teacher. The teacher can ask the student to draw on school knowledge and understanding in creating a project, and the student has the opportunity to select from a range of schoolwork those facts, skills, and concepts desirable for the project. Quality of work is assessed in terms of the criteria of a particular genre; this may include innovation and imagination, aesthetic judgment and technique, or the execution of a performance such as a puppet show. In terms of communication, the projects go beyond the classroom situation and offer an opportunity for students to communicate with a wider audience, including with peers, teachers, and other adults. Finally, reflection involves monitoring one’s goals, assessing what progress has been made, evaluating how the course can be corrected, and how to make use of knowledge that has been attained. One example of a system that has been developed for assessing student work in language arts is the 6-point scale for written reports that comes from the state of Vermont, as shown in Figure 41-1.

In the Vermont system, several judges score each student essay based on this 6-point scoring system. This is far more time intensive than scoring a multiple-choice test, but it gives a deeper understanding of student learning and achievement. Gardner (1993) states that assessing the portfolio is no easy matter, but that it should be possible to trace various forms of growth in production, perception, and reflection through a judicious examination of the portfolio.

Noting the need to support educators in the use of portfolios, The Northwest Evaluation Association, based in Lake Oswego, Oregon, publishes a Portfolio Assessment Newsletter that is available on a regular basis. Portfolios represent an area of descriptive techniques that is at an early stage, but it promises to expand in importance over the coming years as more people find ways for meaningful implementation and interpretation.

Good descriptive studies use data collection instruments that are reliable and valid. Reliability refers to the ability to provide consistent answers in comparable situations, and validity refers to the ability to provide answers that correspond to what they are intended to measure. One way of maintaining reliability is to make sure that the same research instrument is used for all participants within a particular study and that the research staff is well trained for using the same procedure for collecting the data (Borg & Gall, 1989; J. Converse, 1987; Fowler, 1993). It is also important to attend to the clarity of materials used within
the variables that will eventually be used to measure cause and effect, and at the least can help provide surrounding information that will aid logical interpretations of research questions within the context of a specific situation.

Descriptive research has gained acceptance as a valid form of research in education, and in recent years the number of descriptive studies published in research journals and conference proceedings has increased. This is probably due to a couple of reasons. First, educational researchers have realized that trying to mimic scientific research does not work for educational settings. The purely scientific approach reduces educational research to trivial questions that do not address the important, overarching issues within education. Instead, researchers need to address the questions at hand and be willing to use a variety of methodologies in order to ensure the most appropriate and accurate investigation. Second, the rapidly changing technologies available to educators have everyone scrambling for information. Educators are likely to be observing students using media in new ways and for new purposes. It is extremely difficult to set up control groups that adequately control all of the variables that might affect the outcome of the research while providing the same type of learning experience for all of the students. The new classroom activities lend themselves nicely to descriptive research.

In addition, educators want to know how others are implementing the new mediated technologies, the national information infrastructure, and so on, and are very happy to hear reports that describe what others are doing, as well as what happens as a result of the process. Analysis of descriptive research patterns leads to prescriptions that instructional designers and educators can heed as they consider future direction. Within the realm of cognitive skill building and the constructivist paradigm, these patterns and prescriptions become part of the whole process of education rather than isolated components. If current patterns continue, we are likely to see more acceptance of properly conducted descriptive research in the future.

REFERENCES


