Data Collection and Analysis

*Tamara van Gog and Fred Paas*  
Open University of the Netherlands, Heerlen, the Netherlands

Wilhelmina Savenye  
Arizona State University-Tempe, Tempe, Arizona

Rhonda Robinson  
Northern Illinois University, DeKalb, Illinois

Mary Niemczyk  
Arizona State University-Polytechnic, Mesa, Arizona

Robert Atkinson  
Arizona State University-Tempe, Tempe, Arizona

Tristan E. Johnson  
Florida State University, Tallahassee, Florida

Debra L. O’Connor  
Intelligent Decision Systems, Inc., Williamsburg, Virginia

Remy M. J. P. Rikers  
Erasmus University Rotterdam, Rotterdam, the Netherlands

Paul Ayres  
University of New South Wales, Sydney, Australia

Aaron R. Duley  
National Aeronautics and Space Administration, Ames Research Center, Moffett Field, California

Paul Ward  
Florida State University, Tallahassee, Florida

Peter A. Hancock  
University of Central Florida, Orlando, Florida

* Tamara van Gog and Fred Paas were lead authors for this chapter and coordinated the various sections comprising this chapter.
ABSTRACT

The focus of this chapter is on methods of data collection and analysis for the assessment of learning processes and complex performance, the last part of the empirical cycle after theory development and experimental design. In the introduction (van Gog and Paas), the general background and the relation between the chapter sections are briefly described. The section by Savene, Robinson, Niemczyk, and Atkinson focuses on methods of data collection and analysis for assessment of individual learning processes, whereas the section by Johnson and O’Connor is concerned with methods for assessment of group learning processes. The chapter section by van Gog, Rikers, and Ayres discusses the assessment of complex performance, and the final chapter section by Duley, Ward, Szalma, and Hancock is concerned with setting up laboratories to measure learning and complex performance.

KEYWORDS

Assessment criteria: Describe the aspects of performance that will be assessed.
Assessment of learning: Measuring learning achievement, performance, outcomes, and processes by many means.
Assessment standards: Describe the quality of performance on each of the criteria that can be expected of participants at different stages (e.g., age, grade) based on a participant’s past performance (self-referenced), peer group performance (norm-referenced), or an objective standard (criterion-referenced).
Collective data collection: Obtaining data from individual group members; data are later aggregated or manipulated into a representation of the group as a whole.

Complex performance: Refers to real-world activities that require the integration of disparate measurement instrumentation as well as the need for time-critical experimental control.
Direct process measure: Continuous elicitation of data from beginning to end of the (group) process; direct process measures involve videotaping, audiotaping, direct researcher observation, or a combination of these methods.
Group: Two or more individuals working together to achieve a common goal.
Group learning process: Actions and interactions performed by group members during the group learning task.
Holistic data collection: Obtaining data from the group as a whole; as this type of data collection results in a representation of the group rather than individual group member, it is not necessary to aggregate or manipulate data.
Indirect process measure: Discrete measure at a specific point in time during the (group) process; often involves multiple points of data collection; indirect process measures may measure processes, outcomes, products, or other factors related to group process.
Instrumentation: Hardware devices used to assist with the process of data acquisition and measurement.
Mixed-methods research: Studies that rely on quantitative and qualitative as well as other methods for formulating research questions, collecting and analyzing data, and interpreting findings.
Online/offline measures: Online measures are recorded during task performance, offline measures are recorded after task performance.
Process-tracing techniques: Records performance process data such as verbal reports, eye movements, and actions that can be used to make inferences about the cognitive processes or knowledge underlying task performance.
Correct operationalizations of these issues, it is generally published in peer-reviewed, high-impact journals. Because of the importance and difficulty of finding high-quality studies in your domain (which are generally published in peer-reviewed, high-impact journals), it is usually represented through numbers and results of statistical tests.

Task complexity: Can be defined subjectively (individual characteristics, such as expertise or perception), objectively (task characteristics, such as multiple solution paths or goals), or as an interaction (individual and task characteristics).

INTRODUCTION

Tamara van Gog and Fred Paas

The most important rule concerning data collection and analysis is do not attempt to collect or analyze all possible kinds of data. Unless you are conducting a truly exploratory study (which is hardly ever necessary nowadays, considering the abundance of literature on most topics), the first part of the empirical cycle—the process of theory development—should result in clear research questions or hypotheses that will allow you to choose an appropriate design to study these. These hypotheses should also indicate the kind of data you will need to collect—that is, the data you have hypotheses about and some necessary control data (e.g., time on task), and together with the design provide some indications as to how to analyze those data (e.g., 2 × 2 factorial design, 2 × 2 MANCOVA). But, these are just indications, and many decisions remain to be made. To name just a few issues regarding data collection (for an elaboration on those questions, see, for example, Christensen, 2006; Sapsford and Jupp, 1996): Which participants (human/nonhuman, age, educational background, gender) and how many to use? What and how many tasks or stimuli to present and on what apparatus? What (control) measures to take? What instructions to give? What procedure to use? When to schedule the sessions?

Making those decisions is not an easy task, and unfortunately strict guidelines cannot be given because acceptable answers are highly dependent on the exact nature, background, goals, and context of the study. To give you some direction, it might help to have a look at how these questions have been dealt with in high-quality studies in your domain (which are generally published in peer-reviewed, high-impact journals). Because of the importance and difficulty of finding correct operationalizations of these issues, it is generally advisable to conduct a pilot study to test your data collection and analysis procedures.

In educational research many studies share the common goal of assessing learning or performance, and the chapter sections in this chapter provide information on methods for collecting and analyzing learning and performance data. Even though learning and performance are conceptually different, many of the data collection and analysis techniques can be used to assess both; therefore, we first discuss the differences between the assessment of learning and the assessment of performance before giving a brief overview of the content of the chapter sections.

Assessment of Learning vs. Performance

The definitions of learning and performance have an important similarity, in that they can be used to refer both to an outcome or product and to a process. The term learning is used to refer to the knowledge or skill acquired through instruction or study (note that this dictionary definition ignores the possibility of informal learning, unless this is encompassed by study), as well as the process of acquiring knowledge or skill through instruction or study. The term performance is used to refer to things accomplished (outcome or product) and to the accomplishment of things (process). Performance implies the use of knowledge rather than merely possessing it. It seems that performance is more closely related to skill than to knowledge acquisition (i.e., learning), but an important difference between the definitions of learning and performance is that performance can be, but is not defined as, a result of instruction or study.

The similarities and differences between these terms have some important implications for educational research. First of all, the fact that both learning and performance can refer to a product and a process enables the use of many different kinds of measures or combinations of measures to assess learning or performance. This can make it quite difficult to compare results of different studies on learning or performance, as they might have assessed different aspects of the same concept and come to very different conclusions.

Second, collection and analysis of data about the knowledge an individual possesses can be used to assess their learning but not their performance. That possessing knowledge does not guarantee the ability to use it has been shown in many studies (see, for example, Ericsson and Lehmann, 1996). Nonetheless, for a long time, educational certification practices were based on this assumption: Students received their diplomas after completing a series of courses successfully, and success was usually measured by the amount of knowledge a student possessed. Given that this measure
has no one-to-one mapping with successful performance, this practice posed many problems, both for students and employers, when students went to work after their educational trajectory. Hence, in the field of education it is recognized now that knowledge is a necessary but not sufficient condition for performance, and the field is gradually making a shift from a knowledge-based testing culture to a performance-based assessment culture (Birenbaum and Dochy, 1996).

Finally, because performance is not defined as a result of instruction or study, it can be assessed in all kinds of situations, and when applied in instructional or study settings it may be assessed before, during, and after instruction or study phases. Note though, that in that case, only the difference between performance assessed before and after instruction or study is indicative for learning. One should be careful not to interpret gains in performance during instruction or study as indicators for learning, as these may be artifacts of instructional methods (Bjork, 1999).

**Brief Overview of the Chapter Sections**

The first chapter section, Assessment of Individual Learning Processes by Savenye, Robinson, Niemczyk, and Atkinson, introduces educational technology researchers to the conceptual basis and methods of data collection and analysis for investigating individual learning processes. They discuss the quantitative and qualitative research paradigms and the associated approaches to data collection and analysis. They also point out the benefits of combining quantitative and qualitative approaches by conducting mixed-methods studies.

The second chapter section, Assessment of Group Learning Processes by Johnson and O'Connor, focuses on the study of group learning processes, which is more complex than the study of individual learning processes. They discuss several issues that need to be considered prior to setting up a study of group learning processes, such as holistic vs. collective data collection, direct vs. indirect methods of data collection, aggregation or manipulation of individual data into group level data, and special considerations for setting up a study of group learning processes.

The third chapter section, Assessment of Complex Performance by van Gog, Rikers, and Ayres, discusses data collection and analysis methods for assessment of complex performance. In line with the two-edged definition of performance as a thing accomplished or accomplishing a thing, they distinguish product and process measures and subdivide the process measures further into online (while working on a task) vs. offline (after task completion) measures. They also discuss the opportunities to combine several different measures and the benefits of doing so.

The fourth and final chapter section, Setting Up a Laboratory for Measurement of Complex Performances by Duley, Ward, Szalma, and Hancock, provides insight into the technical setup of laboratories for the assessment of learning processes and complex performance. Rather than providing a list of available hardware, software, and instruments, they have chosen to take the more sensible approach of familiarizing the reader with setting up configurations for stimulus presentation, control options, and response recording, which are relevant for many laboratory studies.

**ASSESSMENT OF INDIVIDUAL LEARNING PROCESSES**

**Wilhelmina Savenye, Rhonda Robinson, Mary Niemczyk, and Robert Atkinson**

It is the goal of this section to introduce educational technology researchers to the conceptual basis and methods of data collection and analysis for investigating individual learning processes, including both quantitative and qualitative research techniques. Learning processes, of course, may involve both individual and group efforts of learners in the form of strategies and activities designed to facilitate their learning. Though this section focuses on individual processes and performances, using a variety of methods, these may be adapted for group use (see the chapter section by Johnson and O’Connor).

Several assumptions guide this work. Although methods can be suggested here, the researcher must be responsible for understanding the foundational ideas of any study. He or she will want to conduct the study with the utmost attention to quality and therefore will want to turn to specific and detailed texts to learn more deeply how to apply research methods. This section will point the researcher to such references and resources.

The objectives of this section are listed below. It is hoped that after reading this chapter, educational technology researchers will be able to:

- Describe methods and techniques for conducting research on individual learning, and compare qualitative and quantitative methods.
- Describe common problems in conducting and evaluating quantitative and qualitative research methods to examine learning processes.
- Consider issues that contribute to the quality of studies using mixed methods.
Rationale for Using Mixed Methods

The terms quantitative and qualitative are commonly used to describe contrasting research approaches. Typically, quantitative research is considered to be more numbers driven, positivistic, and traditional (Borg and Gall, 1989), while qualitative research is often used interchangeably with terms such as naturalistic, ethnographic (Goetz and LeCompte, 1984), subjective, or post-positivistic. We define qualitative research in this section as research that is devoted to developing an understanding of human systems, be they small, such as a technology-using teacher and his or her students and classroom, or large, such as a cultural system. Quantitative and qualitative methods for data collection derive in some measure from a difference in the way one sees the world, which results in what some consider a paradigm debate; however, in assessing learning processes, both approaches to data collection have importance, and using elements from both approaches can be very helpful. Driscoll (1995) suggested that educational technologists select research paradigms based on what they perceive to be the most critical questions. Robinson (1995) and Reigeluth (1989) concurred, noting the considerable debate within the field regarding suitable research questions and methods. Learning processes are complex and individual. Lowyck and Elen (2004) argued that learning processes are active, constructive, self-regulated, goal oriented, and contextualized. In addition, digital technologies are changing the nature of knowledge and of teaching and learning (Cornu, 2004). It is clear then that the methods required to collect and analyze how learning processes work, when they work, and why they work can be drawn from a mixed-method approach. Thus, researchers can investigate carefully and creatively any questions they choose and derive valid data to help understand learning processes using a combination of methods from both perspectives. Although not the main focus of this chapter, it is assumed that researchers will submit all procedures, protocols, instruments, and participation forms to the appropriate human-subjects or ethics review unit within their organizations. In any case, researchers should be specific about how they define the assumptions of the study and why what was done was done—in short, they should be able to enter into the current and upcoming discussions as thoughtful, critical, and creative researchers.

Analyzing Learning Using Quantitative Methods and Techniques

Learning achievement or performance in educational technology research is often the primary outcome measure or dependent variable of concern to the researcher. Learning is often therefore studied using more quantitative measures, including what researchers may call tests, assessments, examinations, or quizzes. These measures may be administered in paper-and-pencil form or may be technology based. If they are technology based, they may be administered at a testing center, with tutors or proctors, or completed on the student’s own. In either format, they may be scored by an instructor or tutor or may be automatically scored by the computer (Savenye, 2004a,b). Issues of concern in selecting and developing tests and test items also are relevant when learning is measured en route as performance on practice items. Completion time, often in conjunction with testing, is another learning process variable that can efficiently be examined using quantitative methods.

Learning achievement on complex tasks may also be measured more quantitatively using numerically based rubrics and checklists to evaluate products and performances or to evaluate essays or learner-created portfolios. Rubrics and checklists are also often used to derive quantitative data for measuring learning in online discussions or to build frequencies of behaviors from observations of learning processes, often used in conjunction with more qualitative methods (discussed later in this section). Many computer-based course management systems now routinely collect course statistics that may be examined to determine how learners proceed through instruction and what choices they make as they go. Self-evaluations and other aspects of learning, such as learner attitudes, are more commonly measured using questionnaires. Selected types of quantitative methods for examining learning are discussed in turn:

- Tests, examinations, quizzes (administered via paper or technology, including self-evaluations)
- Rubrics or checklists to measure learner performance
- Measuring learning processes in technology-mediated communications
- Technology-based course statistics
- Attitude measures such as questionnaires using Likert-type items

Selecting Tests

Educational researchers frequently select existing tests to assess how individual learning processes are impacted by a novel educational intervention. During this process, the researchers must be conversant with a number of important concepts, including validity and reliability. In the following sections, these concepts are
described in greater detail with a specific focus on what researchers need to know when selecting tests.

**Validity**

Arguably, the most critical aspect of a test is its quality or validity. Simply put, a test is considered valid if it measures what it was created to measure (Borg and Gall, 1989). A test is generally considered valid if the scores it produces help individuals administering the test make accurate inferences about a particular characteristic, trait, or attribute intrinsic to the test taker. As an example, researchers exploring the relative impact of several learning environments would consider a test valid to the extent to which it helps them make an accurate determination of the relative quality and quantity of learning displayed by the students exposed to the respective learning environments.

Validity is not a unitary concept; in fact, test developers use several widely accepted procedures to document the level of validity of their test, including content-related, criterion-related, and construct-related validity. **Content-related validity** represents the extent to which the content of a test is a representative sample of the total subject matter content provided in the learning environment. Another type of validity is **criterion-related validity**, which depicts how closely scores on a given test correspond to or predict performance on a criterion measure that exists outside the test. Unlike content validity, this type of validity yields a numeric value that is the correlation coefficient reported on a scale of −1 (perfect, negative relationship) to +1 (perfect, positive relationship). The third type of validity is **construct-related validity**, which refers to the extent to which the scores on a test correspond with a particular construct or hypothetical concept originating from a theory.

Also worth mentioning is a relatively unsophisticated type of validity known as **face validity**, which is based on the outward appearance of the test. Although this is considered a rather rudimentary approach to establishing validity, it is considered important because of its potential impact on the test taker’s motivation. In particular, respondents may be reluctant to complete a test without any apparent face validity.

**Reliability**

Another important concept involved in test selection is reliability. Simply put, reliability refers to the consistency with which a test yields the same results for a respondent across repeated administrations (Borg and Gall, 1989). Assuming that the focus of the test—a particular attribute or characteristic—remains unchanged between test administrations for a given individual, reliability sheds light on the following question: Does the test always yield the same score for an individual when it is administered on several occasions?

**Determining and Judging Reliability**

The three basic approaches to determining the reliability of a test are test–retest, alternate forms, and internal consistency (Borg and Gall, 1989; Morris et al., 1987). Perhaps the simplest technique for estimating reliability is the **test–retest method**. With this approach, a test developer simply administers the test twice to the same group of respondents and then calculates the correlation between the two sets of scores. As a general rule, researchers select tests displaying the highest reliability coefficient because values approaching +1.00 are indicative of a strong relationship between the two sets of respondents’ scores; that is, the respondents’ relative performance has remained similar across the two testing occasions. Specifically, values above .80 are preferable (Chase, 1999).

Another approach to determining reliability is the **alternate forms method**, in which two equivalent forms of a test are administered to a group of respondents on two separate occasions and the resulting scores correlate. As with the test–retest method, the higher the reliability coefficient, the more confidence a test administer can place in the ability of a test to consistently measure what it was designed to measure.

The final method for estimating the reliability of a test is referred to as **internal consistency**. Unlike the two previous methods, it does not rely on testing the same group of respondents twice to estimate the internal consistency of a test. Instead, the reliability of a test is estimated based on a single test administration, which can be accomplished in two ways—either using the **split halves** method or using one of the Kuder–Richardson methods, which do not require splitting a test in half.

**Limits of Reliability**

A number of caveats are associated with reliability. First, it is important to recognize that high reliability does not guarantee validity; in other words, a test can consistently measure what it was intended to measure while still lacking validity. Knowing that a test is reliable does not permit someone to make judgments about its validity. Reliability is, however, necessary for validity, as it impacts the accuracy with which one can draw inferences about a particular characteristic or attribute intrinsic to the test taker. The reliability is impacted by several factors. Test length is the first. All things being equal, shorter tests tend to be less reliable than longer
tests because the latter afford the test developer more opportunities to accurately measure the trait or characteristic under examination. The reliability of a test is also impacted by the format of its items. A general heuristic to remember is that tests constructed with select-type items tend to be more reliable than tests with supply-type or other subjectively scored items.

Evaluating and Developing Tests and Test Items

The construction of learning assessments is one of the most important responsibilities of instructors and researchers. Tests should be comprised of items that represent important and clearly stated objectives and that adequately sample subject matter from all of the learning objectives. The most effective way to ensure adequate representation of items across content, cognitive processes, and objectives is to develop a test blueprint or table of specifications (Sax, 1980). Multiple types of performance measures allow students an opportunity to demonstrate their particular skills in defined areas and to receive varied feedback on their performances; this is particularly important in self-instructional settings, such as online courses (Savenye, 2004a,b). Multiple learning measures in online settings also offer security advantages (Ko and Rossen, 2001).

Tests should also give students the opportunity to respond to different types of item formats that assess different levels of cognition, such as comprehension, application, analysis, and synthesis (Popham, 1991). Different approaches and formats can yield different diagnostic information to instructors, as well; for example, well-developed multiple-choice items contain alternatives that represent common student misconceptions or errors. Short-answer item responses can give the instructor information about the student’s thinking underlying the answer (Mehrens et al., 1998).

Because the test item is the essential building block of any test, it is critical to determine the validity of the test item before determining the validity of the test itself. Commercial test publishers typically conduct pilot studies (called item tryouts) to get empirical evidence concerning item quality. For these tryouts, several forms of the test are prepared with different subsets of items, so each item appears with every other item. Each form may be given to several hundred examinees. Item analysis data are then calculated, followed by assessment of the performance characteristics of the items, such as item difficulty and item discrimination (i.e., how well the item separates, or discriminates, between those who do well on the test and those who do poorly). The developers discard items that fail to display proper statistical properties (Downing and Haladyna, 1997; Nitko, 2001; Thorndike, 1997).

Scores on Numerically Based Rubrics and Checklists

Assessing performance can be done by utilizing numerically based rubrics and checklists. Typically, two aspects of a learner’s performance can be assessed: the product the learner produces and the process a learner uses to complete the product. Either or both of these elements may be evaluated. Because performance tasks are usually complex, each task provides an opportunity to assess students on several learning goals (Nitko, 2001). Performance criteria are the specific behaviors a student should perform to properly carry out a performance or produce a product. The key to identifying performance criteria is to break down the overall performance or product into its component parts. It is important that performance criteria be specific, observable, and clearly stated (Airasian, 1996).

Scoring rubrics are brief, written descriptions of different levels of performance. They can be used to summarize both performances and products. Scoring rubrics summarize performance in a general way, whereas checklists and rating scales can provide specific diagnostic information about student strengths and weaknesses (Airasian, 1996). Checklists usually contain lists of behaviors, traits, or characteristics that are either present or absent, to be checked off by an observer (Sax, 1980). Although they are similar to checklists, rating scales allow the observer to judge performance along a continuum rather than as a dichotomy (Airasian, 1996).

Measuring Learning Processes in Technology-Mediated Communications

Tiffin and Rajasingh (1995) suggested that education is based on communication. Online technologies, therefore, provide tremendous opportunities for learning and allow us to measure learning in new ways; for example, interactions in online discussions within Internet-based courses may be used to assess students’ learning processes. Paulsen (2003) delineated many types of learning activities, including online interviews, online interest groups, role plays, brainstorming, and project groups. These activities, generally involving digital records, will also yield online communication data for research purposes, provided the appropriate ethics and subject guidelines have been followed.

The postings learners create and share may also be evaluated using the types of rubrics and checklists discussed earlier. These are of particular value to learners when they receive the assessment tools early in the course and use them to self-evaluate or to conduct peer evaluations to improve the quality of their work.
(Savenye, 2006, 2007). Goodyear (2000) reminded us that digital technologies add to the research and development enterprise the capability for multimedia communications.

Another aspect of online discussions of value to researchers is that the types of postings students make and the ideas they discuss can be quantified to illuminate students’ learning processes. Chen (2005), in an online course activity conducted with groups of six students who did not know each other, found that learners under a less-structured forum condition posted many more socially oriented postings, although their performance on the task was not less than that of the students who did not post as many social postings. She also found that the more interactions a group made, the more positive students’ attitudes were toward the course.

**Using Technology-Based Course Statistics To Examine Learning Processes**

In addition to recording learners’ performance on quizzes, tests, and other assignments, most online course management systems automatically collect numerous types of data, which may be used to investigate learning processes. Such data may include information about exactly which components of the course a learner has completed, on which days, and for how much time. Compilations of these data can indicate patterns of use of course components and features (Savenye, 2004a).

**Measuring Attitudes Using Questionnaires That Use Likert-Type Items**

Several techniques have been used to assess attitudes and feelings of learners in research studies and as part of instruction. Of these methods, Likert-type scales are the most common. Typically, respondents are asked to indicate their strength of feeling toward a series of statements, often in terms of the degree to which they agree or disagree with the position being described. Previous research has found that responding to a Likert-type item is an easier task and provides more information than ranking and paired comparisons. The advantage of a Likert-type item scale is that an absolute level of an individual’s responses can be obtained to determine the strength of the attitude (O’Neal and Chissom, 1993).

Thorndike (1997) suggested several factors to consider in developing a Likert-type scale, including the number of steps, odd or even number of steps, and types of anchors. The number of steps in the scale is important as it relates to reliability—the more steps, the greater the reliability. The increase is noticeable up to about seven steps; after this, the reliability begins to diminish, as it becomes difficult to develop meaningful anchors. Five-point scales tend to be the most common. Increasing the number of items can also increase reliability. Although there is considerable debate about this, many researchers hold that better results can be obtained by using an odd number of steps, which provides for a neutral response. The anchors used should fit the meaning of the statements and the goal of the measurement. Common examples include continua such as agree–disagree, effective–ineffective, important–unimportant, and like me–not like me.

**Analyzing Learning Using More Qualitative Methods and Techniques**

Although learning outcomes and processes can be productively examined using the quantitative methods discussed earlier, in a mixed-methods approach many qualitative methods are used to build a deeper understanding of what, why, and how learners learn. With the increasing use of interactive and distance technologies in education and industry, opportunities and at times the responsibility to explore new questions about the processes of learning and instruction have evolved. New technologies also enable researchers to study learners and learning processes in new ways and to expand our views of what we should investigate and how; for example, a qualitative view of how instructors and their students learn through a new technology may yield a view of what is really happening when the technology is used.

As in any research project, the actual research questions guide the selection of appropriate methods of data collection. Once a question or issue has been selected, the choice of qualitative methods falls roughly into the categories of observations, interviews, and document and artifact analysis, although others have conceptualized the methods somewhat differently (Bogdan and Biklen, 1992; Goetz and LeCompte, 1984; Lincoln and Guba, 1985). Qualitative researchers have basically agreed that the human investigator is the primary research instrument (Pelto and Pelto, 1978).

In this section, we begin with one approach to conducting qualitative research: grounded theory. We then discuss specific methods that may be called observations, interviews, and document and artifact analysis. As in all qualitative research, it is also assumed that educational technology researchers will use and refine methods with the view that these methods vary in their degree of interactiveness with participants. The
following qualitative methods, along with several research perspectives, are examined next:

- Grounded theory
- Participant observations
- Nonparticipant observations
- Interviews, including group and individual
- Document, artifact, and online communications and activities analysis

**Grounded Theory**

In their overview of grounded theory, Strauss and Corbin (1994, p. 273) noted that it is “a general methodology for developing theory that is grounded in data systematically gathered and analyzed,” adding that it is sometimes referred to as the constant comparative method and that it is applicable as well to quantitative research. In grounded theory, the data may come from observations, interviews, and video or document analysis, and, as in other qualitative research, these data may be considered strictly qualitative or may be quantitative. The purpose of the methodology is to develop theory, through an iterative process of data analysis and theoretical analysis, with verification of hypotheses ongoing throughout the study. The researcher begins a study without completely preconceived notions about what the research questions should be and collects and analyzes extensive data with an open mind. As the study progresses, he or she continually examines the data for patterns, and the patterns lead the researcher to build the theory. The researcher continues collecting and examining data until the patterns continue to repeat and few new patterns emerge. The researcher builds the theory from the data, and the theory is thus built on, or grounded in, the phenomena.

**Participant Observation**

In participant observation, the observer becomes part of the environment, or the cultural context. The hallmark of participant observation is continual interaction between the researcher and the participants; for example, the study may involve periodic interviews interspersed with observations so the researcher can question the participants and verify perceptions and patterns. Results of these interviews may then determine what will initially be recorded during observations. Later, after patterns begin to appear in the observational data, the researcher may conduct interviews asking the participants about these patterns and why they think they are occurring.

As the researcher cannot observe and record everything, in most educational research studies the investigator determines ahead of time what will be observed and recorded, guided but not limited by the research questions. Participant observation is often successfully used to describe what is happening in a context and why it happens. These are questions that cannot be answered in the standard experiment.

Many researchers have utilized participant observation methods to examine learning processes. Robinson (1994) observed classes using Channel One in a Midwestern middle school; she focused her observations on the use of the televised news show and the reaction to it from students, teachers, administrators, and parents. Reilly (1994) analyzed video recordings of both the researcher and students in a project that involved defining a new type of literacy that combined print, video, and computer technologies. Higgins and Rice (1991) investigated teachers’ perceptions of testing. They used triangulation and a variety of methods to collect data; however, a key feature of the study was participant observation. Researchers observed 6 teachers for a sample of 10 hours each and recorded instances of classroom behaviors that could be classified as assessment. Similarly, Moallem (1994) used multiple methods to build an experienced teacher’s model of teaching and thinking by conducting a series of observations and interviews over a 7-month period.

**Nonparticipant Observation**

Nonparticipant observation is one of several methods for collecting data considered to be relatively unobtrusive. Many recent authors cite the early work of Webb et al. (1966) as laying the groundwork for use of all types of unobtrusive measures. Several types of nonparticipant observation have been identified by Goetz and LeCompte (1984). These include stream-of-behavior chronicles recorded in written narratives or using video or audio recordings, proxemics and kinesics (i.e., the study of uses of social space and movement), and interaction analysis protocols, typically in the form of observations of particular types of behaviors that are categorized and coded for analysis of patterns. In nonparticipant observation, observers do not interact to a great degree with those they are observing. The researchers primarily observe and record, using observational forms developed for the study or in the form of extensive field notes; they have no specific roles as participants.

Examples of studies in which observations were conducted that could be considered relatively nonparticipant observation include Savenye and Strand (1989) in the initial pilot test and Savenye (1989) in the subsequent larger field test of a multimedia-based science curriculum. Of most concern during implementation
was how teachers used the curriculum. A careful sample of classroom lessons was recorded using video, and the data were coded; for example, teacher questions were coded, and the results indicated that teachers typically used the system pauses to ask recall-level rather than higher-level questions. Analysis of the coded behaviors for what teachers added indicated that most of the teachers in the sample added examples to the lessons that would provide relevance for their own learners. Of particular value to the developers was the finding that teachers had a great degree of freedom in using the curriculum and the students’ learning achievement was still high.

In a mixed-methods study, nonparticipant observations may be used along with more quantitative methods to answer focused research questions about what learners do while learning. In a mixed-methods study investigating the effects and use of multimedia learning materials, the researchers collected learning outcome data using periodic tests. They also observed learners as they worked together. These observations were video recorded and the records analyzed to examine many learning processes, including students’ level of cognitive processing, exploratory talk, and collaborative processing (Olkinuora et al., 2004). Researchers may also be interested in using observations to study what types of choices learners make while they proceed through a lesson. Klein and colleagues, for instance, developed an observational instrument used to examine cooperative learning behaviors in technology-based lessons (Crooks et al., 1995; Jones et al., 1995; Klein and Pridemore, 1994).

A variation on nonparticipant observations represents a blend with trace-behavior, artifact, or document analysis. This technique, known as read-think-aloud protocols, asks learners to describe what they do and why they do it (i.e., their thoughts about their processes) as they proceed through an activity, such as a lesson. Smith and Wedman (1988) used this technique to analyze learner tracking and choices. Techniques for coding are described by Spradley (1980); however, protocol analysis (Ericsson and Simon, 1984) techniques could be used on the resulting verbal data.

**Issues Related to Conducting Observations**

Savenye and Robinson (2004, 2005) have suggested several issues that are critical to using observations to studying learning. These issues include those related to scope, biases and the observer’s role, sampling, and the use of multiple observers. They caution that a researcher can become lost in the multitudes of observational data that can be collected, both in person and when using audio or video. They recommend limiting the scope of the study specifically to answering the questions at hand. Observers must be careful not to influence the results of the study; that is, they must not make things happen that they want to happen. Potential bias may be handled by simply describing the researcher’s role in the research report, but investigators will want to examine periodically what their role is and what type of influences may result from it. In observational research, sampling becomes not random but purposive (Borg and Gall, 1989). For the study to be valid, the reader should be able to believe that a representative sample of involved individuals was observed. The multiple realities of any cultural context should be represented. If several observers will be used to collect the data, and their data will be compared or aggregated, problems with reliability of data may occur. Observers tend to see and subsequently interpret the same phenomena in many different ways. It becomes necessary to train the observers and to ensure that observers are recording the same phenomena in the same ways. When multiple observers are used and behaviors counted or categorized and tallied, it is desirable to calculate and report inter-rater reliability.

**Interviews**

In contrast with the relatively non-interactive, nonparticipant observation methods described earlier, interviews represent a classic qualitative research method that is directly interactive. Interviews may be structured or unstructured and may be conducted in groups or individually. In an information and communication technologies (ICT) study to investigate how ICT can be introduced into the context of a traditional school, Demetriadis et al. (2005) conducted a series of semi-structured interviews over 2 years with 15 teachers/mentors who offered technology training to other teachers.

The cornerstone for conducting good interviews is to be sure one truly listens to respondents and records what they say rather than the researcher’s perceptions or interpretations. This is a good rule of thumb in qualitative research in general. It is best to maintain the integrity of the raw data and to make liberal use of the respondents’ own words, including quotes. Most researchers, as a study progresses, also maintain field notes that contain interpretations of patterns to be refined and investigated on an ongoing basis.

Many old, adapted, and exciting techniques for structured interviewing are evolving. One example of such a use of interviews is in the Higgins and Rice (1991) study mentioned earlier. In this study, teachers sorted the types of assessment they had named previously in interviews into sets of assessments that were
most alike; subsequently, multidimensional scaling was used to analyze these data, yielding a picture of how these teachers’ viewed testing. Another type of structured interview, mentioned by Goetz and LeCompte (1984), is the interview using projective techniques. Photographs, drawings, and other visuals or objects may be used to elicit individuals’ opinions or feelings.

Instructional planning and design processes have long been of interest to educational technology researchers; for example, using a case-study approach, Reiser and Mory (1991) employed interviews to examine two teachers’ instructional design and planning techniques. One of the models proposed for the design of complex learning is that of van Merriënboer et al. (1992), who developed the four-component model, which subsequently was further developed as the 4C/ID model (van Merriënboer, 1997). Such design models have been effectively studied using mixed methods, including interviews, particularly when those processes relate to complex learning. How expert designers go about complex design tasks has been investigated using both interviews and examination of the designers products (Kirschner et al., 2002).

Problem-based instructional design, blending many aspects of curriculum, instruction, and media options (Dijkstra, 2004), could also be productively studied using interviews. Interviews to examine learning processes may be conducted individually or in groups. A specialized group interview method is the focus group (Morgan, 1996), which is typically conducted with relatively similar participants using a structured or semi-structured protocol to examine overall patterns in learning behaviors, attitudes, or interests.

Suggestions for heightening the quality of interviews include employing careful listening and recording techniques; taking care to ask probing questions when needed; keeping the data in their original form, even after they have been analyzed; being respectful of participants; and debriefing participants after the interviews (Savenye and Robinson, 2005).

Methods for Analyzing Qualitative Data

One of the major hallmarks of conducting qualitative research is that data are analyzed continually, throughout the study, from conceptualization through the entire data collection phase and into the interpretation and writing phases. In fact, Goetz and LeCompte (1984) described the processes of analyzing and writing together in what they called analysis and interpretation.

Data Reduction

Goetz and LeCompte (1994) described the conceptual basis for reducing and condensing data in an ongoing style as the study progresses. Researchers theorize as the study begins and build and continually test theories based on observed patterns in data. Goetz and LeCompte described the analytic procedures researchers use to determine what the data mean. These procedures involve looking for patterns, links, and relationships. In contrast to experimental research, the
qualitative researcher engages in speculation while looking for meaning in data; this speculation will lead the researcher to make new observations, conduct new interviews, and look more deeply for new patterns in this recursive process. It is advisable to collect data in its raw, detailed form and then record patterns. This enables the researcher later to analyze the original data in different ways, perhaps to answer deeper questions than originally conceived. It should be noted that virtually all researchers who use an ethnographic approach advocate writing up field notes immediately after leaving the research site each day. If researchers have collected documents from participants, such as logs, journals, diaries, memos, and letters, these can also be analyzed as raw data. Similarly, official documents of an organization can be subjected to analysis. Collecting data in the form of photographs, films, and videos, either produced by participants or the researcher, has a long tradition in anthropology and education. These data, too, can be analyzed for meaning. (Bellman and Jules-Rosette, 1977; Bogaart and Ketelaar, 1983; Bogdan and Biklen, 1992; Collier and Collier, 1986; Heider, 1976; Hockings, 1975).

Coding Data

Early in the study, the researcher will begin to scan recorded data and to develop categories of phenomena. These categories are usually called codes. They enable the researcher to manage data by labeling, storing, and retrieving it according to the codes. Miles and Huberman (1994) suggested that data can be coded descriptively or interpretively. Bogdan and Biklen (1992) recommended reading data over at least several times to begin to develop a coding scheme. In one of the many examples he provided, Spradley (1979) described in extensive detail how to code and analyze interview data, which are semantic data as are most qualitative data. He also described how to construct domain, structural, taxonomic, and componential analyses.

Data Management

Analysis of data requires continually examining, sorting, and reexamining data. Qualitative researchers use many means to organize, retrieve, and analyze their data. To code data, many researchers simply use notebooks and boxes of paper, which can then be resorted and analyzed on an ongoing basis. Computers have long been used for managing and analyzing qualitative data. Several resources exist to aid the researcher in finding and using software for data analysis and management, including books (Weitzman and Miles, 1995) and websites that discuss and evaluate research software (American Evaluation Association, 2007; Cuneo, 2000; Horber, 2006).

Writing the Research Report

In some respects, writing a report of a study that uses mixed-methods may not differ greatly from writing a report summarizing a more traditional experimental study; for example, a standard format for preparing a research report includes an introduction, literature review, description of methods, and presentation of findings, completed by a summary and discussion (Borg and Gall, 1989). A mixed-methods study, however, allows the researcher the opportunity to create sections of the report that may expand on the traditional. The quantitative findings may be reported in the manner of an experimental study (Ross and Morrison, 2004). The qualitative components of research reports typically will be woven around a theme or central message and will include an introduction, core material, and conclusion (Bogdan and Biklen, 1992). Qualitative findings may take the form of a series of themes from interview data or the form of a case study, as in the Reiser and Mory (1991) study. For a case study, the report may include considerable quantification and tables of enumerated data, or it may take a strictly narrative form. Recent studies have been reported in more nontraditional forms, such as stories, plays, and poems that show participants’ views. Suggestions for writing up qualitative research are many (Meloy, 1994; Van Maanen, 1988; Wolcott, 1990).

In addition to the studies mentioned earlier, many excellent examples of mixed-methods studies may be examined to see the various ways in which the results of these studies have been written. Seel and colleagues (2000), in an investigation of mental models and model-centered learning environments, used quantitative learning measures that included pretests, posttests, and a measure of the stability of learning four months after the instruction. They also used a receptive interview technique they called causal explanations to investigate learners’ mental models and learning processes. In this and subsequent studies, Seel (2004) also investigated learners’ mental models of dynamic systems using causal diagrams that learners developed and teach-back procedures, in which a student explains a model to another student and this epistemic discourse is then examined.

Conclusion

The challenges to educational technology researchers who choose to use multiple methods to answer their questions are many, but the outcome of choosing mixed methods has great potential. Issues of validity, reliability, and generalizability are central to experimental research (Ross and Morrison, 2004) and mixed-
methods research; however, these concerns are addressed quite differently when using qualitative methods and techniques. Suggestions and criteria for evaluating the quality of mixed-methods studies and research activities may be adapted from those suggested by Savenye and Robinson (2005):

- Learn as much as possible about the context of the study, and build in enough time to conduct the study well.
- Learn more about the methods to be used, and train yourself in these methods.
- Conduct pilot studies whenever possible.
- Use triangulation (simply put, use multiple data sources to yield deeper, more true views of the findings).
- Be ethical in all ways when conducting research.
- Listen carefully to participants, and carefully record what they say and do.
- Keep good records, including audit trails.
- Analyze data continually throughout the study, and consider having other researchers and participants review your themes, patterns, and findings to verify them.
- Describe well all methods, decisions, assumptions, and biases.
- Using the appropriate methods (and balance of methods when using mixed methods) is the key to successful educational research.

**ASSESSMENT OF GROUP LEARNING PROCESSES**

*Tristan E. Johnson and Debra L. O'Connor*

Similar to organizations that rely on groups of workers to address a variety of difficult and challenging tasks (Salas and Fiore, 2004), groups are formed in various learning settings to meet instructional needs as well as to exploit the pedagogical, learning, and pragmatic benefits associated with group learning (Stahl, 2006). In educational settings, small groups have been typically used to promote participation and enhance learning. One of the main reasons for creating learning groups is to facilitate the development of professional skills that are promoted from group learning, such as communication, teamwork, decision making, leadership, valuing others, problem solving, negotiation, thinking creatively, and working as a member of a group (Carnevale et al., 1989).

Group learning processes are the interactions of two or more individuals with themselves and their environment with the intent to change knowledge, skill, or attitude. We use the term *group* to refer to the notion of small groups and not large groups characterized as large organizations (Levine and Moreland, 1990; Woods et al., 2000). Interest in group learning processes can be found not only in traditional educational settings such as elementary and secondary schools but also in workplace settings, including the military, industry, business, and even sports (Guzzo and Shea, 1992; Levine and Moreland, 1990).

There are several reasons to assess group learning processes. These include the need to measure group learning as a process outcome and to capture the learning process to provide feedback to the group with the intent to improve team interactions and thereby improve overall team performance. Studies looking at group processes have led to improved understanding about what groups do and how and why they do what they do (Salas and Cannon-Bowers, 2000). Another reason to assess group learning processes is to capture highly successful group process behaviors to develop an interaction framework that could then be used to inform the design and development of group instructional strategies. Further, because the roles and use of groups in supporting and facilitating learning have increased, the interest in studying the underlying group mechanisms has increased.

Many different types of data collection and analysis methods can be used to assess group learning processes. The purpose of this section is to describe these methods by: (1) clarifying how these methods are similar to and different from single learner methods, (2) describing a framework of data collection and analysis techniques, and (3) presenting analysis considerations specific to studying group learning processes along with several examples of existing methodologies.

**Group Learning Processes Compared with Individual Learning Processes and Group Performance**

Traditional research on learning processes has focused on psychological perspectives using traditional psychological methods. The unit of analysis for these methods emphasizes the behavior or mental activity of an individual concentrating on learning, instructional outcomes, meaning making, or cognition, all at an individual level (Koschmann, 1996; Stahl, 2006). In contrast, the framework for group research focuses on research traditions of multiple disciplines, such as communication, information, sociology, linguistics, military, human factors, and medicine, as well as fields of applied psychology such as instructional, educational, social, industrial, and organization psychology.
As a whole, these disciplines extend the traditional psychological perspectives and seek understanding related to interaction, spoken language, written language, culture, and other aspects related to social situations. Stahl (2006) pointed out that individuals often think and learn apart from others, but learning and thinking in isolation are still conditioned and mediated by important social considerations.

Group research considers various social dimensions but primarily focuses on either group performance or group learning. Group learning research is focused in typical learning settings. We often see children and youth engaged in group learning in a school setting. Adult group learning is found in post-secondary education, professional schools, vocational schools, colleges and universities, and training sessions, as well as in on-the-job training environments. A number of learning methods have been used in all of these settings. A few specific strategies that use groups to facilitate learning include cooperative learning, collaborative learning (Johnson et al., 2000; Salomon and Perkins, 1998), computer-supported collaborative learning (Stahl, 2006), and team-based learning (Michaelsen, 2004). General terms used to refer to the use of multiple person learning activities include learning groups, team learning, and group learning. Often, these terms and specific strategies are used interchangeably and sometimes not in the ways just described.

In addition to learning groups, adults engage in groups activities in performance (workplace) settings. Although a distinction is made between learning and performance, the processes are similar for groups whose primary intent is to learn and for those focused on performing. The literature on workplace groups (whose primary focus is on completing a task) offers a number of techniques that can be used to study group learning processes much like the literature on individual learning. Group learning process methods include the various methods typically found when studying individuals and also have additional methodologies unique to studying groups.

**Methodological Framework:**

**Direct and Indirect Process Measures**

When studying group learning processes, three general categories of measures can be employed: (1) the process or action of a group (direct process measures), (2) a state or a point in time of a group (indirect process measure), and (3) an outcome or performance of a group (indirect non-process measure). Direct process measures are techniques that directly capture the process of a group. These measures are continuous in nature and capture data across time by recording the sound and sight of the group interactions. Examples of these measures include recording the spoken language, written language, and visible interactions. These recording can be video or audio recordings, as well as observation notes.

Indirect process measures use techniques that directly capture group processes. These measures are discrete and capture a state or condition of the group processes at a particular point in time, either during or after group activity. These measures involve capturing group member or observer perceptions and reactions that focus on group processes. Examples of these measures involve interviews, surveys, and questionnaires that focus on explicating the nature of a group learning process at a given point in time. These measures focus on collecting group member responses about the process and are specifically not a direct observation of the process.

Indirect non-process measures capture group learning data relating to outcomes, products, performance. These are not measures of the actual process but are measures that might be related to group processes. They may include group characteristics such as demographics, beliefs, efficacy, preferences, size, background, experience, diversity, and trust (Mathieu et al., 2000). These types of measures have the potential to explicate and support the nature of group learning processes. These measures are focused on collecting products or performance scores as well as soliciting group member responses about group characteristics. These measures are not a direct observation of the group learning process. Examples of these measures include performance scores, product evaluations, surveys, questionnaires, interview transcripts, and group member knowledge structures. These measures focus on explicating the nature of a given group’s non-process characteristics.

When considering how to assess group learning processes, many of the techniques are very similar or the same as those used to study individuals. Group learning process measures can be collected at both the individual and group levels (O’Neil et al., 2000; Webb, 1982). Because the techniques can be similar or identical for individuals and groups, some confusion may arise when it is realized that individual-level data are not in a form that can be analyzed; the data must be group-level data (group dataset; see Figure 55.1) for analysis.

When designing a study on group learning processes, various measurement techniques can be used depending on the type of questions being asked. Although numerous possibilities are associated with the assessment of group learning processes, three
elements must be considered when deciding on what techniques to use: data collection, data manipulation, or data analysis.

Data collection techniques involve capturing or eliciting data related to group learning processes at an individual or group level. Data collected at the group level (capturing group interactions or eliciting group data) yield holistic group datasets (Figure 55.1). When the collected data are in this format, it is not necessary to manipulate the data. In this form, the data are ready to be analyzed. If, however, data are collected at the individual level, then the data must be manipulated, typically via aggregation (Stahl, 2006), to create a dataset that represents the group (collective group dataset) prior to data analysis (Figure 55.1). Collecting data at the individual level involves collecting individual group members’ data and then transforming the individual data to an appropriate form (collective group dataset) for analysis (see Figure 55.1). This technique of creating a collective group dataset from individual data is similar to a process referred to as analysis constructed dataset creation (O’Connor and Johnson, 2004). In this form, the data are ready to be analyzed.

**Data Collection and Analysis Techniques**

When considering the different group learning process assessment techniques, they can be classified based on the type of measure (continuous or discrete). The corresponding analytical techniques that can be used are dependent on the collected data. Many techniques have been used to assess groups. The following section presents the major categories of techniques based on their ability to measure group processes directly (continuous measures) or indirectly (discrete measures). Table 55.1 summarizes the nature of data collection, manipulation, and analysis for the three major grouping of measurement techniques: direct process measures and the two variations of indirect process measures.

**Direct Process Data Collection and Analysis**

Direct process measurement techniques focus specifically on capturing the continuous process interactions in groups (O’Neil et al., 2000). These techniques include measures of auditory and visual interactions. Several data collection and data analysis techniques
are related to measuring the group learning processes directly. The two key techniques for capturing actions and language are (1) technology and (2) observation. Using technology to capture group processes can provide researchers with data different from the observation data. Researchers can combine the use of technology and observation simultaneously to capture group processes (O’Neil et al., 2000; Paterson et al., 2003). These data can be analyzed in the captured form or transcribed into a text form.

**Use of Technology to Capture Group Process**

*Spoken Language Processes*

Techniques to capture a group’s spoken language involve either audio recording or video recording (Schweiger, 1986; Willis, 2002) the spoken language that occurs during group interactions (Pavitt, 1998). It can also include the spoken language of group members as they explain their thinking during group processes in the form of a think-aloud protocol (Ericsson and Simon, 1993).

*Written Language Processes*

Group learning processes are typically thought of as using spoken language, but new communication tools are available that allow groups to communicate and interact using written language. Examples include chat boards, whiteboards (although these are not limited to written language), and discussion boards. Also, computer-supported collaborative learning (CSCL) is a computer-based network system that supports group learning interactions (Stahl, 2006).

**Visible Processes**

Techniques to capture a group’s visible interactions include video recording of the behaviors and actions that occur in group interactions (Losada, 1990; Prichard, 2006; Schweiger, 1986; Sy, 2005; Willis et al., 2002).

**Use of Observations to Capture Group Process**

Although the use of technology may capture data with a high level of realism, some group events can be better captured by humans because of their ability to observe more than what can be captured by technology. Observations ideally are carried out with a set of carefully developed observation protocols to help focus the observers and to teach them how to describe key process events. Observers are a good source for capturing various types of information (Patton, 2001), such as settings, human and social environments, group activities, style and types of language used, nonverbal communication, and events that are not ordinary. Observational data, for example, are important for studying group learning process (Battistich et al., 1993; Lindgärde, 2002; Sy, 2005; Webb, 1982; Willis et al., 2002). The type of information typically captured includes location, organization, activities, and behaviors (Battistich et al., 1993; Losada, 1990), as well as the frequency and quality of interactions (Battistich, 1993).

**Direct Process Data Analysis**

Data that are a direct measure of group processes are captured in a holistic format that is ready to be analyzed (Figure 55.1 and Table 55.1). Several analytical

---

TABLE 55.1

**Summary of Measurement Techniques Used to Assess Group Learning Processes**

<table>
<thead>
<tr>
<th>Direct Process Measure Techniques—Holistic Group Dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data collection</td>
<td>Directly capturing group learning processes involves techniques that are used by all group members at the same time.</td>
</tr>
<tr>
<td>Data manipulation</td>
<td>Data manipulation not needed because data is captured at group level (holistic group dataset).</td>
</tr>
<tr>
<td>Data analysis</td>
<td>Continuous process techniques focus on interactions of group members generating qualitative and quantitative findings associated with continuous measures.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indirect Process Measure Techniques—Holistic Group Dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data collection</td>
<td>Indirectly eliciting group learning processes involves techniques that are used by all group members at the same time.</td>
</tr>
<tr>
<td>Data manipulation</td>
<td>Data manipulation not needed because data is captured at group level (holistic group dataset).</td>
</tr>
<tr>
<td>Data analysis</td>
<td>Discrete process techniques are dependent on dataset characteristics (focus on process or performance). They can include qualitative and quantitative data analysis techniques associated with discrete measures.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indirect Process Measure Techniques—Collective Group Dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Data collection</td>
<td>Indirectly eliciting group learning processes involves techniques that are used by each group member separately.</td>
</tr>
<tr>
<td>Data manipulation</td>
<td>Individual data is then aggregated to create a dataset that represents the group data (analysis constructed).</td>
</tr>
<tr>
<td>Data analysis</td>
<td>Discrete process techniques are dependent on dataset characteristics (focus on process or performance). They can include qualitative and quantitative data analysis techniques associated with discrete measures.</td>
</tr>
</tbody>
</table>
techniques are available that can be used for analyzing group data, particularly direct process data. The following list is a representative sample of the analysis techniques applied to spoken or written language, visible interaction, and observational data: sequential analysis of group interactions (Bowers, 2006; Jeong, 2003; Rourke et al., 2001), analysis of interaction communication (Bales, 1950; Qureshi, 1995), communication analysis (Bowers et al., 1998), anticipation ratio (Eccles and Tenenbaum, 2004), in-process coordination (Eccles and Tenenbaum, 2004), discourse analysis (Aviv, 2003; Hara et al., 2000), content analysis (Aviv, 2003; Hara et al., 2000), cohesion analysis (Aviv, 2003), and protocol analysis (Ericsson and Simon, 1980, 1993). Visible interactions techniques also include using a behavior time series analysis (Losada, 1990). This analysis involves looking at dominate vs. submissive, friendly vs. unfriendly, or task-oriented vs. emotionally expressive behavior. For observational data, researchers focus on various qualitative techniques associated with naturalistic observations (Adler and Adler, 1994; Patton, 2001). Some common tasks associated with this type of analysis involve group and character sequence analysis and assertion evaluation (Garfinkel, 1967; Jorgensen, 1989).

**Indirect Process Data Collection and Analysis**

Many data collection techniques are related to measuring the group learning processes indirectly. Indirect group process, characteristic, and product measurement techniques elicit group information at a specific point in time. These discrete measures do not capture group processes directly but elicit data that describe group processes or process-related data such as group characteristics or group outcomes (things that may have a relation to the group processes). The three key types of data related to group learning processes are indirect group process data, group characteristic data, and group product data, within which specific factors can be measured. Indirect group process data describe group processes and can include factors such as group communication (verbal/nonverbal), group actions, group behaviors, group performance, and group processes. Group characteristic data, relating to group processes, include factors such as group knowledge, group skills, group efficacy, group attitudes, group member roles, group environment, and group leadership. The key elicitation techniques for both of these types of indirect data include interviews, questionnaires, and conceptual methods (Cooke et al., 2000). Each technique can be focused on group process or group characteristics. After reviewing methods to analyze group processes, we discuss methods for analyzing group products.

**Interviews**

Interviews are a good technique for collecting general data about a group. The various types of interviewing techniques include unstructured interviews (Lingard, 2002) and more structured interviews, which are guided by a predetermined format that can provide either a rigid or loosely constrained format. Structured interviews require more time to develop but are more systematic (Cooke et al., 2000). Interviews are typically conducted with a single person at a time; however, it is not uncommon to conduct a focus group, where the entire group is simultaneously interviewed. In a focus group, a facilitator interviews by leading a free and open group discussion (Myllyaho et al., 2004). The analysis of interview data requires basic qualitative data analysis techniques (Adler and Adler, 1994; Patton, 2001). Conducting interviews can be straightforward, but the analysis of the data relies tremendously on the interviewer’s interpretations (Langan-Fox, 2000). Key steps to analyzing interviews are coding the data for themes (Lingard, 2002) and then studying the codes for meaning. Each phrase is closely examined to discover important concepts and reveal overall relationships. For a more holistic approach to analysis, a group interview technique can be used to discuss findings and to generate collective meaning given specific questions (Myllyaho et al., 2004). Content analysis is commonly used to analyze written statements (Langan-Fox and Tan, 1997). Other key analysis techniques focus on process analysis (Fussell et al., 1998; Pritchard, 2006), specifically looking at discussion topics, group coordination, group cognitive overload, and analysis of task process. Other group characteristic analysis techniques could include role analysis and power analysis (Aviv, 2003).

**Questionnaires**

Questionnaires are a commonly used technique to collect data about group processes (O’Neil et al., 2000; Sy, 2005; Webb, 1982; Willis et al., 2002). Similar to highly structured interviews, questionnaires can also look at relationship-oriented processes and task-oriented processes (Urch Druskat and Kayes, 2000). Questionnaires can be either closed ended or open ended (Alavi, 1994). Open-ended questionnaires are more closely related to a structured interview; the data collected using this format can be focused on group processes as well as group characteristics. Closed-ended questionnaires offer a limited set of responses. The limited responses involve some form of scale that could be nominal, ordinal, interval, or ratio. Data from
this format have a limited ability to capture group process data, but this is the typical format for collecting data associated with group characteristics such as social space, group efficacy scales, group skills, group efficacy, group attitudes, group member roles, leadership, and group knowledge. Data from questionnaires can be analyzed much like interview data if the items are open ended. If the questionnaire is closed ended, then the instrument must be scrutinized for reliability prior to data analysis. Assuming sufficient evidence of reliability, analyzing data from closed-ended questionnaires involves interpreting a measurement based on a particular theoretical construct. The types of data analysis techniques that are appropriate depend on the type of scale used in a questionnaire (nominal, ordinal, interval, or ratio).

**Conceptual Methods**

Conceptual methods involve assessing individual or group understanding about a given topic. Several data collection techniques are utilized to elicit knowledge structures. A review of the literature by Langan-Fox et al. (2000) found that knowledge in teams has been investigated by several qualitative and quantitative methods, including various elicitation techniques (e.g., cognitive interviewing, observation, card sorting, causal mapping, pairwise ratings) and representation techniques (e.g., MDS, distance ratio formulas, Pathfinder) that utilize aggregate methods.

One of the most common methods for assessing group knowledge is the use of concept maps (Herl et al., 1999; Ifenthaler, 2005; O’Connor and Johnson, 2004; O’Neil et al., 2000). Through concept mapping, similarity of group mental models can be measured in terms of the proportion of nodes and links shared between one concept map (mental model) and another (Rowe and Cooke, 1995). Several researchers believe that group knowledge and group processes are linked. Research has shown that specific group interactions such as communication and coordination mediate the development of group knowledge and thus mediate group performance (Mathieu et al., 2000). Group interactions coupled with group shared knowledge are a predominate force in the construct of group cognition. As teammates interact, they begin to share knowledge, thus enabling them to interpret cues in similar ways, make compatible decisions, and take proper actions (Klimoski and Mohammed, 1994). Group shared knowledge can help group members explain other members’ actions, understand what is occurring with the task, develop accurate expectations about future member actions and task states, and communicate meanings efficiently.

Analyzing knowledge data can certainly involve qualitative methods. These methods tend to offer more detail and depth of information than might be found through statistical analyses (Miles and Huberman, 1994; Patton, 2001). Using qualitative analysis, we obtain greater understanding about the relationships between concepts within the context of the individual mental model. We also gain better insight into the sharedness of understanding between group members. Quantitative data analysis techniques provide researchers with tools to draw inferences on the change in group knowledge as well as statistically proving a change or variation in knowledge structures.

Several methods have been developed to analyze data regarding group knowledge. Most of them include an elicitation and analysis component. Some techniques use mixed methods such as the analysis-constructed shared mental model (ACSMM) (O’Connor and Johnson, 2004), DEEP (Spector and Koszalka, 2004), and social network analysis (Qureshi, 1995). Other methods are quantitative in nature, such as the Stanford Microarray Database (SMD) (Ifenthaler, 2005), Model Inspection Trace of Concepts of Relations (MITOCAR) (Pirnay-Dummer, 2006), multidimensional scaling (MDS), distance ratio formula, and Pathfinder (Cooke et al., 2000).

Group product data are the artifacts created from a group interaction. Group products typically do not capture the process that a group undertook to create the product but is evidence of the group’s abilities. Many research studies that claim to study group processes only capture group product data. This is due in part to the claim that is made regarding the link between group products and group processes and characteristics (Cooke et al., 2000; Lesh and Dorr, 2003; Mathieu et al., 2000; Salas and Cannon-Bowers, 2001; Schweiger, 1986). Although some evidence suggests this relationship in a few areas, more research is required to substantiate this claim.

Analysis of the group product data involves techniques used when analyzing individual products. Analyzing the quality of a product can be facilitated by the use of specified criteria. These criteria are used to create a product rating scale. Rating scales can include numerical scales, descriptive scales, or checklists. Numerical scales present a range of numbers (usually sequential) that are defined by a label on either end. Each item in the questionnaire is rated according to the numerical scale. There is no specific definition of what the various numbers mean, except for the indicators at the ends of the scale; for example, a scale from 1 (very weak) to 5 (very strong) is very subjective but relatively easy to create. Descriptive scales are similar, but focus on verbal statements. Numbers can
be assigned to each statement. Statements are typically in a logical order. A common example of a descriptive scale is “strongly disagree (1), disagree (2), neutral (3), agree (4), and strongly agree (5).” A checklist can be developed to delineate specific qualities for a given criterion. This can provide a high level of reliability because a specific quality is presented and the rater simply indicates whether an item is present or not. The validity of a checklist requires a careful task analysis to ensure scale validity.

**General Considerations for Group Learning Process Assessment**

In assessment of group learning processes, researchers should consider several issues. These issues fall into four categories: (1) group setting, (2) variance in group member participation, (3) overall approach to data collection and analysis, and (4) thresholds.

**Group Setting**

To a somewhat lesser degree than the other three issues, group settings should be considered when determining the best approach and methods for a particular study. Finalizing which techniques to use may depend on whether the groups will be working in a group learning setting or individually and then coming together as a group at various points. Some groups may meet in face-to-face settings or other settings that allow for synchronous interactions; however, distributed groups have technology-enabled synchronous and asynchronous tools or asynchronous interactions only. These variations in group setting can influence the selection of specific group learning process assessment methods.

**Variance in Group Member Participation**

When collecting multiple sets of data over time, researchers should consider how they will deal with a variance in group member participation (group members absent during data collection or new members joining the group midstream). There are benefits and consequences for any decision made, but it is necessary to determine whether or not all data collected will be used, regardless of the group members present at the time of data collection. Researchers who choose not to use all data might consider using only those data submitted by group members who were present during all data collection sessions (O’Connor and Johnson, 2004). If data analysis will be based on a consistent number of group members, it will be necessary to consider how to handle data from groups that may not have the same group members present in each measure point. Also, with fluctuations in group compositions, it is important to consider the overall group demographics and possible influences of individual group members on the group as a whole.

**Overall Approach to Data Collection and Analysis**

In a holistic approach, individual group members work together and one dataset represents the group as a whole; however, the processes of group interaction naturally changes how individual group members think. The alternative is to capture individual measures and perform some type of aggregate analysis methods to represent the group; however, researchers should consider whether or not the aggregate would be a true representation of the group itself.

**Thresholds**

When using indirect measures that require an aggregation or manipulation of data prior to analysis, researchers will have to consider such issues as similarity scores. These scores define the parameters for deciding if responses from one individual group member are similar to the responses from other group members (O’Connor and Johnson, 2004; Rentsch and Hall, 1994; Rentsch et al., in press); for example, will the rating of 3.5 on a 5-point scale be considered similar to a rating of 4.0 or a rating of 3.0? When aggregating individual data into a representation of the group, will the study look only at groups where a certain percentage of the group responded to measures (Ancona and Caldwell, 1991)? How will what is similar or shared across individual group members be determined? Will the analysis use counts (x number of group members) or percentage of the group (e.g., 50%)? What level of similarity or sensitivity will be used to compare across groups (O’Connor and Johnson, 2004)—50%? 75%? What about the level of mean responses in questionnaires (Urch Druskat and Kayes, 2000)? Many different thresholds that must be considered when assessing group learning processes and analyzing group data are not concerns when studying individuals.

**Conclusion**

Assessment of group learning processes is more complex than assessment of individual learning processes because of the additional considerations necessary for selecting data collection and analysis methods. As in most research, the “very type of experiment set up by researchers will determine the type of data and therefore what can be done with such data in analysis and interpretation” (Langan-Fox et al., 2004, p. 348).
Indeed, it is logical to allow the specific research questions to drive the identification of data collection methods. The selection of research questions and subsequent identification of data collection methods will naturally place limitations on suitable data analysis methods. Careful planning for the study of group learning processes, from the selection of direct or indirect assessment measures to considering the possible influences group characteristics may have on group learning processes, is essential.

Because of the many possible combinations of methods and techniques available for studying group learning processes, some feel that research has not yet done enough to determine the best methods for studying groups (Langan-Fox et al., 2000). Many group learning process studies consider only outcome measures and do not directly study group learning processes (Worchel et al., 1992). Others look only at portions of the group process or attempt to assess group learning processes through a comparison of discrete measures of the group over time. Still other methods for data collection and analysis of group data are being developed as we speak (Seel, 1999). No one best method has been identified for analyzing group learning process data, so we suggest that studies should consider utilizing multiple methods to obtain a more comprehensive picture of group learning processes. If we are to better understand the notion of group learning processes and utilize that understanding in design, implementation, and management of learning groups in the future, then we must address the basic issues that are related with conceptualization and measurement (Langan-Fox et al., 2004).

**ASSESSMENT OF COMPLEX PERFORMANCE**

Tamara van Gog, Remy M. J. P. Rikers, and Paul Ayres

This chapter section discusses assessment of complex performance from an educational research perspective, in terms of data collection and analysis. It begins with a short introduction on complex performance and a discussion on the various issues related to selecting and defining appropriate assessment tasks, criteria, and standards that give meaning to the assessment. Although many of the issues discussed here are also important for performance assessment in educational practice, readers particularly interested in this topic might want to refer, for example, to Chapter 44 in this Handbook or the edited books by Birenbaum and Dochy (1996) and Segers et al. (2003). For a discussion of laboratory setups for data collection, see the section by Duley et al. in this chapter.

Complex performance can be defined as performance on complex tasks; however, definitions of task complexity differ: Campbell (1988), in a review of the literature, categorized complexity as primarily subjective (psychological) or objective (function of objective task characteristics), or as an interaction between objective and subjective (individual) characteristics. Campbell reported that the subjective perspective emphasized psychological dimensions such as task significance and identity. On the other hand, objective definitions consider the degree of structuredness of a task or of the possibility of multiple solution paths (Byström and Järvelin, 1995; Campbell, 1988). When the process of task performance can be described in detail *a priori* (very structured), a task is considered less complex; in contrast, when there is a great deal of uncertainty, it is considered highly complex. Similarly, complexity can vary according to the number of solutions paths possible. When there is just one correct solution path, a task is considered less complex than when multiple paths can lead to a correct solution or when multiple solutions are possible.

For the interaction category, Campbell (1988) argued that both the problem solver and the task are important. By defining task complexity in terms of cognitive load (Chandler and Sweller, 1991; Sweller, 1988; Sweller et al., 1998), an example of this interaction can readily be shown. From a cognitive load perspective, complexity is defined by the number of interacting information elements a task contains, which have to be simultaneously handled in working memory. As such, complexity is influenced by expertise (i.e., subjective, individual characteristic); what may be a complex task for a novice may be a simple task for an expert, because a number of elements have been combined into a cognitive schema that can be handled as a single element in the expert’s working memory. Tasks that are highly complex according to the objective definition (i.e., lack of structuredness and multiple possible solution paths) will also be complex in the interaction definition; however, according to the latter definition, even tasks with a high degree of structuredness or one correct solution path can be considered complex, given a high number of interacting information elements or low performer expertise.

In this chapter section, we limit our discussion to methods of assessment of complex performance on cognitive tasks. What is important to note throughout this discussion is that the methods described here can be used to assess (improvements in) complex performance both during training and after training, depending on the research questions one seeks to address. After training, performance assessment usually has the goal to assess learning, which is a goal of many studies.
in education and instructional design. If one seeks to assess learning, one must be careful not to conclude that participants have learned because their performance improved during training. As Bjork (1999) points out, depending on the training conditions, high performance gains during training may not be associated with learning, whereas low performance gains may be. It is important, therefore, to assess learning on retention or transfer tasks, instead of on practice tasks. Selection of appropriate assessment tasks is an important issue, which is addressed in the next section.

Assessment Tasks

An essential step in the assessment of performance is the identification of a collection of representative tasks that capture those aspects of the participant’s knowledge and skills that a study seeks to address (Ericsson, 2002). Important factors for representativeness of the collection of assessment tasks are authenticity, number, and duration of tasks, all of which are highly influenced by the characteristics of the domain of study.

Selecting tasks that adequately capture performance often turns out to be very difficult. Selecting atypical or artificial tasks may even impede learners in demonstrating their true level of understanding. Traditional means to evaluate the learners’ knowledge or skills have been criticized because they often fail to demonstrate that the learner can actually do something in real life or in their future workplace with their knowledge and skills they have acquired during their training (see, for example, Anderson et al., 1996; Shepard, 2000; Thompson, 2001).

The argument for the use of authentic tasks to assess the learners’ understanding has a long tradition. It started in the days of John Dewey (1916) and continues to the present day (Merrill, 2002; van Merriënboer, 1997); however, complete authenticity of assessment tasks may be difficult to realize in research settings, because the structuredness of the domain plays a role here. For structured domains such as chess and bridge, the same conditions can be reproduced in a research laboratory as those under which performance normally takes place; for less or ill-structured domains, this is difficult or even impossible to do (Ericsson and Lehmann, 1996). Nonetheless, one can always strive for a high degree of authenticity. Gulikers et al. (2004) defined authentic assessment as a five-dimensional construct (i.e., task, social context, physical context, form/result, and criteria) that can vary from low to high on each of the dimensions.

The number of tasks in the collection and the duration are important factors influencing the reliability and generalizability of a study. Choosing too few tasks or tasks of too short duration will negatively affect reliability and generalizability. On the other hand, choosing a large number of tasks or tasks of a very long duration will lead to many practical problems and might exhaust both participants and researchers. In many complex domains (e.g., medical diagnosis), it is quite common and often inevitable to use a very small set of cases because of practical circumstances and because detailed analysis of the learners’ responses to these complex problems is very difficult and time consuming (Ericsson, 2004; Ericsson and Smith, 1991). Unfortunately, however, there are no golden rules for determining the adequate number of tasks to use or their duration, because important factors are highly dependent upon the domain and specific context (Van der Vleuten and Schwirnth, 2005). It is often easier to identify a small collection of representative tasks that capture the relevant aspects of performance in highly structured domains (e.g., physics, mathematics, chess) than in ill-structured domains (e.g., political science, medicine), where a number of interacting complex skills are required.

Assessment Criteria and Standards

The term assessment criteria refers to a description of the elements or aspects of performance that will be assessed, and the term assessment standards refers to a description of the quality of performance (e.g., excellent/good/average/poor) on each of those aspects that can be expected of participants at different stages (e.g., age, grade) (Arter and Spandel, 1992). As Woolf (2004) pointed out, however, the term assessment criteria is often used in the definition of standards as well. Depending on the question one seeks to answer, different standards can be used, such as a participant’s past performance (self-referenced), peer group performance (norm-referenced), or an objective standard (criterion-referenced), and there are different methods for setting standards (Cascallar and Cascallar, 2003). Much of the research on criteria and standard setting has been conducted in the context of educational practice for national (or statewide) school tests (Hambleton et al., 2000) and for highly skilled professions, such as medicine, where the stakes of setting appropriate standards are very high (Hobma et al., 2004; Van der Vleuten and Schwirnth, 2005). Although formulation of good criteria and standards is extremely important in educational practice, where certification is the prime goal, it is no less important in educational research settings. What aspects of performance are measured and what standards are set have a major impact on the generalizability and value of a study.
The degree to which the domain is well structured influences not only the creation of a collection of representative tasks but also the definition of criteria, setting of standards, and interpretation of performance in relation to standards. In highly structured domains, such as mathematics or chess, assessing the quality of the learner’s response is often fairly straightforward and unproblematic. In less structured domains, however, it is often much more difficult to identify clear standards; for example, a music student’s interpretation of a piano concerto is more difficult to assess than the student’s technical performance on the piece. The former contains many more subjective elements (e.g., taste) or cultural differences than the latter.

Collecting Performance Data

No one best method for complex performance assessment exists, and it is often advisable to use multiple measures or methods in combination to obtain as complete a picture as possible of the performance. A number of methods are described here for collecting performance outcome (product) and performance process data. Methods are classified as online (during task performance) or offline (after task performance). Which method or combination of methods is the most useful depends on the particular research question, the possible constraints of the research context, and the domain. In ill-structured domains, for example, the added value of process measures may be much higher than in highly structured domains.

Collecting Performance Outcome (Product) Data

Collecting performance outcome data is quite straightforward. One takes the product of performance (e.g., an electrical circuit that was malfunctioning but is now repaired) and scores it along the defined criteria (e.g., do all the components function as they should, individually and as a whole?). Instead of assigning points for correct aspects, one can count the number of errors, and analyze the types of errors made; however, especially for assessment of complex performance, collecting performance product data alone is not very informative. Taking into account the process leading up to the product and the cognitive costs at which it was obtained provides equally if not more important information.

Collecting Performance Process Data

Time on Task or Speed

An important indication of the level of mastery of a particular task is the time needed to complete a task. According to the power law of practice (Newell and Rosenbloom, 1981; VanLehn, 1996), the time needed to complete a task decreases in proportion to the time spent in practice, raised to some power. Newell and Rosenbloom (1981) found that this law operates across a broad range of tasks, from solving geometry problems to keyboard typing. To account for the power law of practice, several theories have been put forward. Anderson’s ACT-R explains the speed-up by assuming that slow declarative knowledge is transformed into fast procedural knowledge (Anderson, 1993; Anderson and Lebiere, 1998). Another explanation suggested that speed-up is the result of repeated encounters with meaningful patterns (Ericsson and Staszewski, 1989); that is, as a result of frequent encounters with similar elements, these elements will no longer be perceived as individual units but will be perceived as a meaningful whole (i.e., chunk). In addition to chunking, automation processes (Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977) occur with practice that allow for faster and more effortless performance. In summary, as expertise develops equal performance can be attained in less time; therefore, it is important to collect time-on-task data to assess improvements in complex performance.

Cognitive Load

The same processes of chunking and automation that are associated with decreases in the time required to perform a task are also responsible for decreases in the cognitive load imposed by performing the task (Paas and van Merriënboer, 1993; Yeo and Neal, 2004). Cognitive load can be measured using both online and offline techniques. The cognitive capacity that is allocated to performing the task is defined as mental effort, which is considered to reflect the actual cognitive load a task imposes (Paas and van Merriënboer, 1994a; Paas et al., 2003). A subjective but reliable technique for measuring mental effort is having individuals provide self-ratings of the amount of mental effort invested. A single-scale subjective rating instrument can be used, such as the nine-point rating scale developed by Paas (1992), or a multiple-scale instrument, such as the NASA Task Load Index (TLX), which was used, for example by Gerjets et al. (2004, 2006). As subjective cognitive load measures are usually recorded after each task or after a series of tasks has been completed they are usually considered to be offline measurements, although there are some exceptions; for example, Ayres (2006) required participants to rate cognitive load at specific points within tasks.

Objective online measures include physiological measures such as heart-rate variability (Paas and van Merriënboer, 1994b), eye-movement data, and secondary-task procedures (Brünken et al., 2003). Because
they are taken during task performance, those online measures can show fluctuations in cognitive load during task performance. It is notable, however, that Paas and van Merriënboer (1994b) found the heart-rate variability measure to be quite intrusive as well as insensitive to subtle fluctuations in cognitive load. The subjective offline data are often easier to collect and analyze and provide a good indication of the overall cognitive load a task imposed (Paas et al., 2003).

*Actions: Observation and Video Records*

Process-tracing techniques are very well suited to assessing the different types of actions taken during task performance, some of which are purely cognitive, whereas others result in physical actions, because the “data that are recorded are of a pre-specified type (e.g., verbal reports, eye movements, actions) and are used to make inferences about the cognitive processes or knowledge underlying task performance” (Cooke, 1994, p. 814). Ways to record data that allow the inference of cognitive actions are addressed in the following sections. The following options are available for recording the physical actions taken during task performance: (1) trained observers can write down the actions taken or check them off on an a priori constructed list (use multiple observers), (2) a (digital) video record of the participants’ performance can be made, or (3) for computer-based tasks, an action record can be made using screen-recording software or software that logs key presses and coordinates of mouse clicks.

*Attention and Cognitive Actions: Eye-Movement Records*

Eye tracking (Duchowski, 2003)—that is, recording eye-movement data while a participant is working on a (usually, but not necessarily computer-based) task—can also be used to gather online performance process data but is much less used in educational research than the above methods. Eye-movement data give insights into the allocation of attention and provide a researcher with detailed information of what a participant is looking at, for how long, and in what order. Such data allow inferences to be made about cognitive processes (Rayner, 1998), albeit cautious inferences, as the data do not provide information on why a participant was looking at something for a certain amount of time or in a certain order. Attention can shift in response to exogenous or endogenous cues (Rayner, 1998; Stelmach et al., 1997). Exogenous shifts of attention occur mainly in response to environmental features or changes in the environment (e.g., if something brightly colored would start flashing in the corner of a computer screen, your attention would be drawn to it). Endogenous shifts are driven by knowledge of the task, of the environment, and of the importance of available information sources (i.e., influenced by expertise level) (Underwood et al., 2003). In chess, for example, it was found that experts fixated proportionally more on relevant pieces than non-expert players (Charness et al., 2001). In electrical circuits troubleshooting, van Gog et al. (2005a) also found that participants with higher expertise fixated more on a fault-related component during problem orientation than participants with lower expertise.* Please note that this is not an exhaustive overview and that we have no commercial or other interest in any of the programs mentioned here. Haider and Frensch (1999) used eye-movement data to corroborate their information-reduction hypothesis, which states that with practice people learn to ignore task-redundant information and limit their processing to task-relevant information.

On tasks with many visual performance aspects (e.g., troubleshooting technical systems), eye-movement records may therefore provide much more information than video records. Some important problem-solving actions may be purely visual or cognitive, but those will show up in an eye-movement record, whereas a video record will only allow inferences of visual or cognitive actions that resulted in manual or physical actions (van Gog et al., 2005b). In addition to providing information on the allocation of attention, eye-movement data can also give information about the cognitive load that particular aspects of task performance impose; for example, whereas pupil dilation (Van Gerven et al., 2004) and fixation duration (Underwood et al., 2004) are known to increase with increased processing demands, the length of saccades (i.e., rapid eye movements from one location to another; see Duchowski, 2003) is known to decrease. (For an in-depth discussion of eye-movement data and cognitive processes, see Rayner, 1998.)

*Thought Processes and Cognitive Actions: Verbal Reports*

Probably the most widely used verbal reporting techniques are concurrent and retrospective reporting (Ericsson and Simon, 1993). As their names imply, concurrent reporting is an online technique, whereas retrospective reporting is an offline technique. Concurrent reporting, or thinking aloud, requires participants to verbalize all thoughts that come to mind during task

---

*Note that the expertise differences between groups were relatively small (i.e., this was not an expert–novice study), suggesting that eye-movement data may be a useful tool in investigating relatively subtle expertise differences or expertise development.*
performance. Retrospective reporting requires participants to report the thoughts they had during task performance immediately after completing it. Although there has been considerable debate over the use of verbal reports as data, both methods are considered to allow valid inferences to be made about the cognitive processes underlying task performance, provided that verbalization instructions and prompts are carefully worded (Ericsson and Simon, 1993).

Instructions and prompts should be worded in such a way that the evoked responses will not interfere with the cognitive processes as they occur during task performance; for example, instructions for concurrent reporting should tell participants to think aloud and verbalize everything that comes to mind but should not ask them to explain any thoughts. Prompts should be as unobtrusive as possible. Prompting participants to “keep thinking aloud” is preferable over asking them “what are you thinking?” because this would likely evoke self-reflection and, hence, interfere with the cognitive processes. Deviations from these instructional and prompting techniques can change either the actual cognitive processes involved or the processes that were reported, thereby compromising the validity of the reports (Boren and Ramey, 2000; Ericsson and Simon, 1993). Magliano et al. (1999), for example, found that instructions to explain, predict, associate, or understand during reading influenced the inferences from the text that participants generated while thinking aloud. Although the effect of instructions on cognitive processes is an interesting topic of study, when the intention is to elicit reports of the actual cognitive processes as they would occur without intervention, Ericsson and Simon’s (1993) guidelines for wording instructions and prompts should be adhered to.

Both reporting methods can result in verbal protocols that allow for valid inferences about cognitive processes; however, the potential for differences in the information they contain must be considered when choosing an appropriate method for answering a particular research question. According to Taylor and Dionne (2000), concurrent protocols mostly seem to provide information on actions and outcomes, whereas retrospective protocols seem to provide more information about “strategies that control the problem solving process” and “conditions that elicited a particular response” (p. 414). Kuusela and Paul (2000) reported that concurrent protocols contained more information than retrospective protocols, because the latter often contained only references to the effective actions that led to the solution. Van Gog et al. (2005b) investigated whether the technique of cued retrospective reporting, in which a retrospective report is cued by a replay of a record of eye movements and mouse/keyboard operations made during the task, would combine the advantages of concurrent (i.e., more action information) and retrospective (i.e., more strategic and conditional information) reporting. They found that both concurrent and cued retrospective reporting resulted in more action information, as well as in more strategic and conditional information, than retrospective reporting without a cue.

Contrary to expectations, concurrent reporting resulted in more strategic and conditional information than retrospective reporting. This may (1) reflect a genuine difference from Taylor and Dionne’s results, (2) have been due to different operationalizations of the information types in the coding scheme used, or (3) have been due to the use of a different segmentation method than those used by Taylor and Dionne (2000).

An explanation for the finding that concurrent reports result in more information on actions than retrospective reports may be that concurrent reporting occurs online rather than offline. Whereas concurrent reports capture information available in short-term memory during the process, retrospective reports reflect memory traces of the process retrieved from short-term memory when tasks are of very short duration or from long-term memory when tasks are of longer duration (Camps, 2003; Ericsson and Simon, 1993). It is likely that only the correct steps that have led to attainment of the goal are stored in long-term memory, because only these steps are relevant for future use. This is why having participants report retrospectively based on a record of observations or intermediate products of their problem-solving process is known to lead to better results (due to fewer omissions) than retrospective reporting without a cue (van Gog et al., 2005b; Van Someren et al., 1994).

Possibly, the involvement of different memory systems might also explain Taylor and Dionne’s (2000) finding that retrospective protocols seem to contain more conditional and strategic information. This knowledge might have been used during the process but may have been omitted in concurrent reporting as a result of the greater processing demands this method places on short-term memory (Russo et al., 1989). Although this explanation is tentative, there are indications that concurrent reporting may become difficult to maintain under high cognitive load conditions (Ericsson and Simon, 1993). Indeed, participants in van Gog et al.’s study who experienced a higher cognitive load (i.e., reported investment of more mental effort) in performing the tasks indicated during a debriefing after the experiment that they disliked concurrent reporting and preferred cued retrospective reporting (van Gog, 2006).
Neuroscientific Data

An emerging and promising area of educational research is the use of neuroscience methodologies to study (changes in) brain functions and structures directly, which can provide detailed data on learning processes, memory processes, and cognitive development (see, for example, Goswami, 2004; Katzir and Paré-Blagoev, 2006). Methods such as magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), electroencephalography (EEG), magnetoencephalography (MEG), positron-emission tomography (PET), and single-photon emission computed tomography (SPECT) provide (indirect) measures of neuronal activity. The reader is referred to Katzir and Paré-Blagoev (2006) for a discussion of these methods and examples of their use in educational research.

Data Analysis

Analyzing performance product, time on task, and mental effort data (at least when the subjective rating scales are used) is a very straightforward process, so it is not discussed here. In this section, analysis of observation, eye movement, and verbal protocol data is discussed, as well as the analysis of combined methods/measures.

Analysis of Observation, Eye Movement, and Verbal Protocol Data

Observation Data

Coding and analysis of observation data can take many different forms, again depending on the research question. Coding schemes are developed based on the performance aspects (criteria) one wishes to assess and sometimes may incorporate evaluation of performance aspects. Whether coding is done online (during performance by observers) or offline (after performance based on video, screen capture, or mouse-keyboard records), the use of multiple observers or raters is important for determining reliability of the coding. Quantitative analysis on the coded data can take the form of comparison of frequencies, appropriateness (e.g., number of errors), or sequences of actions (e.g., compared to an ideal or expert sequence) and interpreting the outcome in relation to the set standard.

Several commercial and noncommercial software programs have been developed to assist in the analysis of action data;* for example, Observer (Noldus et al., 2000) is commercial software for coding and analysis of digital video records; NVivo (Bazeley and Richards, 2000) is commercial software for accessing, shaping, managing, and analyzing non-numerical qualitative data; Multiple Episode Protocol Analysis (MEPA) (Erkens, 2002) is free software for annotating, coding, and analyzing both nonverbal and verbal protocols; and ACT Pro (Fu, 2001) can be used for sequential analysis of protocols of discrete user actions such as mouse clicks and key presses.

Eye-Movement Data

For analysis of fixation data it is important to identify the gaze data points that together represent fixations. This is necessary because during fixation the eyes are not entirely motionless; small tremors and drifts may occur (Duchowski, 2003). According to Salvucci (1999), the three categories of fixation identification methods are based on velocity, dispersion, or region. Most eye-tracking software allows for defining values for the dispersion-based method, which identifies fixation points as a minimum number of data points that are grouped closely together (i.e., fall within a certain dispersion, defined by pixels) and last a minimum amount of time (duration threshold). Once fixations have been defined, defining areas of interest (AoIs) in the stimulus materials will make analysis of the huge data files more manageable by allowing summaries of fixation data to be made for each AoI, such as the number of fixations, the mean fixation duration, and the total time spent fixating. Furthermore, a chronological listing of fixations on AoIs can be sequentially analyzed to detect patterns in viewing behavior.

Verbal Protocol Data

When verbal protocols have been transcribed, they can be segmented and coded. Segmentation based on utterances is highly reliable because it uses pauses in natural speech (Ericsson and Simon, 1993); however, many researchers apply segmentation based on meaning (Taylor and Dionne, 2000). In this case, segmentation and coding become intertwined, and the reliability of both should be evaluated. It is, again, important to use multiple raters (at least on a substantial subset of data) and determine the reliability of the coding scheme. The standard work by Ericsson and Simon (1993) provides a wealth of information on verbal protocol coding and analysis techniques. The software program MEPA (Erkens, 2002) can assist in the development of a coding scheme for verbal data, as well as in analysis of coded data with a variety of quantitative or qualitative methods.

* Please note that this is not an exhaustive overview and that we have no commercial or other interest in any of the programs mentioned here.


Combining Methods and Measures

As mentioned before, there is not a preferred single method for the assessment of complex performances. By combining different methods and measures, a more complete or a more detailed picture of performance will be obtained; for example, various process-tracing techniques such as eye tracking and verbal reporting can be collected and analyzed in combination with other methods of assessment (van Gog et al., 2005a). Different product and process measures can easily be combined, and it can be argued that some of them should be combined because a simple performance score* ignores the fact that, with expertise development, time on task and cognitive load decrease, whereas performance increases.

Consider the example of a student who attains the same performance score on two comparable tasks that are spread over time, where cognitive load measures indicate that the learner had to invest a lot of mental effort to complete the task the first time and little the second. Looking only at the performance score, one might erroneously conclude that no progress was made, whereas the learner actually made a subtle step forward, because reduced cognitive load means that more capacity can be devoted to further learning.

The mental efficiency measure developed by Paas and van Merriënboer (1993) reflects this relation: Higher performance with less mental effort invested to attain that performance results in higher efficiency. This measure is obtained by standardizing performance and mental effort scores, and then subtracting the mean standardized mental effort score ($z_E$) from the mean standardized performance score ($z_P$) and dividing the outcome by the square root of 2:

$$z_P - z_E \sqrt{2}$$

When tasks are performed under time constraints, the combination of mental effort and performance measures will suffice; however, when time on task is self-paced, it is useful to include the additional time parameter in the efficiency measure (making it three-dimensional) (Paas et al., 2003; Tuovinen and Paas, 2004):

$$z_P - z_E - z_T \sqrt{3}$$

---

Discussion

Much of the research into learning and instruction involves assessment of complex performances of cognitive tasks. The focus of this chapter section was on data collection and analysis methods that can be used for such assessments. First, the important issues related to selecting an appropriate collection of assessment tasks and defining appropriate assessment criteria and standards were discussed. Then, different ways for collecting performance product and process data, using online (during task performance) or offline (after task performance) measurements, were described. Analysis techniques were discussed and, given the lack of a single preferred method for complex performance assessment, ways to combine measures were suggested that will foster a more complete or more detailed understanding of complex performance.

This chapter section aimed to provide an overview of the important issues in assessment of complex performance on cognitive tasks and of available data collection and analysis techniques for such assessments, rather than any definite guidelines. The latter would be impossible when writing for a broad audience, because what constitutes an appropriate collection of tasks, appropriate criteria and standards, and appropriate data collection and analysis techniques is highly dependent on the research question one seeks to address and on the domain in which one wishes to do so. We hope that this overview, along with other chapter sections, provides the reader with a starting point for further development of rewarding and informative studies.

---

SETTING UP A LABORATORY FOR MEASUREMENT OF COMPLEX PERFORMANCES

Aaron R. Duley, Paul Ward, and Peter A. Hancock

This chapter section describes how to set up laboratories for the measurement of complex performance. Complex performance in this context does not exclusively refer to tasks that are inherently difficult to perform; rather, the term is used here in a broader sense to refer to the measurement of real-world activities that require the integration of disparate measurement instrumentation as well as the need for time-critical experimental control. We have assumed that our primary readership is comprised of graduate students and research faculty, although the chapter addresses issues relevant to all who seek a better understanding of behavioral response.
The central theme of this section relates to laboratory instrumentation. Because instrumentation is a requisite element for complex performance measurement, a common problem encountered by researchers is how to overcome the various technical hurdles that often discourage the pursuit of difficult research objectives. Thus, creating a testing environment suitable to address research questions is a major issue when planning any research program; however, searching the literature for resources relating to laboratory instrumentation configurations yields a surprisingly scant number of references and resources that address these issues. Having made just such an attempt for the purposes of this section, the ability to articulate a general-purpose exposition on laboratory setup is indeed a challenging endeavor. This pursuit is made more difficult by addressing a naturally ambiguous topic such as complex performance; nevertheless, our section looks to provide the bearings needed to resolve such questions. In particular, we cover stimulus presentation and control alternatives, as well as hardware choices for signal routing and triggering, while offering solutions for commonly encountered problems when attempting to assemble such a laboratory. Some portions of this section are moderately technical, but every attempt has been made to ensure that the content is appropriate for our target audience.

**Instrumentation and Common Configurations**

Psychology has a long legacy of employing tools and instrumentation to support scientific inquiry. The online Museum of the History of Psychological Instrumentation, for example, has illustrations of over 150 devices used by early researchers to visualize organ function and systematically investigate human psychological processes and behavior (see http://www.chss.montclair.edu/psychology/museum/museum.htm). At this museum, one can view such devices as an early Wundt-style tachistoscope or the Rotationsapparatus für Komplikations-Versuche (rotary apparatus for complication studies). Titchener, a student of Wundt, continued this tradition in his own laboratory at Cornell University and described the building requirements and the costs associated with items needed for establishing the ideal psychological laboratory (Titchener, 1900, pp. 252–253):

> For optics, there should be two rooms, light and dark, facing south and north respectively, and the later divided into antechamber and inner room. For acoustics, there should be one large room, connected directly with a small, dark, and (so far as is possible without special construction) sound-proof chamber. For haptics, there should be a moderately sized room, devoted to work on cutaneous pressure, temperature, and pain, and a larger room for investigations of movement perceptions. Taste and smell should each have a small room, the latter tiled or glazed, and so situated that ventilation is easy and so does not involve the opening of doors or transom-windows into the building. There should, further, be a clock-room, for the time-registering instruments and their controls; and a large room for the investigations of the bodily processes and changes underlying affective consciousness.

Instrumentation is a central component of complex performance measurement; however, the process by which one orchestrates several devices in the broader context of addressing an experimental question is indeed challenging. Modern-day approaches reflect a paradigm shift with respect to early psychological procedures. Traditionally, a single instrument would be used for an entire experiment. Complex performance evaluation, however, often entails situations where the presentation of a stimulus is controlled by one computer, while supplementary instrumentation collects a stream of other data on a second or perhaps yet a third computer. Certainly, an ideal testing solution would allow one to minimize the time needed to set up an experiment and maximize the experimental degree of automation, thus minimizing investigator intervention, without compromising the scientific integrity of the experiment. Nevertheless, the measurement of complex performance is often in conflict with this idyllic vision. It is not sufficient for contemporary researchers simply to design experiments. They are also required to have access to the manpower and the monetary or computational resources necessary to translate a scientific question into a tenable methodological test bed.

**Design Patterns for Laboratory Instrumentation**

Design patterns represent structured solutions for such recurring assessment problems (Gamma et al., 1995). The formal application of design patterns as abstract blueprints for common challenges has relevance for laboratory instrumentation configuration and equipment purchasing decisions. Although research questions vary, experiments will regularly share a comparable solution. These commonalities are important to identify, as the ability to employ a single set of tools has distinct advantages compared to solutions tailored for only one particular problem. Such advantages include cost savings, instrumentation sharing, instrumentation longevity, and laboratory scalability (e.g., the capacity to run multiple experiments simultaneously).
The purpose of the following section is to provide a level of abstraction for instrumentation configurations commonly encountered in the design of experiments related to complex performance. This approach is favored beyond simply providing a list of items and products that every laboratory should own. We acknowledge the considerable between-laboratory variability regarding research direction, instrumentation, and expertise; therefore, we focus primarily on instrumentation configuration and architecture as represented by design patterns common to a broad array of complex performance manipulations.

Given that an experiment will often require the manipulation of stimuli in a structured way and seeing the impracticality in comprehensively covering all research design scenarios, the following assumptions are made: (1) Stimuli are physically presented to participants, (2) some stimulus properties are required to be under experimental control (e.g., presentation length), (3) measurable responses by the participant are required, and (4) control or communication of secondary instrumentation may also be necessary. These assumptions accommodate a broad spectrum of possible designs and from these assumptions several frameworks can be outlined.

**Stimulus Presentation and Control Model**

Figure 55.2 depicts the simplest of the design patterns, which we term the *stimulus presentation and control* (SPC) model. The SPC model is a building block for more complex configurations. The basic framework of the SPC model includes the *presentation layer* and the *stimulus, control, and response layer*. The presentation layer represents the medium used to physically display a stimulus to a participant (e.g., monitor, projector, speaker, headphones). The stimulus, control, and response layer (SCRL) encapsulates a number of interrelated functions central to complex performance experimentation, such as the experimental protocol logic, and is the agent that coordinates and controls experimental flow and, potentially, participant response. Broadly speaking, SCRL-type roles include stimulus manipulation and timing, instrument logging and coordination, and response logging, in addition to experiment procedure management.

As the SCRL often contains the logic necessary to execute the experimental paradigm, it is almost always implemented in software; thus, the SCRL application is assumed to operate on a computing target (e.g., desktop, portable digital assistant), which is represented by the dashed box in Figure 55.2. As an example implementation of the SPC model, consider a hypothetical experiment in which participants are exposed to a number of visual stimuli for 6 sec each. Each visual stimulus occurs after a fixed *foreperiod* of 1 sec and a subsequent *fixation cross* (i.e., the point at which participants are required to direct their gaze) presented for 500 msec. Each visual stimulus is followed by a 2-sec inter-trial interval (ITI). The only requirement of the participant is to view the visual stimuli for the duration of its presentation. How do we implement this experiment?

This problem has several possible solutions. A monitor (presentation layer) and Microsoft PowerPoint (SCRL) would easily accomplish the task; however, the SPC model is suitable to handle an extensive arrangement of experimental designs, so additional procedural requirements increase the need for added SCRL functionality. Consider an experiment where both a monitor and speakers are required to present the stimuli. This basic pattern still reflects an SPC...
TABLE 55.2
SCRL-Type Applications

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Type</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cogent 2000/</td>
<td>Complete PC-based software environment for functional brain mapping experiments; contains commands useful for presenting scanner-synchronized visual stimuli (Cogent Graphics), auditory stimuli, mechanical stimuli, and taste and smell stimuli. It is also used in monitoring key presses and other physiological recordings from the subject.</td>
<td>Freeware</td>
<td>Windows</td>
</tr>
<tr>
<td>Cogent Graphics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DMX</td>
<td>Win 32-based display system used in psychology laboratories around the world to measure reaction times to visual and auditory stimuli.</td>
<td>Freeware</td>
<td>Windows</td>
</tr>
<tr>
<td>E-Prime</td>
<td>Suite of applications to design, generate, run, collect data, edit, and analyze the data; includes: (1) a graphical environment that allows visual selection and specification of experimental functions; (2) a comprehensive scripting language; (3) data management and analysis tools.</td>
<td>Commercial</td>
<td>Windows</td>
</tr>
<tr>
<td>Flashdot</td>
<td>Program for generating and presenting visual perceptual experiments that require a high temporal precision. It is controlled by a simple experiment building language and allows experiment generation with either a text or a graphical editor.</td>
<td>Freeware</td>
<td>Windows, Linux</td>
</tr>
<tr>
<td>FLXLab (Psychology Experiment Building Language)</td>
<td>New language specifically designed to be used to create psychology experiments.</td>
<td>Freeware</td>
<td>Linux, Windows, Mac</td>
</tr>
<tr>
<td>PsychoPy</td>
<td>Psychology stimulus software for Python; combines the graphical strengths of OpenGL with the easy Python syntax to give psychophysics a free and simple stimulus presentation and control package.</td>
<td>Freeware</td>
<td>Linux, Mac</td>
</tr>
<tr>
<td>PsyScope</td>
<td>Interactive graphic system for experimental design and control on the Macintosh.</td>
<td>Freeware</td>
<td>Mac</td>
</tr>
<tr>
<td>PsyScript</td>
<td>Application for scripting psychology experiments, similar to SuperLab, MEL, or E-Prime</td>
<td>Freeware</td>
<td>Linux, Mac</td>
</tr>
<tr>
<td>PyEPL (Python Experiment-Programming Library)</td>
<td>Library for coding psychology experiments in Python; supports presentation of both visual and auditory stimuli, as well as both manual (keyboard/joystick) and sound (microphone) input as responses.</td>
<td>Freeware</td>
<td>Linux, Mac</td>
</tr>
<tr>
<td>Realtime Experiment Interface</td>
<td>Extensible hard real-time platform for the development of novel experiment control and signal-processing applications.</td>
<td>Freeware</td>
<td>Linux</td>
</tr>
<tr>
<td>SuperLab</td>
<td>Stimulus presentation software with features that support the presentation of multiple types of media as well as rapid serial visual presentation paradigms and eye tracking integration, among other features.</td>
<td>Commercial</td>
<td>Windows, Mac</td>
</tr>
</tbody>
</table>

A number of commercial and freeware applications have been created over the past several decades to assist researchers with SCRL-type functions. The choice to select one application over the other may have much to do with programming requirements, the operating system platform, protocol requirements, or all of the above. Table 55.2 provides a list of some of the SCRL applications that are available for psychological and psychophysical experiments. Additional information for these and other SCRL applications can be found in Florer (2007). The description column is
text taken directly from narrative about the product provided by Florer (2007). A conventional programming language is best equipped for SCRL functionality. This alternative to the applications listed in Table 55.2 may be necessary for experiments where communication with external hardware, interfacing with external code, querying databases, or program performance requirements are a priority, although some of the SCRL applications listed in Table 55.2 provide some varying degrees of these capabilities (e.g., E-Prime, SuperLab).

The prospect of laboratory productivity can outweigh the flexibility and functionality afforded by a programming language; for example, from a laboratory management perspective, it is reasonable for all laboratory members to have a single platform from which they create experiments. Given the investment required to familiarize oneself with a programming language, the single platform option can indeed be challenging to implement in practice. Formulating a laboratory in this manner does allow members to share and reuse previous testing applications or utilize knowledge about the use and idiosyncrasies of an SCRL application.

Despite the learning curve, a programming language has potential benefits that cannot be realized by turnkey SCRL applications. As mentioned, high-level programming languages offer inherently greater flexibility. Although it is important to consider whether the SCRL application can be used to generate a particular test bed, one must also consider the analysis requirements following initial data collection. The flexibility of a programming language can be very helpful in this regard. One might also consider the large support base in the form of books, forums, and websites dedicated to a particular language which can mitigate the problems that may arise during the learning process.

**Stimulus Presentation and Control Model with External Hardware**

Communicating with external hardware is essential to complex performance design. Building upon the basic SPC framework, Figure 55.3 depicts the SPC model with support for external hardware (SPCxh). Figure 55.3 illustrates a scenario where the SCRL controls both monitor and headphone output. The SCRL also interfaces with an eye tracker and a bio-instrumentation device via the parallel port and a data acquisition device (DAQ), respectively. DAQ devices are an important conduit for signal routing and acquisition, and we discuss them in greater detail later. Extensions of the SPCxh and the SPC model are the interface and instrumentation layers. A good argument can be made for another interface layer to exist between the presentation layer and the SCRL, but for our purposes the interface layer specifically refers to the physical connection that exists between the SCRL and the instrumentation layer. Figure 55.3 depicts the stimulus presentation and control model with external hardware support (SPCxh). The SPCxh is derived from the basic
SPC model, with two additional layers: one to represent the external hardware, the second to interface that hardware with the SCRL.

It is important to emphasize that the SPC and SPCxh models are only examples. We recognize that an actual implementation of any one model will most certainly differ among laboratories. The main purpose of illustrating the various arrangements in this manner is to address the major question of how complex performance design paradigms are arranged in an abstract sense. When the necessary components are identified for their specific research objective, then comes the process of determining the specific hardware and software to realize that goal.

It is imperative to understand the connection between a given model (abstraction) and its real-world counterpart. Using the example described above, consider an experiment that requires the additional collection of electrocortical activity in response to the appearance of the visual stimulus. This type of physiological data collection is termed event-related potential, as we are evaluating brain potentials time-locked to some event (i.e., appearance of the visual stimulus in this example). Thus, we need to mark in the physiological record where this event appears for offline analysis. Figure 55.4 depicts one method to implement this requirement. On the left, the SPCxh model diagram is illustrated for the current scenario. A monitor is used to display the stimulus. A programming language called LabVIEW provides the SCRL functionality. Because the bio-instrumentation supports digital input/output (i.e., hardware that allows one to send and receive digital signals), LabVIEW utilizes the DAQ device to output digital markers to the digital input ports of the bio-instrumentation while also connecting to the instrument to collect the physiological data over the network. We are using the term bio-instrumentation here to refer to the hardware used for the collection and assessment of physiological data. The picture on the right portrays the instantiation of the diagram. It should be observed that the diagram is meant to represent tangible software and hardware entities. Although LabVIEW is used in this example as our SCRL application, any number of alternatives could have also been employed to provide the linkage between the software used in our SCRL and the instrumentation.

A number of derivations can also be organized from the SPCxh model; for example, in many cases, the SCRL may contain only the logic needed to execute the experiment but not the application program interfaces (APIs) required to directly control a vendor’s hardware. In these situations, it may be necessary to run the SCRL application alongside (e.g., on another machine) the vendor-specific hardware application. Figure 55.5 depicts this alternative, where the vendor-specific hardware executes its procedures at the same time as the SCRL application. Because the layers for
Data Collection and Analysis

the SPCxh are the same as in Figure 55.3, Figure 55.5 depicts only the example instantiation of the model and not the layers for the SPCxh model. The SPCxh example in Figure 55.5 is a common configuration because hardware vendors do not always supply software interfaces that can be used by an external application. The major difference between the two options, as shown, is that the second option would require a total of three computing targets: one to execute the SCRL application and for stimulus presentation, one to execute the bio-instrumentation device software, and a third to execute the eye tracker device software.

A very common question is how to synchronize the SCRL application and the device instrumentation. As with the previous example, the method of choice for the example is via the DAQ device for communication with the bio-instrumentation and through the parallel port for the eye tracker; however, the option for event marking and synchronization is only applicable if it is supported by the particular piece of instrumentation. Furthermore, the specific interface (e.g., digital input/output, serial/parallel) is dependent on what the manufacturer has made available for the end-user. Given this information, one should ask themselves the following questions prior to investing resources in any one instrument or SCRL alternative. First, what type of limitations will be encountered when attempting to interface a particular instrument with my current resources? That is, does the manufacturer provide support for external communication with other instruments or applications? Second, does my SCRL application of choice support a method to communicate with my external hardware if this option is available? Third, does the device manufacturer provide or sell programming libraries or application program interfaces if data collection has to be curtailed in some particular way? Fourth, what are the computational requirements to run the instrumentation software? Will the software be so processor intensive that it requires a sole execution on one dedicated machine?

**Common Paradigms and Configurations**

A number of common paradigms exist in psychological research, from recall and recognition paradigms to interruption-type paradigms. Although it is beyond the scope of this section to provide an example configuration for each, we have selected one example that is commonly used by many of contemporary researchers; this is the *secondary task paradigm*. Both the SPC and SPCxh models are sufficient for experiments employing this paradigm; however, a common problem can occur when the logic for the primary and secondary tasks is implemented as mutually exclusive entities. An experiment that employs a simulator for the primary task environment can be viewed as a self-contained SPCxh model containing a simulation presentation environment (presentation layer), simulation control software (SCRL application), and simulation-specific hardware (interface and instrumentation layers). The question now is how can we interface the primary task (operated by the simulator in this example) with another SPRL application that contains the logic for the secondary task?
Figure 55.6 contains a graphical representation of a possible configuration: an SPCxh model for the simulator communicating via a network interface that is monitored from the SCRL application on the secondary task side. On the left side of Figure 55.6 is the primary task configuration, and on the right side is the secondary task configuration. It should be noted that the simulator control software, our SCRL application, or the device-specific software does not necessarily need to be executed on separate computers; however, depending on the primary or secondary task, one may find that the processor and memory requirements necessitate multiple computers. On the secondary task side, the diagram represents a fairly complex role for the SCRL application. As shown, the SCRL application has output responsibilities to a monitor and headphones while also interfacing with an eye tracker via the serial port, interfacing the simulator via the network, and sending two lines of digital output information via the DAQ device.

Numerous complex performance test beds require that a primary and secondary task paradigm be used to explicate the relationship among any number of processes. In the field of human factors, in particular, it is not uncommon for an experiment to employ simulation for the primary task and then connect a secondary task to events that may occur during the primary task. A major problem often encountered in simulation research is that the simulators are often closed-systems; nevertheless, most simulators can be viewed as SPCxh arrangements with a presentation layer of some kind, an application that provides SCRL function, and the simulator control hardware itself. If one wishes to generate secondary task events based on the occurrence of specific events in the simulation, the question then becomes one of how we might go about configuring such a solution when there is no ready-made entry point between the two systems (i.e., primary task system and secondary task system). The diagram on the left shows the SPCxh model for the secondary task, which is responsible for interacting with a number of additional instruments. The diagram shows a connecting line between the secondary systems network interface and the network interface of the primary task as controlled by the simulation. Because simulation manufacturers will often make technotes available that specify how the simulator may communicate with other computers or hardware, one can often ascertain this information for integration with the secondary tasks SCRL.

Summary of Design Configurations

The above examples have been elaborated in limited detail. Primarily, information pertaining to how one would configure a SCRL application for external software and hardware communication has been excluded. This specification is not practical with the sheer number of SCRL options available to researchers. As well, the diagrams do not inform the role that the SCRL application plays when interfacing with instrumentation. It may be the case that the SCRL application plays a minimal role in starting and stopping the instrument and does not command full control of the instrument.
Data acquisition devices have traditionally been referred to as A/D boards (analog-to-digital boards) because of their frequent use in signal acquisition. A signal, in this context, loosely refers to any measurable physical phenomenon. Signals can be divided into two primary classes. Analog signals can vary continuously over an infinite range of values, whereas digital signals contain information in discrete states. To remember the difference between these two signals, visualize two graphs, one where someone’s voice is recorded (analog) and another plotting when a light is switched on or off (digital).

The traditional role of a DAQ device was to acquire and translate measured phenomena into binary units that can be represented by a digital device (e.g., computer, oscilloscope). Suppose we need to record the force exerted on a force plate. A DAQ device could be configured such that we could sample the data derived from a force plate at a rate of 1 msec per sample and subsequently record that data to a computer or logging instrument.

One should note that the moniker device as a replacement for board is more appropriate given that modern DAQ alternatives are not always self-contained boards but may connect to a computing target in a few different ways. DAQ devices are available for a number of bus types. A bus, in computing vernacular, refers to a method of transmission for digital data (e.g., USB, FireWire, PCI). The DAQ device pictured in Figure 55.4, for example, is designed to connect to the USB port of a computer.

Despite their traditional role in signal acquisition (e.g., analog input), most DAQ devices contain an analog output option. Analog output reverses the process of an A/D conversion and can be used to convert digital data into analog data (D/A conversion). Analog output capabilities are useful for a variety of reasons; for example, an analog output signal can be used to produce auditory stimuli, control external hardware, or output analog data to supplementary instrumentation. The primary and secondary task example above illustrates one application for analog output. Recall that in the SPCxh model described earlier the simulator was a closed-system interfaced with the SCRL application via a network interface. Suppose that it was necessary to proxy events as they occurred in the simulation to secondary hardware. A reason for this approach might be as simple as collapsing data to a single measurement file; for example, suppose we want to evaluate aiming variability in a weapons simulation in tandem with a physiological measure. One strategy would require that we merge the separate streams of data after all events have been recorded, but, by using another strategy employing the analog output option,
we could route data from the simulator and proxy it via the SCRL application controlling the DAQ device connected to our physiological recording device.

In addition to analog output, DAQ functionality for digital input/output is a critical feature to solving various complex performance measurement issues. Recall the example above where an experiment required that the SCRL application tell the physiological control software when the visual stimulus was displayed. This was accomplished by the SCRL application sending a digital output from the DAQ device to the digital input port on the bio-instrument. Figure 55.7 depicts this occurrence from the control software for the bio-instrument. The figure shows one channel of data recorded from a participant’s scalp (i.e., EEG); another digital channel represents the onset and offset of the visual stimulus.

When configuring digital signals, it is also important that one understand that the event can be defined in terms of the leading or trailing edge of the digital signal. As depicted in Figure 55.7, the leading edge (also called the rising edge) refers to the first positive deflection of the digital waveform, while the trailing edge (also called the falling edge) refers to the negative going portion of the waveform. This distinction is important, because in many cases secondary instrumentation will provide an option to begin or end recording from the leading or trailing edge; that is, if we mistakenly begin recording on the trailing edge when the critical event occurs on the leading edge, then the secondary instrument may be triggered late or not at all. Another term native to digital events is transistor–transistor logic (TTL), which is often used to express digital triggering that operates within specific parameters. TTL refers to a standard where a given instrument’s digital line will change state (e.g., on to off) if the incoming input is within a given voltage range. If the voltage supplied to the digital line is 0 then it is off, and if the digital line is supplied 5 volts then it is on.

Event triggering is an extremely important constituent of complex performance experimentation. Knowing when an event occurs may be vital for data analysis or for triggering subsequent events. The example here depicts a scenario where we are interested in knowing the onset occurrence of a visual stimulus so we can analyze our EEG signal for event-related changes. The channel with the square wave is the event that tells when the event occurred, with the leading edge representing the onset of the visual stimuli (5 volts) and the falling edge reflecting its offset (0 volts). A setup similar to that demonstrated in Figure 55.4 could easily be configured to produce such an example. Although this example only shows two channels, a real-world testing scenario may have several hundred to indicate when certain events occur. One strategy is to define different events as different channels. One channel may represent the visibility of a stimuli (on or off), another may represent a change in its color, and yet another may indicate any other number of events. An alternative solution is to configure the SCRL application to send data to a single channel and then create a coding scheme to reflect different events (e.g., 0 volts, stimulus hidden; 2 volts, stimulus visible; 3 volts, color changed to black; 4 volts, color changed to white). This approach reduces the number of channels and maximizes the number of digital channels that one can control.
The utility of digital event triggering cannot be overstated. Although its application for measurement of complex performance requires a small degree of technical expertise, the ability to implement event triggering affords a great degree of experimental flexibility. Under certain circumstances, however, analog triggering may also be appropriate. Consider an experiment where the threshold of a participant’s voice serves as the eliciting event for a secondary stimulus. In this case, it is necessary to determine whether or not the SCRL application and DAQ interface might support this type of triggering, because such an approach offers greater flexibility for some design configurations.

**Purchasing a DAQ Device**

The following questions are relevant to purchasing a DAQ device for complex performance research. First, is the DAQ device going to be used to collect physiological or other data? If the answer is yes, one should understand that the price of a DAQ device is primarily a function of its resolution, speed, form factor, and the number of input/output channels supported. Although a discussion of unipolar vs. bipolar data acquisition is beyond the scope of this chapter, the reader should consult Olansen and Rosow (2002) and Stevenson and Soejima (2005) for additional information on how this may affect the final device choice.

Device resolution, on the other hand, refers to the fidelity of a DAQ device to resolve analog signals; that is, what is the smallest detectable change that the device can discriminate? When choosing a board, resolution is described in terms of bits. A 16-bit board can resolve signals with greater fidelity than a 12-bit board. By taking the board resolution as an exponent of 2, one can see why. A 12-bit board has \(2^{12}\) or 4096 possible values, while a 16-bit has a \(2^{16}\) or 65,536 possible values. The answer to this question is also a function of a few other factors (e.g., signal range, amplification). A complete understanding of these two major issues should be established prior to deciding on any one device. As well, prior to investing in a higher resolution board, which will cost more money than a lower resolution counterpart, one should evaluate what is the appropriate resolution for the application. If the primary application of the DAQ device is for digital event triggering, then it is important to purchase a device that is suitable for handling as many digital channels as are necessary for a particular research design.

Third, does the design require analog output capabilities? Unlike digital channels, which are generally reconfigurable for input or output, analog channels are not, so it is important to know in advance the number of analog channels that a DAQ device supports. Fourth, does the testing environment require a level of timing determinism that cannot be provided by software in less than 1 msec? For these scenarios, researchers might want to consider a DAQ device that supports hardware timing. For additional information about A/D board specifications and other factors that may affect purchasing decisions, see Staller (2005).

**Computers as Instrumentation**

Computers are essential to the modern laboratory. Their value is evident when one considers their versatile role in the research process; consequently, the computer’s ubiquity in the science can account for significant costs. Because academic institutions will often hold contracts with large original equipment manufactures, computing systems are competitively priced and warranties ensure maintenance for several years. Building a system from the ground up is also a viable option that will often provide a cost-effective alternative to purchasing through an original equipment manufacturer. Although the prospect of assembling a computer may sound daunting, the process is really quite simple, and numerous books and websites are dedicated to the topic (see, for example, Hardwidge, 2006). Customization is one of the greatest advantages to building a machine, and because only necessary components are purchased overall cost is usually reduced. On the other hand, a major disadvantage to this approach is the time associated with reviewing the components, assembling the hardware, and installing the necessary software.

Most new computers will likely be capable of handling a good majority of laboratory tasks; however, one should have a basic understanding of its major components to make an informed purchasing decision when planning complex performance test beds. This is important given that one can potentially save considerable resources that can be allocated for other equipment and be assured that the computer is adequate for a given experimental paradigm.

The following questions should be considered whether building or buying a complete computing system. First, does the number of expansion ports accommodate input boards that may be needed to interface instrumentation? For example, if an instrument interfaced an application via a network port and we wanted to maintain the ability to network with other computers or access the Internet, it would be important to confirm that the motherboard of the computer had a sufficient number of slots to accommodate this addition. Furthermore, because DAQ devices are often sold as input boards this same logic would apply.

Computers have evolved from general-purpose machines to machines with specific aptitudes for particular tasks. A recent development is that vendors now
market a particular computing system for gaming vs. video-editing vs. home entertainment purposes. To understand the reasons behind these configurations, we strongly advocate developing a basic understanding of how certain components facilitate particular tasks. Although space prevents us from accomplishing this within this chapter, it is important to realize that computing performance can alter timing determinism, particularly in complex performance environments.

**Discussion**

The challenge of understanding the various technical facets of laboratory setup and configuration represents a major hurdle when the assessment of some complex performance is a central objective. This section has discussed the common problems and design configurations that one may encounter when setting up such a laboratory. This approach, abstract in some respects, was not intended to illustrate the full range of design configurations available for complex performance evaluation; rather, the common configurations discussed here should only be viewed as general-purpose architectures, independent of new technologies that may emerge. After developing an understanding of the various design configurations, one must determine the specific hardware and software that are required to address the research question. The purpose of providing a few design configurations here was to emphasize that, in many complex performance testing environments, one must specify what equipment or software will fill the roles of presentation, stimulus control and response, instrumentation, and their interfaces.

**CONCLUDING REMARKS**

Setting up laboratories for the measurement of complex performances can indeed be a challenging pursuit; however, becoming knowledgeable about the solutions and tools available to aid in achieving the research objectives is rewarding on a number of levels. The ability to identify and manipulate multiple software and hardware components allows quick and effective transitioning from a research question into a tenable methodological test bed.

**REFERENCES**


803


* Indicates a core reference.