TIMOTHY A. SLOCUM, Utah State University

*Evaluation of Direct Instruction Implementations**

Objectives

After studying this chapter you should be able to

1. Describe four broad roles of measurement in Direct Instruction.
2. Explain the importance of assessing students with placement tests and responding thoughtfully to this information by placing students in appropriate Direct Instruction programs and lessons.
3. Explain the importance of using lesson performance as an assessment of whether students are placed in an appropriate program and lesson.
4. Describe how students' oral and written responses during lessons provide important information for evaluation and adjustment in the implementation of a Direct Instruction program.
5. Describe the role of mastery tests and checkouts in formative evaluation of the implementation of Direct Instruction programs.
6. Explain why it is important to make summative evaluations of Direct Instruction implementations.
7. Explain the basic question that is addressed by measurement validity.
8. Describe what is meant by "evidence based on test content" and how it relates to the general concept of measurement validity.
9. Describe the three main kinds of concerns that help determine the targets of evaluation of Direct Instruction implementations.
10. Describe how we judge the validity of a measure in an evaluation of whether the implementation of a Direct Instruction program (a) achieves the objectives of the program, (b) increases student performance in broad areas such as reading comprehension and math problem solving, and (c) increases students' performance on state or national standards.
11. List three basic evaluation designs and describe their critical features.

Overview

Direct Instruction is about producing measurable improvements in student performance. Measurable improvement in performance is the overarching goal of Direct Instruction, and every element of Direct Instruction is designed and arranged to contribute to this

Journal of Direct Instruction, Vol. 3, No. 2, pp. 111–137.

* This article is coordinated with the textbook, *Introduction to Direct Instruction*, and can be seen as supplemental material for the text.

goal. Since *measurable* progress is the goal, it is not surprising that measurement is integral to all aspects of Direct Instruction. We can define four broad roles of measurement in Direct Instruction.

First, as we learned in Chapter 2, Direct Instruction programs are built on a base of research-validated instructional practices. Many of the strategies, tactics, and particular practices that are embodied in Direct Instruction programs have extensive research support. The findings of this research are built into the teaching procedures in Direct Instruction programs.

Second, as we also learned in Chapter 2, the development of new Direct Instruction programs includes cycles of field testing and revision before final versions are published. Thus, these programs have been subjected to systematic evaluation during the development process. In addition, as older programs are revised and new editions are published, the authors take into account extensive information on how the programs work in the classroom and any aspects that need further refinement.

Third, Direct Instruction programs employ extensive assessment and teacher decision making to adjust the delivery of programs to the specific needs of individuals and groups of students. Direct Instruction programs include (a) placement tests for initial program placement; (b) group and individual oral responses, and daily written activities for immediate adjustments to instruction; and (c) mastery tests and checkouts for identifying any needs for remediation. All of these assessments are built into the program. Proper implementation of a Direct Instruction program *requires* that the results from these assessments be used in making thoughtful instructional decisions.

Fourth, the outcomes of Direct Instruction programs must be measured and evaluated according to standards of schools, districts, states, and the nation. The strong research base that supports specific Direct Instruction procedures, and the integration of powerful ongoing progress monitoring strategies are not enough. The programs themselves must be evaluated as whole programs to demonstrate their effectiveness.

As we learned in Chapter 2, there is an impressive body of literature indicating that, *when implemented properly*, Direct Instruction programs can be extremely effective. Unfortunately, even this extensive data base cannot guarantee the outcomes from a new implementation of a Direct Instruction program. We know that Direct Instruction *can be* effective but that does not mean that it *will be* effective as it is used in a specific school. The programs must be implemented properly. Proper implementation requires every implementation to take advantage of the assessments that are infused into the programs and to make informed decisions about necessary adjustments. In addition, the broad outcomes of the program must be measured and evaluated to determine the effectiveness of an implementation for enabling a particular group of students to meet the goals set by the school, district, state, and nation. This sort of evaluation requires measures from outside the programs—measures such as nationally normed standardized tests, tests of state standards, and others.

This chapter will describe how to make effective use of assessments that are integrated into Direct Instruction programs and how to plan and implement evaluation of the broad outcomes from Direct Instruction programs.

Assessment, Evaluation, and Validity

In order to see how the data collection and decision making processes relate to Direct Instruction, it will be useful to introduce some technical terms and make several important

distinctions. Therefore, it will be useful to distinguish between assessment and evaluation.

Assessment

Assessment is a process of collecting information to answer a question or to inform a decision. In the current context, we focus on assessments that answer questions and inform decisions about student learning in Direct Instruction programs. Assessment techniques include asking oral questions of students, posing written tasks in daily work, using mastery tests and checkouts, as well as using standardized tests and many others. Thus, assessment includes a very broad spectrum of informal and formal procedures for probing student performance.

Evaluation

Evaluation is a process of using assessment information to make a judgment or decision. In this case, we are concerned with the judgments and decisions related to implementation of Direct Instruction programs. These judgments include relatively narrow decisions about a particular teaching adjustment (such as a correction of an error) as well as very broad decisions about continuing use of a program. Thus, assessment is a process of information gathering that forms a basis for evaluation judgments.

Evaluations ask about effectiveness and suggest changes that should be made as a result. There are two kinds of evaluation—formative and summative. *Formative evaluation* examines short-term outcomes and suggests small-scale adjustments within the program in order to make it more effective. In contrast, *summative evaluation* examines longer-term outcomes and suggests large-scale changes. For example, assessment of daily reading accuracy can be used in formative evaluation to direct error correction and additional practice targeted to weak skills. On the other hand, summative evaluation might use information from standardized test assessments and contribute to decisions about whether to make fundamental changes in the implementation or even to select a different program. Both forms of evaluation have the goal of improving student outcomes; formative evaluation does this by informing decision making within lessons and within the program; summative addresses this need by informing annual or longer-term decisions about major changes in programs.

The first section of this chapter will discuss the role of formative evaluation and the ways in which formative evaluation is built into Direct Instruction programs. The second part of the chapter will describe the role of summative evaluation and ways to carry out summative evaluation of Direct Instruction implementations.

Validity

Teachers and administrators are constantly making educational decisions. These decisions include very short-term decisions about whether to make a correction, repeat a set of items, or move ahead in a program; medium-term decisions about where to place students in programs and the organization of groups; and long-term decisions about which programs to use. At all these levels, making good decisions that maximize student learning depends on having information that is relevant to the decision. Therefore, we must be very concerned about the relevance and quality of the information that we are using to make decisions. This concern about the relevance and quality of information is the topic of *validity*. Validity refers to the adequacy of information as a basis for making a decision (American Educational Research Association, American Psychological Association, & National Association on Measurement in Education, 1999; Messick, 1993). If the information gives us a good basis for making a decision, then we would say that the information is quite valid for that purpose. If the information does not give us a solid basis for this decision making, we would say that the information is less valid for that purpose. Of course, different educa-

tional decisions demand different kinds of information. Hearing a student say an incorrect answer in a group unison response is a valid basis for deciding to make a group correction. But this same information would not be a valid basis for deciding whether the program is working and should be continued next year. Thus, we will be interested in finding information that is valid for each of the different kinds of educational decisions that we must make. The topic of validity will be discussed in greater detail later in this chapter.

Formative Evaluation

Direct Instruction programs include extensive means of making frequent formative evaluations. Each Direct Instruction program includes (a) a placement test, (b) frequent oral and written responses that enable the teacher to evaluate progress during the lesson, and (c) mastery tests that provide useful assessments of students' mastery of objectives. Programs also include specific guidelines for using information from each of these assessments to make instructional decisions. Thus, Direct Instruction integrates an extensive and sophisticated formative evaluation system within each program.

Assessment and Decision Making for Proper Placement

Direct Instruction programs are designed to be effective with relatively homogenous groups of students who are placed at a correct instructional level. That is, each student in a Direct Instruction group should have relatively similar skills and relatively similar instructional needs. In addition, each group should be placed in a program and a lesson that addresses their instructional needs. Initial design of programs is based on an assumption that students will be properly placed at their instructional level. Field tests of programs refine standards for proper placement and help assure that students who are placed at an appropriate level will progress successfully.

Research on high-quality Direct Instruction implementations provides indications of the kinds of outcomes that can be expected given proper placement of students. However, if students are not in an appropriate level of the program or are not on an appropriate lesson, this careful design, field testing, and research may be irrelevant. Even a flawlessly designed instructional program cannot successfully teach students who are placed at a level that does not match their skills.

Incorrect placement, whether students are placed in too high or too low of a level, can cause serious problems of learning and behavior. Students who are placed at a point too high in a program will not have the prerequisite skills necessary to succeed. They will likely make many errors and require frequent corrections. This pattern of frequent errors and corrections makes for very inefficient instruction. It slows the progress of the entire group. It will be very difficult for the teacher to bring the whole group to a mastery criterion on the critical tasks of each lesson, and this lack of mastery may seriously compromise their learning. When students are placed too high and fail to reach mastery on lessons, it is common for them to show little gain in an entire academic year. These problems are deepened by the emotional and behavioral consequences of poor placement. If, even with maximum effort, students experience frequent failure, they may put forth less effort in the future and compound the problem. Students who are misplaced often develop patterns of signal errors such as answering too quietly, slightly after others in the group, or not at all. The need to correct these signal errors further deepens the problems.

Placing students too low in a program also causes serious problems. Students receiving instruction on content that has already been mastered are wasting valuable learning time. In addition, students who are placed too low in programs may complain of boredom and develop behavior problems.

Each Direct Instruction program includes a placement test that helps determine an appropriate starting place where students have prerequisite skills but have not yet mastered the material. Direct Instruction placement tests are very brief focused assessments of specific skills. Placement tests are designed to indicate the program level and lesson that is most likely to be appropriate for the individual student. These tests are validated based on the logical analysis and experience of the program authors. Each of the chapters on particular Direct Instruction programs (Chapters 3–8) has provided information on the proper use of placement tests in that area. Administering placement tests and grouping students accordingly are the first steps in forming appropriate groups. However, grouping based on placement test results is only the first step. After initial placement into groups at particular program levels, placements must be further refined based on students' performance on lessons. Additional information about placement is gained from the students' performance each day. Students who make excessive errors and who regularly require extensive correction and repetition of formats to reach mastery should be considered for re-placement in earlier lessons or in a more basic program. Students who make few errors may be candidates for re-placement in later lessons.

Careful use of placement tests is an important form of assessment that is included in each Direct Instruction program. The programs are designed under the assumption that this kind of assessment takes place, that the results are used to make initial placements, and that placements are revised based on student performance in lessons. If this level of assessment and evaluation is not in place, the potential power of the programs will be severely compromised and students may make little progress.

Assessment and Decision Making Within Lessons

Direct Instruction lessons are designed to provide teachers with frequent, detailed, and relevant assessments of student learning during each lesson. Each group unison oral response provides information on each student's skill level on the particular task being taught. These group responses are probably the most efficient data collection system in all of education. Teachers are made immediately aware of the current performance level of each student on a highly relevant task. In addition to group unison responses, interspersed individual oral responses provide more definite information about the skill level of specific students. These oral responses are highly valid assessments of students' skills for the purpose of making immediate instructional decisions.

Based on what they hear in each response, teachers make several decisions. Typically, for a correct response they confirm the accuracy and perhaps fluency of the answer and move on to the next item in the set. For an error, they diagnose the error, make a particular correction depending on the type of error, and typically repeat the item then return to the beginning of the item set. Depending on the pattern of errors, teachers may depart from these typical responses. This interplay between student and teacher creates a highly dynamic lesson in which the program is adapted to the specific needs of the group. For example, teachers adjust the amount of instruction to the needs of the group as they correct errors in individual and group responses. If the students demonstrate the need for more instruction, teachers provide it through corrections or by repeating an entire instructional task. The amount of practice is also adjusted to the needs of the group. If students demonstrate they need more practice (e.g., through errors and hesitant responses), teachers provide this practice by repeating a set of items until student responses are firm.

Most Direct Instruction programs also include written work. If teachers circulate and check answers as students work, they can obtain very immediate feedback on student learning and make corrections during the lesson. If teachers wait until after the lesson to check written work, they can obtain very comprehensive information about the performance of their students, but they cannot remediate until the next day. This direct assessment of students' oral and written responses provides the information for powerful immediate decision making within lessons—a key element in proper implementation of Direct Instruction programs. Direct Instruction programs can be considered well implemented only if this active assessment and decision making process is in place.

As we have emphasized previously, effective decision making depends on valid information. But effective decision making also depends on teachers having appropriate guidance in identifying effective adjustments based on that information. Direct Instruction programs are unusual in the specificity of the guidance that they give teachers in the process of instructional decision making based on student performance within lessons. Correction routines are specified for various kinds of errors in each program. The link between the specific error pattern and the appropriate correction is clear and explicit. In addition, Direct Instruction programs specify particular critical points in the program at which teachers should provide practice until students reach a mastery criterion. Thus, these programs provide teachers with the necessary information, and they also provide guidance in responding to the information.

In many ways, all other assessments and all other levels of evaluation are dependent upon this foundation. If teachers use the information afforded by oral and written responses within Direct Instruction lessons to make effective adjustments, then the higher levels of assessment are most likely to show positive results. However, if this foundational level of data is not used effectively, higher levels of assessment are unlikely to show successful results.

Assessment and Decision Making With Mastery Tests

Direct Instruction programs include regular mastery tests. These tests systematically represent all the critical skills that are being taught in a particular segment of a program. These mastery tests provide an additional check on student learning. They check whether the daily instruction and decision making has been effective for assuring that the students have mastered the skills. Mastery tests provide critical information that must be used thoughtfully to adjust teaching if Direct Instruction programs are to be effective. For example, *Reading Mastery Plus Level 3* includes checkouts on oral reading rate and accuracy every 5 lessons and a written mastery test on the content, skills, and vocabulary every 10 lessons. Checkouts and mastery tests include specific criteria for adequate performance. Mastery tests and checkouts are accompanied by specific guidelines for decision making and providing remedies for students who score below criterion. For example, if students do not achieve the rate and accuracy standards for a checkout, the teacher is advised to increase the amount of oral reading practice through several variations on the repeated reading procedures (Engelmann & Hanner, 2002). If students score below standards on the mastery test, teachers analyze student error patterns including which items each student missed and how many students had problems in each area. Based on this information, teachers may organize additional practice for individuals or the group. The teacher's guide for each level of each program includes specific remedial procedures for students who have common error patterns. In addition, many programs (such as the *Reading Mastery Plus* series) include firming tables along with each mastery test. These firming tables list the lesson in which each item in

the mastery test was introduced. If several students miss a particular item, teachers can consult the firming table to find the lesson that includes the specific format that should be repeated. Teachers then reteach and provide practice on this specific item or concept. The programs provide very elaborate support for teacher decision making.

In addition to their primary use for adjusting instruction to student needs, mastery tests are also a useful source of information for supervision and more formal evaluations. If mastery tests are used effectively, they can serve as key indicators for supervisors monitoring whether the program is being implemented effectively. Mastery tests can be used to pinpoint groups quickly who may need to be provided with additional support and monitored more closely. (See Chapter 10 for more information on supervision and coaching.) Mastery test results are also a valuable component of annual evaluation. When evaluators consider the broad outcomes from a Direct Instruction program, they are often interested in how these broad outcomes are related to the specific goals for the program. Mastery tests are excellent indicators of the degree to which students achieved the specific goals of the program.

Summative Evaluation

Summative evaluation addresses the broad questions of whether an implementation of Direct Instruction has succeeded in enabling students to achieve important outcomes including the outcome goals of the program as well as outcomes identified as important by the school, district, state, and nation. Summative evaluation requires careful thought about many issues including (a) what tests are appropriate, (b) what kinds of scores most clearly show the results, (c) what comparisons are most relevant, and (d) how to summarize and present results.

Measurement and Validity

An implementation of Direct Instruction has many outcomes. There are outcomes that involve students' basic skills, complex skills, specific knowledge, attitudes toward specific content areas and about school in general, and others. Evaluators face a set of critical decisions about which outcomes should be measured and what assessment techniques should be used to measure them. Any method of assessment will focus on some outcomes at the expense of others. The choice of which outcomes to target determines much of what the evaluation will show, what questions will be answered clearly, and what questions will not be addressed. The quality of the measure determines whether we have clear information that can contribute to good decisions about improving or terminating programs, or misleading information that may contribute to poor decisions.

Measurement Validity

The topic of measurement validity has been an important one in education, psychology, and other fields for a long time. Many careful researchers have dedicated substantial thought to the issue of how to judge how well a measure informs a decision. The study of measurement validity can be very complex and intimidating because of the many technical and abstract terms and mathematical analyses, but the basic issues of measurement validity are simple, direct, and critically important for understanding the effectiveness of an educational program. Measurement validity is a judgment of how well a measure informs our decisions (American Educational Research Association et al., 1999; Messick, 1993). Measurement validity is important when we consider any source of information that might help us make decisions. It is important whether we are considering curriculum-based measurements, standardized tests, portfolio assessments, teacher judgments, or any other source of information. In this chapter, we are concerned with making decisions about the

effectiveness of Direct Instruction implementations and then taking action to adjust these implementations, expand them, or end them. So in this context, a valid measure is one that provides excellent information for making our decisions and taking action within Direct Instruction implementations. An invalid measure is one that provides poor information for making these decisions. Whether a particular measure is a valid measure depends on the specific decisions we are making. For example, oral reading rate tends to be a highly valid measure for making decisions about a student's reading skill (Deno, Mirkin, & Chaing, 1982; Fuchs, Fuchs, & Maxwell, 1988), but it is, of course, a highly invalid measure for decisions regarding students' math skills. If we are concerned about finding out how well a program works, we *must* be concerned about measurement validity.

The first and most important step in thinking about validity is to clarify exactly what we want to know and what decisions we want to make. We cannot begin talking about validity until this is established. Trying to judge validity without a very clear definition of what we want to know and what we want to do as a result would be like judging the accuracy of an archer without identifying the target at which he is aiming. The concept of accuracy makes no sense without reference to a target. Similarly, the concept of measurement validity makes no sense without reference to specific information that we want and specific uses of that information.

Once we have established what we want to know and what we want to do as a result, then we can begin evaluating how well a given measure might help us accomplish these things—that is, its validity. We would want to know what evidence exists about whether a particular measure will do the job. There are many sources of information on how well a measure will provide information and inform a decision. The most obvious source of information is careful examination of the measure and asking whether the particular items, style of questions, and so on correspond with what we want to know. This source of evidence is termed *evidence based on test content* (American Educational Research Association et al., 1999). Evidence based on test content is concerned with the alignment of the test with the information that we want to know.

Evidence based on content will be very important as we consider the types of tests that might be useful in evaluation. This content evidence is not limited to a superficial analysis of what the question may appear to test. We have to examine the various influences on what may make an item easy or difficult. For example, a reading comprehension item may be quite difficult for students who do not have specific background knowledge, but may be very easy for students who have that knowledge. Test items that are very unlike anything that students have been taught may be highly influenced by general intelligence and less influenced by the specific skills that we are attempting to assess. Many multiple-choice items can be easier for students who have good test-taking skills. In each of these examples, an item may have been designed as a test of a specific skill, but a careful examination may suggest that it may be substantially influenced by other factors. Figure 1 illustrates the relationship between validity and various influences on test scores. This figure indicates that validity is highest when test scores are strongly influenced by the target of testing (e.g., math problem solving, reading fluency, written composition skills) and minimally influenced by other factors such as other skills learned in school or outside of school, test taking skills, and intelligence. When these other factors have a stronger influence on test scores, the test is less valid. That is, when test scores are influenced by these extraneous factors, the scores are not as good a basis for decision making.

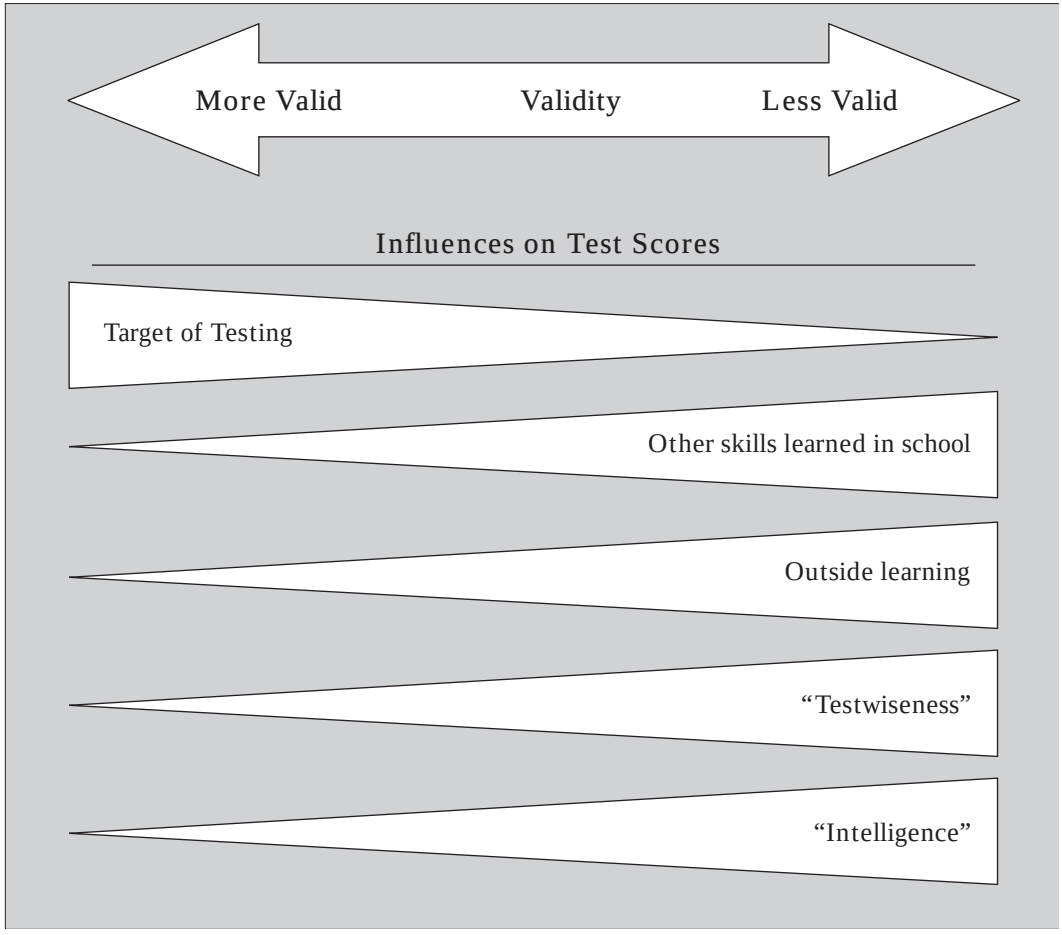
A second source of evidence about validity is how well test scores correlate with scores from other tests. For example, if oral reading rate is

highly correlated with scores from a standardized test of reading comprehension, this correlation is evidence to suggest that oral reading rate may be a fairly valid measure of reading comprehension. So we may be concerned with whether one test correlates with other tests that are accepted as valid tests for the decisions you are making. These correlations are important because they are another way of judging whether the test is strongly influenced by the target. If test scores correlate with results from other tests of the same target, this is some evidence that they are being influenced by that target.

As we mentioned above, the topic of measurement validity is very technical. There is a lot more to know about validity. (For a more complete introduction to measurement validity see Gall, Gall, & Borg, 2003, or Cohen & Swerdlik, 2001. If you want to understand these issues more deeply, study American

Figure 1

Validity and various influences on test scores.



Educational Research Association et al., 1999, and Messick, 1993.) But it is important to avoid getting distracted by the technical aspects and losing sight of the main question addressed by validity—whether an assessment gives us the information that we need to make decisions. At this introductory level, it is important to focus on evidence by carefully examining the test's content and evidence from correspondence with other tests.

Evaluation Questions and Measures

The critical decisions regarding what kinds of outcomes to target depend on the purposes and questions that drive the evaluation. In this chapter, we focus on academic outcomes. Broadly, there are three kinds of concerns that may determine our academic targets. First, we may be concerned with how well the implementation of Direct Instruction programs achieves the objectives of the programs. In this case, we are interested in measuring the specific skills and generalizations that are taught in the program. Second, we may be concerned with how well the implementation achieves goals such as building reading comprehension or math problem-solving skills. In this case, our interest is not defined by what was taught in the program; instead, our interest is in these areas of achievement. The main difference between this concern and the earlier one is that here "reading comprehension" may include aspects that were not taught in the Direct Instruction program and there may be topics in the Direct Instruction program that are not included within our definition of "reading comprehension." Third, we may be concerned with how well the implementation achieves state or national standards. In this case, we are interested in a set of skills and complex performance that are not defined by the Direct Instruction program, but rather are defined by a set of standards developed on a state or national level. These standards may or may not correspond closely with the skills and complex performances that are taught in a particular Direct Instruction program. These three concerns that drive evaluation decisions are each distinct; however, they do overlap. The question of how much each of these concerns overlaps with the others and how much each overlaps with the content for the Direct Instruction program that is being evaluated is very important. It is a question that we will return to later in the chapter.

Mastery of program objectives. Our evaluation concerns may include questions about whether students actually achieve mastery of the objectives of the Direct Instruction program. In this case, we would want to find a measure that assesses exactly what the program attempts to teach. The targets that we want to measure are the intended outcomes of the Direct Instruction program. We would judge the validity of a measure by how well it is aligned with the objectives and activities of the program. For example, at the end of *Reading Mastery I*, we would be interested in whether students can decode the words taught in the program and if other new words can correctly be decoded by applying the strategies taught in the program. We would also be interested in whether students can correctly answer the types of questions that appear in the program.

When we evaluate evidence from test content, we would examine the alignment between the measure and the program. Evaluating alignment includes three basic steps. First, we would list the objectives and main activities of the program. Second, we would list the topics and specific tasks that make up the test. Finally, we would compare the two lists. Where the two lists correspond, the measure is assessing relevant content. Where the Direct Instruction program teaches content but that content is not reflected on the measure, the measure is under-representing the content of the program. That is, there are important outcomes from the programs that are not reflected in the assessment. Where the measure assesses content that is not targeted by the program, the measure reflects irrelevant issues. The more that the

measure assesses relevant content, the more we would judge it to be valid. The more that the measure under-represents the program and includes irrelevance, the more we would judge it to be invalid.

Our evaluation of the content evidence of validity is concerned with the relationship between the program's intended outcomes and the content of the measure. We might consider using the mastery tests and checkouts from the end of the program to assess these outcomes. This use of mastery tests would make sense because the program authors designed the tests for a very similar purpose. Nonetheless, we would examine these measures in comparison to the program's objectives to assure that we are measuring all the important objectives. A mastery test or checkout may or may not correspond with our evaluation needs. For example, it may focus on skills that are taught near the end of the program and neglect skills that were taught earlier. If we were considering using an existing test that was developed outside of Direct Instruction, we would have to be very careful in our analysis of how well the test content corresponds to the program content. In addition to this content evidence, we would consider evidence of how scores on the test compare to other relevant scores. The scores most relevant to our target (the intended outcomes from the program) would include scores on mastery tests, checkouts, and daily work. If we were examining a possible new measure of program outcomes, that measure would be judged most valid if its results corresponded closely with student performance on these other measures. This correspondence would be evidence that scores from the measure would be a good basis for making decisions regarding student outcomes on the objectives of the program.

Evaluation of broad skill areas. A second concern that could drive our evaluation is how well a Direct Instruction implementation teaches a particular outcome such as math problem solving or writing. In this case, the target of evaluation is not defined by the specific content or goals of the Direct Instruction program, but rather by an outcome that may or may not correspond with the program. In order to focus our judgment of validity of a test for this purpose, we must begin with a clear definition of the skills to be targeted. We would seek a definition from the literature that carefully describes and defines the content area. For example, the National Reading Panel recently examined and analyzed the content area of reading and described the critical components of phonological skills, decoding, fluency, vocabulary, and comprehension (National Reading Panel, 2000). The National Reading Panel's report would be a good starting point for defining these areas. We would seek more detailed analyses of each of these broad areas. So we would examine the area of decoding and describe the most important decoding skills. This process would result in a list of skills and tasks that embody what we mean by "reading."

Once we have defined exactly what we want to measure, we proceed in the same way that we did above. We would examine the content of a test and compare it to the components of our target area. We would identify areas in which the measure assesses relevant content; we would look for aspects of the target content that are not assessed by the measure, that is, content under-representation; and we would look for aspects of the measure that are not part of the target content, that is, content irrelevant aspects. We would judge measures most relevant that included the most content relevance and minimized content under-representation and irrelevance. We would also be interested in how scores from the measure correspond to scores from other measures that reflect this target. For example, if we were measuring reading outcomes, we would be interested in how the scores from the measure correspond with results from other measures that we believe measure reading outcomes. If there was a high level of correspondence, then we would consider the measure to be more

valid. That is, we would have more evidence that indicates the measure would be useful in decision making with respect to reading.

Evaluation of state and national standards. The third kind of concern that could drive an evaluation of a Direct Instruction implementation is whether the implementation enables students to meet state or national standards. In this case, our judgment of validity of a measure would be based on how well it corresponds to these standards. Content evidence would come from examining the relationship between the items in the standards and the content of the measure. The measure would be judged most valid if it had a high degree of overlap with the standards (content relevance) and low amounts of under-representation and irrelevance. In addition, if there are existing measures that we believe are highly related to the standards, we would be interested in how closely the results of these measures correspond.

Summary. The most important points in this discussion are that judgments about validity of a measure are based on how well it informs decisions. Evaluations of Direct Instruction are usually concerned with decisions about adjusting the way the programs are being implemented, expanding the implementation, or reducing it. To a large extent, these decisions depend on how well the implementation is achieving academic goals for students. The validity of our measures is a matter of how well the measures reflect the important academic outcomes. The problem is that a measure is influenced by many factors—the target of measurement is just one of them. So in evaluating validity, we attempt to find and evaluate evidence about the influence of the target on test scores as compared to the influence of other factors such as incidental learning outside of school, test taking skills, and so on. Tests that most clearly reflect the target of measurement and are not overly affected by other factors give us the best basis for decision making and therefore are considered to be most valid.

Scores for Describing Results

Many different kinds of scores are used to describe test results. In this section we will describe some of the more common kinds of scores and their meaning.

Raw Scores

The most basic and simple way to express a test result is the raw score. Raw scores give the number of items correct or the rating given in an assessment. However, knowing students' raw scores does not tell us very much about their skill. For example, knowing that a student got 10 items correct on a test is not very useful by itself. Was that 10 correct out of 10 items or 10 correct out of 50 items? Was that the highest score in the class or was it the lowest? Is that score above our criteria for mastery of the material or does it imply a need for further teaching? In order to make any sense out of a raw score, we need to provide some kind of context or comparison.

Percentage Scores

One important context for our score of 10 is the highest possible score on the test. Percentage scores combine the raw score with the number possible. They express student performance relative to the best possible performance. So, rather than reporting that a student got 10 items correct, we would say that he got 90% correct—that is, 90% of the possible points. Percentage scores are often (but not always) more meaningful than the raw score by itself.

Rate

In some skill areas, the speed of responses is very important. For example, rapid decoding is a critical goal of reading instruction, and saying basic math facts quickly is an important outcome of teaching math. Reading 10 words correctly in a minute is very different from reading 10 words correctly in 6 seconds. When rate of performance is important, the raw score can be made more meaningful by converting it to a

rate of performance. In order to be consistent, we usually express rates as number of responses per minute. Thus, rather than reporting 10 correct words in 6 seconds, we would convert that to 100 correct words per minute. Expressing the score as 100 correct words per minute tells us about his speed of response. The student's rate of performance is not reflected in raw scores and percentage scores.

Criterion-Referenced Scores

Another important way of giving meaning to test results is to describe them in comparison to criteria for excellence, mastery, acceptable performance, or some other category. Criterion-referenced scores describe student performance in comparison to standards for performance. For example, a teacher may decide that 95% correct on a particular test constitutes "excellent performance" and is awarded an "A." Thus, rather than reporting that a student achieved 10 items or 90% correct, she may report that the student's performance was "excellent" and was "A" work. This score of "excellent" or "A" is a criterion-referenced score because it describes the performance in relation to a particular criterion. In this case, the teacher set the criterion. Many tests, including most state-developed tests, yield criterion-referenced scores. On these tests, a panel of experts in the particular area of testing sets the criteria. The criteria based on state standards are often described as "mastery," "near mastery," and so on. These criteria represent a judgment about the levels of performance that *should* be achieved by students at a given grade level. Criterion-based scores are, of course, only as meaningful as the criteria that have been set. A panel could set a low standard, and the test results would reveal that the vast majority of students meet "grade-level standards," or the panel could set a very high standard and the results would indicate that few students meet "grade-level standards." Nonetheless, if particular standards are meaningful to important decision makers, they become important to teachers and evaluators. However, it is very important to remember that criterion standards are someone's judgment of what is good enough. Criterion-referenced scores are an attempt to make student performance more understandable by giving them labels such as "A," "excellent," or "mastery," that are more meaningful than "10 items correct." The criterion is the standard for giving the label. Even if we are not very familiar with the test, we can understand something about the performance if it is described as "mastery."

Norm-Referenced Scores

Norm-referenced scores describe student performance by comparing it to the performance of other students. For example, we might say that a student's performance was "the best in the class" or "typical for third graders." These statements give meaning to performance by describing it in comparison to the rest of the class or "typical third graders." Norm-referenced scores get their meaning by describing the student performance in relation to how other students performed on the same test and under similar conditions. Norm-referenced scores are based on a comparison of the student's score to a large group of other students who have taken the test—the norm group.

Grade and Age Equivalent Scores

Grade equivalent scores are norm-referenced scores that describe student performance, not as a raw number of items correct, but as the grade at which this number of items would be the average. So for our example student who got 10 items correct, we would find the grade level at which 10 items correct was the average score on this test, and report that grade level as the grade equivalent (GE) score. If students in the 3rd month of second grade had an average raw score of 10 on this test, then the grade equivalent of 10 is 3.2. We could also describe a student's raw score as an age equivalent—the age at which the student's performance would be average. Rather than saying 10 items correct, we could report that the stu-

dent got the raw score that is average for students who are 8 years and 5 months old (AE = 8.5). Age and grade equivalents are an attempt to make test scores more understandable by describing how a particular student's performance compares to students of various ages or grades. Age and grade equivalent scores are often misinterpreted as indicating the "level of work" that the student is doing. This interpretation is incorrect. For example, a second grader could have very strong addition/subtraction facts. On a test of these facts, she may do as well as the average 10th-grade student. This, of course, does not mean that the second grader is doing 10th-grade math. She is doing math facts as well as most 10th graders, but probably does not know nearly as much as most 10th graders about many math topics.

Percentiles

Percentiles are norm-referenced scores that express student performance compared to that of other grade-level students. (Some tests include norms based on ages. In that case, percentiles may be defined in comparison to other students of the same age as the target student.) A percentile is the percentage of the norm group in a particular student's grade who scored the same as or below that student. If a third grader had a raw score of 10, her percentile score would be the percentage of third graders in the norm group who scored 10 or less. If 56% of these comparable students scored 10 or less, then our student's score is in the 56th percentile. The percentile score may be more meaningful than a raw score because it indicates where the student's performance ranks relative to comparable students (i.e., those who are in the same grade). Even if we are not very familiar with the test, we can interpret a score of the 56th percentile by saying that it is fairly typical of students in her grade and that it is neither extremely high nor extremely low. We can interpret a 90th percentile as quite a high score for students in that grade. Percentile scores are between 1 and 99.

Technical Problems With Grade/Age Equivalent and Percentile Scores

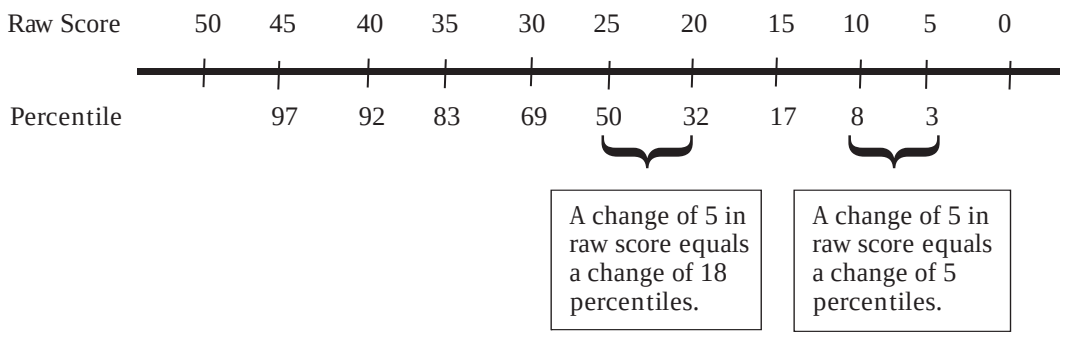
Grade/age equivalent scores and percentiles have some technical problems that are important to understand in evaluating an implementation of Direct Instruction. The problem refers to a basic assumption that we make when we add numbers. When we add or subtract, we treat the units added or subtracted as if they are all the same. This assumption of equivalent units is why "adding apples and oranges" is not considered proper. If we say that 3 + 2 = 5, we assume that each of the 3 things and the 2 things are, in some sense, the same. However, statisticians question whether percentiles are equal units. Figure 2 shows how raw scores relate to the percentile scale on an imaginary test. (On other tests, the raw scores would, of course, be different, but the unequal relations between raw scores and percentiles would almost always be present.) On a test, the difference between a raw score of 20 and 25 (a difference of 5 points) could move a percentile rank 18 points. However, the difference between raw scores of 5 and 10 (also a difference of 5 items) could move the percentile rank only 5 points. In this sense, percentiles are not equal sized units. From this perspective, percentiles (and age and grade equivalent scores that have the same problem) should not be added, subtracted, or averaged. We may ask if this is a technical problem that does not really matter in the "real world." There are certainly some situations in which averaging percentiles or grade equivalents would give us a very misleading result. Whether our particular result will be substantially misleading if we ignore this problem depends on many factors.

Normal Curve Equivalent (NCE) Scores

Normal curve equivalent (NCE) scores are similar to percentiles in that they range between 1

Figure 2

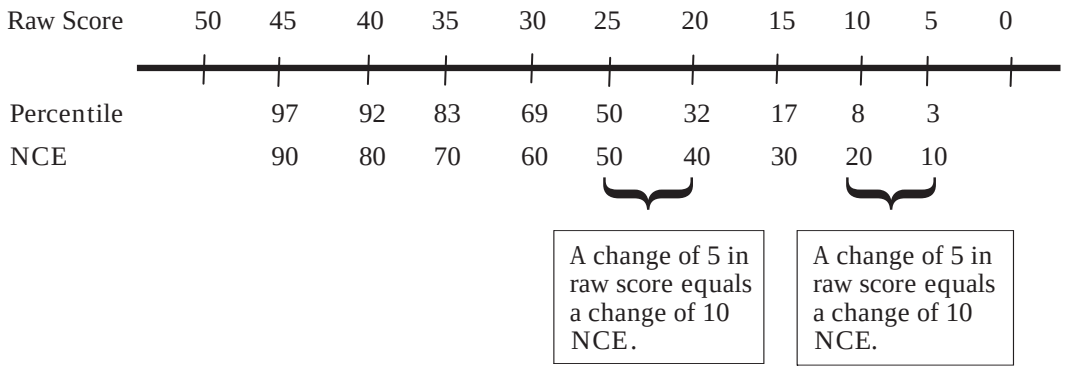
Comparison of raw scores and percentile scores.



and 99 with a mean of 50. However, the scale is organized into equal size units. Figure 3 shows how the NCE scale relates to raw scores and percentiles. If the difference between raw scores of 25 and 30 corresponds with 10 NCE units, then any other difference of 5 raw score points will also correspond with 10 NCE units. In this sense, NCE units are "equal interval." Therefore, NCE scores may be added, subtracted, averaged, and subjected to other manipulations without the problems inherent in percentiles. The main limitation of NCEs is that they do not have an obvious interpretation for most people. Knowing that a student's score is in the 35th NCE is not very meaningful. It does *NOT* indicate that the score is above or equal to 35% of other scores at that grade level. Therefore, best use of NCE is for mathematical computation. After mathematical computations are completed, it is usually best to convert NCE scores into percentiles.

Figure 3

Comparison of raw scores, percentile scores, and NCE scores.



Standard Scores

Standard scores are another kind of norm-referenced score. Standard scores describe the student's performance in terms of number of standard deviations away from the mean. Standard deviation is a measure of the spread of a group of scores. If scores in a set are all close to the same, the standard deviation is small; if scores in a set are spread over a wide range, the standard deviation is large. There are many varieties of standard scores. The most common type of standard score is the kind used for IQ tests. In this type of standard score, the average score for students at a given grade level is 100 and a standard deviation is 15. Thus, a student who scored exactly at the mean for her grade level would have a standard score of 100. A student whose raw score was one standard deviation above the mean would have a standard score that of the mean (100) plus one standard deviation (15), that is, 115. A student whose raw score was one standard deviation below the mean would have a standard score of the mean (100) minus one standard deviation (15), or 85. As a further example, we can return to our student with a raw score of 10. If the test had a mean of 6 and a standard deviation of 2, her raw score of 10 would be two standard deviations above the mean of 6. Thus, her standard score would be the mean of standard scores (100) plus two standard deviations ($2 \times 15 = 30$) or 130.

Standard scores are similar to NCE scores in that they can be manipulated by adding, subtracting, and so on. Also, like NCE scores, standard scores can be difficult to understand. Parents, teachers, and administrators who do not have extensive experience with standard scores may not readily understand the meaning of a standard score of 120.

Limitations of Interpreting Scores

Of course, no score can tell us *why* students have strong, typical, or weak skills in a particular area. A score in the 3rd percentile on a reading comprehension test could be a result of the instructional materials that were used, how well the programs were implemented, or most realistically, a combination of many factors. In order to begin to understand the effects of an implementation of Direct Instruction on students, we have to go beyond the scores themselves and consider the design of evaluation.

Evaluation Designs

Thus far, we have focused on ways to describe the results of Direct Instruction implementations. But it is very difficult to interpret these results unless we have some kind of comparison. For example, suppose that at the end of second grade, students in a Direct Instruction implementation scored at the 55th percentile on a test of math problem solving. Is that good? Is it a sign of success or a sign of failure? In order to answer these questions, we need some kind of comparison. If we know that at the beginning of the year, the students scored at the 23rd percentile, then finishing in the 55th percentile is a substantial victory. If we know that another class with similar students that did not use Direct Instruction scored at the 70th percentile, then this is a major problem. Evaluation design is concerned with arranging the evaluation to create important and meaningful comparisons that allow for strong conclusions about the success (or lack of success) of a program.

Three main kinds of comparisons are used in evaluation of Direct Instruction implementations. First, outcomes may be compared to the performance of the same students at earlier points in time—a *one-group pretest–posttest design*. Second, outcomes may be compared to those of a comparable group of students who did not experience Direct Instruction—a *treatment–comparison group design*. Third, the two methods of comparison mentioned above (pretest–posttest and treatment–comparison group) can be combined by using a pretest and posttest for two groups, one that experiences Direct Instruction

and one that experiences some alternative—*a pretest–posttest–comparison group design*. There is no single evaluation design that will unerringly reveal the truth. Each design is intended to help identify the effects of the intervention. No matter which design is used, we must always carefully consider the question of whether something other than the intervention could have caused the outcomes. All designs must be used thoughtfully with careful attention to possible sources of bias. The following section describes each of these designs, briefly discusses some of the common problems with each, and suggests important considerations for minimizing the problems.

Pretest–Posttest Design

In the basic one-group pretest–posttest design, the students take a pretest, experience Direct Instruction, and then take a posttest. We evaluate learning as the gain from pretest to posttest (see Figure 4). The pretest and posttest may have been given for reasons outside the evaluation. For example, if students are tested each spring, the tests from the year before the implementation may provide a convenient pretest, and the tests at the end of the year of implementation may be a useful posttest. Of course, evaluators must consider whether these tests are valid for the purposes of the evaluation.

Students' skills tend to improve as the students get older. Thus, if we see a gain in skill from the beginning of the year to the end of the year, we may not be able to attribute that change to Direct Instruction. Perhaps the change between the pretest and the posttest was simply the kind of growth that we expect over the course of a school year. One solution to this problem is to use norm-based scores such as percentiles, NCE scores, or standard scores. These scores reflect the student's rank relative to other students. If students gain skills at the same rate as typical students in the norm group, their percentile rank, NCE scores, and standard scores would not change. That is, if a student scores in the 36th percentile at the beginning of the year, then makes normal growth during the year, we would expect her to score at approximately the 36th percentile at the end of the year. If a student moves up from the 36th percentile at the beginning of the year to the 60th percentile at the end of the year, this change represents more growth than was typical in the norm group.

In this way, percentile, NCE, and standard scores account for typical growth in skills. These scores change only if the student changes her standing in comparison to the norm group. This property makes percentile, NCE, and standard scores particularly useful in evaluation. However, this use of scores depends on the assumption that the students participating in the evaluation are similar to those in the norm group and would be expected to make the same growth. This assumption, however, may not be warranted. Students in the implementation could differ from those in the norm group in important

Figure 4

Pretest–posttest design.

Pretest(s) (may include several pretests)	Implementation of Direct Instruction	Posttest(s) (May include several posttests)

ways. For example, there may be other things happening in the school or neighborhood that give the students an advantage or disadvantage. Students in the Direct Instruction implementation may be more socially advantaged than the norm group and therefore may be expected to show greater growth. This problem of not knowing whether the norm group is sufficiently similar to the Direct Instruction group is why one-group pretest–posttest designs are somewhat weak and are often combined with treatment–comparison group designs to increase their strength (see discussion of these designs below). However, when a comparison group is not feasible, a one-group pretest–posttest design can be very useful. One way of strengthening this design is to examine test results from several years before the implementation. For example, if *Corrective Reading Decoding* is implemented with a group of sixth-grade students, it would be more powerful to compare their sixth-grade outcomes to their test results in third, fourth, and fifth grades. With this information, we can compare sixth-grade performance to the pattern of performance across 3 previous years.

Treatment–Comparison Group Design

In a treatment–comparison group design, posttest scores from students who participated in the implementation of Direct Instruction are compared to posttest scores from a similar group of students who did not experience Direct Instruction (see Figure 5). This kind of comparison can be made on any scale. A teacher could compare a Direct Instruction group and a non-Direct Instruction group within his own classroom. A school could compare a Direct Instruction classroom to a non-Direct Instruction classroom. At the district level, a school that implements Direct Instruction could be compared to one that does not use the programs. Students from a single Direct Instruction classroom are usually compared to students in a single non-Direct Instruction classroom, and students in a school in which Direct Instruction is used schoolwide are usually compared to those at a single non-Direct Instruction school. However, if appropriate scores are available, students in a Direct Instruction classroom could be compared to those in several different non-Direct Instruction classrooms. For example, if a school has four third-grade classrooms and one of these classrooms is implementing Direct Instruction, the results from the Direct Instruction classroom could be compared to each of the three non-Direct Instruction classrooms. This kind of comparison is useful because it creates a contrast between the Direct Instruction classroom and several different non-Direct Instruction classrooms. We can

Figure 5

Treatment–comparison group design.

Information on comparability of groups	Implementation of Direct Instruction	Posttest
	Alternative to Direct Instruction	Posttest

see how the outcomes from the Direct Instruction implementation compare to those from a range of non-Direct Instruction possibilities. This kind of comparison can also be made at the school and district levels. The results from a school that is implementing Direct Instruction could be compared to all (or several) other schools in the district. In some situations, classroom teachers and other personnel do not want to be singled out for a possibly embarrassing comparison. In this case, it may be useful to compare a Direct Instruction classroom (or school) to an average of several other classrooms (or schools). This arrangement gives individuals in the non-Direct Instruction groups some greater anonymity.

Since all groups in the evaluation are tested at approximately the same time (posttest), the problem of typical growth across a year that is so difficult in one-group pretest–posttest designs is not a problem in treatment–comparison group designs. Instead, the major concern in this kind of design is whether the treatment and comparison groups were comparable before the intervention began. If one group had stronger skills or a better ability to learn at the beginning, then we cannot attribute differences at the end to the Direct Instruction implementation. Thus, we must seek evidence about whether the groups were comparable at the beginning. This evidence may include information such as their performance on tests, the percentage of students who qualify for free or reduced-cost lunch, the percentage of students for whom English is a second language, and so on. Test scores are particularly useful. If, at the beginning of the study, students are similar on academic skills that are related to the target area, this similarity provides some evidence that, in the absence of an intervention, they are likely to be similar on the specific skill that is to be measured at the end of the study. (If we have pretest scores on the same test that will be used for a posttest, then we can use the pretest–posttest–comparison group design that is described in the next section.)

Groups must also be comparable in terms of which students are tested. If lower (or higher) performing students are eliminated from either group, then the comparability of the groups is reduced. Students who are likely to present special instructional challenges are sometimes subtly excluded due to absences during testing, because English is not their native language, because they are in remedial programs, and similar reasons. The critical point is that the rules for inclusion or exclusion from groups must be the same for Direct Instruction and comparison groups. Therefore, it is very helpful to have specific written rules defining which students are to be included in the evaluation.

In addition to differences in student skills, we must also be attentive to differences in the way the groups are treated and tests are administered. For the group outcomes to be comparable, tests must be delivered in comparable ways. If one group had more advantageous testing conditions, of course, differences in outcomes could not be attributed to Direct Instruction.

Pretest–Posttest–Comparison Group Design

We generally get a stronger evaluation if we are able to combine the two designs described above. We identify two (or more) groups, pretest them, implement Direct Instruction with one (or more) of the groups, then posttest all the groups (see Figure 6). The pretest provides a reasonable basis for deciding whether the groups were comparable at the outset. The comparison group(s) provides an alternative to which the Direct Instruction group can be compared. However, even with these strong features, there are many issues that demand careful thought if the evaluation is to be useful.

First, even though the pretest gives us strong evidence about the initial comparability of the groups on an important skill, we must be aware that the groups could differ on other

academic skills and the ease with which they learn the new material. Also, if some students who are pretested are not posttested, bias may be introduced. For example, if one group loses students who are predominantly lower performers, this difference in attrition would create an illusion of improvement in group performance. On the other hand, if one group loses students who are predominantly higher performers it would create a false worsening of the group's performance. Second, if tests are not administered similarly, results may not be comparable. Therefore, it is important to take steps to assure that test administration is as similar as possible for all groups and between the pretests and the posttests.

The field of evaluation design (or experimental design) is very large and complex. We have provided a brief introduction to designs that are most frequently useful for small-scale evaluations. We have not discussed the issue of random assignment, which is necessary for true "experimental designs." For more complete information please consult Martella, Nelson, and Marchand-Martella (1999), or for a very detailed in-depth treatment see Shadish, Cook, and Campbell (2002).

Interpreting Results

Once results have been collected from an evaluation, the results must be displayed in a way that clarifies the important comparisons. The field of data analysis and statistics is very large and complex. Of course, it is much too involved to discuss here. However, graphs of results can be much simpler to produce and understand, and they can provide powerful insight into the results from an evaluation. In this section, we will describe several forms of graphs that are often useful in evaluations of Direct Instruction implementations.

Results from a pretest–posttest design can be displayed with either a column graph, as shown in Figure 7, Panel A, or a line graph, as shown in Panel B. The column graph represents the pretest and posttest scores by the height of a column. The line graph shows the same information in the height of the dots. If the results are measured in percentile, NCE, or standard scores, then a posttest that is higher than the pretest represents student growth above that which would be expected due to normal learning over a year.

Figure 6

Pretest–posttest–comparison group design.

DI Group	Pretest(s)	Implementation of Direct Instruction	Posttest(s)
Comparison Group 1	Pretest(s)	Alternative to Direct Instruction	Posttest(s)
Comparison Group 2 (optional)	Pretest(s)	Alternative to Direct Instruction	Posttest(s)

If we have several sets of scores from before the implementation, these can be shown in simple variations on the column graph (Figure 8, Panel A) or the line graph (Figure 8, Panel B). In the column graph, all the columns that represent performance are filled with the same color to emphasize that they are all related, and the column that represents performance after the implementation is shown in a different color to indicate that this column is distinct. On the line graph, the dot that represents performance after the implementation is not connected to the others; this gap in the line visually represents the fact that this dot is different.

We are not limited to showing a single point for each group. If we want to indicate how the implementation impacted students with different levels of academic skills, we can show performance at several levels. One convenient way to do this is to count the number of students who score at or above the 25th percentile. In the norm group, 75% of the students score at this level. If an implementation of Direct Instruction results in 90% of the students scoring above the 25th percentile, this pattern would indicate that they were improving the performance of their very low performing students. Their very low performers score higher than would be expected based on the test norms. If we see more students scoring at or above the 25th percentile on posttests than we did on pretests, this pattern would indicate that very low performers scored higher after the intervention than they did before it.

We can display similar results for the percentage of students who score above the 50th percentile or the 75th percentile. The percentage of students who exceed the 50th percentile reflects performance of students who would normally be expected to score below average on the test. The percentage of students who score above the 75th percentile indicates performance of middle to upper level performers. Of course, we can choose any levels of performance to display on this kind of graph. If we were particularly interested in how the Direct Instruction implementation impacts high performing students we could display the percentage of students who score above the 95th percentile. If the percentage of students who score at this level increases from pretest to posttest, this pattern would indicate that higher performers are scoring even higher after the intervention. Understanding and interpreting this kind of graph requires some thought and study, but it is worth the effort as this kind of display can give insight into the important issue of the effects of Direct Instruction on students at various performance levels. Figure 9, Panel A shows this kind of graph for pretest and posttest scores. Panel B

Figure 7

Graphics for simple pretest–posttest design.

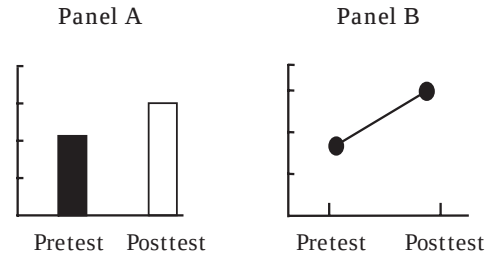
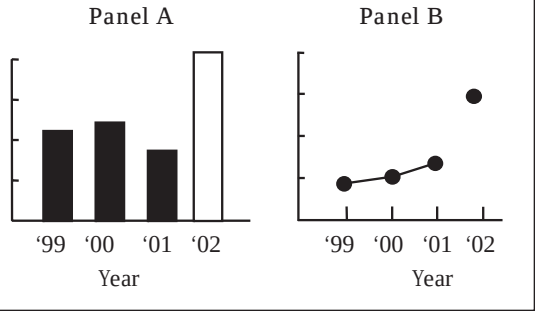


Figure 8

Graphics for multiple pretest–posttest design.



of this figure shows a similar graph that depicts performance for 2 years before an implementation (2000 and 2001) and at the end of a year of implementation (2002).

To display results from a treatment–comparison group design, graphs similar to those described above can be used. Column graphs can be used to compare results from a Direct Instruction group to those from a non-Direct Instruction group (see Figure 10, Panel A). If the evaluation includes several non-Direct Instruction groups, this also can be shown in a simple column graph (Panel B). Performance of students at various academic levels in Direct Instruction groups can be compared to performance in non-Direct Instruction groups using graphs that are similar to those that are used for one-group pretest–posttest designs.

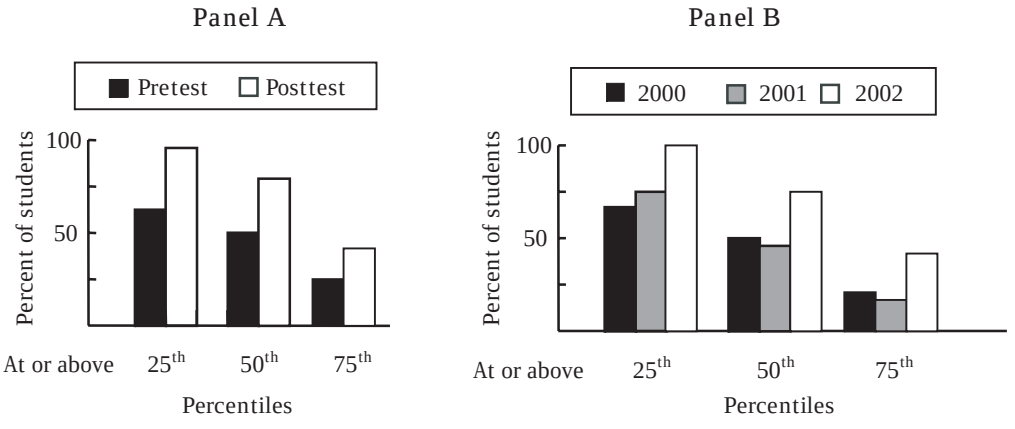
Pretest–posttest–comparison group designs can take advantage of the same types of graphs that are used for treatment–comparison group designs. Often, pretest scores are examined to judge the comparability of groups at the outset of the evaluation. Then, results from the posttest are displayed just as they are for treatment–comparison group designs. Thus, the graphs shown in Figure 11 are also applicable to this design. In addition, it is sometimes informative to show pretest and posttest performance for all groups on a single graph. Figure 12, Panel A shows a comparison of pretest and posttest performance for a Direct Instruction group and a non-Direct Instruction group. Panel B of that same figure shows a graph of results with three groups (two non-Direct Instruction and one Direct Instruction) each tested at three points before the implementation then again after the 1st year of the implementation.

Statistical Summaries

Graphic displays can convey detailed and subtle information that can be viewed simply, while displaying powerful results. A well-constructed graphic can often present more information than a table of complex statistics, and it can do so in a way that does not require extensive specialized training. In addition to graphics, it is also useful to summarize group performance in numbers.

Figure 9

Graphics showing multiple levels of achievement with pretest–posttest designs.



Summarizing Performance of a Group

The *mean* or *average* (sum of all the scores divided by the number of scores) is the most common and generally the most understandable summary of a group's performance. Computing the mean is simple and valid for raw scores, percentages, rates, NCE scores, and standard scores.

However, as we discussed earlier, some scores such as grade/age equivalents and percentiles should not be added together, and thus, should not be averaged. One solution to this problem is to convert these scores to NCE or standard scores that can be added, subtracted, and averaged. Table 1 is a conversion table that gives the NCE and standard score that is equivalent to each percentile score.

Once scores have been converted to NCE or standard scores, they can be averaged. The result, of course, will be the average NCE or the average standard score. If we want to know the percentile of the average score, we can convert the average NCE or standard score back into a percentile using Table 1. To summarize this procedure, in order to derive the percentile of the average score (a) convert all scores to NCE, (b) average the NCE, and (c) convert the average NCE back into a percentile.

Conversion from grade/age equivalent to NCE is different for each standardized test. There is no single table such as Table 1 that can give these conversions. The conversion for the Woodcock Reading Mastery Test is different from the conversion for the Stanford Achievement Test. Thus, converting

Figure 10

Graphics for treatment–comparison group designs.

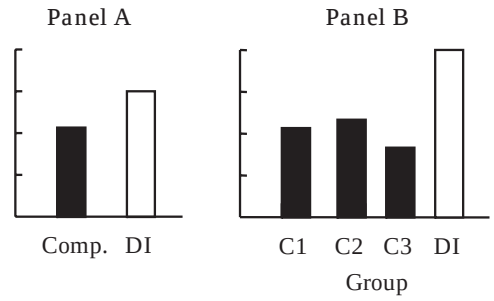
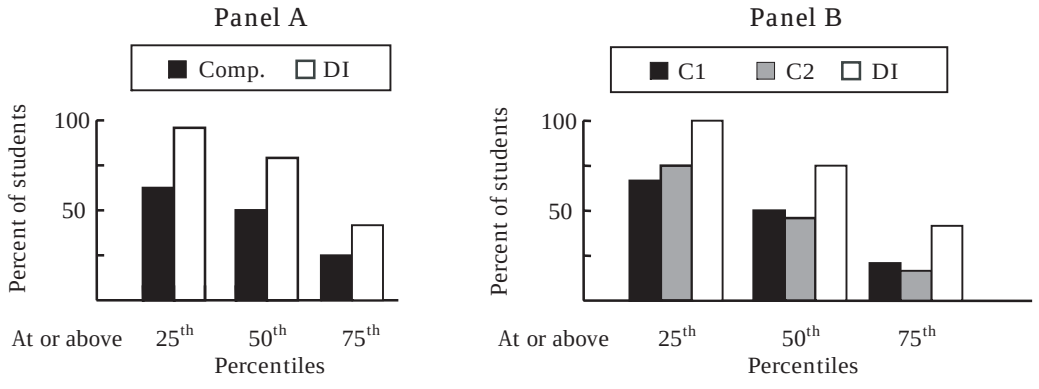


Figure 11

Graphics showing multiple levels of achievement in treatment–comparison group designs.



grade/age equivalent scores to other forms requires tables that are specific to the particular test from which the scores were derived. For some tests, these tables can be found in the test manuals. Other tests do not give these tables, but allow you to derive NCE or standard scores directly from raw scores. In addition, virtually all tests that include computer scoring software will print out NCE or standard scores for each student.

However, the mean is not always the best summary. As noted above, computing the mean can be difficult if we have grade/age equivalent or percentile scores. In addition, scores that are far from the main body of scores influence the mean. A small number of scores that are extremely high or extremely low can have a great influence on the group's mean. For these reasons, we often use the *median* as a summary of how well a group performed. The median is the middle score in a group of scores that are arranged in numerical order. To find the median of a group of scores, we arrange the scores from lowest to highest, and find the middle score. If we have 21 scores in order from lowest to highest, the 10th score is the median. If we have an even number of scores, then there is no single middle score. In this case, we average the two scores that are closest to the middle. For example, if we have 20 scores, the median is the average of the 10th and 11th scores.

In summary, if we are working with grade/age equivalent scores or percentiles, we should (a) summarize group performance with the median, or (b) convert all scores to NCE before averaging. If we are working with raw scores, percentages, rates, NCE scores, or standard scores, we may use either the mean or the median to describe a group's performance. The decision would depend on whether we want our summary to reflect the influence of scores that are at the extremes.

Summarizing the Difference Between Groups

We often want to make comparisons between two groups or between the pretest and posttest performance of a single group. In this

Figure 12

Graphics for pretest–posttest–comparison group designs.

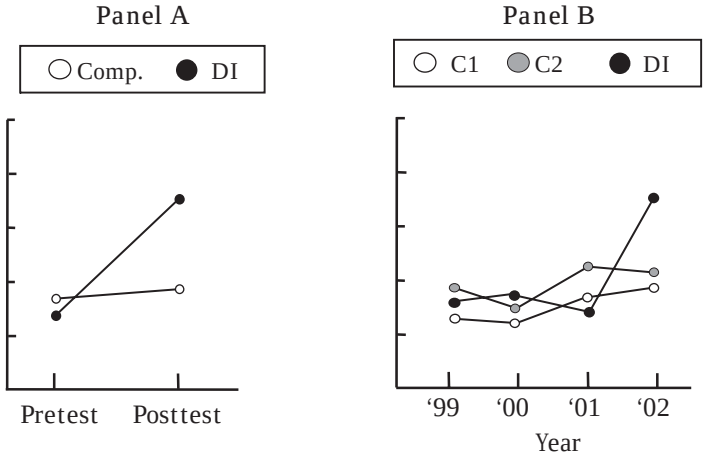


Table 1

Percentile, NCE, and Standard Scores

%tile	NCE	Stand. Score	%tile	NCE	Stand. Score	%tile	NCE	Stand. Score
1	1.0	65	34	41.9	41.9	67	59.8	107
2	6.7	69	35	42.5	42.5	68	60.4	107
3	10.4	72	36	43.0	43.0	69	61.0	108
4	13.1	74	37	43.6	43.6	70	61.7	108
5	15.4	75	38	44.1	44.1	71	62.3	109
6	17.3	77	39	44.7	44.7	72	62.9	109
7	18.9	78	40	45.2	45.2	73	63.5	110
8	20.4	79	41	45.7	45.7	74	64.2	110
9	21.8	80	42	46.3	46.3	75	64.9	111
10	23.0	81	43	46.8	46.8	76	65.6	111
11	24.2	82	44	47.4	47.4	77	66.3	112
12	25.3	82	45	47.9	47.9	78	67.0	112
13	26.3	83	46	48.4	48.4	79	67.7	113
14	27.2	84	47	48.9	48.9	80	68.5	113
15	28.2	84	48	49.5	49.5	81	69.3	114
16	29.1	85	49	50.0	50.0	82	70.1	114
17	29.9	86	50	37.1	37.1	83	70.9	115
18	30.7	86	51	50.5	101	84	71.8	116
19	31.5	87	52	51.1	101	85	72.8	116
20	32.3	87	53	51.6	102	86	73.7	117
21	33.0	88	54	52.1	102	87	74.7	118
22	33.7	88	55	52.6	102	88	75.8	118
23	34.4	89	56	53.2	103	89	77.0	119
24	35.1	89	57	53.7	103	90	78.2	120
25	35.8	90	58	54.3	103	91	79.6	121
26	37.1	37.1	59	54.8	104	92	81.1	122
27	37.7	37.7	60	55.3	104	93	82.7	123
28	38.3	38.3	61	55.9	105	94	84.6	125
29	39.0	39.0	62	56.4	105	95	86.9	126
30	39.6	39.6	63	57.0	105	96	89.6	128
31	40.2	40.2	64	57.5	106	97	93.3	131
32	40.7	40.7	65	58.1	106	98	99.0	135
33	41.3	41.3	66	58.7	107	99	59.8	107

case, we usually compare the two means and are interested in whether the difference is relatively small or relatively large. The simplest approach would be to find the difference between the means of the two groups. If the Direct Instruction group has a mean of 60 and the non-Direct Instruction group has a mean of 50, we could summarize the difference by saying that the Direct Instruction group scored 10 points above the non-Direct Instruction group.

This is a good start, but it is limited. The basic problem is that it is difficult to judge whether 10 points define a large difference or a small difference. Of course, we should label the score to indicate that the difference is 10 items in a raw score, or 10 responses per minute, or 10 NCE points. But even with a label, we must be very familiar with the particular test and/or the type of score in order to understand whether the difference of 10 points is large or small.

Researchers use a statistic called *effect size* to describe the size of a difference between two means. Effect size is simply the difference between the means (10 in the example above) divided by the standard deviation of the comparison group or pretest. (Note: There are many statistics that describe effect size. The effect size statistic described in this chapter is a standardized mean difference statistic known as Glass' Delta. See Martella et al., 1999, for a more complete discussion of effect sizes.) In practical situations, we will have to get the standard deviation from a computer printout of statistics describing the groups with which we are working. For example, suppose that a Direct Instruction group had a mean of 12 items correct on a test, a non-Direct Instruction group had a mean of 10, and the standard deviation of the non-Direct Instruction group was 4. The effect size would be computed by finding the difference between the means (12 - 10 = 2) and dividing that difference by the standard deviation (2 ÷ 4 = 0.50). The difference between these groups would have an effect size of 0.50. In another example, suppose that a Direct Instruction classroom had a pretest mean NCE of 45, a pretest standard deviation of 15, and a posttest mean of 50. The effect size would be computed by finding the difference between pretest and posttest means (50 - 45 = 5) and dividing that by the pretest standard deviation (5 ÷ 15 = .33) for an effect size of 0.33.

Researchers and evaluators often compute statistics that describe the statistical significance (*p* values) of the differences between groups. Correctly interpreting statistical significance requires extensive technical background. However, two facts about statistical significance are crucial. First, statistical significance gives the probability of getting differences this large (or larger) by chance alone. Second, statistical significance *does not* directly describe the size or educational importance of a result. Effect size is the best measure of the size of a difference between groups. Judging educational importance depends on effect size, statistical significance, and an understanding of how large a difference is important in your particular situation. We should not use statistical significance (*p* values) as our primary measure of whether differences are large enough to be meaningful. Effect sizes should be the basis of this decision.

References

American Educational Research Association, American Psychological Association, & National Association on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cohen, R. J., & Swerdlik, M. E. (2001). *Psychological testing and assessment: An introduction to tests and measures* (5th ed.). Boston: McGraw-Hill.

Deno, S. L., Mirkin, P. K., & Chaing, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36–45.

Engelmann, S., & Hanner, S. (2002). *Reading Mastery Plus Level 3 teacher's guide*. SRA/McGraw Hill: Columbus, OH.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*, 20–28.

Gall, M. D., Gall, J. P., & Borg, W. R. (2003). *Educational research: An introduction* (7th ed.). Boston: Allyn and Bacon.

Martella, R. C., Nelson, J. R., & Marchand-Martella, N. E. (1999). *Research methods: Learning to become a critical research consumer*. Boston: Allyn and Bacon.

Messick, S. (1993). Validity. In R. L. Linn (Ed.), *Educational measurement* (4th ed., pp. 13–103). New York: Macmillan.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Retrieved November 15, 2002, from http://www.nationalreadingpanel.org

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.