Definitions for the *No Child Left Behind Act of 2001:* Scientifically-Based Research

Judith Wilde, PhD
National Clearinghouse for English Language Acquisition and
Language Instruction Educational Programs
The George Washington University
Washington, DC

January 2004

The National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs (NCELA) is funded by the U.S. Department of Education's Office of English Language Acquisition, Language Enhancement and Academic Achievement for Limited English Proficient Students (OELA) and is operated under contract No. ED-03-CO-0036 by The George Washington University, School of Education and Human Development. The contents of this publication do not necessarily reflect the views or policies of the Department of Education, nor does the mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Table of contents

List of Exhibits

No Child Left Behind and Scientifically-Based Research

On December 13, 2001, the 107th Congress passed the *No Child Left Behind Act of 2001* (NCLB), the latest reauthorization of the *Elementary and Secondary Education Act of 1965* (ESEA); President George W. Bush signed the legislation in January 2002. Within NCLB and related policy initiatives, policymakers, researchers, and practitioners are beginning to focus more strongly on the nature, purpose, and value of "scientific research" and "evidence-based" practices in education. There are specific statutory provisions regarding how research is to be conducted and, to some extent, what is to be studied in that research. Within these statutory provisions is one major goal for the federal Department of Education: to identify and disseminate conclusive information about "what works" in education. Through this mandate, the U.S. Department of Education is to identify the instructional input (programs, curricula) that highly qualified teachers will provide to students in order to improve student outcomes in academic achievement and language proficiency. Thus in current and future applications for discretionary and formula funds, local schools, districts, and state departments of education will need to justify their choice(s) of instructional and professional development activities through sound, scientific evidence – through scientifically-based research.

Phrases such as "evidence-based decisions" and "scientifically-based research" occur 111 times in NCLB. The statutory provisions describe "scientifically-based research" as that which,

(i) at minimum, employs systematic, empirical methods;

(ii) involves rigorous data analyses that, when relevant to the line of inquiry or purpose of the investigation, are adequate to test a stated hypothesis and to justify general conclusions drawn;

(iii) relies on measurements or observational methods that provide reliable and valid data from the investigators and observers involved in the study, and provides reliable and valid data from multiple measurements used, and observations made in the study; and

(iv) uses every opportunity to conduct experimental or quasi-experimental designs in which individuals, entities, programs, or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest (§9101(37)).

What does this really mean to schools, school districts, and state departments of education? The U.S. Department of Education provides a great deal of support to education each year through formula funding (e.g., Title I, Title III services to students) and discretionary grants (e.g., Title III professional development grants, Safe Schools/Healthy Students projects). However, because evaluations of these projects, and many educational programs, often are not as systematic as might be expected, it may be difficult to determine whether the monies spent have resulted in increased long-term academic achievement for students. The purpose of scientifically-based research within NCLB is to identify and disseminate conclusive information on "what works" in education, with an emphasis on determining *what* instructional input (curricula, instructional techniques) *works* to increase student outcomes such as academic achievement and language proficiency. Scientific research, such as that used in medical research and other "hard" sciences, offers rigorous methodologies for studying *what works,* for determining whether a particular curriculum or instructional technique will result in improved academic performance (or increased language proficiency) for a majority of the children for whom it was designed.

The purpose of this document is to provide a definition of *scientifically-based research*, and the major elements necessary to conduct such research. These definitions, and those in the NCELA document *Definitions for No Child Left Behind: Assessment,* also provide background for two others in the series: *Research and Evaluation that Works within NCLB Standards* and *Criteria for Evaluating Evidence-Based Research* – all of which are available on the NCELA website.

Introduction

When researchers conduct a study, they have a specific question they want to answer, a decision to make, or a hypothesis to test. In order to determine the answer, they must collect information relevant to the question. Sometimes the information may be anecdotal and not quantified; in this case there are no statistical methods used, but they will interpret the information by describing and synthesizing it. In most cases, however, the researcher(s) will collect information that either already is quantified or is quantifiable. Those data then are analyzed statistically, leading to the formulation of an answer to the question, finalizing a decision, or determining the legitimacy of a hypothesis.

The information collected in research is called *data*, especially if that information is numerical. Suppose, as an example, we stood on the corner of two main streets in your town from 9 a.m. to 10 a.m. each day for a week. We watch the people who walk into the building on the northeast corner of the intersection. We record whether they are male or female and whether they carry a briefcase of some type, rate their clothes on a scale of "neatness" from 1 to 4, and note the color of their hair. At the end of the week, we know how many men and women walked into the building, how many had briefcases, how neat they looked, and how many blondes, red-heads, and brunettes there were. These are all numeric data except for the color of hair, which we could code as 1 for blondes, 2 for red-heads, and 3 for brunettes.

This example of research probably does not seem very "researchy" simply because we do not know the original purpose of the study – the decision or question or hypothesis that generated the study. Without this focus, the data are just numbers that we do not know how to analyze or report. Unfortunately, there are many times when this happens in the "real world" of research: data are collected and then the researcher looks for a question that the data will answer. So, research must begin, always, with a carefully formulated purpose. Next come some precisely defined questions, hypotheses, or decision points. Only then are data gathered, analyzed, interpreted, and reported. In the end, the analyses are useful and the data are relevant only if the original purpose is clear.

In providing the various definitions that are the majority of this document, it will be helpful to give examples to help understand what can be difficult concepts. Rather than select different examples, we describe here a hypothetical situation that we might want to study. Then, throughout the document, we will refer back to this example to provide explanations and definitions.

A potential research study

Reading is a "hot topic" now. Let us assume that our K-7 school district has been using Method A, from the ABC Publishing Company as the curriculum for reading throughout the grade levels. Teachers like the curriculum and are familiar with it, it has readiness activities for younger students, and students are performing reasonably well, scoring neither extremely high nor extremely low on the state-mandated Stanford Achievement Test, 9th Edition (SAT-9) subtest on reading/language arts. Method Z has just been introduced by the XYZ Publishing Company with the claim that it is better for diverse students (e.g., those living in poverty, English language learners [ELLs]). Exhibit 1 provides an overview of our district.

Exhibit 1: Overview of district demographics

School figures

- 1 elementary schools, 2 middle schools
- 3 charter schools
- 5,085 students enrolled

Enrollment figures

- By grade: between 650 and 700 at each elementary grade, between 700 and 750 at each middle school
- By ethnicity: 47% white, 40% Hispanic, 5% Asian, 4% African-American, 2% American Indian, 1% Filipino and Pacific Islander, and 2% other responses
- Special programs: 51% Free/reduced price meals, 38% ELLs, 63% Compensatory education
- ELLs (as percent of total enrollment): 34% Spanish-speakers, others are Lao, Khmer, Vietnamese, and "others"

Certificated staff

- 33 administrators, 29 pupil services (counselors, nurses, etc.), and 334 teachers
- Teachers' certificates: 97% full credential, 2% emergency, 1% university intern, and <1% waiver
- Teachers' ethnicity: 89% white, 8% Hispanic, 1% each Asian and African-American, and 2% "other"

Our district is lucky to have a good relationship with the local university. That university has professors who specialize in K-7 education and are interested in research; these professors have students who need to do research. We will approach them and ask that they conduct a study to determine whether we should change the district from Method A to Method Z; we need to know whether students do substantially better using Method Z so that the cost to purchase the curricula, provide appropriate professional development activities for teachers, and develop supporting materials is justifiable.

Hypotheses

The term "scientific method" refers to many attitudes and procedures that characterize empirical investigation of problems in the social sciences. One aspect of scientifically-based research is the formulation and testing of hypotheses. The role of statistical inference is to provide hypothesis-testing procedures so stringent that when conclusions are drawn from the statistics that test the hypotheses, they are drawn with some certainty. For us, this means that we will know whether it is prudent to change from Method A to Method Z to teach reading to K-7 students in our district.

Hypotheses are the backbone of research. They can be written as statements or as questions (in which case they usually are referred to as "research questions"). The major hypothesis (or set of hypotheses) is referred to as the "null" hypothesis, indicating that we anticipate no difference in the observations or scores on tests; the "alternate" or "research" hypothesis proposes a relationship among the data studied. For instance, ***there will be no difference between K-7 students' reading comprehension whether they are taught with Method A or Method Z , as measured by scores on the reading subtest of the SAT-9 or appropriate performance-based assessments*** is the null hypothesis and ***there will be a difference between reading comprehension whether K-7 students are taught with***

Method A or Method Z, as measured by scores on the reading subtest of the SAT-9 or appropriate performance-based assessments is the research hypothesis. The null hypothesis will be scrutinized through standard procedures and statistically analyzed. Only if what we find is quite rare in light of the null hypothesis do we reject the null hypothesis in favor of the alternate hypothesis.

The two hypotheses, null and alternate, must explain all possible outcomes of the study. In the example above, the two methods of reading instruction result in scores that are either different or the same – those are the only possible outcomes. However, we may have strong reason to believe that one method will result in higher scores. If we want to anticipate only that outcome, then the null hypothesis would be that *K-7 students taught reading with Method Z will score the same or lower than K-7 students taught with Method A, as measured by the reading subtest of the SAT-9 and appropriate performance-based assessments* and the alternate hypothesis would indicate that *K-7 students taught with Method Z will have higher reading comprehension than K-7 students taught with Method A, when measured by the reading subtest of the SAT-9 and appropriate performance-based assessments*. Again, the two hypotheses include all possible outcomes: Method Z results in the same or lower comprehension than Method A (the null hypothesis), or Method Z will result in higher reading comprehension scores than Method A (the research hypothesis). Note that when we say the scores will be "the same," we do not mean identical, but that there will not be measurable, important, statistically significant, differences.

Research questions are similar but usually are posed as simple questions rather than made as statements. For instance, *will there be a difference in reading comprehension if K-7 students are taught with Method A rather than Method Z, when measured by the reading subtest of the SAT-9 and appropriate performance-based assessments?*

Any hypothesis (null or alternate) or research question should be stated clearly to indicate

- who is involved in the study (in our example, K-7 students),
- what is being studied (2 different curricula),
- what is being measured (reading comprehension), and
- how it is being measured (the reading subtest of the SAT-9 and performance-based assessments).

Based on this information, a person with statistical knowledge should be able to define the type(s) of statistics that will be used to determine whether the null hypothesis will be rejected or not. (Note that technically, a null hypothesis can never be accepted, we can only (1) reject it (in which case we have anticipated the outcome of the study correctly) or (2) fail to reject it (in which case the study did not show the differences in test scores that we had anticipated).

Using multiple measurements

The NCLB statutes specifically indicate that "multiple measurements" should be used; in the two potential studies described in this section, students will be tested with at least two different assessments (SAT-9 and appropriate performance-based assessments). Why? No test is perfect; no test will measure a student's knowledge, skills, or abilities with absolute clarity. Much has been made in the past about errors found in nationally developed, standardized tests – errors that may have affected the outcomes for many students. In-class, teacher-made, or district-developed assessments generally have greater potential for problems since they are developed with smaller resources. Therefore, it is important to ***triangulate*** results. Triangulation refers to having several differing sources of information that converge to provide a good, solid "picture" of how well a student is doing. Multiple measures should include at least two of the following:

- performance-based assessments that will help to determine whether students actually can use the knowledge and skills that they have been taught,
- state-mandated standardized assessments to determine whether students are meeting the state's content and performance standards and to allow a comparison of local students with state and national expectations,
- teacher and/or student reflections that indicate their understanding of students' knowledge and skills, and
- other forms of assessment that may add to knowledge about a particular content area.

The importance of being able to assess students' knowledge and skills and report them with certainty cannot be stressed enough. Not only are tests not perfect, but on any given day a particular student may not feel well or just may not perform optimally. Multiple assessments can help. If two types of assessment give one "picture" of a student and a third assessment gives a different "picture," we can, with some certainty, decide to eliminate the results of the third assessment from consideration. However, if that were the only assessment we had given, we would be unaware that the assessment is giving us bad information.

Longitudinal and Cross-Sectional Studies

The best way to determine whether something "works" is to study it over a period of time. By identifying one group of students and following them for several years, it is possible to study their persistence, change, growth, and/or development. Such research is referred to as ***longitudinal***. A major strength of longitudinal research is that it allows us to begin to discover causal factors – what causes attitudes, skills, behaviors, or knowledge to change. A major weakness of longitudinal research is that students change schools, drop out, and decide to leave the program, any one of which can compromise the study. Exhibit 2 is a longitudinal study of reading comprehension when students are taught with one of two different methods.

Exhibit 2: A longitudinal design to study reading comprehension

- We are going to assume that we can demonstrate that schools in our district are fairly similar – about equal size, achievement levels, and socio-economic status, with the same language groups, and numbers of ELLs at each school. (While we recognize that this is rarely the case, this will simplify greatly our discussion of research.)
- Select half of the K classes and half of the 3rd grade classes to receive instruction with Method A; the remaining K and 3rd grade classes will receive Method Z.
- The curriculum for each method is appropriate for K-7 (with K and 1 focusing on readiness activities, 2-7 focusing on developing independent readers). Our study will allow us to look at students who are just learning to read (kindergarten) as well as students who have some experience reading (grade 3).
- Keep these students together for at least 4 years (i.e., K students complete 3rd grade, 3rd grade students complete 6th grade) – longer if possible.
- Use the same performance-based assessment(s) district-wide to measure reading skills in September, January, and May for each year of the study; use the state-mandated test of content and achievement standards annually.
- If students move from one school to another within the school district, be sure they attend the appropriate reading class so they can continue participating in the study.
- Analyze the data to determine (1) the difference in reading comprehension (do students using Method Z read differently from students using Method A?) and (2) the pattern of change in reading skills (does children's reading comprehension increase linearly – in a straight and predictable fashion -- or does it move up and down in different grade levels?)
- The null hypothesis for this study: Students' reading comprehension will be the same when measured by performance-based and state mandated assessments regardless of
    - Curriculum used (Method A from ABC Publishers or Method Z from XYZ Publishers),
    - Length of time in the reading program (up to 4 years – K-3 or 3-6), and/or
    - Combinations of curriculum and time (e.g., the pattern for K students' achievement will be the same as the pattern for 3rd grade students).

Because we generally do not have the necessary time (or funds) to conduct longitudinal research[1], we often use ***cross-sectional*** research instead. Cross-sectional research has been likened to a snapshot in time because it collects data (e.g., test scores) only one or two times during the brief study, providing a quick picture of how well students are doing at that time. A major strength of

[1] Although these are typical reasons for not doing longitudinal research, Slavin (2002) has pointed out that one well-designed longitudinal study may provide much better and much more information than several cross-sectional studies – which may make the longitudinal study more cost-effective in the long run.

cross-sectional research is that a great number of participants can be studied in a relatively short period of time. A major weakness is that causality is more difficult to demonstrate because of the short-term nature of the study. Exhibit 3 is a cross-sectional study of reading comprehension when students are taught with one of two different methods.

Exhibit 3: A cross-sectional design to study reading comprehension

- We are going to assume that we can demonstrate that schools in our district are fairly similar – about equal size, achievement levels, and socio-economic status, with the same language groups, and numbers of ELLs at each school. (While we recognize that this is rarely the case, this will simplify greatly our discussion of research.)
- Randomly select 6 elementary schools and 1 midschool. Randomly decide which group of 7 schools will receive instruction with Method A; the remaining 7 schools will receive instruction with Method Z.
- The curriculum for each method is appropriate for K-7 (with K and 1 focusing on readiness activities, 2-7 focusing on developing independent readers).
- Use the same performance-based assessment at each grade level to measure reading skills in September, January, and May and the state-mandated assessment of content and performance standards, the SAT-9, annually.
- Analyze the data to determine whether there area differences in reading comprehension at each grade level based on whether students are instructed with Method A or Method Z.
- Our null hypothesis for the study: Students' reading comprehension will be the same when measured by performance-based and state-mandated assessments regardless of
    - Curriculum used (Method A from Publisher ABC or Method Z from Publisher XYZ,
    - Grade level of student (K-7), and/or
    - Combinations of curriculum and grade level (e.g., 3rd grade students using Method A will score similarly to 6th grade students using Method Z).

For our purposes, the basic differences between longitudinal and cross-sectional research are three-fold:

- longitudinal research studies fewer individuals for a longer period of time while cross-sectional research studies many individuals for a shorter period of time and
- longitudinal research involves one set of individuals who age with the project while cross-sectional research involves several different sets of individuals who represent each age group of interest, so therefore
- longitudinal research allows some causal interpretations of data while cross-sectional research allows differences at specific points in time to be reported.

Three years of reading achievement data from a possible study are presented in Exhibit 4. The data are the percentage of students scoring above the 50th percentile on the state-mandated SAT-9, the percentage who are scoring above average. If students are doing well, then more students should be achieving at this level across time. Depending on whether the data are read as columns, rows, or diagonals, the information presented is somewhat different.

Exhibit 4: Interpreting data - % "all" students scoring above the 50th percentile in reading achievement, SAT-9

Year	2nd grade	3rd grade	4th grade	5th grade	6th grade	Cross-sectional question answered
2000	48	46	45	48	50	Are there differences across the grade groupings by year (e.g., the students in the year 2000)?
2001	48	51	50	46	54	
2002	51	49	52	49	50	
Cross-sectional question answered	Are the cohorts of students (e.g., the 3 groups of 2nd graders) different from one another?					

Longitudinal questions answered:
Has the reading achievement of students changed as they moved from 2nd to 4th grade?
Has the reading achievement of students changed as they moved from 3rd to 5th grade?
Has the reading achievement of students changed as they moved from 4th grade to 6th grade?

Interpretation of data: the numbers presented are the percentage of all students in the district who scored above the 50th percentile. We would anticipate that half of the students would score above, and half below, the 50th percentile, so when this number is larger than 50% it indicates that more students are scoring higher, when this number is below 50% it indicates that more students are scoring lower.

Are there differences across grade groupings by year? This cross-sectional analysis indicates that in the year 2000, nearly half the students scored near the 50th percentile; in 2001, students had a broader range of scores, with 45-54% of students scoring above the 50th percentile; and in the year 2002, just about exactly half the students scored above the 50th percentile. Across these three years, students' varied very little. Looking at the specific scores, the 6th grade students of 2001 had the highest scores with 54% of them scoring above the 50th percentile; the 4th grade students in 2000 scored the lowest, with 45% scoring above the 50th percentile. We cannot make many generalizations about these comparisons because each of these years is made up of different students, at different grade levels.

Are the cohorts of students different from one another? Students in 2nd, 4th, and 5th grades scored highest in 2002, with about half scoring above the 50th percentile; 3rd and 6th graders scored highest in 2001, with just over 50% of the students scoring above the 50th percentile. Each cohort does score somewhat differently, but those differences are not great. We cannot make many generalizations about these comparisons because the cohorts are made up of different students in different years in different grade levels.

Has reading achievement of students changed as they moved from one grade to the next? Longitudinal data are better, more valid and more reliable, as they extend across more time. Therefore there are three sets of data to interpret, as indicated by color in the table. Students who began 2nd grade in 2000 showed a fairly steady increase in their scores, with 48% scoring above the 50th percentile the first year, and 52% scoring above the 50th percentile the last year. Students who began 3rd grade in 2000 increased their scores most dramatically between 3rd and 4th grade, with a slight decrease as they completed 5th grade. Students who began 4th grade in 2000 showed a slight increase in scores for 5th grade, and a larger increase as they completed 6th grade. It would appear that for "all" students in the district, the reading instruction curriculum is successful. Students are increasing their reading achievement scores as they move from grade to grade.

Sampling from a Population

A major element in designing any research study is determining how to select the "subjects" – the people who will participate in the study and on whom conclusions will be based. Two definitions must be provided immediately: population and sample.

- The ***population*** refers to all possible people (or classrooms, or whatever is to be studied) who could be involved in the study. The population might be defined narrowly or broadly, whichever makes more sense to the study. For instance, the population might be "all 3rd graders who are left-handed and have red hair" or the population might be "all ELL 3rd graders." What is the purpose of the study?
- The ***sample*** refers to members of the population (people, classrooms, etc.) who are selected to take part in the study. How the sample is selected is important for determining an appropriate method for analyzing the data and for the generalizations that can be made based on the study. Again, what is the purpose of the study?
- A third definition describes an individual within the population or sample. Every individual (whether the "individual" is a person, a classroom, or something else) is referred to an ***element*** of the population.
- ***Variable*** refers to a property of members of a group that differ from one another – e.g., gender is a variable with the properties of male and female; grade is a variable with the properties of kindergarten, 1st grade, and so on. If a study involves only one property of the members (if we study only girls), then it is not a variable and our results must be limited to the group studied.

It is rarely possible to include an entire population in a study, regardless of how narrowly it is defined. Therefore we must select a sample from the population; the more closely that sample matches, or "looks like" the population, the better we are able to generalize the information back to the population and the more sure we are that the results would occur under other circumstances.

Sampling procedures are either ***random*** or ***nonrandom***. Within these two general rubrics, there are several methods for selecting the sample; some are quite sophisticated, some are very simple. Exactly how the population is selected affects the statistics we use to analyze the data and the types of general statements we can make about the study outcomes. If we select a particular 4th grade class-room because we know the teacher and it is convenient, then the study results can only be applied to that 4th grade classroom and cannot be generalized further. If, however, we randomly select a 4th grade classroom, the results can be generalized to the population from which that classroom was selected. So, if the 4th grade classroom was randomly selected from within a school district, the results can be generalized to all 4th grade classrooms in that district; if the 4th grade classroom was randomly selected from all 4th grade classrooms in a state, then the results can be generalized to all 4th grade classrooms in the state. It is obvious, then, that how we identify our sample is important to the study and how we interpret the results of the study. The methods for sampling described below are divided into *random sample techniques*, those that allow generalization, and *nonrandom sampling techniques*, those that do not allow generalization past that particular group of subjects.

Random sampling techniques

The formal definition of a random sample is that "each member of the population has an equal (and non-zero) probability of being selected." By selecting participants randomly, there should be no biases in the sample. As an example, if the population is students from our school district, then there should be boys and girls, special education and gifted education students, English learners and English proficient students, and representatives of each of our ethnic, linguistic, and socioeconomic groups in about the same proportion as present in the entire school district population. A sample including only girls would be considered biased because it is not representative of the entire

population of children – girls and boys. There are various methods for randomly selecting the sample; three commonly used techniques are described below.

Simple random sample

The simple random sample typically is used with a relatively small population and is the least sophisticated method. The classic example of how to select a simple random sample is to give each element in the population a number, put all the numbers in a "hat," mix them well, and then select as many subjects as needed. Many computer programs have random number generators and statistical texts have random number tables to help select a random sample. A simple random sample should be representative of the entire population. An example of drawing a sample from our school district to participate in the reading study is provided in Exhibit 5.

Systematic random sample

It may not be convenient to use a simple random sample, often because the population is so large that identifying and numbering each possible subject is virtually impossible. A systematic random sample may be used if there is a list of the population (e.g., a phone book or an enrollment list). After determining the number of individuals needed and selecting a random "start point," then selecting every n^{th} element (*n* referring to a specific set of elements, such as every 10^{th} element or every 33^{rd} element) until the sample is large enough. Most statisticians consider this a random sampling technique because the start point is randomly selected; others disagree because of the systematic specific pattern selection follows thereafter. This method also should result in a sample that is representative of the population from which it was selected.

Stratified random sample

The most sophisticated technique is the stratified random sample. This procedure requires prior knowledge about the numbers of individuals in the population who come from specific strata. The sample then is selected based on this information. Typically, the sample is proportional to the numbers in each stratum. There also are times when it is appropriate to select a sample with equal numbers from each stratum in the population – e.g., from each ethnic group. If there has been little research on one or more of the strata, and the lack of research is based on the small number of people in that stratum, then equal numbers from each group can be selected. This planned equal sampling of all groups can allow the researcher to study them in more detail and give equal weight to their knowledge, skills, and/or beliefs. The researcher then can make generalizations that usually could not be considered because of the few number of people on whom the generalizations would be based.

Cluster sampling

The stratified random sample typically involves only one source of information (e.g., ethnicity OR language proficiency). Other data sources, such as gender, age, language group, or socio-economic status, also could be included as part of the selection criteria – when using several data sources, the technique may be referred to as cluster sampling, rather than stratified sampling. This random sampling technique should ensure an extremely representative sample.

Exhibit 5: Selecting a random sample

Any of these random sampling techniques could be used for our reading comprehension study. The following are examples of how we could use each one. In all examples, we will assume that we want a sample of 300 students from the 5,085 students in our district.

Simple random sample

- Obtain a computerized list of all the students who are enrolled in the district; make sure that each student has a unique identification number.
- The computer should have a *random number generator* in its software. Ask the software to generate 300 random numbers between 1 and 5,085. OR
- Go to a *random number table* that can be found in virtually any statistics textbook. Decide whether to read numbers up, down, left, or right. Drop a pencil on the table to find a start point. From there, read the numbers up, down, left, or right (as already decided) in groups of 4 that are no larger than 5,085. For instance, the following numbers are from a random number table:
  43 44 09 42 72 00 41 86 79 79 68 47 22 00 20 35 55 31 51 51 00 83 63 22 55 40 76 66
  Assuming that we randomly identified the first number as our start point, we would then select students with the following ID numbers: 4344 942 7200 (skip-too big) 4186 7979 (skip-too big) 6847 (skip-too big) 2200 2035 5531 (skip-too big) 5151 (skip-too big) 83 6322 (skip-too big) 5540 (skip-too big) 7666 (skip-too big). Continue through the table following the predetermined patterns until we have 300 students.

Systematic random sample

- Get a list of all students who are enrolled in the district. They can be listed alphabetically, by grade level, or any other way.
- Determine the percentage of the population that we need to use in the study: 300/5085=5.9%, or about 6%.
- A 6% sample is 1 student in every 17 students, so randomly select a number between 1 and 17 to be the start point.
- If we use the random number table above (reading the numbers in pairs and selecting the first number that is 17 or less), our first student is number 9 on the list.
- Beginning with the 9th student, select every 17th student. Our sample consists of students numbered 9, (9+17=)26, (26+17=)43, 60, 77, 94, 111, and so on.

Stratified random sample (proportional to the population)

- We want to ensure that our sample matches the population with regard to ethnicity, the stratum we will study.
- Looking at Exhibit 1, we see the percentages for each ethnic group in the district. Get a list of students separated by ethnic group.
- Randomly select the appropriate number of students from each ethnic group so that we have a total sample of 300.
- Our sample will consist of 140 whites (47% of 300), 120 Hispanics (40% of 300), 15 Asians, 11 African-Americans, 6 American Indians, 3 Asian/Pacific Islanders, and 5 "other" individuals.

Stratified random sample (equalized samples)

- We note that we have very few Asians, African-Americans, American Indians, Asian/Pacific Islanders, and "others" in our district and that some of them have lower reading comprehension scores. We want to ensure that the reading curriculum helps these students. Therefore we want these students to have equal representation in the sample.
- There are 6 ethnic groups plus "others," a total of 7 groups. Divide 300 by 7 and select 42 from each ethnic group and 6 "others" (42 * 7 = 294 + 6 = 300).

Continued ...

Exhibit 5, continued

Cluster sampling

- Rather than only investigating the effects of the reading curriculum with respect to ethnicity, we can add other information as well.
- Determine what other information might be of interest – perhaps language proficiency and school attended as well as ethnicity.
- Get a list of students grouped by all three variables: girls who go to School A and speak only English; boys who go to School A and speak only English; girls who go to School A and are fluent English proficient with another home language; and so on.
- Determine the proportion of students in each of the groupings.
- Randomly select the appropriate percentage of students from each grouping, so that the total number in the sample is 300.

Nonrandom sampling procedures

There may be times when the population is poorly defined, when the researcher merely wants to explore a topic without making generalizations, or for some reason it is not possible or practical to select the sample randomly. There are two types of sampling techniques that often are used, but are not considered random.

Purposeful sample

The purposeful sample is selected because it meets specific criteria – for instance, our red-headed, left-handed, 3rd grade students who were selected from all possible 3rd grade students. The results of this study can only be used to talk about red-headed, left-handed, 3rd grade students, further generalizations (i.e., discussions related to right-handed students or students with other hair colors, or in other grades at school) cannot be made.

Fortuitous sample

An instructor who decides to use her/his classroom as a study site because it is easy, or someone who asks for volunteers to participate in a study has used fortuitous sampling. This term refers to any sample that is selected for convenience rather than randomly. Generalizations past the group of individuals actually studied cannot be made.

Assigning to different conditions

There are two aspects to a study that must be random: random selection of subjects and random assignment of those subjects to the different "treatment" conditions – for our study, random assignment to Method A or Method Z. The purpose is to continue assuring ourselves, and those who read about our study, that we have done nothing that might contaminate our study, nothing that might ensure that only "good" readers were taught with Method Z and "poor" readers were taught with Method A, thus biasing the outcome of the study. Random assignment can be easily accomplished. Assuming there are only 2 "treatments,"

- Decide that the first treatment is number "1" and the second treatment is number "2,"
- Randomly select a subject – that subject goes in the first group,
- Randomly select a second subject – that subject goes in the second group,
- Randomly select a third subject – that subject goes in the first group,
- And so on.

As an example, we decide for our study that Method A is the first treatment and Method Z is the second treatment. Then we randomly select the first classroom to be taught by Method A, the second classroom to be taught by Method Z, the third to be taught by Method A, and so on.

Research Designs

Research design has one purpose: to provide a framework for planning, conducting, and completing the study. In order to design a study, the researcher must carefully consider many aspects of the project and then control as many extraneous factors as possible. The overall question is: What is happening (description); is there a systematic effect (cause); and why or how is it happening (process or mechanism)? Feuer, Towne, and Shavelson further state that "although no universally accepted description of the principles of inquiry exists, we argue nonetheless that all scientific endeavors

- Pose significant questions that can be investigated empirically,
- Link research to relevant theory,
- Use methods that permit direct investigation of the questions,
- Provide a coherent and explicit chain of reasoning,
- Yield findings that replicate and generalize across studies, and
- Disclose research data and methods to enable and encourage professional scrutiny and critique" (2002, p 7).

There are many different research designs that meet these principles, but a relative few are used frequently. Some are used more often in the so-called hard-sciences (e.g., biology, medicine) and some are used more often in the so-called soft-sciences (e.g., sociology, education). Those that are most frequently used in education, and which can meet the mandates of the NCLB legislation and the above principles are described briefly below.

Descriptive research

Descriptive research is used to answer the question "What is happening now?" Rather than trying to determine differences between groups, or whether one educational method is more successful than another, this type of research collects information about the current status of events, and creates a report without doing any statistical analyses. The report will include appropriate numbers (e.g., percentages of individuals in certain groups or average scores on an assessment). Most research should include an aspect of descriptive research in order to provide a context for the study.

Survey research

Survey research generally answers questions such as "What would [the respondents] do if … ? What is their attitude about … ? How do they think they would change if … ?" Surveys also can be used to collect background information about respondents and their current circumstances. The number of surveys distributed, and the number returned (the "response rate") should be documented. Although surveys are powerful, a limitation on their generalizability and on their worthiness is the response rate – a low response rate makes interpretation of the results difficult.

Surveys can be highly structured (specific questions with a set group of response options from which to choose) to unstructured (general questions with the respondent providing whatever responses s/he feels appropriate); surveys can be sent through the mail, completed in-person, or used as an interview. However, the information gathered is only as good as the questions on the survey instrument. It can be difficult to interpret the results if the questions are open to interpretation or if the structured response options do not allow the respondent a full range of options. (For instance, consider this question: "Is the program staff sensitive to culture, language, and gender issues?" If the answer is "no," does this mean that they are not sensitive in any of the three areas, or in one or more of the areas? In which area[s] are they sensitive?) Similarly, a simple yes/no response option does not make it possible to measure the degree of sensitivity the staff shows. It is difficult to de-

sign a complete research study using only survey methodology, but survey(s) can be an important part of several different types of research.

Quantitative research

The term "quantitative research" covers a broad group of research designs all of which collect numeric data that are analyzed using statistics of varying levels of sophistication. Three major types are described briefly here.

True experimental research

True experimental designs are those most often used in the hard sciences. They are used to study cause-and-effect relationships; that is, did the reading curriculum cause students to learn to read and increase their reading comprehension? This is the most powerful research design, but is restricted by two requirements:

(1) participants (referred to as "subjects") must be randomly selected from the population and then randomly assigned either to the control group (the "regular" reading program, Method A) or the treatment group (the new curriculum, Method Z) and

(2) the program being studied must be controlled carefully with no other students receiving its benefits.

A true experiment is considered to be the strongest quantitative methodology because it does allow a clear determination of whether the program studied caused the students' knowledge, skills, and/or behaviors to change. However, the first condition (random selection and random assignment) is especially difficult for education programs – it is rarely possible to randomly assign students since the very existence of a particular program may be predicated on a demonstrated need of students for the program (e.g., schools have ESL programs because they have students who need such assistance in order to increase their English proficiency so they can perform in an English-only school environment).

Experimental designs require some type of pretest (an assessment administered before the educational program begins) followed by a posttest (an assessment administered after the program ends) to determine whether students have increased their knowledge and skills. It is desirable to have a *control group* of some type (students who were not in the education program and received none of its benefits) so that the researcher can say (1) students in the program changed their knowledge, skills, and/or behavior and (2) students in the program changed their knowledge, skills, and/or behavior at a greater rate than did students not in the program.

How are control groups selected? ***True control group(s)*** are students who are randomly selected from the school and randomly assigned to the control group. This type of control group is essential for a true experimental design. A subset of the true control group is the "wait list" design. In this case, students are randomly assigned to either the new educational program now, or they are assigned to wait a semester (or other reasonable time period) before starting the program. In this way, we have a true control group but we do not make them miss the new program completely. The wait list design may be helpful in studies where we do not feel we can ignore the educational needs of the child. However, in cases such as our reading comprehension study, we are not ignoring student needs when we assign them to the current curriculum that we know "works" reasonably well and the purpose of the study is to determine whether a new curriculum works even better.

Quasi-experimental research

The quasi-experimental design is somewhat less restrictive than true experimental research. The design is similar to the true experimental design except that subjects are neither randomly selected

from the population nor randomly assigned to the treatment group. These designs offer greater flexibility and greater potential for generalization to a "real" educational setting. The researcher still should control as many other elements that may influence the educational program as possible.

A pretest and posttest is required in quasi-experimental research, but the control group is somewhat different. The ***nonproject comparison group(s)*** are subjects who are similar to those in the educational program being studied, but are not identical to them; they generally have neither been randomly selected nor randomly assigned. Statistical procedures will help to determine how similar the two groups of students are; the greater the similarity, the stronger the statements about the final outcomes of the study. Another type of control group that can be used is the ***norm group comparison***. When it is impossible to find students for a comparison group who are similar to the students in the experimental group, then a "live" comparison group might be replaced by a norm group. These norm group students are (1) the average score from the norm group from a nationally-developed norm-referenced test or (2) a test score such as a school district average or state average used to represent the norm group. This type of comparison is appropriate if the research question seeks to show that the program students are becoming more similar to mainstream students.

Correlational research

Correlational research does not seek to identify a cause-and-effect relationship but rather seeks to determine the extent of a possible relationship between variables or to predict an outcome based on knowing about the input information. The data are all numerical. For instance, height is related to weight and high school grade point average can predict a college entrance exam score. Correlational research answers questions such as "Is there a relationship between English language arts achievement scores and math achievement scores?" or "Do language proficiency scores predict English language arts achievement scores?" We cannot say that English language arts achievement *causes* math achievement, but there may be a *relationship* such that those who do well in one content area tend to do well in the other. And, while English language proficiency does not *cause* English language arts achievement, proficiency may predict achievement – one needs some level of proficiency in order to study for achievement.

A correlation value ranges from +1.00 (the two variables are virtually identical) through zero (there is no relationship at all between the two variables) to -1.00 (the two variables are opposites). For instance, height and weight are positively related – those who are taller tend to be heavier (numeric values for both variables, height and weight are either both larger or both smaller); countries' gross national product values and infant mortality are negatively related – wealthier countries tend to have lower infant mortality rates (values for wealth are larger when infant mortality is lower and values for wealth are smaller when infant mortality is higher).

Qualitative research

Employing such methods as naturalistic, pluralistic, and ethnographic, qualitative research takes place in the natural world, uses multiple methods that are interactive and humanistic, is emergent rather than tightly prefigured, and is fundamentally interpretive (Marshall & Rossman, 1999). All these techniques are based on ethnographic methodologies developed by anthropologists. They can provide in-depth information about individuals, groups, or institutions as they naturally occur. They are regarded as "responsive" because they take into account and value the positions of multiple audiences. These studies tend to be more extensive (not necessarily centered on numerical data), more naturalistic (based on program activity rather than program intent), and more adaptable (not constrained by experimental or preordained designs). A major feature of many naturalistic studies is the observer who collects, filters, and organizes the information; this person's biases (both for and against the program) can have an impact on the outcome(s) of the study. Naturalistic inquiry differs from surveys and experimental or quasi-experimental designs in that usually a relatively small num-

ber of learners are studied in greater depth. And, while average numbers and percentages may be reported for descriptive purposes, generally statistical analyses are not performed in qualitative research.

Researchers involved in ethnographic research can (and some say "should") become so intimately involved in the community that they virtually loses their identify; Marshall and Rossman refer to this as the "individual lived experience" (1999, p. 61). The research itself takes a great deal of time, which must be followed by intensive review of observations, interviews, and so on. Once the data are available, there again must be much time in order to generate the theories and supporting findings. While these are not true experimental studies, they are detailed, time-consuming, and can provide a great deal of valuable information. Thus qualitative research can meet nearly all of the statutory provisions of the NCLB legislation.

Statistical Considerations

When developing a research study, there are several elements that must be considered in conjunction with sampling technique, research design, and how long the research will follow the subjects and collect data. How the study is designed will have an effect on the generalizations that can be made, how believable the results are, and how much weight we give the findings in the overall "picture" of this area of study. Some of the considerations include the reliability and validity of the study's data, the type(s) of analyses used, and the generalizations made.

Reliable and valid data

Reliability and validity refer to two distinct areas of the study:

1. the assessment used to collect data and
2. the overall "goodness" of the study.

The former is discussed at greater length in the NCELA document *Definitions for No Child Left Behind: Assessment.* Suffice it to say here that assessments are valid when they actually measure what they purport to measure and the results can be used legitimately to make decisions about students (e.g., program placement); assessments are reliable when a test's scores are consistent across time and within groups of students.

When considering a research study, *reliability* refers to whether the study could be replicated with similar results and *validity* refers to whether the study accurately reflects what is being studied. Both reliability and validity are somewhat subjective in this context, but are highly desirable for any study to be considered important. Some of the features, or elements, of a valid and reliable study include:

- research questions that are understandable and backed by solid theory;
- a well defined population to be studied;
- a relatively large sample selected randomly from the population;
- background information showing that the sample matches, or reflects, the population;
- multiple data collection instruments (assessments, observations, logs, and so on) that are technically sound (see NCELA's publication *Definitions for No Child Left Behind: Assessment*);
- a long-term, in-depth study;
- appropriate data analyses;
- realistic interpretations of data;
- generalizations that are realistic and appropriate;
- findings that build upon previous research; and
- a well-written report.

All of these elements are typical of studies that are scientifically-based.

Rigorous data analyses

Analyses form an important element of all studies. Analyses that are done poorly or incorrectly can invalidate an otherwise strong study. Analyses that are done well but are described poorly can make it difficult or impossible for readers to understand the results.

Qualitative research, though generally thought of as nonmathematical, still includes numbers. Qualitative research should provide information about the numbers of subjects involved; high, low,

and average numbers for any event that can be counted; and percentages of people who responded to surveys and other instruments.

The data for qualitative research often includes open-ended statements and answers from participants or lists of observations made by the researcher. Rather than report all of these verbatim, qualitative analyses consist of categorizing the information, then reporting the number of responses that fit into each category and some of the more typical or unusual responses. As a follow-up, how these responses support and expand previous research in the area also is important to discuss as well as how these data support and expand theories in the area.

Quantitative research involves mathematical analyses. Some of the analyses will be descriptive, providing background or supporting information about the subjects and the context of the study. Following this, statistical analyses are performed. The statistical analyses are designed to help answer the research questions. For instance, the research question for our reading study is "Does the reading comprehension of K-7 students differ when taught with Method A as opposed to Method Z, when measured by performance-based assessment(s) and the SAT-9?" Just looking at the data and seeing that Method Z students' scores are higher than Method A students' scores is not enough to answer the question. In statistical terminology, we must determine whether the scores are "significantly" different and whether the difference is large enough to be considered "important." Whether or not the scores are significantly different is based on the size of the sample, the difference in the scores, how homogenous the scores for the two groups of students are, the type of analyses performed, and the degree of significance we decide will be large enough to impress us. Given these factors, the Method A test scores and the Method Z test scores may be different, but the difference may not be large enough to be statistically significant; if the scores are not statistically significant, the students using both curricula are said to perform equally.

A peculiarity with quantitative analyses is that if the sample is large and/or the scores within each group are homogeneous, a small difference in average scores for each group may be mathematically significant; if the sample is small and/or the scores are heterogeneous, a large difference in average scores for each group may not be mathematically significant. How can this be interpreted? A further mathematical outcome, one that many researchers do not calculate, is the *percent variance accounted for*, sometimes referred to as an *effect size*. To think about this outcome, let us consider our example. We want to explain reading achievement. If we could explain all the causes of reading achievement, we would know everything, 100%, about reading achievement. It is unlikely we will ever know everything about reading achievement because of the number of events in our lives that can have an effect on our reading abilities. For instance, a partial list of things that might affect reading ability includes

- age when reading began,
- current age,
- gender,
- amount and type of reading material in the home,
- amount that parents and other family members read,
- difficulty of the reading material,
- the language of the reading material and assessment (a child who speaks English will not perform well on a French reading test after one semester of French),
- topics of the reading material (i.e., if a child is uninterested in a topic, his/her reading ability will test lower than if the topic is of interest),
- type of assessment (standardized, norm-referenced or performance-based),

- the technical qualities of the assessment used to measure reading comprehension, and
- many other known and unknown variables.

In general, an effect size of 10% or greater, indicating that we can explain 10% of the causes of what we're studying (e.g., reading comprehension), is considered quite good – regardless of whether the statistical analyses are significant.

Justifying general conclusions

General conclusions for any study must be developed carefully and must consider

- the scope of the research,
- the subjects who participated,
- the subjects who declined to participate or dropped out of the study,
- the type(s) of data collected,
- the statistical outcome(s), and
- whether the study supports the findings of other studies in the area (and if not, why not).

For instance, our study of reading comprehension involved K-7 students whose home languages were English, Spanish, and a few Southeast Asian languages. We would have trouble generalizing to 8th grade students or those whose home languages were from the former Soviet Union.

Regardless of the type of research, there is one common mistake that many people make: they refer to research as "proving" their stance. For instance, they say that the "research proved that Method A is better than Method Z." This is not, and cannot be, the case. Research can "support," "indicate," and "find," but it cannot "prove" because there are too many extenuating circumstances that could come to play. Note, for instance, the number of published studies on what appear to be the same topic but that have different, and often contradictory, results.

Synthesizing Research

When a new study is under consideration, or when decisions must be made about educational programs, it often is helpful to synthesize past research in order to guide the decisions that must be made. For instance, suppose a professor is called upon to testify before the local School Board as to whether smaller classrooms (i.e., fewer students in each classroom) are worth the additional cost of more teachers and more physical space. Or perhaps a policymaker faces the challenge of restructuring the public schools. A Department of Education program officer is asked to describe the success of different language instruction educational programs. How do they determine what is the "right" course of action? According to Hunter, Schmidt, and Jackson, there are "two steps to the cumulation of knowledge: (1) the cumulation of results across studies to establish facts and (2) the formation of theories to place the facts into a coherent and useful form" (1982, p 10). The "cumulation of results" is a determination of an overall pattern of results from earlier studies, evaluations, research, and projects. Further, there is an underlying assumption that new studies, or projects, will incorporate and improve upon the lessons learned in earlier work. The synthesis is the intermediate step between past and future work. The professor, policymaker, and program officer will rely upon such evidence to make their decisions about education in the future. Generally, these individuals will rely on one of three methods to synthesize the findings of studies, research, and evaluation.

Narrative review

In the ***narrative review***, the researcher collects published and unpublished studies, then provides an overall description of the findings. Frequently, this amounts to combining the conclusions sections of various studies. The resulting information is presented serially, with an overall synopsis. For instance, "Study A indicates that … Study B found that … Study C identified … [and so on to] Study Y reported … In summation, …" Rarely are critical comments made in a narrative review.

The narrative review is fairly easy to do. The researcher usually chooses only some of the many possible studies to review. The researcher must have enough statistical knowledge to have an overall feel for what each study purports to find, and to determine whether this finding is realistic.

There are several disadvantages to the narrative review. The primary one is that the review is fairly subjective with the outcomes highly dependent on the selection of studies included in the review. There are few formal rules for doing a narrative review. It also is an inefficient way to extract useful information, especially when the number of studies being reviewed is large. Finally, it may be difficult for the researcher to juggle the relationships among studies, the findings in each study, and to develop a meaningful synthesis.

As an example, Munsinger examined a group of studies on children who were adopted and concluded that environmental effects are small. As he stated, "Available data suggest that under existing circumstances heredity is much more important than environment in producing individual differences in IQ;" later Kamin reviewed the *same* set of studies and reached the *opposite conclusion* (in Light & Pillemer, 1984).

Vote-counting method

The ***vote-counting method*** is a refinement of the narrative review. In this case, the researcher collects the studies, then lists each study and its conclusions. The results are then tabulated as *positive result, no result,* or *negative result*. For instance, if the researcher is looking at studies of differences in achievement between students using a new reading curriculum and students using the reading curriculum that has been used for the last five years, the results might indicate that achievement scores are *higher in the new curriculum (Method Z) classroom, higher in the old curriculum (Method A) classroom,* or that there are *no differences between the two curricula used*. The researcher then counts the number of studies supporting each of the three possible outcomes and

chooses as "successful" the outcome (new curriculum better, old curriculum better, no difference) with the most "votes."

The vote-counting review generally is easy to do and provides an overall picture of the results for several studies. The vote-counting method also allows for the combination of different foci of various studies. On the other hand, this method does not consider some of the important research features such as how great the differences were between the groups, whether the design of the study was appropriate, how many participants were involved, and so on. This method also still is dependent on which studies are selected for review.

Meta-analysis

Meta-analysis often is referred to as the "analysis of analyses" or the "statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating findings." In addition to looking at the actual numeric value of specific results (e.g., what was the actual difference in the average score between students taught using a new reading curriculum, Method Z, and those using the older curriculum, Method A?), meta-analysis allows the researcher to analyze statistically such features as the "goodness" of the research design (e.g., was this a true experimental design or a correlational study?, were data longitudinal or cross-sectional?), assessment instruments (e.g., how old?, reliable?), and the size of the sample (20 students or 20 classrooms of students); the size of the differences in outcomes between groups of participants; the year of the study, the type of study (published, dissertation, etc.); and other features of the overall study.

When creating a meta-analysis, the effect size is important. If the original research did not report one, the meta-analysis researcher can use the data provided to calculate (or estimate) the effect size. While individual studies look for differences in test scores based on characteristics of the group(s) studied, a meta-analysis looks for differences in effect sizes based on characteristics of the group(s) in the original studies and the characteristics of the studies themselves.

Meta-analysis allows the researcher to consider many facets of each study being reviewed. It is very objective; each study is, in essence, a "participant" in the research project. Indeed, meta-analyses are considered to be research projects unto themselves and require a great deal of time and energy. On the negative side, there still is some dependence on the studies selected for inclusion, but now the choices must be specifically defended and explained within the methodology of the meta-analysis. The researcher must have a great deal of research and statistical knowledge and abilities. Some of the statistical procedures still are being refined and developed and, because actual statistical numbers are used in the review, no qualitative studies can be included.

As an example, Baker and de Kanter (1981) reviewed 28 studies on "bilingual education" using the narrative review method. They concluded that the case for bilingual education was weak. Willig (1985) conducted a meta-analysis of the same studies, eliminating 5 because they studied programs that did not fit her definitions of "bilingual education." When statistical controls for methodological inadequacies were employed, participation in bilingual education programs consistently produced small to moderate effect sizes favoring bilingual education in the areas of reading, language, mathematics, writing, social studies, listening comprehension, and attitudes toward school or self. Programs characterized by instability and/or hostile environments showed lower effects. More recently, Slavin and Cheung (2003) performed a synthesis of experimental research focused on the issues of policy and practice regarding the language in which a student is first taught to read (native language or English) and the instructional strategies used (bilingual or English-only). They selected studies that dated back to 1976, based on whether they used true experimental designs and found strong support for (1) using bilingual strategies and techniques and (2) teaching children to read in their home language in order to enhance English reading achievement.

Conclusions

Under NCLB, the focus is on well-designed research studies, with somewhat of a preference for true experimental design, also referred to as a randomized controlled trial design. Other types of research also are acceptable although with varying support. Those writing many types of grants now are required to provide research support for the program(s) they intend to implement. Several of the grant applications suggest that the following criterion be used, with the designators of IA (the highest level, "best" research) to IV (the lowest level research that only carries weight when other types of research also support the findings). Exhibit 6 provides the criteria for each of these types of studies. As an example, a IA study, the strongest type of research, can be described as

- a randomized controlled design – random selection and random assignment of participants with as much control of other factors as possible, with
- statistically significant positive effects – statistical analyses show a significant increase in participants' knowledge or skills that can be attributed to the "treatment" or educational program,
- positive effect sustained for at least one year post intervention – with further testing a year later (and no further "treatment") still showing the increase in participants' knowledge or skills, and
- post effect replicated in one or more settings and/or populations – with other studies now completed and showing similar findings.

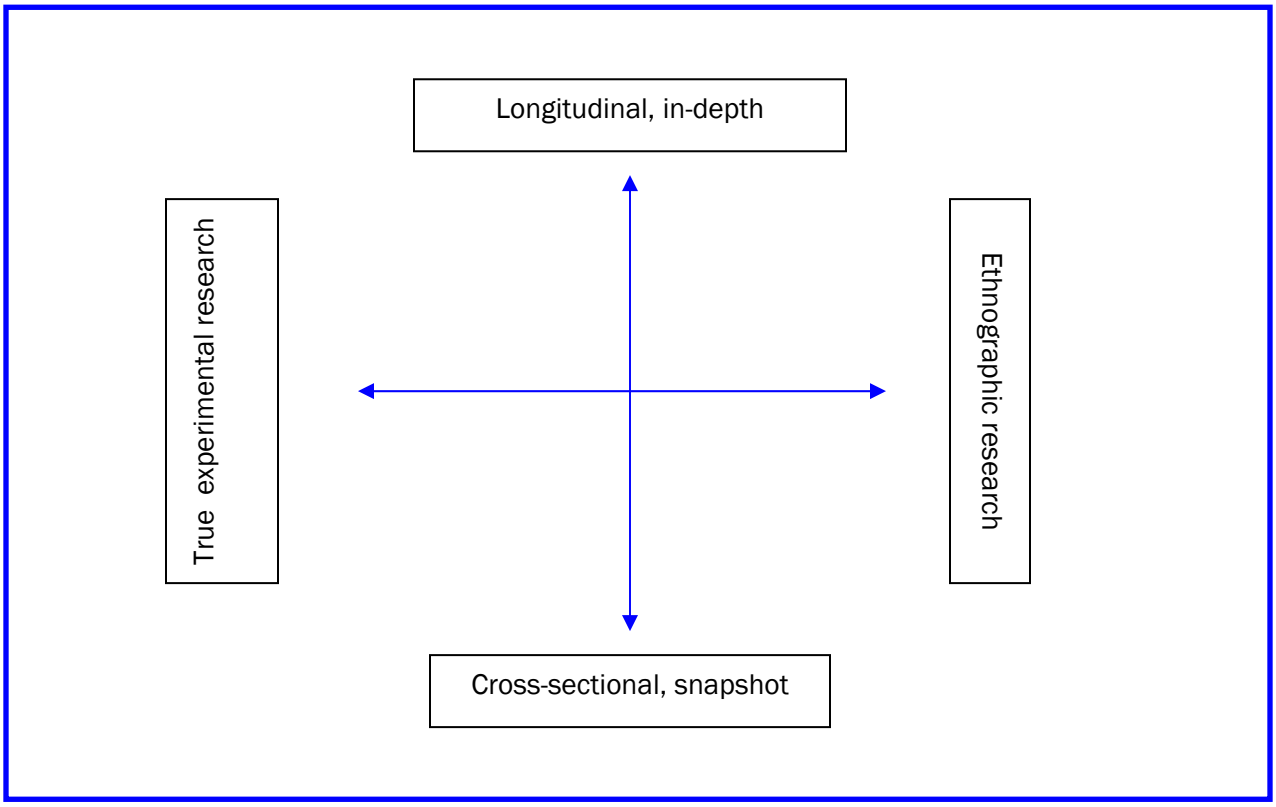
Exhibit 6: Types of scientific research and levels of evidence

Criterion	I-A	I-B	II-A	II-B	III	IV
Randomized controlled trial design	X					
Quasi-experimental controlled design		X	X	X		
Statistically significant positive effect	X	X	X	X	X	
Positive effect sustained for at least 1 year post intervention	X	X		X		
Post effect replicated in 1 or more settings and/or populations	X	X	X		X	
Opinions of respected authorities						X

There are differences among these types of research just as there are differences among different types of courtroom evidence. It is much easier to convict a defendant based on eyewitness testimony ("I saw the defendant commit the crime, can identify the defendant from among several similar individuals, and can say with conviction 's/he did it'") than to convict a defendant based solely on circumstantial evidence ("we can show that the defendant was in the area at the time, disliked the deceased, and owned a gun"). The analogy holds with type IA research being similar to having specific witnesses to the events and type IV research being similar to having only circumstantial evidence.

It also may be helpful to think of these research designs not on a "best" to "worst" scale, but on a scale based on the type of data, of evidence, that are collected. In addition, it is important to note the depth of the study – longitudinal, long-term, in-depth data collection as opposed to one-shot, quick, more surface-like data collection. Exhibit 7 is considered in more detail in the NCELA document *Definitions for No Child Left Behind: Research and Evaluation that Meets NCLB Mandates*. It shows a two-dimensional figure that involves two continua: one for the type of research and one for the length and depth of the research. The "best" research is that which can be charted above the horizontal line: in-depth research that carefully follows a solid research design that is either true experimental or true ethnographic. "Best" on the horizontal axis is based on the needs of the research and the end purpose of the research.

Exhibit 7: Research design matrix



Summary

The purpose of this document has been to provide definitions for terms and techniques used in educational research. It is not meant to be judgemental, but to provide these definitions for those who wish to understand the current mandates and guidances of the federal Department of Education but who have little knowledge of research techniques. For further information, see the two other NCELA documents that have been referenced throughout this documents; one provides more details on assessment and one provides further information on how to read, understand, and judge research. In addition, there are many good statistical and research design textbooks that may be helpful.

Reference llst

Baker, K. & de Kanter, A. (1981). *Effectiveness of bilingual education: A review of the literature.* Washington, DC: US Department of Education.

Feuer, M.J.; Towne, L.; & Shavelson, R.J. (2002). Scientific culture and educational research. *Educational Research, 31*(8), 4-14.

Hunter, J.E.; Schmidt, F.L.; & Jackson, G.B. (1982). *Meta-Analysis: Cumulating Research Findings across Studies.* Beverly Hills: Sage.

Light, R.J.; & Pillemer, D.B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University.

Marshall, C, & Rossman, G.B. (1999). *Designing Qualitative Research,* 3rd ed. Thousand Oaks, CA: Sage.

Slavin, R.E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21.

Slavin, R.E. & Cheung, A. (2003). *Synthesis of research on beginning reading programs for English language learners.* Report for the Institute of Educational Sciences, US Department of Education (Grant #OERI-R-117-40005).

Willig, A.C. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55,* 269-317.