

Guidelines for Evaluating Research and Research Claims

Martin Kozloff

Here's a big idea.

When research produces **valid** information (findings), or when journal articles that advocate a certain educational method are based on **valid** information (findings), then the **findings** and the **claims** made by authors (e.g., that a method **works** or that teachers should use the method) are **credible** (believable). And when findings and claims are credible, they may help to make **educational decisions**; for instance, to TRY an innovative method.

But when research does **not** produce valid information (findings), or when journal articles that advocate a certain educational method are **not** based on valid information (findings), then the findings and the claims made by authors (e.g., that a method works or that teachers should use the method) are **not** credible (not believable). And when findings and claims are not credible, they should **NOT be used** to make educational decisions; for instance, to TRY an innovative method. To use (on many school children) an innovation that is not backed up by a **great deal of valid scientific research, is as unethical as using (on many children) a medication that is untested.**

Please make sure to look up in the "Vocabulary" the terms that are unfamiliar to you.

Following are some things to look for when you examine research and publications. These are the things that contribute to validity.

1. What is the **purpose** of the research study or of the journal article that advocates an education method? There are at least four possibilities.

a. **Persuasion. Nonresearch Claims.** The intention is to advocate a method or innovation; to persuade readers to believe what the writer believes. References to other persons' work are used to support the writers' beliefs, but there are few references to work that **challenges** the writer's beliefs. Concepts (variables) are not precisely defined. The writer may call what he or she did "research," but *what the writer really did was collect information that supported what he or she already believed*. Also, the writer interpreted the "data" ONLY in a way that supported what he or she believed. *There are no controls on the collection or analysis or interpretation of the data*; that is, no one else observed the same thing and interpreted the data so that the extent of agreement between different observers could be determined. *Too few persons (e.g., students) were in the sample*. Therefore, there is no way to tell if the "findings" would have applied to students and schools outside of the few students the researcher studied. *In other words, the "research" or the article is biased to support one set of beliefs without TESTING them.* This is NOT scientific research.

b. **Basic research and Pilot-tests.** This is *level 1 research*, as described by Grossen in "What Does it Mean to be a Research-Based Profession?"

(1) In **basic research**, the researcher is NOT trying to prove that his or her beliefs are correct. Rather, based on past research, the researcher may have a hunch about how things **work** or about how things are **connected** (for example, that a teaching method improves learning), and the researcher wants to **check to see if the hunch has any merit**---any support.

So, the researcher might collect data (field observations, achievement scores) under specific conditions (e.g., when teachers use a certain method, such as peer tutoring) to **see IF there is any**

relationship between using peer tutoring and student learning. The **observations are done carefully**. **Quantitative** data (e.g., scores on achievement tests) are analyzed carefully. This basic research may not be an experiment (with different groups) but it is serious research.

Perhaps the researcher **does** find that when teachers use peer tutoring (for example, more advanced students are taught to use a certain instructional routine to coach less advanced students), the less advanced students make significant progress. However, the **researcher is cautious when drawing conclusions**. She does NOT advocate peer tutoring. She concludes that peer tutoring is worthy of a **more formal TEST**.

- (2) **Pilot tests often follow basic research**. The pilot test begins with a **research hypothesis**---a statement of what the researcher thinks will happen. “Students who receive peer tutoring (the intervention) will make significantly higher gains in math achievement (the outcome) than students who receive the usual remedial instruction.” *But the researcher does NOT collect data merely to **show** that the research hypothesis is supported*. Instead, **the researcher tries to show that the research hypothesis is FALSE!** The researcher states the **null hypothesis**: “Students who receive peer tutoring (the intervention) will NOT make significantly higher gains in math achievement (the outcome) than students who receive the usual remedial instruction.” *The researcher then designs the research so that data may show that **indeed the null hypothesis is supported**---*students who received peer tutoring did NOT make higher gains than students who received the usual remedial instruction. *If the NULL hypothesis is supported, then the RESEARCH hypothesis is FALSE*. This kind of test (collect data to see if the null hypothesis is supported and the research hypothesis is false) would help to ensure that students are

not harmed by faddish innovations that were never tested to see if they did NOT work.

The researcher working on peer tutoring might conduct a *small* experiment with **20 students selected at random** (*random sampling*) from the three third grade classes in an elementary school. The **20 students are randomly assigned to two groups:** peer tutoring (*experimental group*) and the usual remedial instruction provided by the teachers (*control group*). Because students were selected at random from the three classes, the students in the study are likely to be a **small version of the entire three classes**. That is, about the same distribution by sex, ethnic group, math ability, parental support, age. Also, because the 20 students were randomly **assigned** to the two groups, the **two groups are likely to be very similar in composition** (sex, ethnic groups, math ability, parental support, age). If the two groups are similar on all these *extraneous factors* (which could influence math achievement), and if the **ONLY** big difference between the two groups is that one received peer tutoring and the other received the usual remedial instruction, then (logically), if the peer tutoring group makes significantly higher gains in math achievement, this difference is likely to be the result of the peer tutoring---since that is the only way that the two groups were different (that the researcher knows of).

The researcher then uses **standardized and validated instruments to pre-test the students in the two groups** to see what their math achievement is so far. [In other words, the researcher carefully **DEFINES** and **MEASURES** the outcome variable---math achievement.] **Two persons examine the test data** to ensure that the tests were scored **accurately**. [In other words, the researcher checks the **reliability** (*sameness*) of the scores.]

Next, 20 high-achieving students from upper grades are selected and trained to follow an instructional routine to assess math skills, provide tutoring, and to monitor weekly progress. These peer tutors must demonstrate proficiency in the routine before they are allowed to tutor. [In other words, the researcher carefully **DEFINES** and **MEASURES** the intervention variable---peer tutoring.]

Then the two groups receive instruction---peer tutoring vs. usual remedial instruction. *The peer tutors are observed to ensure that they are using the tutoring protocol as they were trained.* After two months, the two groups are given a different version of the same standardized math test (**a post-test**). Data from the post-tests (where the students ended up) are compared with data from the pre-tests (where the students started) to see if there are any **significant differences** in the achievement gains (the outcome variable) between the experimental and control groups.

If there are significant differences in gains, the researcher draws the **cautious conclusion** that peer tutoring **MAY** be useful, and **MAY** be more effective than the usual remedial instruction. However, the researcher says that the research **MUST** be **replicated** (done again) several times before anyone can be confident that peer tutoring is useful and more effective.

- c. ***Demonstration research and Replication Research.*** This is **level 2 research**. It is usually conducted in a more natural setting, such as a whole classroom.
- (1) In **replication research**, the researcher who did a pilot test that showed that peer tutoring yielded significantly higher math gains than the usual remedial instruction), **does the research *again*, in a different setting (a whole class) and with different students.** Or, other researchers who learned of the pilot test do the replication research. The purpose is to see if the results of the pilot test were a fluke (wouldn't happen again) or were limited to students in an

experiment (which by itself makes students feel special and may affect achievement) or to the kinds of students in the sample but not in the larger school or district population.

The replications must be done as carefully (scientifically) as the pilot test. The more often the replication studies yield the same findings, the more confident researchers and readers can be of the findings.

(2) **Demonstration research** might be done after a series of replication studies. With each replication study, researchers find something new. For example, they might find that peer tutoring works best when boys are tutored by either boys or girls, but that girls are best tutored by girls. Another replication study may show that it is better to have tutoring sessions at least three times per week. And another replication study may show that it is best to keep the sessions under 20 minutes. Using all of these findings, researchers may put the whole thing together into a ***whole program and show (demonstrate) how it works in a class***. This research must have the same rigor (the same features) as the pilot and replication research.

d. **Larger-scale Evaluation Research.** This is level 3 research. This research should follow pilot, replication, and demonstration research. *It would be unethical to use on a large scale a method or innovation that had not been thoroughly evaluated on a large scale.* An example of level 3 research would be evaluating the peer tutoring program in a whole school and at different schools in a whole district. ***The school and district provide a larger and more diverse sample of students, teachers, administrators, and school organization.*** Therefore, the evaluation research may find that peer tutoring works best under certain conditions; e.g., schools where the principal is “on top of things” and makes sure all projects are done according to protocol;

with teachers who are already proficient at highly focused and explicit instruction.

Model: “Watch me...”

Lead: “Now do it with me.”

Test/check: “Your turn.”

Additional Things to Look for

2. Is the research question, research hypothesis, problem, or interest stated precisely enough that things to measure **objectively** can be derived from it?

Poor. “Do students learn more when tasks are **authentic**? [What does ‘authentic’ mean? What does ‘learn more’ mean? Amount? Speed? In other words, the terms (concepts, variables) are vague. This means that you can’t tell if the researchers used valid measures, because you don’t even know **WHAT** they are measuring. If an article begins with vague ideas, you might as well not bother reading any more of it.]

Better. “Is there a difference in the number of learning trials required for students to correctly decode words that are in common use (such as fun, car, cat, is, run) vs. pseudowords (such as fis, nif, ris)?”

3. Did the article or research review provide extensive coverage of the topic at hand, or only narrow and perhaps biased coverage?

Poor. A new theory of reading is presented. The only literature presented was written by the author and by persons sympathetic to the author’s position---that a new theory is needed. The author criticizes other approaches to reading instruction (except his own), but presents none of the extensive research that **DOES**

support other approaches and/or that criticizes the author's. In other words, the author is misleading readers into thinking that his approach is the best one.

Better. A new theory of reading is presented. The author identifies the major **other** approaches to reading instruction, and presents a representative sample of the research on each approach—both research that supports and that challenges each approach. The author presents research that supports and that challenges his own approach. The author then identifies **gaps** in the research--- unanswered questions---and says that his research is designed to answer these.

4. Did the literature review examine literature that is both directly relevant to the topic at hand **and** larger relevant issues? For example, did a literature review on effective mathematics instruction include mathematics research **and** research on instruction in general?

Poor. The literature review focuses narrowly on the topic.

Better. The literature review focuses on the topic and other relevant issues. Therefore, the reader can see that the topic at hand is important on both a small and larger scale.

5. Did the writer summarize the literature with a set of generalizations stated as propositions and facts, and diagrams? Did the writer identify possible gaps in what is known; e.g., additional factors (besides instruction) that affect motivation.

Poor. The literature review is merely a **cascade of citations** that gives the appearance of serious intent to conduct unbiased work. The literature does not lead obviously to particular research questions.

Better. The literature review ends with a summary of what is known, partly known but not confirmed, and what is unknown. These findings are stated so that it is clear **exactly what the writer is talking about.**

Not good. “The more **engaged** students are the more they **participate.**”
[Engaged and participate mean the same thing. So how can engagement—which MEANS participation---CAUSE participation?]

Good. “The higher the rate of opportunities to respond, the longer is student’s attention during instruction.”

6. Is the project do-able? Can sufficient data be collected and analyzed?

Poor. “We will determine **the most effective** method for teaching comprehension.” [It is impossible to determine what IS most effective because you cannot study everything that IS and will be.]

Better. “We plan to determine **which of three methods** for teaching comprehension is **associated** with the highest gains in scores of comprehension on the Especially Fine Test of Reading Comprehension.”

Poor. “Our objective is to determine whether violent images in the media **cause** violent behavior in schools. [You may find out if students who commit more violent acts in school **also watch** more violent events on TV, but you can never learn if seeing those events **causes** the violence. Besides, there are so many other factors in students’ lives that you could never study **all the possible “causes” of violence.**]

Better. “Our objective is to determine whether students who commit more violent acts in school **also watch** more violent events on TV.”

7. Is the design of the research (survey, experiment, field observations) appropriate to the purpose? [Please see number 1.]
- a. **Opinions.** If you want to find out what people think and feel, ask them---using interviews or questionnaires. This information may be useful in making certain decisions. For example, teachers using a new reading program may say that they need more assistance. **But opinions cannot be used to judge whether a program is effective.** Effectiveness should be measured in terms of student achievement by teachers who used the program properly.
 - b. **Effectiveness, What Works.** Let’s say the question has to do with which method works better (is associated with better outcomes), or whether some new method works at all. For example, is one reading program more effective than another? Does a new math program produce skilled math students? What kind of research is proper?

Poor. *Do a survey of teachers.* Ask their opinions. [All you are measuring is teacher perception. You SHOULD measure student achievement. And that requires that you measure what students DO, not merely what teachers think. Would you give a new medication to thousands of children just because 10 physicians said “We think it works”?]

Poor. *Have teachers make up their own instruments to measure progress and outcomes.* [How will you know whether these instruments measure what they are supposed to measure? Teachers may make instruments that are so vague, and have such easy criteria for passing, that children who have NOT learned

much will APPEAR to have learned a lot. Then, principals might adopt the new method and find out that their students don't learn. *Instruments and measurements should be rigorous.* Would you give a new medication to thousands of children based on instruments that were **never validated**; that is, you don't know if the instruments actually measure what they are supposed to measure?]

Poor. *Conduct observations in classes.* Observe student-teacher interaction and students working together and individually. [This may be interesting information, but it does NOT tell you WHAT math or reading skills students learned. Therefore, it doesn't say anything about effectiveness. Would you give a medication to thousands of children based on observations of how persons TOOK the medication, but not on whether it worked?]

Poor. *Use instruments that do NOT DIRECTLY measure the skills that define math achievement.* For example, measure students' **attitudes** towards math, or how often they do math on their own. [If you are testing whether an innovation is effective (that is, it yields beneficial behavior change) you have to measure **behavior CHANGE**, and not just whether students like it. Liking a math program is an important variable, but change (effectiveness) is the most important. Would you give a medication to thousands of children based on other children saying they liked how it tasted—but without knowing whether it worked?]

Here is what you DO want to see in research that has to do with what works and what doesn't.

- a. Variables (e.g., reading; reading achievement) are clearly defined so (1) it is clear what is being talked about; and (2) VALID measures can be selected.

Conceptual definitions should cover all relevant features of a concept (or variable), and should exclude irrelevant features. [See Vocabulary.]

Poor. Reading is defined as a psycholinguistic guess game. [Therefore, researchers will measure student guessing as evidence of good reading. Is guessing what words say an example of reading?]

Poor. Reading is defined as making sense of text--comprehension. [Therefore, researchers will only measure comprehension. If a teacher is not yet working on comprehension (but IS working on **other** reading skills) the teachers' students will by that narrow definition be considered nonreaders.]

Better: Reading is a cognitive routine for accurately and rapidly decoding written text into words and connected statements, and then comprehending the definitions and propositions communicated by the text. [This conceptual definition covers all the reading skills (the correspondence between letters and sounds, sounding out words, reading words and sentences fluently, and knowing vocabulary and comprehension strategies). Research based on this definition would be obliged either to study **all** of what is meant by reading or to **explicitly limit** the research to certain subskills.]

Operational definitions (*examples* of the concepts or variables) should be derived from conceptual definitions. In addition,

operational definitions should provide clear examples that cover the range of what is implied by the conceptual definition, and operational definitions should exclude what is not relevant. [See Vocabulary.]

Poor. Reading is conceptually defined as a cognitive routine for accurately and rapidly decoding written text into words and connected statements, and then comprehending the definitions and propositions communicated by the text. However, *the operational definition of reading includes how children handle books* (upright, turn pages), name the parts of a book, and recognize environmental print. These may be important behaviors, but they are NOT part of the conceptual definition of reading. Imagine if children's book handling were measured, and most children did well. By this definition, children who can't read a single word would be considered good readers.

Again, the operational definition must give examples that are consistent with the conceptual definition, or else *you will be collecting data on something else*.

Better. The researchers state that they will **not** measure **all** aspects of reading. They are interested (in this study) only in students accurately and rapidly decoding words. **They operationally define decoding words** this way: "By decoding words, we mean the student (1) says each sound in a word, (2) does not stop between the sounds, and (3) says the word as a unit (blends sounds into a whole)." The variable (decoding) is operationally defined **clearly** enough and **precisely** enough that you can easily think of how to **measure** it. Just show students words from a list, and score whether for **each** word they do the three things in the

operational definition. You will end up knowing how many students read 100%, 90%, 80%, etc., of the set correctly.

- b. **Measures should be objective** (something *anyone can see and hear*) and *measures should be derived from the operational definitions*.

Poor measures. Students' "enjoyment of reading" is measured by how many times they smile as they read. Students' "appreciation of literary genres" is measured by saying they like poetry but don't like plays. Smiling and stating preferences are not objective measures. You can't always tell what a smile means. You can't be sure what the word "like" means.

Better measures.

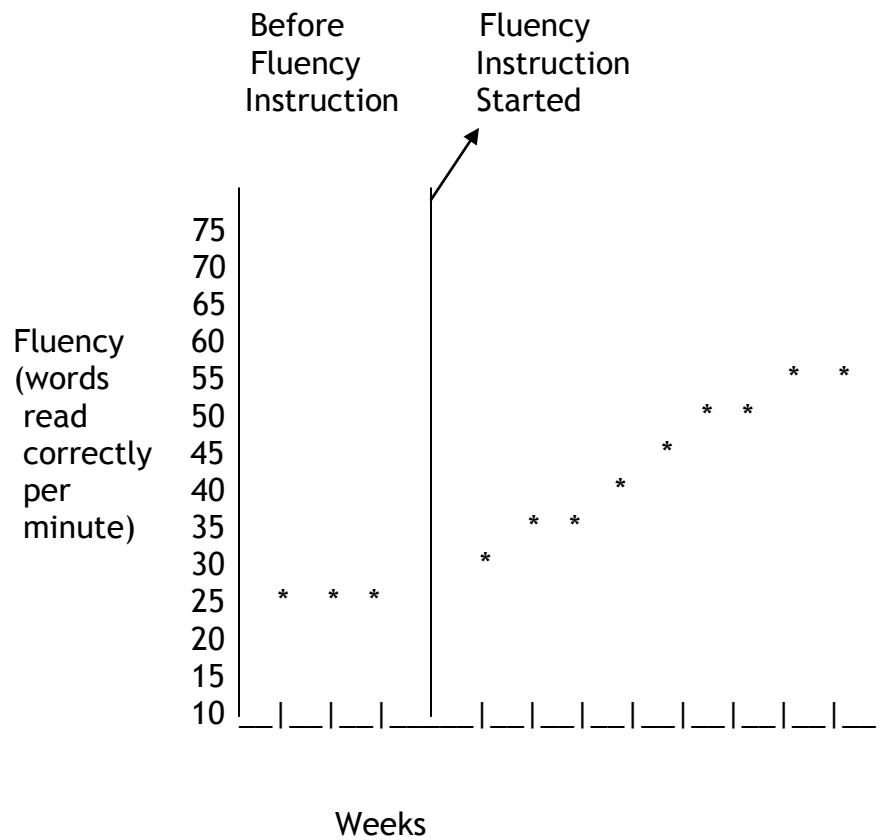
- (1) Enjoyment of reading might be measured (indirectly) by HOW MANY books or how much time students read on their own.
- (2) Students' appreciation of literacy genres might be measured objectively by how many samples of literacy genres (fiction, biography, poetry) students correctly name.

- c. **When possible, outcome variables should have multiple measures.**

For example, changes in math skill might include (1) accuracy (percentage correct); (2) fluency (number correct per minute); (3) percentile ranking in a population; (4) retention (percentage correct several months after instruction. If all of these measures signify high skill then the researcher and reader can be more confident that students' skills were in fact high. This is called **triangulation**.

- d. **Causal time order.** The researcher should have evidence that any **change in the outcome variables** (the expected effect or result; e.g., change in achievement) **came after change in the intervention variables** (e.g., an innovation in instruction). Otherwise, students

might have begun to make achievement gains **before** the new method was introduced. This requires a *pre-test* to see how much students know **BEFORE** the new method is introduced.



- d. The researcher should collect data on the outcome variables (e.g., math achievement) *and* the intervention variables (e.g., how proficiently tutors used the peer tutoring protocol). Otherwise, if data show that achievement with peer tutoring is *low*, researchers might conclude (wrongly) that peer tutoring is *ineffective* when in fact the tutors did not tutor properly.

Imagine that you are testing a new medication (intervention) that is supposed to lower blood pressure (the outcome). But it does NOT seem to work on 50% of the persons in the study. Do you conclude that it does NOT have reliable effects? But what if most of the persons

whose blood pressure did NOT go down, did NOT take the medication properly? This shows you that you have to measure the intervention (did teachers use the new method correctly) AND the expected outcome.

- e. **The sample (of locations and participants) should be representative of the population to which you may want to generalize the findings.** However, a pilot project (to see if a new method has any merit---level 1 research) may have a small sample. [If a method may NOT work, you would not want to try it with a lot of persons.] But research on program effectiveness---level 3 research (that could be used by many principals and districts to select curriculum materials) should have large and varied samples, and should be replicated (done many times) to try to ensure that the results are reliable (not chance and not applicable only to special settings and persons).
- f. Researchers should describe the **sampling plan**. They should describe:
 - (1) The **population** for which the research is **relevant**: what kinds of students, schools, families, communities, with certain **characteristics**?
 - (2) The **population pool** from which the **sample will be drawn**. Researchers seldom draw a sample from the whole population of persons, classes, schools, or districts. The pool may be defined by certain **characteristics**. For example, the **population** for which the research is **relevant** may be **all** students, schools, and districts, but the **pool** (the part of the population) from which the sample is drawn may be children from poor communities. The idea is that if the new method works with these children, it is almost certain to work with children in affluent communities.
 - (3) **How the sample will be drawn from the pool**. For example, will

persons, classes, schools or districts be selected at random from a list of persons, classes, schools or districts in the pool (**simple random sampling**)? Will persons, classes, schools or districts be selected in a purposive way (**purposive sampling**)? For example, researchers make sure that important groups are included (poor, comfortable; rural, urban, large districts, small districts). are selected at random from a list of all persons, classes, schools or districts in the assigned pool. Will persons, classes, schools or districts be selected because they are convenient (**convenience sample**). For example, districts that are very willing to participate.

Each kind of sample has advantages and disadvantages. A simple random sample means that all characteristics in the pool (age, sex, ethnicity, talent) have a chance to be in the sample. However, since the sampling is random, characteristics that are infrequent (e.g., children of a small minority group) may not end up in the sample. A purposive sample can ensure that important groups are included, but the sampling is not random. There may be some bias in the selection. A convenience sample is easiest to obtain, but the sample may not be representative of the population.

- (4) If comparison groups will be used (as in an experiment), **how will participants be assigned to groups?** For example, will the researcher use: (a) **random allocation** (e.g., 50 participants are allocated to a 25 member experimental group vs. a 25 member control group based on a coin toss); (b) **matching** (e.g., you make sure that each group is matched on certain characteristics (e.g., each groups has equal proportions of males and females, religions, and other variables that you think are important)); (c) **existing groups** (e.g., one math class is the experimental group and the other is the control group).

- g. The researcher should report on the validity of: (1) instruments (Do they measure what they are supposed to? Have they been validated?); and(2) data collection (Is it clear how data were collected?)
- h. The researcher should consider the possible influence of other factors (called **extraneous variables**) that could accounted for at least part of the findings. For example, students may make large gains after a new math program is introduced, but these gains could in part be the result of:
- (1) The fact that many students in the group were especially talented. The results would not be as good with a more usual sample.
 - (2) Parents of some students gave math instruction at home.
 - (3) Test results were unusual. They were flukes. If tested again, students would not score as highly.
 - (4) Maturation accounts for some of the progress.
 - (5) Students who would NOT have benefited much from the program dropped out; or students who were not making progress were moved out. Because the low scores of these students were not included, the program looks more effective than it is.
 - (6) Teachers did not use the materials exactly as they were supposed to. They added instruction to make up for weaknesses. This made the program look more effective than it is.
- i. **Data analysis and conclusions.** Did the researcher describe how data were analyzed? Are the conclusions (“Our method is effective.”) justified and credible (can you be confident?) given the size of the sample, the variables considered, the definitions, and the instruments?

Guidelines for Research

1. The data must measure what they are supposed to measure. For example, skill at sounding out words (man mmmmaaaan) is measured by having students sound out a list of words, not by having students repeat words from memory. Repeating words from memory is not sounding out.
2. You must measure all of what you are supposed to measure. Reading involves five skills. You cannot measure just two of these skills and use those two measures to rate the quality of a school's **whole** reading program, any more than you can take a person's blood pressure and temperature and make statements about their overall health.
3. Measurement should be direct. What you want to measure must be seen or heard. It must be objective. For example, accurate reading of connected text (sentences) is objectively what students SAY. A direct measure of this might be counting the number of words that students correctly say (read) in a 100 word text.

Accurate reading of text is NOT objectively how students FEEL about how well they read. And a direct measure of accuracy is NOT how teachers RATE students' reading: very good satisfactory not good. Because these ratings are of how the TEACHER feels, not about the student's accuracy.

4. The samples with which a method or program is tested must have the same characteristics as the students with whom the method or program is later to be used.
5. Measures must be reliable. That is, they must yield the same result when done again or when done by different observers.
6. You must measure all subgroups. For example, you cannot measure math achievement of middle class children only, and then report on the achievement of the whole school.
7. You must not collect information on skill (for example) only at the end of using a method or curriculum (post-test only). This will not tell you if there was any change.
8. You must collect information on skill (for example) before, during, and at the end of using a method or curriculum (pre- and post-test). This is the only way to show change.
9. You must use experimental groups (that are using a new method, for example) and control groups (that are **not** using the new method) to see if the new method is associated with better results. Without a control group, you cannot say if the method did anything. A control group might have done just as well.
10. If you use control groups, they must be as similar as you can make them to the experimental groups. Otherwise, if the experimental groups and control groups are different (e.g., in social class, the amount of help families give children, children's background knowledge), then THESE factors may be

responsible for different outcomes in the experimental and control groups, and not the methods used.

11. There are two ways to try to make experimental and control groups as similar as possible. *Random allocation* means that you assign students or schools to the groups at random. Therefore, all factors that apply to each person or school have an equal chance of being in the groups. *Matching* means that you already know some ways you want the groups to be equivalent (for example, the ratio of boys and girls, or social class). So, you make sure that for every boy in the experimental group, there is a boy in the control group; for every student that is middle class in the experimental group, there is a student who is middle class in the control group.
12. Research that is used to test a method or program must be repeated (**replicated**) with the **same** kinds of groups (to see if the results are repeated or a fluke).
13. Research that is used to test a method or program must be repeated (**replicated**) with the **different kinds of groups** (e.g., social classes, ages, degrees of background knowledge) to see if THESE factors make a difference. For example, you may find out that a reading program does not work well with students who do not already know the sounds that go with the letters. Therefore, if you know what kinds of students you have, you can select the right program---the program that has been shown to work with your students.
14. Research that is used to test a method or program must be done over an extended period (**longitudinal**) to see how long it takes for there to be effects (if any) and how long effects last (if at all). If you plan to use a

program with children as they move from grades one to grade four, then the research that tested the program must be done with groups of students over those four years.

15. Research that is used to test a method or program must **use several measures of the same thing**. If all measures say much the same thing (“85% of the students now read on a grade six level”) then you can be pretty confident that the data are valid. But if only one measure says that 85% of the students now read on a grade six level, it could be that the one measure is wrong.
 16. The features of a program or method must be supported by strong research. [This is called “research based.”] For example, strong research shows that frequent, short practice and review sessions are important for students to become more firm on a skill and to retain it. So, if a program or method does NOT involve frequent, short practice and review, then it is NOT consistent with research.
 17. It is not enough for a program or method to have features that are supported by strong research. The program or method as a whole must be tested. [This is called “field testing” or “evaluation research.”]
-
1. The language of a claim (in a book, article, conference presentation, website) suggests accountability, honesty, precision. It appeals to your intellect – the language addresses serious questions that you have about effectiveness and

about how evidence was collected and analyzed, because you as a consumer and a person responsible for children in your care.

The language is not flowery and vague. It does not use feel-good words that distract you from the serious questions.

2. The claims are reasonable. They state limitations. They do not make promises; they do not sound too good to be true; they do not sound *exaggerated* and puffed up.
3. **The thing being tested** (e.g., curriculum) is **described fully**. You know exactly what's in it and how it is used. Also, the **effects of the thing being tested are defined clearly and concretely**. The words are understood and the words (definitions) POINT to things (effects?) that can be observed; that is, defined and measured *objectively*---that is, defined by what students DO (e.g., how many problems they solve correctly), not teachers' opinions, impressions, and perceptions.
4. Researchers look not only for beneficial effects but also for harmful effects. If you don't look for harmful effects, you won't find them.
5. **Research** was done in settings where the materials or methods (being tested) will be used---*empirical research*.
6. The research used **LARGE enough samples** of persons and classrooms and schools (not just 20 students in one class) so that the researchers could see if there were **different kinds and degrees of effects**. [You can't tell if an instructional method, for instance, has a whole **range** of effects if you test it with only five students. Five students can't show much **VARIATION** in effects.]
6. The research was done **again and again** (this is called *replication*), to see if the effects were **RELIBALE** (dependable, can be counted on)?
7. The replications of the research **used samples** of students and teachers, schools, and even districts, **that cover a range of differences that are found**

in the larger population (*representative samples*)---small and large schools, affluent and poor students, males and females, ethnic groups.

8. The research in SOME replications was done **over a long period** (*longitudinal research*) to see the long term effects.
9. The research was conducted (or the claims were made) by independent researchers or persons who had no stake in the outcomes.

In later units we will add crite

1. The language of a claim (in a book, article, conference presentation, website) suggests accountability, honesty, precision. It appeals to your intellect – the language addresses serious questions that you have about effectiveness and about how evidence was collected and analyzed, because you as a consumer and a person responsible for children in your care. The language is not flowery and vague. It does not use feel-good words that distract you from the serious questions.
2. The claims are reasonable. They state limitations. They do not make promises; they do not sound too good to be true; they do not sound *exaggerated* and puffed up.
3. **The thing being tested** (e.g., curriculum) is **described fully**. You know exactly what's in it and how it is used. Also, the **effects of the thing being tested are defined clearly and concretely**. The words are understood and the words (definitions) POINT to things (effects?) that can be observed; that is, defined and measured *objectively*---that is, defined by what students DO (e.g., how many problems they solve correctly), not teachers' opinions, impressions, and perceptions.
4. Researchers look not only for beneficial effects but also for harmful effects. If you don't look for harmful effects, you won't find them.
5. **Research** was done in settings where the materials or methods (being tested) will be used---***empirical research***.

6. The research used LARGE enough samples of persons and classrooms and schools (not just 20 students in one class) so that the researchers could see if there were **different kinds and degrees of effects**. [You can't tell if an instructional method, for instance, has a whole **range** of effects if you test it with only five students. Five students can't show much **VARIATION** in effects.]

6. The research was done **again and again** (this is called *replication*), to see if the effects were **RELIBALE** (dependable, can be counted on)?
7. The replications of the research **used samples** of students and teachers, schools, and even districts, **that cover a range of differences that are found in the larger population** (*representative samples*)---small and large schools, affluent and poor students, males and females, ethnic groups.
8. The research in SOME replications was done **over a long period** (*longitudinal research*) to see the long term effects.
9. The research was conducted (or the claims were made) by independent researchers or persons who had no stake in the outcomes.

In later units we will add criteria to our Baloney Detector. For example, we will want to know how **samples were created**? [Did researchers select students with whom a curriculum was **LIKELY** to work?] We also want to know if **the measures used** (e.g., of effectiveness) **are good measures**---relevant. [Some persons have used children's skill at turning pages and at naming parts of a book as measures of reading skill. Is this the information you really need to see if a curriculum teaches children to read?] And we want to know if the researchers **RULED OUT the possible effects of other factors**. [For example, students might read more skillfully at the end of the year **IN PART** because their parents taught them. If **THIS** factor is not ruled out, then how can researchers make any claims about the effects of the curriculum in school?]

ria to our Baloney Detector. For example, we will want to know how **samples were created?** [Did researchers select students with whom a curriculum was LIKELY to work?] We also want to know if **the measures used** (e.g., of effectiveness) **are good measures---**relevant. [Some persons have used children's skill at turning pages and at naming parts of a book as measures of reading skill. Is this the information you really need to see if a curriculum teaches children to read?] And we want to know if the researchers **RULED OUT the possible effects of other factors.** [For example, students might read more skillfully at the end of the year IN PART because their parents taught them. If THIS factor is not ruled out, then how can researchers make any claims about the effects of the curriculum in school?]

Type of research

Field observations

Interview

Experiment

1. Comparison groups? If yes,
 - a. Random or matched samples?

- b. "Treatments" defined well enough to tell if what was done (e.g., "phonics based") matches the label?

2. Pre-tests and post-tests?

3. Quantitative measures? If yes,

- a. Do measures clearly operationalize the concepts? For example, correct wpm is a good measure of reading fluency. Scores on Woodcock-Johnson Word Attack is a good measure of "phonemic awareness.")

If not, what's wrong?

- b. Standardized, widely used instruments/measures? Or just made up?

- c. Observers trained?

- d. Observers' reliability checked?

Sample size

Unsupported or ill-supported conclusions? If yes,

- a. Alternative explanations for the findings? For example: teachers really didn't use the named approach; children matured; other forms of instruction happening at other times; measurement error.
- b. Findings themselves are meagre in relation to the claims. For example, it is claimed that a technique is effective, but it appeared to positively affect only a minority of students.

Type of research

Field observations

Interview

Experiment

1. Comparison groups? If yes,
 - a. Random or matched samples?

 - b. "Treatments" defined well enough to tell if what was done (e.g., "phonics based") matches the label?

2. Pre-tests and post-tests?

3. Quantitative measures? If yes,
 - a. Do measures clearly operationalize the concepts? For example, correct wpm is a good measure of reading fluency. Scores on Woodcock-Johnson Word Attack is a good measure of "phonemic awareness.")

If not, what's wrong?

 - b. Standardized, widely used instruments/measures? Or just made up?

 - c. Observers trained?

 - d. Observers' reliability checked?

Sample size

Unsupported or ill-supported conclusions? If yes,

- a. Alternative explanations for the findings? For example: teachers really didn't use the named approach; children matured; other forms of instruction happening at other times; measurement error.
- b. Findings themselves are meagre in relation to the claims. For example, it is claimed that a technique is effective, but it appeared to positively affect only a minority of students.

**Reporting the Results of Your Study:
A User-Friendly Guide for Evaluators of Educational Programs and Practices**

2

October 2005

This publication was produced by the Coalition for Evidence-Based Policy, in partnership with the What Works Clearinghouse, under a contract with the U.S. Education Department's Institute of Education Sciences (Contract #ED-02-CO-0022). The views expressed herein do not necessarily reflect the views of the Institute of Education Sciences.

This publication is in the public domain. Authorization to reproduce it in whole or in part for educational purposes is granted.

3

Purpose: To provide clear, practical advice on reporting the results of an evaluation of an educational program or practice ("intervention").

Specifically, this is a guide for researchers, and those who sponsor and use research, to reporting the results of "impact" studies - that is, studies which evaluate the effectiveness of an intervention by comparing participants' educational outcomes (e.g., reading or math skills) with:

(i) those of a control or comparison group that does not receive the intervention, or

(ii) participants' pre-intervention ratings on these outcome measures. The Guide suggests key items to include in the study report so as to give the reader a clear understanding of what was evaluated, how it was evaluated, and what the evaluation found.¹

Overview: The following is an outline of the Guide, showing the key items we suggest

you include in each of the sections that typically comprise a study report.

Section of the Study Report: Key Items To Include:

- 1. Title and Abstract**
 - A. A clear, informative title.
 - B. A "structured abstract," including identification of the research design.
- 2. Background and Purpose**
 - A. Background information on the intervention being studied.
 - B. Purpose of the study, including the research question(s) it seeks to answer.
- 3. Methods**
 - A. Description of the study setting (e.g., place and time it was conducted).
 - B. Description of the study sample (including number of sample members and how they were recruited into the study).
 - C. Concrete details of the intervention, and how it differed from what the control/comparison group received.
 - D. Description of how the study sample was allocated to intervention and control/comparison groups.
 - E. Description of how and when outcomes were measured (including evidence that the tests/instruments used to measure are reliable and valid).
 - F. Statistical methods used to compare outcomes for the intervention and control/comparison groups (or outcomes before and after the intervention).
- 4. Results**
 - A. Indicators of whether the study was successfully carried out (e.g., amount of sample attrition).
 - B. Any descriptive data on how the intervention was actually delivered in the study (e.g., extent to which participants completed the intervention).
 - C. Estimates of the intervention's effect on all outcomes measured.
 - D. Any estimates of its effect on subgroups within the study sample.
 - E. If analyzed, any estimates of its effect on those who received greater versus lower "doses" of the intervention.
- 5. Discussion**
 - A. Interpretation: what the results say about the intervention's effectiveness.
 - B. Extent to which the results may be generalizable to others who receive or could receive the intervention.
 - C. Significance of the results to educators, policymakers, and researchers.
 - D. Factors that may account for the intervention's effect (or lack thereof).
 - E. Any study limitations (e.g., small study sample, sample attrition).

4

A. A clear, informative title (Illustrative example: "Randomized Controlled Trial of a Peer-Tutoring

Program for Second Graders: Effect on Math Achievement Two Years Later”).

B. An abstract of the study (1-2 pages), which:

□ **Should follow the “structured abstract” format suggested by the U.S. Education**

Department’s Institute of Education Sciences (see appendix for an example).²

□ **Should identify the type of research design used in the study.** Common examples include:

o **Randomized controlled trial** - a study that measures an intervention’s effect by (i)

randomly assigning individuals (or other units, such as classrooms or schools) to a group

that participates in the intervention, or to a control group that does not; and then (ii)

comparing outcomes for the two groups.

o **Comparison-group study (also known as a “quasi-experimental” study)** - a study that

measures an intervention’s effect by comparing outcomes for intervention participants

with outcomes for a comparison group, chosen through methods other than random

assignment. We suggest you also indicate if the comparison-group study is one of the

following two types, generally regarded as among the stronger comparison-group designs:

- A comparison-group study with equating - a study in which statistical controls and/or

matching techniques are used to make the intervention and comparison groups similar

in their pre-intervention characteristics.

- A regression-discontinuity study - a study in which individuals are assigned to intervention or comparison groups solely on the basis of a “cutoff” score on a preintervention

measure (e.g., students scoring at or below the 25th percentile on the Iowa Test of Basic Skills in math are assigned to the intervention group, and those

scoring above the 25th percentile are assigned to the comparison group).

o **Pre-post study** - a study that examines whether intervention participants are better or

worse off after the intervention than before, and then associates any such improvement or

deterioration with the intervention. [Note: The What Works Clearinghouse does not

consider this type of study to be capable of generating valid evidence about an intervention’s effect. This is because it does not use a control or comparison group, and so

cannot answer whether the participants' improvement or deterioration would have occurred anyway, even without the intervention.]

5

A. Background information on the intervention being studied, including such items as:

- A brief description of the intervention, including the problem it seeks to address and the population for which it is intended. (This would be a general description, with the concrete details provided in 3C, below.)**
- The theory or logic of how it is supposed to improve educational outcomes.**
- The intervention's history and the extent of its current use.**
- A summary of the results of any previous rigorous studies of the intervention or closely related interventions.**

B. Purpose of the study, including:

- A clear statement of the research question(s) it seeks to answer (e.g., "Did the XYZ reading program for first graders increase student reading achievement, and reduce the number of students retained in-grade or placed in special education classes? If so, by how much?").**
- An explanation of why the study is, or should be, important to educators and/or policymakers.**

A. A concise description of the study setting (e.g., five public elementary schools in central Philadelphia, during the period 2001-2004).

B. A description of the study population (i.e., "sample"), including:

- How they were recruited into the study, including (i) the eligibility requirements for participation in the study, and (ii) how those eligible were invited or selected to join the study sample (e.g., the principal of Monroe Middle School requested that all teachers in the school participate in the study of a professional development program or, alternatively, asked for volunteers to participate).³**
- If the sample is intended to be representative of a larger group (e.g., a representative sample of schools participating in a national program), what methods were used to obtain such a sample (e.g., random sampling).**

The total number of sample members allocated - through random assignment or other means - to (i) the intervention group, and (ii) the control/comparison group at the start of the study (or, in a pre-post study, the total number of sample members prior to the intervention). You may also wish to report the results of an analysis showing that the study sample is large enough to provide meaningful answers to the study's research questions ("power analysis").⁴

6

Descriptive statistics on the study sample (e.g., age, race, gender, family income, and preintervention measures of the outcomes the intervention seeks to improve, such as reading or math achievement). You should briefly describe how and when these descriptive data were obtained, and the percentage of sample members from whom they were obtained. You may also wish to discuss the extent to which the sample is typical of the larger population that receives or could potentially receive the intervention.

C. A clear description of the intervention as it was implemented in the study, and how it differed from what the control/comparison group received.

The description should include the concrete details of the intervention that a reader seeking to replicate it would need to understand including, among other things, who administered it, what training or supervision they received, what it costs to implement in a typical school or community setting,⁵ and how it may differ from the model program on which it is based (e.g., the National ABC Science curriculum was used but only lessons 1- 20 out of 40).

You should also describe clearly how the intervention differed from what the control/comparison group received.

Illustrative example: A study of a one-on-one tutoring program for beginning readers should describe such items as -

- o Who conducted the tutoring (e.g., certified public school teachers, paraprofessionals, or undergraduate volunteers);
- o The training they received in how to tutor;

- o The curriculum and materials they used to tutor;
- o The duration of the tutoring sessions, and setting in which they took place (e.g., daily 20-minute sessions held after school, over a period of six-months);
- o Whether the tutors were supervised or monitored and, if so, how;
- o Any unusual events that substantially affected delivery of the tutoring program (e.g., school closing for two weeks due to a major snowstorm);
- o The cost of the tutoring program per student (excluding costs of research or program development that would not be incurred in replicating the program);
- o The reading instruction or other services received by the students in the control/comparison group (e.g., the school's usual reading program); and
- o Where the reader can obtain additional information on the tutoring program (e.g., a website or program manual).

D. A description of how the study sample was allocated to intervention and control/comparison groups, including:

- Whether the study allocated *individuals* (e.g., students), or *clusters of individuals* (e.g., classrooms or schools), to the intervention and control/comparison groups.** If the study allocated clusters, you should also describe any steps taken to ensure that the placement of individuals within the clusters (e.g., placement of students in classes) was unaffected by whether the clusters were in the intervention or control/comparison group (e.g., students

7

were assigned to classes *prior to* the allocation of classes to the intervention and control/comparison groups).

- Whether the random assignment or formation of comparison groups was “blocked”** (e.g., students were identified as (i) high-achieving, (ii) average-achieving, or (iii) low-achieving based on prior test scores, and then allocated *within each of the three achievement levels* to the intervention and control/comparison groups).

- Whether the ratio of those allocated to the intervention versus control/comparison group (e.g., 60:40) was held constant over time, and within blocks.**

- Whether all those originally allocated to the intervention and control/comparison groups were retained in their group for the duration of the study - even:**

- o Intervention participants who failed to participate in or complete the intervention

(retaining them in the intervention group is known as an “intention-to-treat” approach).

o Control/comparison group members who may have participated in or benefited from the intervention (these are known as “cross-overs” or “contaminated” members of the control/comparison group).

In a randomized controlled trial, the random assignment process used (e.g., coin toss, lottery, or computer program), including:

o Who administered it (e.g., the researchers or school staff); and

o Any steps taken to protect against intentional or unintentional manipulation of the process

(e.g., concealing from researchers, school staff, and study sample members any information they could use to predict in advance who would be assigned to the intervention versus control group).

In a comparison-group study, how the comparison group was formed (e.g., from students

in a neighboring school who had achievement levels and demographic characteristics similar to

intervention participants). Among other things, this description should include:

o Whether the comparison group was formed before or after the intervention was

administered;

o Any “matching” techniques used to increase the initial similarity of intervention and

comparison group members in their observable characteristics (e.g., propensity score

matching); and

o In a regression-discontinuity study (described in 1B, above), what cutoff score was used to

form the intervention and comparison groups, how sample members’ scores were distributed around this cutoff score, and any evidence that the cutoff score was used with

few or no exceptions to determine who received the intervention.

E. A clear description of how and when outcomes were measured, including:

What tests or other instruments were used to measure outcomes (e.g., widely-used

standardized tests, such as the Stanford Achievement Test; tests designed by the research

8

team for purposes of the study; structured interviews with study sample members;

observations of classroom behavior; school records).

Any evidence that these instruments:

- o Are “reliable” (i.e., yield similar responses in re-tests or with different raters);
 - o Are “valid” (i.e., accurately measure the true outcomes that the intervention is designed to affect); and
 - o Were applied in the same way to the intervention and control/comparison groups (e.g., the same test of reading skills was administered in comparable settings).
 - Who administered these instruments, and whether any of the following conditions applied which might affect the objectivity of their measurements:
 - o They knew the study sample members (e.g., were their teachers);
 - o They might have a stake in the study outcome (e.g., were the developers or implementers of the intervention being studied); and/or
 - o They were kept unaware of who was in intervention versus control/comparison group (i.e., were “blinded”).
 - When the instruments were administered (e.g., just prior to the start of the intervention, so as to obtain “baseline” data”; and at 12-month intervals thereafter for three years, so as to obtain outcome data at the one-year, two-year, and three-year follow-ups).
- F. The statistical methods used to compare outcomes for the intervention and control/comparison groups (or, in a pre-post study, outcomes before and after the intervention), including:**
- The regression or other model used to estimate the intervention’s effects, including - in a regression - the variables that indicate whether a sample member is in the intervention or control/comparison group, and any covariates used.
 - Any techniques used to adjust statistically for initial differences between the intervention and control/comparison groups, or improve the precision of the estimated effects (e.g., analysis of covariance).
 - If the study randomly assigned or compared clusters (e.g., classrooms), rather than individuals, whether the study accounted for such clustering in estimating statistical significance levels.
 - If the random assignment or formation of comparison groups was “blocked,” whether the study accounted for such blocking in estimating statistical significance levels.

If the study sample is intended to be representative of a larger group (e.g., a representative sample of schools participating in a national program), whether the study accounted for this in estimating statistical significance levels. (i.e., used a “random effects” model).

9

A. Indicators of whether the study was successfully carried out, including:

The results of statistical tests assessing the extent to which the intervention and control/comparison groups were equivalent in key characteristics prior to the intervention

- especially whether they were equivalent in pre-intervention measures of the outcomes the intervention seeks to improve, such as reading or math achievement. (Exception: regression discontinuity studies - described in 1B above - do not seek to create equivalent intervention and comparison groups.)

The percentage of individuals originally allocated to (i) the intervention group, and (ii) the control/comparison group, for whom outcome data could not be obtained at each follow-up point (i.e., the degree of “sample attrition”).

An analysis of whether sample attrition created differences between the intervention and control/comparison groups (e.g., whether a greater number of students with low initial test scores were lost from the intervention group as opposed to the control/comparison group, creating a difference between the two groups). You should also preferably include an analysis of whether (i) the original study sample and (ii) those for whom outcome were obtained, differ in their preintervention characteristics, so as to gauge whether attrition altered the nature of the study sample.

The extent to which any control/comparison group members participated in the intervention, or otherwise benefited from it (e.g., by borrowing techniques or materials from intervention group members).

B. If available, descriptive data on how the intervention was actually delivered in the study, including such items as:

The extent to which the intervention group members completed the intervention, and what intervention services or treatment they actually received (e.g., in a study of a teacher professional development program, the average number of training sessions the teachers participated in, the length of each session, and the extent to which the trainers covered the key items in the training curriculum).

Data on any related (non-intervention) services or treatment provided to the intervention group and/or control/comparison groups (e.g., participation in other professional development activities).

C. Estimates of the intervention’s effect on all outcomes measured (not just those for which there are positive effects), including:

Whether the effects are statistically significant at conventional levels (usually the .05 level);

and

The magnitude of the effects, reported in “real-world” terms that enable the reader to gauge their practical importance (e.g., report an improvement in reading comprehension of a half grade-level, or a reduction in the percentage of students using illicit drugs from 20 to 14 percent, rather than only reporting items such as “standardized effect sizes” or “odds ratios”).

10

The sample size used in making each estimate, and the standard error of the estimate.

D. Any estimates of the intervention’s effects on subgroups within the study sample, including:

A precise description of each subgroup (e.g., first-grade male students scoring in the highest 25th percentile on teacher-rated aggressive behavior).

A brief rationale for why the subgroup was chosen to be analyzed (e.g., reasons why the intervention might have a different effect on the subgroup than on the overall study population).

A complete listing of all subgroups analyzed (not just those for which there are positive effects), and the effects found for each.

□ The sample size used in making each estimate, and the standard error of the estimate.

E. If analyzed, any estimates of the intervention’s effect on those who received greater versus lower “doses” of the intervention (e.g., those who successfully completed all sessions of a teacher training program versus those who completed few or no sessions). [Note: Such “doseeffect” analyses can sometimes yield useful information to supplement the results for the full sample. However, if intervention participants *self-select* themselves into the higher and lower dose groups, the What Works Clearinghouse would not regard such analyses as producing valid estimates of the intervention’s effect (because self-selection would plausibly create differences between the two groups in motivation levels and other characteristics, leading to inaccurate results).]

A. Interpretation of the study results: what they say about the effectiveness of the intervention in the context of other rigorous studies of this or related interventions.

B. The extent to which the results may be generalizable to others who receive or could potentially receive the intervention.

C. Significance of the results to educators, policymakers, researchers, and others.

D. Factors that may account for the intervention’s effect (or lack thereof).

E. Any study limitations (e.g., small study sample, sample attrition).

11

Abstract

Citation: Ricciuti, A.E., R.G. St.Pierre, W. Lee, A. Parsad, and T. Rimdzius. Third National Even Start

Evaluation: Follow-Up Findings From the Experimental Design Study. U.S.

Department of Education,

Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Washington, D.C., 2004.

Background: The Even Start Family Literacy Program has provided instructional services to lowincome children and their parents since 1989. A previous randomized controlled trial in the early

1990s did not show this program to have positive impacts.

Purpose: To assess the effectiveness of Even Start in a group of grantees around the country. An

earlier report from this study presented impact findings based on pretest and posttest data at the start

and end of a school year. No program impacts were found. The purpose of the current report is to

present impact analyses of follow-up data collected one year after posttest data.

Setting: 18 Even Start grantees in 14 states that operated in the 1999-2000 and 2000-2001 school years.

Study Sample: 463 families eligible for and interested in participating in Even Start family literacy services.

Intervention: Even Start families were offered family literacy services, defined as (1) interactive parent-child literacy activities, (2) parenting education, (3) adult education, and (4) early childhood education.

Research Design: Randomized controlled field trial in which families were randomly assigned either to Even Start (309 families) or a control group (154 families).

Control or Comparison Condition: Control families could participate in any educational and social services to which they were entitled, but they were not allowed to participate in Even Start for one year.

Data Collection and Analysis: Pretest data on child and adult literacy skills were collected in the fall, posttest data were collected in the spring/summer, and follow-up data were collected the next

spring. Measures included direct assessment of children (Peabody Picture Vocabulary Test, Woodcock-Johnson Battery, Story and Print Concepts), direct assessment of parents (Woodcock-Johnson Battery), teacher report on children (Social Skills Rating System), parent reports on economic and educational status, child literacy-related skills, home literacy environment and activities, parent assessment of children (Vineland Communication Domain), and school records. A longitudinal sample (data at all three waves) of children and parents was created for each outcome measure, and t-tests were conducted to assess differences in gains between Even Start and control groups. The sample size for

the analysis of any given outcome depends on several factors including attrition, age of the child, exclusion of families who were assessed in Spanish, and the need for longitudinal data. For example, the PPVT analysis for children was done with samples of 97 Even Start and 44 control children, and the

12

Woodcock-Johnson analysis for parents was done with samples of 149 Even Start and 65 control parents.

Findings: As was the case at posttest, Even Start children and parents made gains on a variety of literacy assessments and other measures at follow-up, but they did not gain more than children and parents in the control group. It had been hypothesized that follow-up data might show positive effects because (1) Even Start families had the opportunity to participate for a second school year, and (2) change in some outcomes might require more time than others. However, the follow-up data do not support either of these hypotheses.

Conclusion: The underlying premise of Even Start as described by the statute and implemented in the field was not supported by this study.

13

1 Our suggested list of key items is drawn, to a large extent, from the following authoritative sources:

Evidence Standards of the Institute of Education Sciences' What Works Clearinghouse, at www.w-wc.org/reviewprocess/standards.html; the Standards of Evidence of the Society for

Prevention Research, in Brian R. Flay et. al., "Standards of Evidence: Criteria for Efficacy, Effectiveness, and Dissemination,"

Prevention Science, forthcoming, September 2005 (available online at <http://www.preventionresearch.org/commlmon.php#SofE>); the Consolidated Standards of Reporting Trials

(CONSORT) Statement, in David Moher, Kenneth F. Schulz, and Douglas Altman, "The CONSORT Statement:

Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials," JAMA, vol. 285, no. 15, April 18, 2005, pages 1987-1991.

2 The structured abstract format suggested by the Institute of Education Sciences is based on the concept

proposed in Frederick Mosteller, Bill Nave, and Edward J. Miech, "Why We Need a Structured Abstract in

Education Research,” *Education Researcher*, vol. 33, no. 1, January/February 2004, pp. 29-34.

3 Preferably, you should also indicate the extent to which those eligible for the study, and those invited or selected, actually became members of the study sample.

4 If a power analysis is undertaken, you should report not only its results, but also its statistical assumptions (e.g., desired power, minimum detectable effect size, intra-class correlation coefficient, and fixed or random effects).

5 We suggest you report here only the budgetary cost of the intervention (e.g., in the tutoring example, the cost of program materials, and paying the tutors and other program staff). If the intervention increases or decreases the cost of other services (e.g., reduces the number of students referred to special education classes), these effects should be reported in the Results section of the report (section 4).