U.S. Department of Education
June 2014

Going public: Writing about research in everyday language

Mark Dynarski
Pemberton Research

Ellen Kisker
Twin Peaks Partners, LLC

REL 2014–051

The National Center for Education Evaluation and Regional Assistance (NCEE) conducts unbiased, large-scale evaluations of education programs and practices supported by federal funds; provides research-based technical assistance to educators and policymakers; and supports the synthesis and the widespread dissemination of the results of research and evaluation throughout the United States.

June 2014

This report was prepared for the Institute of Education Sciences (IES) by Decision Information Resources, Inc. under Contract ED-IES-12-C-0057, Analytic Technical Assistance and Development. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Dynarski, M, & Kisker, Ellen. (2014). *Going public: Writing about research in everyday language* (REL 2014–051). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. Retrieved from http://ies.ed.gov/ncee/edlabs.

Contents

Explaining Complex Research Concepts

Communicating complex concepts to practitioners, policymakers, and other nontechnical readers is a challenge that all policy researchers face.

Research in education uses many concepts from methodology and statistics. If researchers want to communicate their findings to an audience of other researchers, they can safely assume that their audience is familiar with these concepts. Research terminology, or *jargon*, is efficient to use because it communicates concepts quickly.

Complex concepts are typically explained within the main text of a report, which includes its executive summary and the chapters that are the body. Many readers will read only the executive summary. For this reason alone, an executive summary should be free of jargon.

However, if the objective is to communicate findings to an audience of practitioners, policymakers or interested readers who are not researchers, researchers should not assume that their audience is familiar with research concepts. For this audience, communicating with research jargon is inefficient. To various degrees, depending on backgrounds and training, readers will have to decipher the jargon and guess at its meaning. They may not decipher it correctly, or they may get the meaning wrong, or they may simply stop reading.

Whatever the case, jargon gets in the way of disseminating research findings to lay audiences.

The body of a report that follows the executive summary can include some jargon because it is nearly impossible to write chapters of technical research without it. A central principle of science is replication, and if other researchers cannot understand or replicate what a study did, they are trained to be skeptical about its findings. However, even in the chapters, using less jargon will contribute to greater readability. Appendixes to a report can be full of jargon and details. Readers expect that appendixes are for specialists who desire that level of detail.

The concepts in this brief are from the three main areas of an empirical research report: *study design*, *measurement*, and *data analysis*. Not all concepts in a study fit neatly into those three areas, but many do. Some concepts also belong both in design and in analysis because the design phase of many studies includes the consideration of methods for analyzing the data. Concepts related to *program* design and development are beyond the scope of this brief.

The brief is not a guide on how to write or structure a research report. Many sources have covered this ground. The focus here is on how to explain complex concepts for readers who are not expected to know what they are. This brief provides guidelines for communicating with these audiences.

Following a discussion of these guidelines is a glossary of research terms (concepts) that are frequently used in empirical research on education. The glossary defines the terms, provides examples of usage that may be difficult

for lay audiences to interpret, and includes alternate wordings that are simpler but retain the essential meaning. The terms are grouped into the areas of *design*, *measurement*, and *analysis*.

Guidelines for Explaining Research Concepts

The translation of research concepts shares some aspects with the translation of one language to another. Research concepts are the 'source language.' Putting them into everyday terms is like translating them into the 'target language.' The translator needs to create an expression in the target language that retains the meaning of the expression in the source language. The languages here are just different forms of English, but anyone who has picked up a research journal in a field outside their expertise quickly recognizes that researchers can have their own language.

In this brief, research concepts in technical language are "translated" into nontechnical language. Doing so involves applying the following guidelines:

- **Simpler is better.**
- **Focus on what readers need to know.**
- **Reduce possible misinterpretations.**

Simpler is better

Einstein said "a scientific theory should be as simple as possible, but no simpler." Analogously, concepts should be explained as simply as possible while keeping their meaning. The glossary that follows these guidelines explains complex concepts in simple terms while striving to retain their meaning. The caution that Einstein conveyed is that simplification can go too far, to a point where concepts lose meaning. Getting the balance right is the challenge.

Concepts should be explained as simply as possible while keeping their meaning.

A researcher may write "The study used ordinary least squares to estimate the effect of the new curriculum on reading test scores." Lay readers might be puzzled by the technical term "ordinary least squares." But a researcher wanting to simplify the term faces a dilemma. It has a history going back hundreds of years, and each of the three words conveys part of its meaning. Making the term simpler may change its meaning.

Can the researcher use other words and yet have readers grasp what was done? "The study used an accepted statistical approach to estimate the effect of the new curriculum on reading test scores." This is easier to grasp.

To the lay reader, "an accepted statistical approach" in place of "ordinary least squares" has not changed the meaning. Readers are likely to be more interested in whether the study's estimation method is accepted than in its exact specification. And using both works too. "The study used an accepted statistical approach (ordinary least squares regression) to estimate the effect of the new curriculum on reading test scores." General readers will know it's an accepted approach and technical readers will know which of the possible approaches it was.


A researcher may write, "The study's counterfactual was business as usual." Lay readers might stop at the term "counterfactual." It is literally translated as "contrary to fact," it is a fundamental concept in evaluation, and it is a state of the world that is never observed. Readers not trained to know what counterfactuals are will probably struggle to gather meaning from context.

A longer but simpler expression will work: "Individuals who did not participate in the intervention continued to use the current curriculum, and their outcomes provide an estimate of what would have happened to participants if they had not received the intervention."

Focus on what readers need to know

A common writing principle is that authors should prefer the concrete to the abstract. But "concrete" and "abstract" are perceptions of readers. Knowing your audience means writing for your readers.


"An accepted statistical approach" may seem more abstract than "ordinary least squares." Lay readers might think that "an accepted statistical approach" is concrete and "ordinary least squares" is abstract. Researchers might think the opposite.

Focusing on what readers need to know helps clarify direction.

If lay readers know that the estimation method is an accepted one, it conveys that researchers use it regularly. If they know only that the estimation method was ordinary least squares, it conveys little about whether researchers use the method, and some readers may think it is unusual or unconventional.


Another example: a researcher may write that a study used "random assignment." To other researchers, random assignment conveys a concrete detail of the study. To lay readers, random assignment may be abstract and raise questions. Didn't the study have to use a specific approach to assign participants? How can assignment be random? Can a study just choose an assignment approach at random? To a researcher, random assignment is a highly controlled process. To a lay reader, "random" suggests fuzzy, out of control, at the whim of natural forces.

Substituting for the term may help. "The study created groups by assigning random numbers to study participants." But the sentence still poses difficulties for lay readers. They may wonder what a random number is or how one is created.

What do readers need to know? Nearly everyone will know or have played games of chance. The analogy of flipping a coin or playing a lottery fits random assignment well. "Groups were created by using an approach analogous to coin-flipping." The next sentences can be, "This process yielded groups that were similar on average. Statistical analyses confirmed that the groups were similar."

Do readers need to know more? Random assignment has a crucial feature: by the "law of large numbers," groups created by random assignment will have similar characteristics on average, even characteristics the study did not or could not measure. This similarity is a distinguishing feature of experiments, and it makes sense for readers to know it. The researcher can add a fourth sentence: "Because this approach was used, unobserved characteristics of groups also will be similar."

This last sentence requires readers to trust that unobserved characteristics will be similar. Do lay readers need to know why this is true? Trying to explain every concept in a research study will generate a lot of text, and much of it may be challenging for lay readers. (For example, it will take a lot of text to explain the three terms in ordinary least squares.) Leaving out an explanation of why unobserved characteristics are similar makes sense.

In considering what lay readers need to know, the researcher may find it useful to imagine that readers will need to explain to another person what they learned from the study.

The person may not be fictional: the reader may have been asked to read the study to inform a policy deliberation or decision. For readers to convey information to others means that they need to come away from their reading with a broad understanding of the study's methods and findings. Returning to the example, a lay reader will find it easy to tell another person that the study used an accepted statistical method but less easy to tell another person that the study used ordinary least squares.

It is likely that lay readers do not have a lot of time to read a report thoroughly and absorb its nuances. Most will read only the executive summary, evidence of their desire for speed. The researcher recognizing this desire for speed will avoid text that hinders progress. Jargon hinders progress because readers need to decipher it. Vague antecedents and references hinder progress in understanding a report because readers have to pause to think about meaning.

Building on the preceding example, suppose that the researcher had written "Because this approach was used, unobserved characteristics of groups also will be similar (Smith 1996)." What does "Smith 1996" add? Different readers may be puzzled in different ways about the citation. Did Smith 1996 prove that unobserved characteristics are similar? Or is the study following the example of Smith 1996, who also asserted that groups will be similar? Or perhaps Smith 1996 lays out the methods for experiments and encourages authors to

mention that unobservable characteristics will be similar? Because readers have to guess what "Smith 1996" is doing in the sentence, reading it slows them down.

What the reader needs to know and can grasp quickly is this: "Because this approach was used, unobservable characteristics of groups also will be similar (Smith 1996 discusses the theory supporting why unobserved characteristics will be similar)." The example now has more words, but readers know what they should understand about "Smith 1996."

Reduce possible misinterpretations

Researchers can make misinterpretations less likely by being clear about what they are and are not saying—in other words, writing simply and clearly. Researchers cannot prevent readers from misinterpreting a study, but they should strive to write the findings to limit misinterpretations.

Authors should ask themselves how a sentence, paragraph, or section could be misinterpreted, and write so that possible misinterpretations are minimized.

Writing "Findings from the study show that the new curriculum had no effect on reading scores" invites misinterpretation. Some readers may conclude that scores were the same when students were tested in the fall and the following spring, suggesting that reading skills did not improve at all. But the researcher may have used "no effect" as shorthand for "equal amount of growth in the treatment and control groups."

To avoid misinterpretation, the researcher could write "The study found that students using the new curriculum increased their reading scores by the same amount as students using the current curriculum." The sentence now conveys the two parts of the finding: students using the new curriculum increased their reading scores, but the increase was as large as for students using the current curriculum.

Here is another example that invites misinterpretation: "The study found significant effects." Lay readers may wonder what "significant" means in this sentence. Were effects large? Were they statistically significant? Both? Maybe effects were small but the authors judged them important?

Rewriting the sentence to reduce misinterpretation can take different routes, as in the following examples, depending on what the study found:

- "The study found small effects that were statistically significant."
- "The study found an effect that was statistically significant. Though small, the size of the effect is relevant for policy. It was equivalent to moving a student from the fiftieth percentile to the fifty-third percentile on a standardized test."

In the latter case, the researcher presumably has indicated how the paper gauged whether effects are "relevant for policy." If educators or policymakers are interested in reading curricula that improve on the current curriculum, a small, positive, and statistically significant effect is relevant. Or the relevance may have been established in the study's design phase when the study sample size was set to detect some effect size, usually by referring to what previous studies have found. Or researchers could use their own criteria to establish reasons for why a finding is relevant, such as that study findings show that the intervention would close part of the black-white achievement gap, or that findings are similar in size to findings from the Tennessee class-size experiment. Readers can judge for themselves whether they share one or more of these criteria.

Another example invites misinterpretation: "The study's findings have limited generalizability." The example has two issues. One is that the term "generalizability" is jargon, and the lay reader may not know it. To a researcher it means that a study's findings apply to other populations and settings. To a lay reader, generalizing may mean making a broader statement based on narrower evidence, which may be what the reader wants to do after reading the study. The second issue with the example is that the limits are unknown, and a reader may sense a warning that they need to interpret.

Writing in concrete terms will reduce misinterpretations. The study may have been done in a few schools in one region. The researcher could write "The study included five schools in one region, and its findings apply to similar schools and regions." Or the study may have been done in one urban school district. The researcher could write "The study was conducted in one urban school district, and its findings apply in urban districts of comparable size."

Writing in concrete terms will reduce misinterpretations.

Writing about generalizability can be challenging, as the last example suggests. Do findings apply to urban districts of different sizes, suburban districts of the same size, or to similar students who attend rural schools? Ultimately, readers need to judge generalizability. The concept inherently is about going beyond a study's findings to ask in what other settings the findings apply. But the answer is only a prediction until the intervention operates in these other settings. The answer to the question "will it work for me?" can be predicted but not known with confidence.

Researchers can provide readers with useful guidance by adopting the perspective of readers who want to apply its findings in their settings. What is known about this kind of program or intervention? If it is known that reading curricula often have different results when students are English Language Learners (ELLs), a study that tested a new reading curriculum can point out that few students in its sample were ELLs. Readers who work in schools or districts with many ELLs now know that the findings may not apply in their settings. A study of a new science module might indicate that the module was designed for a science curriculum developed to suit one set of standards and that its effectiveness if used with other curricula is not known.

Concluding Thoughts

Using these three guidelines may lead to clearer communication of findings from researchers to interested readers. Applying them also may improve communication of findings from researchers to researchers. However, researchers writing for journal articles and academic publications can rely more heavily on jargon. Indeed, journal pressure for shorter papers creates an incentive to use jargon.

Researchers should not be surprised that policymakers read few academic papers.

The intent in this guide is to support researchers who want policymakers to know, and possibly respond to, a study's findings. A research paper that proposes a new theory or takes issue with an existing theory is likely to be part of a conversation among researchers. In contrast, a research paper that reports on the effectiveness of an intervention or program is part of a conversation among researchers and potential users of the intervention, policymakers, or educators. In this conversation, jargon is inefficient and unnecessary. The following glossary offers language and approaches to communicating key research concepts clearly to inform policymakers and other lay readers about study findings. The glossary is not an exhaustive list, but researchers can apply the simplification of the terms and concepts to communicate more effectively to lay audiences.

A Glossary of Research Terms

The intent of the glossary is to illuminate how complex concepts can be made simpler. The concepts are grouped into design, measurement, and analysis. The glossary shows how a concept might be used as jargon, explains what the concept is, and shows how the concept might be written in simpler language. The examples are likely to be found in contemporary education evaluations. The glossary is not intended to be a dictionary of research concepts. For such a reference, see Paul Vogt's *Dictionary of Statistics and Methodology* (2005). It also is not intended to be complete. Empirical research encompasses many approaches and techniques, and the glossary could include hundreds of examples. The three guidelines of simpler is better, focusing on what the reader needs to know, and reducing possible misinterpretations can be applied to all glossary examples.

Design

Concept	Examples of Technical Usage	Explanation	Revised Usage
Causal, causality, causal inference	"The study used an experimental design to support causal inferences."	"Causality" is the relationship between cause and effect—being able to say that a variable X causes an outcome Y. It is fundamental to research and evaluation. A study of a policy, program, intervention, or approach is investigating whether the policy caused outcomes to change.	The study used an experimental design to learn whether the intervention improved outcomes.
Internal validity	"An experimental design was chosen for its internal validity."	"Internal validity" is a property of causal inferences, which are statements such as "intervention X causes outcome Y to improve." Study designs yield internally valid causal inferences when they rule out other causes of improvement except the intervention. For example, an experimental study of a reading intervention may yield evidence that the *intervention* increased reading skills more than the existing reading curriculum. By its design, the experiment, if it is conducted correctly, rules out other possible causes of the improvement so that the inference that the intervention caused a larger increase in reading skills than the existing reading curriculum is internally valid. An internally invalid design is one in which other possible causes are not ruled out. Quasi-experimental designs, for example, generally do not yield internally valid inferences, because they do not eliminate the possibility that factors other than the intervention being studied caused differences in outcomes.	An experimental design was chosen for its ability to support statements that the intervention caused improvements.
External validity	"The study's external validity is limited because only a small number of schools participated."	"External validity" refers to whether a study's findings apply more broadly. It often is used as a synonym for "generalizability."	The experiment showed that the intervention caused an increase in scores. Whether the finding applies more broadly is not known because only a small number of schools participated.

Concept	Examples of Technical Usage	Explanation	Revised Usage
Counterfactual	"The counterfactual in this study was business as usual." "The counterfactual was the curriculum currently used for math."	The "counterfactual" is what would have happened if study participants had not received the intervention. Because participants cannot simultaneously receive and not receive an intervention, the counterfactual cannot be observed. It can be estimated by using various evaluation designs, such as a randomized controlled trial. Outcomes of the control group estimate the "counterfactual" outcomes of the treatment group.	The study compared outcomes for participants in the intervention to outcomes of similar individuals who did not participate. Outcomes of individuals who did not participate in the intervention provide an estimate of what would have happened to participants without the intervention.
Confound, confounding factor	"Results of the study were confounded because teachers could not be separated from schools." "The adoption of new state standards created a confound with the study's findings."	"Confounds" arise when two or more explanations for a study's findings cannot be separated. For example, a study that assesses the effects of a new reading program by implementing it in one teacher's classroom and comparing reading outcomes to those in another teacher's classroom will "confound"—that is, not be able to separate—the effects of the intervention from the effects of the one teacher relative to the other. Similarly, in a study of a supplemental math curriculum, if the regular curricula in a treatment classroom were different from the regular curricula in comparison classrooms, the effect of the supplemental curriculum will be confounded with the effect of the main curriculum.	The estimated impacts of the intervention could have arisen either from the intervention or from differences in teachers.

Concept	Examples of Technical Usage	Explanation	Revised Usage
Confirmatory and exploratory analyses	"The study conducted a confirmatory analysis of the first hypothesis." "Results of the confirmatory analysis provide support for the hypothesis."	"Confirmatory analyses" refers to analyses conducted to address the primary research questions defined at the beginning of the study. For example, for a study of a new math curriculum, a typical study will have as its primary research question whether the curriculum improves math test scores. Analysis of the curriculum's effect on math scores was a confirmatory analysis. A study can have multiple primary research questions and associated confirmatory analyses.	The study focused on its main question of whether the intervention increased test scores. It also considered secondary questions such as whether the effects were the same for boys and girls.
	"The exploratory analysis found that effects were evident in some subgroups."	"Exploratory analyses" refers to analyses of additional research questions that are of interest but are not the primary focus of the study. Exploratory analyses often help researchers decide on the primary research questions that they will investigate in future studies. In a study of whether an intervention improved math scores, for example, whether the intervention had different effects for high or low skilled students, or for boys and girls, may be a secondary question to be addressed with an exploratory analysis.	
Effectiveness trial; efficacy trial	"The study was designed as an effectiveness trial."	An "effectiveness trial" asks "will it work?" and tests an intervention under realistic conditions, such as implementation by typical teachers in schools and districts that agree to participate. In contrast, an "efficacy trial" asks "can it work?" and tests an intervention under ideal or desired conditions.	The study was designed to test the intervention under realistic conditions typical of its expected use.
	"The study was designed as an efficacy trial."	For example, testing how a software application improves reading skills in a regular school classroom is an effectiveness trial. Testing how the software application improves reading skills when used by researchers with students in a lab setting is an efficacy trial.	The study was designed to test the intervention under ideal conditions.

Concept	Examples of Technical Usage	Explanation	Revised Usage
Randomized controlled trial	"The study was designed as a randomized controlled trial." "Randomized controlled trials are considered a gold standard for research evidence."	A "randomized controlled trial" is a study in which eligible study participants are assigned to groups by a probability process that acts like a coin flip or game of chance, which is known as "random assignment." One group receives the intervention, and the other group does not. A central feature of a randomized controlled trial is that because of the random process by which study participants are assigned to groups, both observed and unobserved characteristics of the groups will be similar on average. More complex randomized controlled trials may use more than two groups. For example, a trial might randomly assign one group of students to one curriculum, a second group to another curriculum, and a third group to another curriculum.	The study (a "randomized controlled trial") was designed to estimate intervention impacts by creating equivalent groups, one of which was able to receive the intervention and the other of which was not.
"'Quasi-experimental design" or "quasi-experiment"	"The study used a quasi-experimental design to measure impacts."	A "quasi-experimental design" or "quasi-experiment" compares outcomes of intervention participants with outcomes of a "comparison" group. Unlike a randomized controlled trial, in a quasi-experiment, participants first choose to participate in the intervention and then researchers identify a comparison group for the study. Because participants chose to participate and members of the comparison group did not, it is always possible that the characteristics or circumstances of participants that led them to participate are a cause of any differences in outcomes between participants and comparison group members. A quasi-experiment cannot assure that unobserved differences in the characteristics of the two groups are similar on average, and because some differences may be unmeasured, they cannot be controlled in the impact analyses.	The study compared outcomes of intervention participants to outcomes of students who were similar in terms of their demographic and socioeconomic characteristic but may have differed in ways that were not measured in the study.
Regression discontinuity design	"The study used a regression discontinuity design to measure impacts, with a student's reading test score as the forcing variable."	A study using a "regression discontinuity" design uses a cutoff value of a variable—the "forcing variable"—to assign eligible participants to receive the intervention or not. Participants on one side of the cutoff receive the intervention, and participants on the other side of the cutoff do not. The study then compares outcomes of participants within a certain range on both sides of the cutoff—that is, the "bandwidth."	The study used a design that compared outcomes for students on each side of the cutoff score.

Measurement

Concept	Examples of Technical Usage	Explanation	Revised Usage
Validity	"'The scale was chosen for its demonstrated validity."	In measurement, a scale or measurement instrument is "valid" if theory and evidence support its proposed use. A scale that is valid for one purpose may be invalid for another. For example, a scale may be valid for measuring personality dimensions—theory and evidence show that the scale identifies different personality dimensions—but invalid for measuring latent criminality (no theory or evidence indicates that the scale correctly predicts that a person will commit crimes).	The scale was chosen based on theory and evidence that support its use as a measure of the outcome.
Reliability	"The scale has been shown to be reliable in previous research."	In testing and measurement, an instrument is considered "reliable" if it yields similar results in similar conditions. Reliability can be assessed in different ways. If two observers rate the same teacher, the observation process is considered reliable if observer ratings are close ("inter-rater reliability"). If a student takes a test twice within a short time span, the test is reliable if the two scores are close ("test-retest reliability"). If individuals answer similar scale items in similar ways, the items are considered reliable ("internal consistency reliability"). Reliability is related to a measure's variability rather than to its average. A measure that consistently yields the wrong answer (such as a miscalibrated thermometer that measures the temperature) is invalid but reliable. A measure that inconsistently yields the right answer—sometimes much too high and sometimes much too low—is valid but not reliable. A reliable and valid measure is consistently accurate.	The scale yields consistent values when it is administered in similar conditions.

Concept	Examples of Technical Usage	Explanation	Revised Usage
Pre-test	"The two groups had similar average pre-test scores."	In education research, a "pre-test" typically is a test administered to participants before the intervention is delivered, ideally before they are assigned to study groups. In some studies, the pre-test is administered a short time after intervention begins because the study creates its treatment and control groups and begins the intervention before students can be tested (for example, when groups are assigned during the summer and the intervention begins when school begins). In some studies, researchers use the most recent state achievement test administered before the beginning of the intervention as the pre-test. A pre-test sometimes is called a "baseline" test.	The two groups had similar average scores on the test administered before the first intervention session.
Post-test	"The study measured effects by using post-test scores for the two groups."	A post-test is the analog to the pre-test, but it is administered during or after the intervention period or at some future point.	The study measured effects by comparing average scores of the two groups on the spring assessment.
Imputation	"Missing data were imputed by using the "hot deck" method.	Imputation is filling in a value for a missing value. A variety of approaches can be used to fill in missing values. Each uses the logic that some values are more likely than others to be closer to the true value. The approaches differ in how "likely" is defined.	Missing data were filled in by using the most likely values according to a state-of-the-art method.

Analysis

Concept	Examples of Technical Usage	Explanation	Revised Usage
Regression model, regression analysis	"'To determine if the differences between the treatment and control groups were statistically significant, a regression model was estimated."	A "regression model" can refer to a wide array of analytic techniques and approaches. Generally, it is an analysis of the relationship between an outcome of interest, such as reading test score, and a set of variables related to the outcome, such as the test score from the previous year, participation in a program to boost reading scores, and a term that represents the influence of random variables not in the model.	To measure whether the program improved reading scores, the study assumed that reading scores were related to a set of variables, including program participation, and estimated the relationship between participation and reading scores.
	"For teacher outcomes, regression models were used to estimate program impacts."	Regression models have the useful property that the measured effect of one variable is separated from effects of other variables; this property is sometimes referred to as "adjusting for" or controlling for other variables.	To measure whether the program improved teaching, the study assumed that its measures of teaching were related to a set of variables, including program participation, and estimated the relationship between participation and teaching.
Covariate	"'The regression model of the outcome included age, race, and gender as covariates."	A "covariate" is a variable that the researcher believes is correlated with the outcome but which is not a treatment indicator. For example, a study of reading growth may include age, sex, free-lunch status, and maternal education level as covariates for the reading test score, along with the treatment indicator. The model's precision for estimating the treatment effect is greater if the covariates explain a larger fraction of the test score's variance.	The regression model of the outcome included age, sex, and race to adjust for their influence on reading scores.
Hierarchical linear model	"Because students were clustered in classrooms, a hierarchical linear model was estimated."	A hierarchical linear model is a type of regression model for data that are "nested" in levels ("hierarchies"). For example, a study of a reading approach might examine student test scores by first choosing districts, then schools, then classrooms. In this example, students can be viewed as nested in a classroom, the classroom as nested in a school, and the school as nested in a district. Hierarchical linear models allow for variance created by nesting. The extent to which outcomes are correlated can be estimated by the "intracluster correlation coefficient."	The regression model accounted for the way in which students, classrooms, and schools were related to each other.

Concept	Examples of Technical Usage	Explanation	Revised Usage
Intracluster correlation coefficient	"Initial calculations gave indications of a significant value of the intracluster correlation coefficient, so variances were adjusted for clustering."	The intracluster correlation coefficient (ICC) is a measure of the degree to which outcomes of individuals or units within groups or "clusters" are correlated. It can range from zero to one. When it is zero, outcomes of individuals within clusters are not correlated. When it is one, outcomes of individuals within clusters are perfectly correlated. The outcome has the same value for the entire cluster. The larger the ICC, the more clusters a study needs to reach the same statistical power. Analyzing the data as if individuals are not clustered in groups will underestimate variances and can lead to incorrect conclusions about effects being statistically significant when they are not.	Initial calculations indicated that outcomes of individuals within clusters were correlated, so variances were adjusted for the correlation.
Treatment effect	"The analyses found significant treatment effects."	A "treatment effect" is the amount by which the average outcome of the treatment group differed from the average outcome of the control (or comparison) group. For example, a study may find that after using a new reading program, reading scores of the treatment group were 15 points higher than reading scores of the control group. The treatment effect is 15 points. When regression models are used, the treatment effect is the estimated coefficient for the treatment indicator variable. Regression models commonly include other variables related to outcomes, such as pre-test scores, age, gender, socioeconomic status, and so on. Commonly, these are called covariates. In a randomized controlled trial, these other variables are uncorrelated with the treatment indicator, and their role in the model is to explain part of the variance of the outcome, which enables a more precise estimate of the treatment effect. In a quasi-experiment, these other variables may be correlated with the treatment indicator, and their role in the model is both to explain part of the variance of the outcome and to adjust for the correlation with the treatment indicator.	The analysis found that average outcomes of the treatment group were higher than average outcomes of the control group. The differences were statistically significant.

Concept	Examples of Technical Usage	Explanation	Revised Usage
Effect size	"Effect sizes between 0.10 and 0.20 have been reported for the intervention."	An "effect size" is a measure of the size of the treatment effect relative to a benchmark. Different benchmarks can be used. A common one is the standard deviation of the outcome (either for the full sample or the control group). When the standard deviation of the outcome is used, an effect size of 0.20 means that the treatment effect equals 20 percent of the outcome's standard deviation. For IQ tests, which commonly are designed to have an average of 100 and a standard deviation of 15, this effect size means that the treatment increased IQ by 3 points. For an achievement test reported in "normal curve equivalent" units, the standard deviation by design is 21.06, and an effect size of .20 is an increase in NCE units of about 4. Because an effect size is measured relative to a benchmark, it does not have a unit, and effect sizes of different outcomes can be compared. Reporting that an intervention increased reading scores by an effect size of 0.20 and reduced behavior problems by an effect size of 0.10 means that the intervention had larger effects on reading than on behavior. Effect sizes can be combined for different studies of the same intervention or class of interventions, a property that accounts for their central role in meta-analyses (methods that focus on contrasting and combining results from different studies, in the hope of identifying patterns among study results, sources of disagreement among those results, or other interesting relationships that may come to light in the context of multiple studies).	When treatment effects were measured relative to a benchmark (the standard deviation of the outcome), they ranged in size from 10 to 20 percent of a standard deviation. In education research, treatment effects of this size are usually considered meaningful.

Concept	Examples of Technical Usage	Explanation	Revised Usage
Subgroup	"The study reported no effect overall, but several subgroup effects were statistically significant."	A "subgroup" is part of the overall study sample that is defined by a baseline variable or characteristic. For example, a study may be interested in whether intervention effects differ for boys and girls. Its sample would be divided into boys and girls and effects in these two subgroups would be contrasted. In evaluation studies, subgroups for which intervention effects are desired should be based on characteristics of the groups at baseline to ensure that comparisons within subgroups are internally valid (though with a smaller sample size). Basing subgroups on characteristics that can change because of the intervention (such as a reading score after one year of intervention), disrupts the internal validity of subgroup effects. The number of baseline data items determines how many subgroups a study can create. However, creating many subgroups yields many effects. Because each effect has a probability of being statistically significant due to chance, creating many subgroups increases the possibility that one or more subgroup effects are statistically significant due to chance. This is an example of the "multiple comparisons" problem.	The study reported no effect overall, but effects for boys differed significantly from effects for girls.
Multiple comparisons	"A multiple comparisons adjustment was used because the study examined a large number of outcomes."	The "multiple comparisons" problem arises when statistical tests of more than one outcome are considered together. Because each test is constructed to have a Type I error rate of, say, 5 percent, when tests are combined, the probability that at least one test will be significant by chance is greater than 5 percent. For example, if an experimental study of a new reading approach used six different tests to assess reading outcomes, the probability that at least one treatment effect is statistically significant is greater than 26.4 percent (found as $[1-.95^6]$), even when the true effect is zero.	The study examined a large number of outcomes and used a statistical adjustment to reduce the likelihood of finding a significant result for one or more outcomes by chance.

Concept	Examples of Technical Usage	Explanation	Revised Usage
Implementation fidelity	'The intervention was implemented with high fidelity.'	'Implementation fidelity' is the degree to which (1) staff implemented a program or curriculum as it was designed to be implemented, (2) staff used techniques or strategies to deliver the program or curriculum that are prescribed by the developer or meet outside benchmarks, (3) participants received the prescribed exposure to the program or curriculum, and (4) participants responded to or were engaged by the program or curriculum. Information about implementation fidelity can be important for interpreting impact evaluation results, suggesting why a program or curriculum did or did not have favorable impacts, and providing information useful for program improvement.	Staff implemented the intervention as designed, using the prescribed techniques, and most participants attended all sessions and were actively involved in session activities.
Dosage	"Program dosage was lower than anticipated."	"Dosage" is the amount of exposure to a program or curriculum. The "intended dosage" is the exposure to the program or curriculum prescribed by the developer or required by funders. The "offered dosage" is the exposure to a program or curriculum that staff deliver. The "received dosage" is the exposure to a program or curriculum that participants actually get.	Although staff offered the intended number of program sessions, the number received by participants was lower because of high absence rates.
Implementation quality	"A significant positive relationship existed between quality of implementation and academic performance."	"Implementation quality" refers to how well the program or curriculum was delivered and received. Implementation quality can be assessed by using benchmarks or assessment tools provided by the developer or by using criteria or assessment tools that reflect best practices. While implementation fidelity refers to the extent to which a program or curriculum was implemented as designed, implementation quality refers to the skill with which staff implemented the planned program or curriculum.	When the curriculum was implemented well (the teacher was well-prepared, presented the lessons clearly, responded accurately to student questions, and established good rapport with students), students' academic performance was stronger, on average.

Concept	Examples of Technical Usage	Explanation	Revised Usage
Sensitivity analysis	"Sensitivity analyses indicated that the results differed, depending on how missing values were imputed." "The sensitivity analysis showed that including additional years of data did not affect the findings."	A sensitivity analysis uses a variety of approaches to examine whether a study's results vary if other assumptions were made or other approaches were used. For example, different approaches can be used to impute missing data. A sensitivity analysis would estimate effects by using approaches not used in the main part of the study and would assess whether the assumption mattered for the findings. Similarly, sensitivity analyses can enter different variables into models, or use different types of data, such as income from wage records versus income from self-reports, to assess whether findings vary when assumptions vary. Findings are "robust" if the sensitivity analyses determine that findings are not affected by varying assumptions.	Varying how the missing data were imputed did not affect the study's findings. Adding more years of data did not affect the study's findings.