HAS INQUIRY MADE A DIFFERENCE?
A SYNTHESIS OF RESEARCH ON THE IMPACT OF INQUIRY SCIENCE INSTRUCTION ON STUDENT OUTCOMES

Technical Report 6: Operationalizing the Coding of Research Rigor, Context, and Study Findings

The Inquiry Synthesis Project

Center for Science Education

Education Development Center, Inc.

Final Update JUNE 2009

Please use the citation below when referring to this document:

The Inquiry Synthesis Project, Center for Science Education, Education Development Center, Inc. (EDC) (2009, June). *Technical report 6: Operationalizing the Coding of Research Rigor, Context, and Study Findings*. Retrieved [date retrieved], from http://cse.edc.org/products/inquirysynth/pdfs/technicalReport6.pdf.

This research is supported by a grant from the National Science Foundation (NSF) (#ESI-0101766). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Copyright © 2009 Education Development Center, Inc.

INTRODUCTION

This technical report is the sixth in a series of reports that introduce and explain structures and processes used for the Inquiry Synthesis Project, which is addressing the research question: What is the impact of inquiry science instruction on student outcomes? *Technical Report 1: Generating the Synthesis Sample of Studies* describes **Phase I** of the project (Report Collection) and the criteria used for conducting the search for research reports to include in the synthesis. **Phase II** of the project is the Coding Process, which has three stages:

1. Inclusion/exclusion coding—*Technical Report 3: Operationalizing the Inclusion/Exclusion Coding Process*
2. Inquiry-science-instruction coding—*Technical Report 5: Operationalizing the Inquiry Science Instruction Coding Process.*
3. Research rigor, context, and findings coding

The third stage of coding captured information about the context of the study, methodological integrity of the research, and study findings. This stage of coding is described in this technical report. There are two other technical reports in this series. *Technical Report 2* discusses the structure for describing inquiry science instruction used in this synthesis and *Technical Report 4* describes the issues associated with the unit of analysis for the synthesis.

PREVAILING ISSUES

We choose the term "synthesis" intentionally to refer to our work because the inclusion of both qualitative and quantitative studies *and* because we used both quantitative and qualitative methods in our own synthesis activities. Throughout our work, we were always conscious of the paradigmatic tension between positivist and constructivist approaches to research and made efforts to reconcile this tension so that we could represent the entire research landscape as a coherent whole body of work. In some instances, we reconciled this paradigmatic tension by coding studies with the same protocols (e.g., inquiry instruction), and in other instances, we made parallel assessment items that were framed within each paradigm (e.g., rigor and findings). We also collected both narrative and numerical data on all studies in our sample regardless of their own design type.

Future dissemination efforts will further specify our approach to paradigmatic reconciliation, but we wanted to address at this juncture what we did not do in this synthesis. Within both paradigms there are synthesis techniques that are typically used. In the positivist approach, these techniques are meta-analysis (Lipsey & Wilson, 2001), or integrative research review (Cooper & Hedges, 1994); and in the constructivist approach, it is qualitative synthesis (Barbour & Barbour, 2003), meta-synthesis (Paterson, Thorne, Canam, & Jillings, 2001), or meta-ethnography (Noblit & Hare, 1988). Since each of these methodological techniques is entrenched within one paradigmatic tradition, we did not use these techniques in this synthesis because it was designed to reflect the cross-paradigmatic body of work addressing our research question. As a result, while we see the utility of delving more deeply into the studies in our dataset with these more-refined paradigm-specific analysis techniques, this was not done for the current project.

As is typical of any kind of research synthesis work, there are numerous challenges related to the nested nature of the data that are necessary to capture. In the inquiry synthesis dataset, there are six levels of data: research document, research report, individual study, instructional treatments in a study, data sources used in a study, and findings generated from a study. The challenges related to identifying "a unique study"—our unit of analysis for this synthesis—from research documents and reports are articulated in *Technical Reports 1* and *4*. Once a study has been identified, however, there are still four levels of data to be captured in the third stage of coding that relate to the study research design and methodology, data sources, contextual information about instructional treatments, and findings. Capturing each of these levels of data will be discussed in this report; however, to give the reader a sense of the scope of the issue, the synthesis dataset contains 10 design types, and within a study, a maximum of 8 instructional treatments, 12 data sources, and 12 findings were coded.

Copyright © 2009 Education Development Center, Inc.

NATURE OF THE STUDIES INCLUDED IN STAGE 3 CODING

To have a more conceptually consistent database of studies to include in the analysis, we decided to exclude a number of studies that had passed through the previous two stages of coding but still presented interpretive challenges for our analysis.

Threshold for Descriptive Clarity of the Instructional Treatment

We recognized from the previous two stages of coding that it was essential for a study to exhibit sufficient descriptive clarity of the treatment to understand its impact on student outcomes and, thus, to warrant further investment in staff coding time. As a result, we determined that a study had to provide sufficient information for a coder to clearly determine the presence or absence of at least three (a majority) of the five components of instruction to meet this threshold for descriptive clarity of the instruction to warrant its inclusion in the analysis. Thus, studies that reported on two or fewer of the five components of instruction were excluded. We also excluded studies from the analysis when the multiple instructional treatments administered in the study could not be distinguished from one another based on our coding protocols and, thus, did not allow us to connect our independent and dependent variables of interest in these studies.

Focusing the Dependent Variable

Increasing students' science content knowledge is of high interest for the field of education given the current policy context; therefore, we decided to focus on the studies in our dataset that had the following dependent variables: student understanding or retention of science facts, concepts, or principles and theories in physical science, life science, or earth/space science. Studies that did not have any explicit instruction in either physical, life, and/or earth/space science content, as indicated on the inquiry coding, were eliminated.

Within the dataset there were a number of descriptive studies that were not intended to measure changes in student understanding as a result of instruction but, rather, their intention was to describe how students use a particular curriculum, what they do with it, and how they interact with it and their classmates and/or teachers. These studies were not excluded earlier in the coding process because we accepted exploratory questions as valid for this synthesis. However, we excluded studies with such exploratory questions at this point because they addressed a different research question than we are currently investigating.

Focusing on Generalizability

Along with refining the inclusion criteria for the instruction and dependent variables, we also were mindful of the context diversity within which the interventions described in the studies in our sample took place. To have studies in our analysis that were most relevant for the K–12 classroom, we excluded studies that had limited interpretive validity for that environment. These included studies conducted in museum contexts and case studies of individual students. Single-subject studies offered such a close look at instruction and student outcomes that it would be difficult, because of their narrow focus, to apply their findings in a meaningful way to studies of larger groups of students. This decision is similar to our earlier decision to set aside studies that were so large in scale that it was impossible to portray the nature of the instruction and, thus, apply their findings to our analysis. At the same time, we recognize that single-subject studies offer detail and perspective that could add richness to our synthesis of larger-scale studies. As a result, we will set the single-subject studies aside for now, and we can turn to them in the future, as we can turn to the larger studies as additional sources of data on the impact of inquiry science instruction on student outcomes. The final sample included in the stage 3 coding and analysis was 138 studies.

GENERAL CODING PROCESS FOR STAGE 3

Prior to the beginning of the coding process, the codebook was tested and revised. During this phase, all members of the research team used the codebook to code the same study and then met as a group to review ratings and make revisions that addressed and accommodated differences in understandings. The development of definitions and examples for each item in the codebook was a critical component of the testing and refinement process.

After three rounds of codebook testing and refinement, it was determined that the coding process needed to proceed by consensual coding teams until appropriate levels of inter-rater agreement were achieved. Each study was coded independently by two researchers on a coding team whereby the coders clearly documented

Copyright © 2009 Education Development Center, Inc.

their evidence to support their coding choice in the study text itself and on the coding sheet. Each coding pair then met to consensually code and to reconcile any differences in ratings. If differences could not be reconciled through discussion, a senior researcher was called in to resolve the question based on the evidence presented. Once agreement was reached, reconciled data were entered into the dataset.

Approximately 82 studies were coded by teams of coders in the manner described above. An additional 56 studies were coded independently by coders who had reached an inter-rater agreement rate of 85%. Any questions regarding how specific items should be coded were reviewed together before a decision was reached. Every fifth study was coded commonly and then consensually coded to maintain consistency.

CODING FOR THE RESEARCH DESIGN

One of the major objectives of this synthesis was to intentionally include a diversity of research designs so that we could develop a cross-paradigmatic understanding of our research question. Therefore, we coded for three research designs—experimental, quasi-experimental and non-experimental—which categorized *how* the subjects of study were selected and *when* data is collected in relation to the educational intervention under study. Some of the specific design types that we used within these broader categories are not typically addressed in classification schemes (Pedhazur & Schmelkin, 1991) because they are considered to be confounded. However, to be inclusive, we did not apply a standard for specific designs that would be included in the synthesis. Rather, we chose to describe the designs we encountered in the field and assess the methodological rigor appropriately for each type of design that was employed in a study. This allowed us to present the most complete picture of the state of research that was considered relevant to our research question. Thus, the general scheme we used to discriminate research designs was customary distinction between the presence or absence of random assignment to treatment groups.

An *experimental design* was defined as a study that uses quantitative data that entails manipulation of an independent variable and random assignment. These studies met the following criteria:

- At least one variable is manipulated (e.g., the type of instruction or the amount of instruction)
- Subjects/schools/classes are randomly assigned to the different levels or categories of the manipulated variable, thus constituting at least one experimental group and at least one comparison/control group; randomization can occur at any level to qualify for this design
- Can be one or multiple occasions of measurement

Within the synthesis dataset are three specific types of experimental designs: (1) equivalent control group design—pre-post measurement (also includes more than one post-test measurement); (2) equivalent control group design—post-only measurement; and (3) within subject cross-over design—subjects in the different treatment groups are randomly assigned the order in which they receive the treatments and, thus, serve as their own matched comparison (see Figure 1 below).

Figure 1. Within Subject Cross-over Design

X_1= intervention of interest
X_2= comparison intervention
O= point of data collection

Randomly assigned order of treatment:

group 1: O X_1 O X_2 O
---
group 2: O X_2 O X_1 O

A *quasi-experimental design* was defined as a study that uses quantitative or qualitative data that entails manipulation of an independent variable, and a comparison/control group is present but randomization is absent at all levels. These studies met the following criteria:

- At least one variable is manipulated (e.g., the type of instruction or the amount of instruction)

Copyright © 2009 Education Development Center, Inc.

- Subjects/schools/classes are **NOT** randomly assigned to the different levels or categories of the manipulated variable, thus constituting at least one experimental group and at least one comparison/control group
- Can be one or multiple occasions of measurement

Within the synthesis dataset are three specific types of quasi-experimental designs: (1) non-equivalent control group design—pre-post measurement; (2) non-equivalent control group design—post-only measurement; and (3) qualitative time-series design—multiple pre-treatment and post-treatment measures of the dependent variable were made across multiple students in different classrooms. The different classrooms constituted a multiple treatment group comparison, thus categorizing this as quasi-experimental.

A *non-experimental design* was defined as a study that uses either quantitative or qualitative data, does not have a comparison group, has no level of randomization, and the type or amount of instruction as the independent variable is documented but may not be formally manipulated. Within the synthesis dataset are four specific types of non-experimental designs: (1) single treatment group design—pre-post measurement; (2) single treatment group design—post-only measurement; (3) single treatment group design—multiple measurement points (e.g., teacher reports of classroom activities across a number of students); and (4) single treatment group design—one measurement point during the intervention.

CODING FOR THE CONTEXTUAL INFORMATION ABOUT INSTRUCTIONAL TREATMENTS

To add explanatory power to our study, we captured additional information about the context within which the treatments took place for any given study. This information allowed us to explore our findings to see if there were any patterns related to differences in the students or the instructional conditions. Though we had captured some information about the context of reports in our inclusion/exclusion coding, we did not capture this information at the study and treatment level until this third stage of coding. For each treatment group in a study, we coded for:

- who provided the instructional intervention (e.g., researcher, teacher, guest scientist, textbook, computer simulation);
- the instructional provider's experience using the pedagogical approach in the treatment prior to implementing it for the research study (e.g., none, up to three years, more than three years);
- who conducted the research (e.g., researcher, teacher as researcher, joint collaboration);
- the kind of instructional setting in which the study took place (e.g., regular classroom, science classroom, outdoor setting, indoor informal setting, artificial research setting);
- the duration of the treatment (e.g., number of classes, days, weeks, months, years); and
- demographic information about students in the sample (e.g., number of students, age, ethnicity composition, socio-economic status composition, location of the students—urban, rural, suburban, learning status composition, gender composition).

Another category of variables that we captured were the factors that influenced the instructional context. These factors covered the broader educational and policy context within which the instruction took place, and were much more commonly reported with qualitative data. We rated each of the factors below as influencing the instruction either positively, negatively, or neutrally (i.e., a factor was part of the context that was explored and shown to have no influence on the instructional treatment):

- Demographics of the teacher (e.g., race, gender)
- Physical classroom environment
- Classroom culture (e.g., from the beginning of the year, the teacher developed a culture that encouraged self-confidence, engagement in discussions, ability to ask relevant questions)
- Professional development of the instructor
- Principal engagement
- Science program support (e.g., access to instructional support from science program leaders at the school or district level, meetings planned by the science program for science teachers to address topics of interest)
- Accountability measures (e.g., regulations that directly affect the amount of science instruction or the nature of science instruction that science teachers provide)
- Financial resources

Copyright © 2009 Education Development Center, Inc.

- Parental involvement
- External partnerships
- Teacher content knowledge
- Other (e.g., instructor enthusiasm for the subject matter of the intervention)

CODING FOR METHODOLOGICAL RIGOR

Since the synthesis database contains both qualitative and quantitative studies, we made efforts to develop coding protocols that are universally applicable to research in either paradigm. In developing the stage-3 protocol, we referenced a number of resources on research rigor (see the reference section of this report for a full list). We felt it was imperative to develop a coding system that captured the methodological rigor with which each study was designed and conducted. This would ensure the integrity of our own findings and allow us to explore the relationship between level of methodological rigor and findings in the existing knowledge-base of science education.

We developed a set of items for this stage of coding that could describe three aspects of rigor and, thus, provide a guide for us in interpreting our findings based on their composite trustworthiness. These three aspects are descriptive clarity, data quality, and analytic integrity. The descriptive-clarity items capture the amount and clarity of information that was provided to allow for independent assessment of the research site, sample, treatment activities, data collection strategies, data analyses, and findings. The data-quality items address the technical aspects of the methodology—appropriate sample sizes, reliability, validity of the instruments, and influence of attrition. The analytic-integrity items address the appropriateness of data analysis strategies used, the ability of the design and analysis to account for threats to internal and external validity, and the appropriateness of the data coding procedures used. Not all of the items noted below are appropriate for every design type; however, we ensured that there were items for each aspect of rigor that were appropriate for comparative and non-comparative treatment designs as well as for both qualitative and quantitative methods. This allowed for each aspect of rigor to be coded for every study in our dataset.

Descriptive Clarity

Descriptive clarity relates to the extent to which the author has described the research endeavor thoroughly and clearly. When descriptive clarity is poor, the reader knows little and must make more inferences. When the descriptive clarity is greater, the reader knows more and makes fewer inferences. Descriptive clarity and rigor are directly related, the former enhancing the latter. Clarity can be present in the descriptions of the components of a study, including the research question, the research site, the sample, the sampling strategy, the inquiry treatment, the comparison and/or control treatment (if there is one), the data collection strategy, the data that is collected, and the approach to analysis.

For each of the items listed below, we coded whether the information provided was good, marginal, or poor, or whether no information was provided at all.

- The extent to which the research questions are clearly stated (or not articulated)
- The extent to which the research site is described regarding:
  a) type of setting (e.g., elementary, middle, high, nature center, museum)
  b) type of community (e.g., urban, suburban, rural)
  c) size of school
  d) public or private school
  e) geographic location in the country
  f) economic profile of community/school population (e.g., affluent, professional, working class, high poverty)
- The extent to which the sample is described regarding:
  a) specific ages of students in the sample
  b) race of students in the sample
  c) SES of students in the sample
  d) gender of students in the sample
  e) number of students in the sample
- The extent to which the sampling strategy is described regarding:
  a) how the sample was chosen (e.g., convenience, random, snowball)

Copyright © 2009 Education Development Center, Inc.

b) why the particular strategy was chosen
c) how the sample was identified
d) the degree to which students were (or were not) representative of the sampling frame from which they were selected (e.g., the three students selected from a particular class were the highest achievers of the group)

- The extent to which the data collection strategy is described regarding:
  a) how data was collected (e.g., observation, test, interview, videotapes)
  b) when data was collected in relation to the intervention
  c) the kinds of instruments used to collect the data (e.g., the type of test, the interview protocols if used)
  d) why the data collection strategy was chosen
  e) the researcher's role in the data collection process
- The extent to which the analysis strategy is described regarding:
  a) what the analysis strategy is
  b) why the data analysis strategy was chosen
  c) the timing and process of the steps of the analysis strategy
  d) the use of the available data sources in the analysis
  e) a description of who is conducting the analysis and how they will arrive at conclusions
- The extent to which the treatment of interest and comparison treatments are described, referring to the coder's ability to picture what was done in the classroom

Table 1: Methodological Quality codes assigned to indicators of *descriptive clarity*

Aspects	Indicators*
Extent to which the research question is stated.	-1=no research question or focus is stated 0=only the area of investigation is described 1=the focus or purpose of the research is stated 2=research questions are clearly stated
Extent to which the research site is described.**	• type of setting (e.g. elementary, middle, museum) • type of community (e.g. urban, suburban, rural) • size of school • public or private school • geographic location in the country • economic profile of community/school population (e.g. affluent, professional, working class, high poverty)
Extent to which the sample is described.**	• specific ages of students in the sample • race of students in the sample • SES of students in the sample • Gender of students in the sample • number of students in the sample
Extent to which the sampling strategy is described. (n/a=not a comparison design) **	• how the sample was chosen (e.g., convenience, random, snowball, etc.) • why the particular strategy was chosen • how the sample was identified • the degree to which they were (or were not) representative of the sampling frame from which they were selected (e.g., the 3 students selected from a particular class were the highest achievers of the group)
Extent to which the data collection strategy is described.**	• how the data was collected (e.g., observation, interview, video, etc.) • when data was collected in relation to the intervention it was collected • the kinds of instruments used to collect the data (e.g., the type of test, the interview protocols if used) • why the data collection strategy was chosen • the researcher's role in the data collection process

Copyright © 2009 Education Development Center, Inc.

Aspects	Indicators*
Extent to which the analysis strategy is described**	• what the analysis strategy is • why the data analysis strategy was chosen • the timing and process of the steps of the analysis strategy • the use of the available data sources in the analysis • a description of who is conducting the analysis and how they will arrive at conclusions
Extent to which the treatment of interest is described.	-1=no sense of the instruction that was done in the classroom 0=general sense of what was done in the classroom, but it is very vague 1=understand one aspect of instruction but not another 2=clear picture of what was done in the classroom
Extent to which the comparison treatment(s) is described. (n/a=not a comparison design)	-1=no sense of the instruction that was done in the classroom 0=general sense of what was done in the classroom, but it is very vague 1=understand one aspect of instruction but not another 2=clear picture of what was done in the classroom

*Methodological Rigor Quality Scale: -1= very poor rigor; 0=poor rigor; 1=marginal rigor; 2=good rigor
**-1=no description of the items listed
0=one or two of the items listed are described
1=three of the items listed are described
2=four or more of the items listed are described

Data Quality

Coding for the Reliability and Validity of the Data from Data Sources

Within a study, researchers often used a number of different assessment instruments. We were interested in capturing the information related to the data sources researchers used to determine the effect of instructional interventions on students' science content understanding or retention (dependent variable). We divided the types of data sources into two categories: (1) data collection instruments and protocols (e.g., norm and criterion referenced tests; interview, observation, and focus-group protocols; questionnaires and surveys; student-work collection instruments—concept map templates, performance assessments), and (2) other data sources (e.g., unstructured observation/field notes, student work, school records, science course grades, teacher notes, video observation/notes, interview transcripts, classroom transcript, and unstructured classroom conversations).

The instruments and protocols in *the first category* were subject to further scrutiny to determine the quality of the data that was produced from these items based on the following:

- The status of pilot testing of the instrument (e.g., a new instrument was developed for this study and was **NOT** pilot tested, a new instrument was developed for this study and was pilot tested, or an established instrument was used)
- Whether or not appropriate kinds of instrument reliability was demonstrated for the current sample to which it was being applied (e.g., test-retest reliability, internal consistency, alternate form reliability, inter-rater agreement)
- If reliability was tested, what level was achieved (e.g., low, acceptable or good)
- Whether or not the validity of the instrument was demonstrated for the current sample to which it was being applied (e.g., content, convergent, discriminant, criterion-related)

Coders were then asked to rate the overall instrument quality as indeterminable (none of the bulleted items were addressed sufficiently); poor (one or two of the bulleted items were addressed sufficiently); marginal (three of the bulleted items were addressed sufficiently); or good (all four of the bulleted items were addressed sufficiently). For studies with more than one instrument, the rating to this overall instrument quality item was averaged across instruments and that was the score used in the final rigor calculations.

The data sources in *the second category* were of the nature that made it inappropriate to judge them for traditional forms of instrument reliability and validity to determine the quality of the data that they generated. Therefore, we turned to the qualitative paradigm for guidance (Maxwell, 1992) on determining the factual

Copyright © 2009 Education Development Center, Inc.

accuracy of the researcher's account and, thus, the descriptive validity of the data generated from these types of data sources. Some of the factors that we considered when determining data quality included the following:

- The researcher's presence at the research site
- The data sources the researcher used
- The researcher's experience conducting research
- The researcher's experience with the subject/site of study
- The researcher's acknowledgment of his or her role/perspective in data collection and the potential for bias
- The researcher's use of memoing, peer debriefing/audit—with another researcher, member checking—with the subject

All of the data sources in this second category were considered together for the study's overall rating of qualitative data validity as follows:

- No validity—there is evidence that indicates the researcher's credibility is in question
- Poor validity—credibility of the researcher is evident in *few* of the factors listed above, and this evidence is *weak*; weak evidence, for example, is that which is minimal, inferred, and/or of an uncertain source
- Marginal validity—credibility of the researcher is evident in *some* of the factors listed above, and some of the evidence is *weak and* some is *strong*; strong evidence, for example, is that which is abundant, from multiple sources, and/or explicit
- Good validity—credibility of the researcher is evident in *some or all* of the factors listed above, and the evidence that is present is *strong*

Coding for Design or Measurement Flaws that Effect Data Quality

There were three other items that contributed to the rating for a study's data quality:

- The extent to which sample attrition was addressed
- The extent to which treatment contamination was addressed
- The extent to which pretreatment differences between subjects were addressed

Table 2: Methodological Quality codes assigned to indicators of *data quality*

Aspects	Indicators*
Extent to which attrition of subjects is addressed in the analysis.	-1=not reported -1=attrition is present but NOT addressed in the analysis 2=attrition is present and it is addressed in the analysis 2=no attrition present n/a=no time elapsed between treatment and measurement
Extent to which treatment groups are kept separate for the duration of the study.	-1=not reported -1=treatment groups are NOT kept separate 2=treatment groups are kept separate for the treatment n/a=not a comparison design
Consistency of instructional providers for the treatment groups.	-1=not reported -1=Instructional providers are the same 2=Instructional providers are NOT the same n/a=not a comparison design
Pretreatment differences between subjects in treatment groups	-1=not reported -1=Some differences, impact on findings unclear because pretreatment difference not taken into account in the analyses 2=Some differences, but corrected for in the analyses 2=Negligible differences, non-influential impact on findings n/a=not a comparison design
Level of instrument quality	-1=not reported 0=one or two bullets below apply 1= three bullets below apply 2=all four bullets below apply n/a=no instruments were named

Copyright © 2009 Education Development Center, Inc.

Aspects	Indicators*
	• Pilot testing was done • Appropriate reliability tests were performed • Calculated reliability levels were acceptable • Some form(s) of instrument validity were demonstrated
qualitative data validity	-1=not reported -1=evidence to indicate the researcher's credibility is in question 0=credibility of the researcher is evident in *few* of the factors listed below and this evidence is *weak*. Weak evidence is minimal, inferred, and/or of an uncertain source. 1=credibility of the researcher is evident in *some* of the factors listed below, and some of the evidence is *weak and* some is *strong*. Strong evidence is abundant, from multiple sources, and/or explicit. 2=credibility of the researcher is evident in *some or all* of the factors listed below, and the evidence that is present is *strong*. n/a=not applicable (quantitative methods) • The researcher's presence at the research site; • The data sources the researcher uses; • The researcher's experience conducting research; • The researcher's experience with the subject/site of study; • The researcher's acknowledge their role/perspective in data collection and the potential for bias; • The researcher uses memoing, peer debriefing/audit—with another researcher, member checking—with the subject

*Methodological Rigor Quality Scale: -1= very poor rigor; 0=poor rigor; 1=marginal rigor; 2=good rigor

Analytic Integrity

Analytic integrity is determined by whether or not the researcher employed techniques to investigate rival hypotheses in explaining the results of the analysis (e.g., threats to internal validity, effects of missing data, relationship of non-significant to significant results, relationship of exemplars to exceptions in qualitative data).

The following items were developed and used to determine a study's analytic integrity.

- Was there systematic analysis of quantitative data evidenced by the following:
  a) The unit of *analysis* matching the unit of assignment to treatment groups (only for multiple-comparison-group designs)
  b) Both inferential and descriptive statistics were reported for dependent variables
  c) Inferential statistics were correctly interpreted (e.g., statistical significance interpreted relative to practical significance)
- To what extent did researchers discuss how they handled missing data on the data sources so that an independent judgment could be made about the effect on the results reported?
- Given the inherent limitations of the specific design type in this study, were there fatal threats to internal validity that severely compromised the study's findings, such as the following:
  a) History
  b) Maturation
  c) Testing
  d) Instrumentation
  e) Regression artifacts
  f) Selection bias
  g) Experimental mortality
- In quantitative analyses, has the researcher accurately accounted for the value of his or her statistically non-significant results in his or her interpretation of the results?
- Was there a systematic analysis of the qualitative data evidenced by the following:
  a) Developing matrices (or other categorization schemes) of codes or themes

Copyright © 2009 Education Development Center, Inc.

b) Developing and applying codes to capture theme of interest in the data
c) Creating cognitive maps to capture subjects' information/data

- Was there evidence to support qualitative findings?
  a) Strength of evidence
  b) Extent to which all data are accounted for in the analysis
  c) Individual or sub-samples identified as representative or unique

Table 3: Methodological Quality codes assigned to indicators of *analytic integrity*

Aspects	Indicators*
appropriate and systematic analysis of quantitative data	-1= no, one or more bullets below not addressed 2=yes, all three bullets below are addressed n/a= not applicable—qualitative methods • the unit of *analysis* matches the unit of assignment to treatment groups (only for multiple comparison group designs) • both inferential and descriptive statistics were reported for dependent variables • inferential statistics are correctly interpreted
qualitative systematic analysis	-1=no evidence of systematic analysis as described in bullets below 2=yes there is evidence of systematic analysis described in bullets below n/a= not applicable—quantitative methods • development of matrices (or other categorization schemes) of codes or themes • developing and applying codes to capture theme of interest in the data • creation of cognitive maps by the researcher to capture subjects' information/data
Extent to which researchers discussed how they handled missing data on the data sources so that an independent judgment could be made about the effect on the results reported.	-1=not reported -1=there is missing data and researchers did not discuss how they handled missing data 1=there is missing data and researchers discussed how they handled missing data 2= there is no missing data
quantitative validity threats	-1=not enough information to determine validity threats -1=there are threats which compromise the study's findings (thus, not valid) 2=there are NOT threats that compromise the study's findings (thus, valid) n/a=not applicable (qualitative methods) potential threats to internal validity: • history • maturation • testing • instrumentation • regression artifacts • selection bias • experimental mortality
lack of bias in reporting qualitative findings/results	-1=no, qualitative findings were not supported with evidence 2= yes, qualitative findings were supported with evidence n/a=not applicable—quantitative methods • Strength of evidence was demonstrated • Extent to which all data are accounted for in the analysis • Individual or sub-samples are identified as representative or unique

Copyright © 2009 Education Development Center, Inc.

Aspects	Indicators*
quantitative non-significant (ns) findings reported	0=Non-significant findings are not reported 1=Non-significant findings **are** reported but they **are not** accounted for 2=Non-significant findings **are** reported and they **are** accounted for n/a=Not applicable (qualitative study or no ns findings were found)

*Methodological Rigor Quality Scale: -1= very poor rigor; 0=poor rigor; 1=marginal rigor; 2=good rigor

CODING FOR FINDINGS

Again, due to the inclusive nature of the dataset for the synthesis, we had to develop a coding scheme for findings that could be applied across types of designs and data to capture comparable information to be included in the analysis. Therefore, we developed research-design-appropriate parallel items to capture finding direction and supportive evidence.

Defining Finding Type

The dependent variable in this study, science-subject-matter content knowledge, was divided into six different finding types that could be expressed in physical science, life science or earth/space science:

- Understanding(s) related to science facts and vocabulary
- Understanding(s) related to science concepts
- Understanding(s) related to science principles and theories
- Retention of science facts and vocabulary
- Retention of science concepts
- Retention of science principles and theories

We thought this level of discrimination reflected a meaningful distinction between levels of scientific information and between the initial acquisition of this information and long-term retention of the information. We distinguished facts as discrete, concrete, observable bits of scientific information (e.g., pitch is a characteristic of sound; iron is an element; water comprises of two hydrogen atoms and one oxygen atom). Concepts, however, are more abstract and reflect common characteristics of a whole class of objects, ideas, or happenings (e.g., density, electricity, properties of matter, heredity). Principles are statements that describe relationships between and among concepts; for example, "opposite poles of a magnet attract" is a principle, and within this, the concepts are opposite, poles, magnet, and attract. Another example of a principle is "mass equals force times acceleration," and within this principle, the concepts are mass, force, and acceleration. Theories draw principles together in an effort to explain a more global phenomenon (e.g., gravitation, atomic structure, and, of course, evolution). In the studies in our synthesis, it was not possible to distinguish between principles and theories; thus, they were collapsed into one finding type.

Finding Direction

The second part of the finding information that was coded was the direction of the finding. For non-experimental designs that used qualitative data sources, we coded whether students showed a negative, a positive, a mixed, or no response to the instructional intervention. For non-experimental designs that used pre-post quantitative data, we coded whether students showed a non-significant but negative trend on post-treatment scores, no indicated difference between the pre- and post-treatment scores, non-significant but positive trend on post-treatment scores, or post-treatment scores were significantly better than pre-treatment scores. For quasi-experimental or experimental designs that used quantitative data, we coded which treatment group(s) were statistically significantly better on post-treatment scores compared with other treatment groups, or whether there was no significant difference between treatment-group outcomes. For quasi-experimental or experimental designs that used qualitative data, we coded which treatment group(s) were better on post-treatment scores compared with other treatment groups, or whether there was no difference between treatment-group outcomes.

Identifying Supportive Evidence

For each finding based on qualitative data, we determined whether it was supported by strong evidence or weak evidence. When the study provided a description of students' observable behaviors that supported the findings, this was considered strong evidence. When a study suggested that there was evidence to support a

Copyright © 2009 Education Development Center, Inc.

finding, but there was no further description of the observable behaviors comprised in that evidence, the evidence was considered to be weak. For each finding, we selected one of the five choices below.

- There was *no* evidence to supports this finding
- There was little evidence to support this finding and it was *weak*
- There was a lot of evidence to support this finding but it was *weak*
- There was little evidence to support this finding but it was *strong*
- There was a lot of evidence to support this finding and it was *strong*

For each finding based on quantitative data, we documented the primary types of *statistics* used to demonstrate each finding, which included the following options:

- Means and standard deviations
- Gain scores
- *t*-test
- *F*-test
- chi-square
- Frequency counts
- Proportions
- Probit analysis results
- Odds ratios/cross-product ratio
- Regression coefficients
- Correlation coefficients
- Factor analysis loadings
- Multi-level modeling coefficients (this included statistics generated from structural equation modeling, analysis of covariance structures, hierarchical linear modeling)

Unique Findings

We also documented if any particular findings were unique for particular students and, if so, whether these students experienced any unique conditions in the instructional intervention. This information was recorded in our text dataset to be analyzed using qualitative techniques.

REFERENCES FOR RESOURCES USED IN DEVELOPING THIS STAGE 3 CODING PROTOCOL

Burns, N. (1989). Standards for qualitative research. *Nursing Science Quarterly, 2*(1), 44-52.

Cooper, H. & Hedges, L. (Eds.) (1994). *The handbook of research synthesis*. New York, NY: Russell Sage Foundation.

Khan, K. S., & Kleijnen, J. (2001). Stage II: Conducting the review, Phase 4: Selection of studies. In *Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews* (CRD Report Number 4, 2nd Edition). Retrieved April 30, 2002, from University of York, National Health Service Center for Reviews and Dissemination Web site: http://www.york.ac.uk/inst/crd/crd4_ph4.pdf

Khan, K.S., ter Riet G., Popay, J., Nixon, J., & Kleijnen, J. (2001). Stage II: Conducting the review, Phase 5: Study quality assessment. In *Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews* (CRD Report Number 4, 2nd Edition). Retrieved April 30, 2002, from University of York, National Health Service Center for Reviews and Dissemination Web site: http://www.york.ac.uk/inst/crd/crd4_ph5.pdf

Khan, K.S., & Kleijnen, J. (2001). Stage II: Conducting the review, Phase 6: Data extraction and monitoring progress. In *Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews* (CRD Report Number 4, 2nd Edition). Retrieved April 30, 2002, from University of York, National Health Service Center for Reviews and Dissemination Web site: http://www.york.ac.uk/inst/crd/crd4_ph6.pdf

Copyright © 2009 Education Development Center, Inc.

Kirk, J. & Miller, M. (1986) *Reliability and Validity in Qualitative Research.* Newbury Park: Sage.

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. In *Applied Social Research Methods Series, 49*. Thousand Oaks, CA: Sage Publications.

Maxwell, J. (1992). Understanding and Validity in Qualitative Research. *Harvard Educational Review*, *62*(3), 279–299.

Mishler, E. (1990). Validation in inquiry-guided research: The role of exemplars in narrative studies. *Harvard Educational Review, 60*(4), 415-442.

Onwuegbuzie, A. J. & Daniel, L. G. (2003, February 19). Typology of analytical and interpretational errors in quantitative and qualitative educational research [Electronic version]. *Current Issues in Education, 6*(2). Retrieved March 14, 2003, from Arizona State University, College of Education Web site: http://cie.ed.asu.edu/volume6/number2

Patton, M. (1990). *Qualitative evaluation and research methods, 2nd edition*. Newbury Park: Sage. (Chapter on "Enhancing the Quality and Credibility of Qualitative Analysis.")

Pedhazur, E. J. & Schmelkin, L. Pl. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Sandelowski, M., & Barosso, J. (2002, Winter). Reading qualitative studies. *International Journal of Qualitative Methods*, *1*(1), Article 5. Retrieved from University of Alberta, Canada, International Institute for Qualitative Methodology Web site: http://www.ualberta.ca/~ijqm/

Slavin, R. E. (2003). A reader's guide to scientifically based research. *Educational Leadership,* 12-16.

Spindler, G., & Spindler, L. (1992). Cultural process and ethnography: An anthropological perspective. In M. D. LeCompte, W. I. Millroy, & J. Preissle (Eds.*), Handbook of qualitative research in education*. San Diego, CA: Academic Press.

The Cochrane Collaboration. (2002, April). *Cochrane reviewers' handbook 4.1.5*. Retrieved April 30, 2002, from http://www.cochrane.org/cochrane/hbook.htm

United States Department of Education, What Works Clearinghouse. (2003, January). *Cumulative research evidence assessment device, Version 0.6*. Retrieved from http://w-w-c.org/standards.html

United States Department of Education, What Works Clearinghouse. (2003, January). *Cumulative research evidence assessment device, (CREAD)Version 0.6*. Retrieved March 5, 2003, from http://w-w-c.org/standards.html

United States Department of Education, What Works Clearinghouse. *(*2003, March). *Study design and implementation assessment device (Study DIAD) Version 0.6*. Retrieved March, 5, 2003, from http://w-w-c.org/standards.html

United States Department of Education, What Works Clearinghouse. *(*2003, March). *The process of study DIAD development*. Retrieved March 5, 2003, from http://w-w-c.org/standards.html

Vogt, W.P. (1999). *Dictionary of statistics & methodology: A nontechnical guide for the social sciences*. Thousand Oaks, CA: Sage Publications.

Workshop on Education Research Methods, Division of Research, Evaluation and Communication, National Science Foundation, November 19–20, 1998. (1999, April). *Research Methods in Mathematics and Science Education Research: Report of a Workshop*. Symposium conducted at the meeting of the American Educational Research Association, Montreal, Canada.

ADDITIONAL INFORMATION

For more information on this or other CSE research projects or to view additional technical reports, visit http://cse.edc.org/work/research/

Inquiry Project Staff
Daphne D. Minner, Ph.D., Principal Investigator (dminner@edc.org)
Abigail Jurist Levy, Ph.D., co-Principal Investigator
Jeanne Rose Century, Ed.D., co-Principal Investigator (August 2001–July 2005)
Erica S. Jablonski and Erica T. Fields, Research Associates

Copyright © 2009 Education Development Center, Inc.