CENTER FOR ADVANCING RESEARCH & COMMUNICATION

ARC-REESE Criteria & Guidelines for Rating the Methodological Rigor of Educational Research in STEM

Center for Advancing Research and Communication in
Science, Technology, Engineering, and Mathematics (ARC)
http://arc.uchicago.edu

NORC at the UNIVERSITY of CHICAGO

Contents

ARC was asked by NSF to conduct a pilot project to review the research methodologies employed by a sample of projects funded by the REESE (*Research and Evaluation on Education in Science and Engineering*) program. ARC convened an expert panel in consultation with NSF to develop standards and a rubric for rating the rigor of REESE projects' methodologies, with the ultimate goal of reporting on the methodologies employed in the REESE program overall. Panelists concurred that the guidelines provided in the American Educational Research Association's (2006) *Standards for Reporting on Empirical Social Science Research in AERA Publications* provided a sound basis for developing standards for assessing methodological rigor, although recognizing that there is not a direct a relationship between good reporting and good science the panel was concerned to distinguish the two. Two additional panels commented extensively on preliminary guidelines for raters based on their experiences assessing materials from a sample of 24 completed REESE projects. Accordingly while the following guidelines for raters quote extensively from the AERA *Standards* (in some instances with slight word changes to facilitate meaning and understanding), the panelists recommended numerous modifications and elaborations to the AERA *Standards* to provide guidelines for evaluating the rigor of completed work, and examples have been included to illustrate the types of evidence that should lead a rater to conclude that specific rigor standards have been met.

Please direct any questions or comments regarding this pilot project or the standards, criteria, and guidelines in this document to:

Barbara Schneider, Principal Investigator and Rigor Review Project Director, ARC
bschneid@msu.edu

Kevin Brown, Member, ARC Rigor Review Project Team
brown-kevin@norc.org

This material is based upon work supported by the National Science Foundation under Grant No. 0815295. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

I. Background to the REESE rigor rubric

ARC was asked by NSF to conduct a pilot project to review the research methodologies employed by a sample of projects funded by the REESE (*Research and Evaluation on Education in Science and Engineering*) program. The purpose of this pilot project is to report on the methodologies employed in the REESE program overall, *not* make judgments about specific projects or investigators' work. However, to be able to construct an overall assessment it is critical to be able to judge the rigor of specific projects. These guidelines are designed to assess specific projects so that the subsequent ratings can be aggregated to be used to assess the overall rigor of the REESE program.

ARC's goal is to rate the methodological rigor of *completed* projects in contrast to the more speculative assessment of the potential rigor of projects typically undertaken at the proposal review stage. This is most similar to the process IES uses with the What Works Clearinghouse (WWC), which focuses on synthesizing the effects of intervention studies that meet specific evidence standards. The REESE portfolio includes intervention studies; however, there are other types of projects in the program that are not focused on the effects of interventions. Despite the variation of types of programs in the portfolio the Panel agreed that the scale used by the WWC could be applied to the REESE portfolio. Thus projects could be rated with the scale of: (1) "Meets 'Appropriate Rigor' Standards"; (2) "Meets 'Appropriate Rigor' Standards with Reservation"; or (3) "Does Not Meet 'Appropriate Rigor' Standards".[1]

The WWC is a three point scale which we have adopted. We maintained in the course of this pilot that all standards be rated with this scale. During the process, panelists recognized that some criteria may not be relevant for particular projects. In such cases, the panelists maintained that the obligation is on the investigator to explain why a particular criterion was not met. If raters accept the explanation, then it is appropriate to rate the project as "meets" or "meets with reservation," suggesting in the second instance, that raters may still have some questions regarding the justification.

The criteria used by WWC needed to be expanded and modified to meet the diversification of the REESE portfolio. The expert panels convened for this project included 18 scholars whose methodological backgrounds were quite diverse. All panelists agreed that the American Educational Research Association's (2006) *Standards for reporting on empirical social science research in AERA publications* (the AERA standards) provide a workable framework for establishing rigor standards and rating criteria. Specifically, panelists agreed that methodological rigor and the appropriateness of methods should be judged with consideration for projects': (1) **contributions to knowledge**; (2) **designs**; (3) **sources of data**; (4) **measurement and classification**; (5) **analyses and interpretations**; and (6) **generalization**.

[1] As described in the June 2008 summary of the *WWC Review Process* (retrieved May 11, 2011 from http://ies.ed.gov/ncee/wwc/PDF/WhtPapers/wwcreviewprocess.pdf), a study of an education intervention may be rated "Meets Evidence Standards," "Meets Evidence Standards with Reservations," or "Does Not Meet Evidence Screens" (p. 1).

As 'good reporting' is not necessarily indicative of 'good science', the Panels focused their attention on modifying and elaborating the AERA standards to articulate a metric for rating the methodological rigor of completed projects. For each of the six standards, the Panels agreed upon a **rationale** for employing the standard and recommended **criteria** to apply in deciding whether or not projects meet the standards. Panelists also identified considerations reviewers might find it helpful to take into account in applying the criteria. In doing so, panelists were particularly mindful of the distinct purposes and modes of inquiry associated with four major categories of projects identified in the REESE portfolio which include projects:

1. employing experimental designs to make causal inferences regarding the effects of interventions (***experimental***);
2. using quasi experimental designs to make causal inferences regarding the effects of interventions and descriptive studies that identify associative relationships and can potentially identify workable hypotheses for future studies;
3. seeking to "catalyze discovery and innovation at the frontiers of STEM learning, education, and evaluation" by putting forward "groundbreaking ideas, concepts, theories, and methodological approaches" (NSF 10-586, p.4), (***developmental***); and
4. synthesizing existing knowledge, or conducting conferences or workshops (***other***).

The first panel set the initial standards and applied them to a select group of projects. Using the standards and criteria developed by panel A, panels B and C were asked to rate materials obtained from a purposive sample of 24 completed REESE projects. Based on their experiences, the panelists recommended modifications to streamline the standards and elaborate considerations raters should bear in mind in applying the criteria. The following document incorporates their recommendations including suggestions for the composition of the panels.

To account for the diversification of the REESE portfolio, the composition of the rating panels was constructed and vetted with NSF staff. The three panels participating in this pilot project included expertise in science, mathematics, engineering, and technology. In addition to substantive expertise, members of all three panels had deep knowledge of various methodological approaches. While this process was successful, the panelists recommended that if a project was outside the substantive expertise of the panel, ad hoc raters should be included.

Given that this type of activity is relatively new within the STEM educational research community, we are recommending that in the beginning a standing panel of eight to ten members be selected. The construction of such a panel should facilitate a dialogue that brings consensual meaning of the standards, agreement on acceptable explanations for justifiable variations to standards that may arise, and stability to the review process.

PLEASE NOTE: Recognizing that specific types of studies (e.g., intervention studies, meta-analyses) have more defined criteria for evaluating their effectiveness, the Panel abstracted criteria that could apply to projects across the portfolio, with the exception of workshops and conferences. Additional sources of information that may be helpful to raters in applying the criteria which follow are referenced in footnotes to the standards. For more general discussions of factors affecting assessments of the quality of intervention research see: Shadish, Cook, and Campbell (2002); Valentine and Cooper (2008); What Works Clearinghouse (2008). For more information on factors affecting assessments of the quality of research syntheses, see: Cooper (2009, 1982); Cooper, Hedges, and Valentine (2009); Higgins and Green (2011); Suri and Clark (2009).

II. Criteria and guidelines for rating methodological rigor

For each of the six standards upon which the rigor of methods are judged these guidelines provide: a **rationale** for the standard; the **standard**; and a summary checklist of the **criteria** to apply in deciding whether or not projects meet the standard. Additional guidelines raters may find it helpful to consider in assessing rigor are provided.[2] On the basis of these criteria each project will be rated as "meets", "meets with reservation", or "does not meet" the standard.

1. Contribution to knowledge

Rationale. Projects that meet the 'appropriate rigor' standard with respect to contribution to knowledge will describe the problem(s) they seek to address, and the contribution(s) to knowledge they seek to make. This information is critical as it provides the benchmark(s) against which study design, sources of data, measurement, analysis and interpretation, and generalization can be assessed.

Standard.[3] A project "Meets 'Appropriate Rigor' Standards" with respect to its contribution to knowledge if it:

1. **Clearly states the problem(s) investigated**. A clear statement of the problem is not in itself sufficient to establish a project's contribution to knowledge, but without a clear statement, it is not possible to apply the criteria. In determining your rating, consider whether: the research question(s) are appropriate for the study purposes, and the question(s) were worth investigating.

2. **Situates the problem(s) investigated in a broader context, describing the larger body of knowledge to which the project seeks to contribute**. The theoretical framework is described and a literature review is included. The conceptualization of the study reflects or is based upon extant knowledge. Innovative approaches are clearly justified. The investigator has clarified the larger body of knowledge to which the project seeks to contribute.

3. **Explains how the research is contributing to a resolution of the problem(s) and the accumulation of knowledge in this area**

4. **Situates the problem formulation as it relates to the groups being studied (e.g., historical, linguistic, social, and cultural backgrounds)**

[2] Resources consulted in constructing these guidelines include: Boaz and Ashby (2003); Crowe and Sheppard (2011); Gersten, Fuchs, Compton, Coyne, Greenwood, and Innocenti (2005); Mays and Pope (2000); Powell and Davies (2001); Suri and Clarke (2009); and Valentine and Cooper (2008).

[3] This standard builds upon the "problem formulation" standard provided in the AERA reporting standards (American Educational Research Association, 2006: 34).

Summary checklist for rating projects' contribution to knowledge

This REESE research project …	Meets	Meets, with reservations	Does not meet
Clearly states the problem(s) investigated			
Situates the problem(s) investigated in a broader context, describing the larger body of knowledge to which the project seeks to contribute			
Explains how the research is contributing to a resolution of the problem(s) and the accumulation of knowledge in this area.			
Situates the problem formulation as it relates to the groups being studied (e.g., historical, linguistic, social, and cultural backgrounds)			
Overall assessment with respect to the 'contribution to knowledge' standard			

Considerations in applying the criteria. Many projects have multiple objectives. In deciding whether or not a project "meets", "meets with reservation", or "does not meet" the standard with respect to contribution to knowledge, it is critical to establish whether the project has a single objective, or whether it seeks to address multiple problems/research questions. In the case of complex projects with multiple objectives (e.g., projects with multiple embedded studies and/or employing mixed methods), each problem should be separately evaluated. Projects that do not clearly relate individual problems to a statement of intent articulating the larger, 'umbrella' problem which unifies these strands and provides a frame for assessing how distinct elements combine to address a larger 'contribution to knowledge' goal would not meet the appropriate rigor standard with respect to contribution to knowledge. When providing an overall assessment for this standard, it is important to recognize that "contributions" to the field and society (i.e., broader impacts) will vary depending on the project's phase in the research and development cycle (Rand D cycle). For projects further along the R and D cycle, such as an effectiveness trial of an intervention, broader impacts should be apparent and considered when assessing contribution to knowledge.

Exhibit 1: Example of evidence supporting a rating that a project meets the standard criteria with respect to contribution to knowledge

"Algebraic reasoning stands as a formidable gatekeeper for students in their efforts to progress in mathematics and science, and to obtain economic opportunities (Ladson-Billings, 1998; RAND, 2003). Currently, mathematics education research has focused on algebra in order to provide access and opportunities for more students. There is now a growing awareness that the essential concepts that make up school algebra are accessible to students before secondary-level education, and that earlier introduction could facilitate students' algebraic development (Carpenter, Franke, & Levi, 2003; Kaput, Carraher, & Blanton, 2007; National Council of Teachers of Mathematics [NCTM], 2000, National Research Council [NRC], 1998; RAND, 2003). In order to understand middle school students' transition from arithmetic to algebraic reasoning, and to develop and evaluate effective educational approaches to improve the learning and teaching of increasingly complex mathematics, future efforts need to be grounded in sound theory. This theory needs to encapsulate both how students develop algebraic reasoning and acquire domain knowledge, and the beliefs, knowledge, and existing practices of teachers. The theory must also acknowledge the complexity of this area of study, including its multi-tiered nature, diversity of settings and participants, and the high degree of interconnectedness among important components. For example,

to understand students' algebraic reasoning and development, we need to pay attention to classroom interactions, student preconceptions, teachers' beliefs about mathematics and learning, how teachers' beliefs and instructional practices shape the learning environment, and how teachers themselves learn and change.

In an effort to conduct research along these lines, a team of researchers from the University of Colorado, University of Wisconsin, and Carnegie Mellon University developed a framework that guided a recent IERI-funded project,* Supporting the Transition from Arithmetic to Algebraic Reasoning (STAAR). This framework outlines a comprehensive, systemic research and development program to address several inter-related areas, or *tiers*, that we see as central to this effort—student learning and development, teacher beliefs, knowledge and practice, and professional development (cf., Lesh & Kelly, 2000). … Our approach emphasizes the parallel structures and processes among these tiers, viewing them as distinct but inseparable aspects of a unified system. … The authors…conducted research and development within this multi-tiered and dynamic framework in an attempt to move beyond piecemeal, disconnected insights to reach a deeper appreciation of the conceptual terrain and learning processes to inform instruction, curriculum development and professional development. The foundation on which this multi-tiered structure operates gives a sense of the domain of algebra as it is learned in schools.

*The Interagency Education Research Initiative (IERI) is a federal partnership that includes the US Department of Education-Institute of Education Sciences (IES), the National Institute of Child Health and Human Development (NICHD), and the National Science Foundation (NSF)." (Nathan and Koellner, 2007: 179-180).

SOURCE: Nathan, M.J., and Koellner, K. (2007). A framework for understanding and cultivating the transition from arithmetic to algebraic reasoning. *Mathematical Thinking and Learning, 9*(3): 179-192.

2. Design

Rationale. Design considerations are critical in assessing methodological rigor.

Standard.[4] A project "Meets 'Appropriate Rigor' Standards" when the design:

1. **Makes clear its logic of inquiry**. The information required to clarify the logic may vary depending upon the type of project. For example, **for intervention studies**, the intervention should be clearly described, and comparison conditions should be clear and reasonable. A clear account should be given of subjects' exposure to the intervention (e.g., it should be clear whether 'dosage' varies by group or sub-group and/or treatment conditions, and a persuasive rationale should be given for differential exposures/dosage). A clear account should be given of benchmarks against which fidelity of implementation was to be assessed. **For interventions conducted with sampling**, the minimal sample size that would permit a sufficiently precise estimate of the effect size should be specified. **For quasi-experimental studies**, the eligibility criteria for including subjects should be persuasively argued and constructed. Potential sources of bias should be identified and considered.

[4] This standard builds upon the "design and logic" standard provided in the AERA reporting standards (American Educational Research Association, 2006: 34).

Consideration should be given to the potential impacts of interactions, moderators, or other effect modifiers.

2. **Shows how and why the methods and procedures that were used were appropriate for the problem as formulated.** A persuasive rationale is given for the selection of the approach, design, and methods employed. The rationale for selecting any outcome(s), output(s), and/or predictor(s) is persuasively argued. It is clear how implementing this design would yield evidence sufficient to address the study question(s).

3. **When appropriate, clearly describes significant changes or developments in the design**.

Summary checklist for rating projects' designs

This REESE research project …	Meets	Meets, with reservations	Does not meet
Makes clear its logic of inquiry			
Shows how and why the methods and procedures that were used were appropriate for the problem as formulated			
When appropriate, clearly describes significant changes or developments in the design			
Overall assessment with respect to the 'design' standard			

Considerations in applying the criteria. It is important to distinguish between proposed and realized designs. When the two are not aligned, it is imperative that investigators provide the rationale for the differences, enabling raters to assess whether the realized design/design elements were appropriate in this light. In applying the criteria for rating rigor with respect to design, it is critical to retain distinctions between what constitutes sound evidence warranting causal inference and what constitutes sound evidence for yielding 'good science' at each stage of the research and development process. For projects with multiple research questions and/or utilizing a mixture of methods, it is also important to consider separately the rigor of each element of the overall project design.[5] Synthesis projects should also be held to the same standards as other projects regardless of whether they employ a meta analysis or some other strategy for a comprehensive research review. Both approaches have extensive technical literatures that should be reflected in project research designs and products.[6]

[5] For additional information on factors to consider in rating particular types of designs, raters may find it helpful to consider: re **single case designs**, Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, and Shadish (2010); re **experimental, quasi-experimental and observational designs** (including approximating randomized assignment through the use of fixed effects models, instrumental variables, propensity score matching, and regression discontinuity designs) Schneider, Carnoy, Kilpatrick, Schmidt, and Shavelson; re **regression discontinuity designs** specifically Schochet, Cook, Deke, Imbens, Lockwood, Porter, and Smith (2010).

[6] See for example the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, online at http://www.prisma-statement.org/); the American Psychological Association's guidelines for reporting on meta-analyses in the *Journal Article Reporting Standards (JARS): Information Recommended for Inclusion in Manuscripts that Report New Data Collections Regardless of Research Design*, online at http://www.apastyle.org/manual/related/JARS-MARS.pdf); and published protocols available from the Campbell

Exhibit 2: Example A of evidence supporting a rating that a project meets the standard criteria with respect to design

For a project to test "the effectiveness of WestEd's Reading Apprenticeship® teacher professional development with a focus on integrating reading instruction and science instruction to understand its impact on teacher knowledge and skills, instructional practices, and on student achievement in science and reading" (http://www.wested.org/cs/we/view/rstudy/32):

"The principal aim of the study was to test the effectiveness of teacher training in the integration of reading instruction and science content on teacher knowledge and skills, instructional practices, and on student achievement in science and reading. A *true, group-randomized, experimental design* was designed to control for most threats to internal validity (Cook & Campbell 1979, Murray 1998). Schools and the participating teachers within them were randomly assigned to one of two different groups – an experimental group and a wait-listed control group – with a minimum of 25 schools per group…" (p. 13).

"The target population was high school biology teachers and their students in public high schools across California. … [T]he study took place in high schools in California that serve populations of students historically underrepresented in the advanced sciences. The sample consisted of schools with high proportions of these students to better ascertain the impact of integration of literacy instruction with biology course-work for groups of students historically underrepresented in the sciences. Schools, not teachers, served as the unit of randomization to minimize contamination of the control group through teacher interaction. Prior to randomization, participating high schools were pair-matched with similar schools based on the California Department of Education's 2004 School Characteristics Index (SCI) - a composite index representing a school's demographic composition (California Department of Education, 2000).

Schools were randomly assigned to treatment and control groups within each pair of schools. The SCI is based on the following factors: student mobility (percent of students who first attended school in current academic year), ethnicity (percent of students in seven ethnic/race categories), average parental education, percent receiving subsidized meals, percent of teachers fully credentialed, percent of teachers with emergency credentials, percent of English language learners, average class size, and year-round school status. In creating the index, each factor was weighted proportional its relationship to the California's Academic Performance Index, based on a linear regression model. To control for the effect of experience, all teachers were credentialed in biology and had taught for at least 3 years at the initiation of the data collection phase of the study" (pp. 14-15).

"**Data collection**. Several types of data were collected to answer the research questions. These data sources included measures of student achievement and engagement, teacher surveys, analysis of teacher assignments, and observations of classroom practice," (p. 15; NOTE that the data sources are described in detail on pp. 15-19).

"**Retention of Schools and Teachers**. … Overall, 105 biology teachers in 83 schools were recruited, with 56 teachers (43 schools) assigned to the treatment group and 49 teachers (40 schools) assigned to the control group. Note that teachers and schools were recruited and randomized to condition in the spring of 2005, two to three months prior to the scheduled summer professional development institute. Schools and teachers were randomly assigned in batches so that adequate notice could be given to teachers to schedule

Collaboration Library of Systematic Reviews (available online from http://www.campbellcollaboration.org/library.php).

participation in the summer professional development. … 89 percent of treatment teachers and 76 percent of control teachers provided responses on the baseline teacher survey, 79 and 76 percent of treatment and control 20 teachers participated in the 1st-year post-implementation teacher survey, and 59 and 53 percent participated in the 2nd-year post-implementation survey. Return rates for other types of data after the 2nd study year were similar to those for the 2nd-year post-implementation survey. Student longitudinal data and student OTL survey data were secured from approximately 50 percent of randomly assigned teachers, teacher interviews were conducted with 55 percent of teachers, and lesson assignment data were collected from 63 percent of treatment teachers and 45 percent of control teachers. Cross-sectional student test score data were collected from 64 percent of treatment teachers and 51 percent of control teachers. As discussed above, Integrated Learning Assessment (ILA) data were collected as an option from volunteer teachers. Approximately 29 and 20 percent of treatment and control teachers, respectively, participated in ILA data collection. … [A chart documents] similar data return rates as that for teachers" (pp. 19-20).

"**Equivalence of Treatment and Control Groups**. Although data attrition levels were fairly high, attrition patterns were fairly similar for treatment and control schools. Exceptions to this were apparent for the student cross-sectional data, the student OTL surveys, and the lesson assignment data – with higher data return rates exhibited for treatment teachers than for control teachers. To describe treatment/control group equivalence (or lack thereof) at the time of random assignment and at subsequent data collection periods, [the authors] … present school-, teacher-, and student characteristics by data source. … Overall, the randomized and teacher pretest samples show a high degree of similarity, with few meaningful differences in school performance and demographic characteristics. The student OTL sample, which is comprised of about 50 percent of randomized schools/teachers, exhibits more evidence of treatment/control group non-equivalence than the teacher pretest sample, but none of the differences are statistically significant. Treatment schools had about 30 percent more English Learners than control schools (21% vs. 16%), and participating teachers in treatment schools averaged about 1.8 more years of science teaching experience than their control group counterparts (9.3 vs. 7.5 years). … [The authors discuss potential rationale for differences in pre-intervention characteristics of students in treatment and control schools.] No statistically significant differences between treatment and control schools were present, but, as indicated by the longitudinal test score sample, treatment schools had higher proportions of English learners (42% vs. 25%) and Latinos (53% vs. 29%), and lower proportions of white students (16% vs. 33%). Treatment schools also exhibited baseline test scores that were between one-fifth and one-fourth of a standard deviation lower than those in control schools. This provides some evidence that participation of Latino students, English learner students, and students with lower standardized test scores was less likely in control schools than in treatment schools, but these differences could have arisen by chance factors alone." [The authors discuss other potential reasons for differential participation by English language learners.] (p. 22).

SOURCE: Integrating Reading Apprenticeship® and Science Instruction in High School Biology on the WestEd website at http://www.wested.org/cs/we/view/rstudy/32, and retrieved from a hyperlink on the same web page: Greenleaf, C., Hanson, T., Herman, J., Litman, C., Madden, S., Rosen, R., Boscardin, C., Schneider, S., and Silver, D. (2009). Integrating literacy and science instruction in high school biology: impact on teacher practice, student engagement, and student achievement: Final report to the National Science Foundation. Retrieved June 6, 2011 from http://www.wested.org/sli/downloads/nsf-final-report.pdf.

Exhibit 3: Example B of evidence supporting a rating that a project meets the standard criteria with respect to design

For a project to "investigate the mechanisms through which high school opportunity structures and students' figured worlds of STEM are linked on the ground of actual school practice to student choice of STEM major and college destination."

Setting: "Four high schools in the Denver, CO, metropolitan area; four high schools in the Buffalo, NY, metropolitan area."

Research design. "This project has a longitudinal and comparative research design and will generate evidence that is both descriptive [case study, ethnography, observational] and associative/correlational [interpretive commentary]. Original data is being collected from high achieving high school sophomores with interests in science, math, engineering, and technology using school records, assessments of learning, observation [personal observation], and survey research [self-completion questionnaires, structured interviewer-administered questionnaires, and semi-structured interviews].

Instruments or measures being used include student interviews, teacher interviews, principal interviews, counselor interviews, parent interviews; questions from NELS and ELS to construct a descriptive survey that will be administered to all students regarding course selection, use of technology and experiences with counseling for college; and questions from HSLS (2009) to construct a survey of STEM interests and activities that will be used to compare our sample with national samples.

Data will be analyzed qualitatively (via coding and connecting strategies) to (1) construct models of opportunity structures at each school; (2) identify figured worlds at each school; (3) outline student trajectories over time; (4) identify mechanisms by which opportunity structures and figured worlds can be linked to choice of major and college; and (5) to compare opportunity structures, figured worlds, and major and college choices in (a) STEM versus non-STEM focused schools and (b) NY vs. CO."

SOURCE: Structured abstract for the REESE project *Urban High School Opportunity Structures, Figured Worlds of STEM, and Choice of Major and College Destination* (PIs Margaret Eisenhart and Lois Weis) retrieved June 6, 2011 from the ARC website at https://arc.uchicago.edu/reese/projects/urban-high-school-opportunity-structures-figured-worlds-stem-and-choice-major-and-c-0.

3. Sources of data

Rationale. Assessments of methodological rigor must take into account "the data or empirical materials collected or identified to address the research question or problem" (American Educational Research Association, 2006: 35). Because "the role of the researcher and the relationship between the researcher and the participants can influence the data collected," it is important to address this relationship in descriptions of sources of data (American Educational Research Association, 2006: 35).

Standard.[7] A project "Meets 'Appropriate Rigor' Standards" with respect to sources of data if:

1. **The units of study (sites, groups, participants, events, or other units) and the means through which they were selected are adequately described and justified.** It is clear how characteristics of project sites, groups, participants, events, or other units of study bear on the interpretation of outcomes given the phenomena under study.

[7] This standard builds upon the "sources of evidence" standard provided in the AERA reporting standards (American Educational Research Association, 2006: 35).

2. **The collection of data or empirical materials is clearly described, including how and when they were gathered, how often, by whom, for what purposes.**

3. **The description is precise and sufficiently complete to enable another researcher to understand and replicate or reproduce the methods of data collection.**

4. **For studies involving survey research, observations, or interviews, the adequacy of instrumentation** (e.g., with respect to their context validity, cultural appropriateness, and psychometric properties) is indicated.

Summary checklist for rating projects' sources of data

This REESE research project …	Meets	Meets, with reservations	Does not meet
Justifies the selection of all sources of data including the site, group, events, or other units of study			
Clearly describes the collection of data or empirical materials, including how, when, by whom, and how often			
The description is precise and complete so that other researchers could replicate the data collection			
For studies involving survey research and/or observations, the adequacy of instrumentation and the use of appropriate statistical procedures is indicated			
Overall assessment with respect to the 'sources of data' standard			

Considerations in applying the criteria. In this standard it is critical to consider the context of the specific research project or activity before the rater (e.g., in the case of projects with multiple embedded research questions, specific sub-questions). Acceptability does not rest on evidence of literal compliance with every criterion. Instead, the above criteria are intended to offer guidance on the kinds of information essential for assessing sources of data in light of the nature of the project and its research objective(s). For example, for intervention studies, the data collection design should allow for detection of meaningful differences between treatment and control groups taking into account pre- and post-conditions including fidelity of implementation of the intervention. For intervention studies with controlled assignment to treatment groups, results should take into account attrition (explicitly considering the potential impacts of differential attrition), and evaluate the baseline equivalence of treatment and control groups (reporting any statistical adjustments for yielding comparable groups). When projects seek to yield generalizable findings, samples should be representative of the population to which the findings are generalized.

Exhibit 4: Example A of evidence supporting a rating that a project meets the standard criteria with respect to sources of data

For a project to develop a rubric for measuring the quality of mathematics in instruction, to "investigate the mathematical knowledge needed for teaching, and how such knowledge develops as a result of experience and professional learning" (p. 6):

"Sample and Recording of Practice. We recruited ten teachers to participate in our video study based on their commitment to attend professional development workshops and to participate in our study. As such, this is a convenience sample—but one that we hoped would represent a wide range of mathematical knowledge for teaching. Our teachers taught various grades from 2nd to 6th, although the 6th grade teacher was moved to 8th grade in the second year of taping. Seven of the teachers taught in districts serving families from a wide range of social, economic, and cultural backgrounds, including many non-native English speakers. For example, one elementary school within one district enrolled students speaking over 50 different languages. The three other teachers taught in the same school in a small, upper-class, primarily Caucasian district.

Teachers were taped three times in the spring of 2003 prior to a week-long mathematics-intensive professional development, three times in the fall of 2003, and three times again the following spring of 2004. The professional development offered five additional days of follow-up sessions in the fall of 2003. Because these teachers had all registered early for the professional development, they might be considered unusually motivated to improve their mathematics teaching, however their scores on our measures reflect a large range (22nd to 99th percentile) in their mathematical knowledge for teaching at the beginning of the professional development.

The videotaping was done by LessonLab using high-quality professional equipment, including a separate microphone for the teacher, boom microphone for the students, and a custom-designed stand that allowed for fluid movement of the camera around the classroom. Following every lesson, teachers were interviewed about the lesson and these interviews were also videotaped. All videotapes were then transcribed by LessonLab. Following the first wave of videotaping, we realized that having copies of the curriculum the teachers used in preparing the lessons would be an important resource for analysis, so for Waves 2 and 3, curriculum materials were collected from the teachers for each of the lessons. Finally, teachers completed our pencil-and-paper measures at the beginning of the study and, for the most part, after their participation in professional development" (pp. 13-14).

SOURCE: Learning Mathematics for Teaching (2006). *A Coding rubric for Measuring the Quality of Mathematics in Instruction* (Technical Report LMT1.06). Ann Arbor, MI: University of Michigan, School of Education. Available via http://isites.harvard.edu/icb/icb.do?keyword=mqi_training.

Exhibit 5: Example B of evidence supporting a rating that a project meets the standard criteria with respect to sources of data

From a project examining female and minority students' participation in engineering in public universities

"Interview and focus group transcripts were coded in an Atlas-ti database using a codebook developed for the entire study. Atlas-ti is a qualitative software device that organizes transcripts and allows researchers to sort interview transcripts by codes. Codes are words or phrases that correspond to key ideas integral to the research, for example, if an interview were to discuss their experiences as a woman or minority a researcher would assign that segment of text to the code 'experiences of women and minorities.' Atlas-ti can then query all documents for segments of text assigned to this code. …[M]embers of the analysis team coded segments of transcripts corresponding to specific topics and with the central topic of a given chapter. For example, chapter four examines student preparation, development of interest in engineering, and classroom pedagogy. Analysts coded interviews and focus group data to 'background' and 'curriculum/pedagogy' codes. The background code included sub-codes for 'engineering interest' and 'preparation'. The curriculum/pedagogy code included sub-codes 'course work' and 'professors.' Chapter authors utilized Atlas-ti to select text segments included in these sub-codes and then analyzed

them to uncover themes, or discernable patterns in the interviews. For example, if all students reported dissatisfaction with PowerPoint lectures that would constitute a theme for pedagogy. We employ a mixed methods research design in which qualitative analysis occurs first, with subsequent survey analyses occurring at the data interpretation stage (Leech & Onwuegbuzie, 2009). Based on the themes uncovered by the qualitative component, we identified items from the survey that were closely related to facilitate comparison of results from both analyses. We used information obtained from analysis of selected survey items to complement and clarify themes derived from the qualitative analysis." (Chapter 1, pp. 14-15).

"Analyses in this chapter are taken from semi-structured individual interviews conducted by our research team at four engineering programs: University of South Florida, Florida International University, University of Florida, and Florida Agricultural & Mechanical University-Florida State University Colleges of Engineering with a sample of switchers interviewed at USF. Participants were ensured confidentiality and data was transcribed, coded, and entered in the manner described in chapter one. This chapter examines interviews with switchers, administrators, and staff.

To address what seemed to be underreporting of factors that lead students to switch out of engineering among persisters, we sampled all 288 students who declared engineering as their major or pre-major at USF between May 2002 and August 2006 and who were still enrolled at USF in other majors in the spring of 2008. Switchers typically existed engineering during the first or second year of college after completing lower-level prerequisites and before starting upper-level coursework. We used e-mail and recruitment advertisements in the university daily newspaper to solicit potential respondents. We conducted 17 retrospective interviews with current USF students who had switched from USF Engineering into another major. Switcher interviews include 7 women and 10 men. Among the women are 4 Black females, 1 Hispanic female, 1 White female, and 1 Native American female. Among the men are 4 Black males, 1 Hispanic male, 1 Asian male, and 4 White males. Face-to-face interviews were conducted by members of the research team using protocols similar to interviews with current students to allow researchers to contrast responses and determine how these differences may have influenced switching," (Chapter 3, pp. 57-58).

SOURCE: Borman, K.M., Halperin, R.H., and Tyson, W. (2010). Introduction: The scarcity of scientists and engineers, a hidden crisis in the United States. In Kathryn M. Borman, Will Tyson, and Rhoda H. Halperin (Eds.). *Becoming an Engineer in Public Universities: Pathways for Women and Minorities* (pp. 53-80). New York, NY: Palgrave Macmillan. AND Tyson, W., Smith, C.A.S., and Ndong, A.N. (2010). To stay or to switch? Why students leave engineering programs. In Kathryn M. Borman, Will Tyson, and Rhoda H. Halperin (Eds.). *Becoming an Engineer in Public Universities: Pathways for Women and Minorities* (pp. 53-80). New York, NY: Palgrave Macmillan.

4. Measures and classification schemes

Rationale. The Panel concurred that "the validity of empirical studies depends, in part, on the claim that classifications and measurements preserve important characteristics of the phenomena they represent" and noted the importance of distinguishing "key data elements that are crucial to the logic and interpretation of the outcomes" (American Educational Research Association, 2006: 36).

Standard.[8] A project "Meets 'Appropriate Rigor' Standards" with respect to measures and classification schemes when:

1. **The measures used are appropriate to the phenomena under study**

2. **Classification schemes represent the range of the data that the researcher is trying to accommodate**

3. **A rationale for the relevance of measures and classifications used to capture characteristics of groups studied (e.g., gender, race, ethnicity, SES, linguistic) is provided**

Summary checklist for rating projects' measures and classification schemes

In this REESE research project …	Meets	Meets, with reservations	Does not meet
Measures used are appropriate to the phenomena under study			
Classification schemes represent the range of the data that the researcher is trying to accommodate			
A rationale for the relevance of measures and classifications used to capture characteristics of groups studied (e.g., gender, race, ethnicity, SES, linguistic) is provided			
Overall assessment with respect to the 'measures and classification schemes' standard			

Considerations in applying the criteria. Metrics and measures should be assessed on the basis of their psychometric qualities, and that metrics, measures, and classification schemes should fit with the question(s) being addressed. The investigator should provide evidence that any schemas or symbols that represent discourse, actions, or interactions in transcriptions of audio- or video-recordings are clear, and offers a rationale for their use.

Exhibit 6: Example A of evidence supporting a rating that a project meets the standard criteria with respect to measures and classification

For a project to compile "intraclass correlation values of academic achievement and related covariate effects that could be used for planning group-randomized experiments in education" (p. 60):

Some interventions are designed to be compensatory. Experimenters investigating such interventions might choose only schools within a particular context that have low mean achievement or large numbers of low-SES students to evaluate the intervention. We operationalized low achievement by ordering, on average achievement, the entire sample of schools in a setting and selecting the lower 50% of the schools, omitting the upper 50% of the schools. We operationalized low SES by ordering, on the proportion of students eligible for free or reduced-price lunch, the entire sample of schools in a setting and selecting the upper 50% of the schools, omitting the bottom 50% of the schools. One might argue for a more extreme definition of low-SES or low-achievement schools (e.g., the lower 30% of schools). We chose the lower 50% of schools to achieve a balance between the construct definition (low achievement or low SES) and

[8] This standard builds upon the "measurement and classification" standard provided in the AERA reporting standards (American Educational Research Association, 2006: 36).

sufficient sample size to obtain sufficiently precise estimates of the parameters of interest. The choice we made yields some standard errors that are on the order of .02, corresponding to a 2-*SE* band on either side of the estimate (a very crude 95% confidence interval) of width .08. Because even this range is large enough to have important substantive consequences, we judged that restricting the proportion of schools in the definition of the low-SES or low-achievement sample (which would decrease sample sizes of those groups) would lead to unacceptable impreciseness" (p. 63).

SOURCE: Hedges, L.V., and Hedberg, E.C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1): 60-87.

Exhibit 7: Example B of evidence supporting a rating that a project meets the standard criteria with respect to measures and classification

For a project that "evaluates a unique distance learning, inquiry-based professional development course in physical science developed to meet the needs of central Appalachian middle school teachers" (p. 173):

" A survey instrument was developed and administered to teachers before and after they completed the distance learning course to assess their perceptions of the effectiveness of the course in helping them develop their understanding of temperature and heat concepts, their views toward teaching physical science, and their views toward participating in the distance learning temperature and heat course. Additional questions were included to identify what components of the course contributed the most to their learning of the concepts. Course participants rated statements concerning their perceptions of content knowledge and views toward physical science based on a four-point Likert-type scale. Specific answer options for survey items are discussed later…. Frequencies for positive responses were combined and calculated for each survey item to determine the pervasiveness of the attitudes expressed. In addition, chi square analyses were used to determine statistical significance of observed differences in pre and post responses" (p. 179).

SOURCE: Krall, R.M., Straley, J.P., Shafer, S.A., and Osborn, J.L. (2009). Hands-on at a distance: Evaluation of a temperature and heat distance learning course. *J Sci Educ Technol, 18*: 173-186.

5. Analyses and interpretations

Rationale. Central to providing "evidence that the outcomes and conclusions are warranted and that disconfirming evidence, counter-examples, or viable alternative interpretations have been appropriately considered" is the provision of "a clear statement of the process and outcomes of data analysis and a discussion of how they address the research questions or problems" (American Educational Research Association, 2006: 36).

Standard.[9] A project "Meets 'Appropriate Rigor' Standards" with respect to analysis and interpretation if it:

1. **Justifies all analytic procedures, clarifying how they are related to the problem stated and the data collection design**

[9] This standard builds upon the "analyses and interpretations" standard provided in the AERA reporting standards (American Educational Research Association, 2006: 37).

2. **Describes analytic procedures precisely and transparently, specifying assumptions**. Clarity in modeling is critical for purposes of replication of statistical methods and the reproduction of results.

3. **Clarifies how the analyses and presentation of outcomes support claims or conclusions drawn in the research**

4. **Clarifies circumstances (intended or unintended) that impact the interpretation of outcomes (e.g., limiting applicability, compromising validity)**

5. **Connects research findings to the problem(s), question(s), and issue(s) addressed, and the larger body of research to which the study contributes**

6. **Considers appropriately: disconfirming evidence; counter-examples; or viable alternative interpretations**

Summary checklist for rating projects' analyses and interpretations

This REESE research project …	Meets	Meets, with reservations	Does not meet
Justifies all analytic procedures, clarifying how they are related to the problem stated and the data collection design			
Describes analytic procedures precisely and transparently, specifying assumptions			
Clarifies how the analyses and presentation of outcomes support claims or conclusions drawn in the research			
Clarifies circumstances (intended or unintended) that impact the interpretation of outcomes (e.g., limiting applicability, compromising validity)			
Connects research findings to the problem(s), question(s), and issue(s) addressed, and the larger body of research to which the study contributes			
Considers appropriately: disconfirming evidence; counter-examples; or viable alternative interpretations			
Overall assessment with respect to the 'analyses and interpretations' standard			

Considerations in applying the criteria. Because "the processes of analysis tend to follow somewhat different paths in quantitative and qualitative methods" (American Educational Research Association, 2006: 36), specific standards should be applied to each, in addition to the more general standards described above. For example, studies reporting statistical results should include an "index of the quantitative relation between variables (an effect size of some kind such as a treatment effect, a regression coefficient, or an odds ratio)", or, "for studies that principally describe variables, an index of effect that describes the magnitude of the measured variable; an indication of the uncertainty of that index of effect (such as a standard error or a confidence

interval)"[10]; the test statistic(s) and associated significance level(s) employed in any hypothesis testing; a "qualitative interpretation of the index of the effect that describes its meaningfulness in terms of the questions the study was intended to answer." When making causal inferences, investigators should have pursued plausible alternative hypotheses and disconfirming evidence.

Studies analyzing qualitative data should detail the processes used in analysis and interpretation of data generated using qualitative methods sufficient for others to "trace the logic of inquiry", making transparent the processes used to develop "the descriptions, claims, and interpretations". All studies should present evidence sufficient to warrant each claim made, with appropriate consideration for qualifications, conditions, and "significant counter examples" documenting efforts "to search for disconfirming evidence and alternative interpretations of the same evidence". Issues to be addressed in interpretive commentary might include: "how and why the patterns described may have occurred; the social, cultural, or historical contexts in which they occurred; how they relate to one another; how they relate to (support or challenge) theory and findings from previous research; and what alternative claims or counter-claims were considered."

Exhibit 8: Example of evidence supporting a rating that a project meets the standard criteria with respect to analyses and interpretation

From a project to "develop a three-year learning progression focusing on complex thinking about biodiversity" (p. 2):

"Analytical method. We recognize the multi-level nature of our data, e.g., students are nested within classes/teachers. Accordingly, student characteristics are considered at level one and teacher characteristics are considered at level two. Moreover, as we implemented the curricular program through the teachers and subsequently entire classes received an identical treatment dose, we considered the effect of the curricula program to be a level two treatment. We did not pursue a third level of the hierarchy nor fixed school effects since we have, on average, one to two teachers per school and thus can not accurately partition the variance that is uniquely due to school characteristics. Although the number of groups at level two is small (n=22), prior research has concluded that maximum likelihood in multilevel models with a small number of groups still provides unbiased estimates of fixed effects such as our treatment (Browne & Draper, 2000; Van Der Leeden, Busing & Meijer, 1997).

Model. In constructing our hierarchical model of achievement, we focus solely on a random intercept model. Although hierarchical models allow within school (level one) independent variables to vary randomly as well, we constrain these additional random effects to be zero as our primary interest rests on the average effect of the program. We centered all independent level one and level two variables around their respective grand means save the effect of the program and standardized the outcome (mean=0, SD=1). Our level two model exclusively models the average biodiversity achievement adjusted for students' academic and social backgrounds (Appendix A). With our fully unconditional model, we estimate approximately 13%, 12.6% and 12.5% of the variance in overall biodiversity, complex, and standardized achievement, respectively, can be uniquely attributed to the teacher level.

…

[10] The following resources may be interesting for raters interested in considering more closely evidence investigators provide regarding methods for estimating treatment effects, interpreting effect sizes, and/or addressing missing data problems. For information on estimating treatment effects for clustered randomized trials, see Schochet (2009). For information on factors to consider in interpreting effect sizes, see Valentine and Cooper (no date). For a discussion of implications of missing data in group randomized controlled trials see Puma, Olsen, Bell, and Price (2009).

"*Missing Data*. Rather than remove those students or teachers that have incomplete data, we employed the multiple imputation procedure to impute missing values (e.g., Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001). Although no data were missing on our teacher level variables, up to thirty percent of our student sample had at least one missing data point resulting from student mobility and/or absenteeism. Although unverifiable directly, our data suggest that student attrition was unrelated to achievement and program percent completed. Using multiple imputation, we generated five separate, multiply imputed, student level data sets. In an additional effort to increase the robustness of our inferences, we based the imputations on all available variables measured at both the student and teacher level (Peugh & Enders, 2004). Table 3 provides descriptive statistics on the raw and imputed data for student variables presented in subsequent analyses" (pp. 16-18).

SOURCE: Songer, N.B., Kelcey, B., and Gotwals, A. (2009) When and How Does Complex Reasoning Occur? Empirically Driven Development of a Learning Progression Focused on Complex Reasoning about Biodiversity. *Journal of Research in Science Teaching. (46)*6: 610-631.

Exhibit 9: Example of a discussion of analytic approaches that meets the standard criteria as established for analysis and interpretation

From a description of methods for analyzing data derived from video records:

"Video can be rich with interactional phenomena, including eye gaze, body posture, content of talk, tone of voice, facial expressions, and use of physical artifacts, as well as between-person processes such as the alignment and maintenance of joint attention (Barron, 2003). Because this complexity makes it easy to become lost in detail, explicit strategies for focusing the attention of the analysts are needed. Strategies are also needed for establishing the content of the tapes and making decisions about how to represent the phenomena included within them. Erickson (2006) provides three sets of guidelines, each reflecting fundamentally different approaches to inquiry. Briefly, he describes: 1. a whole-to-part inductive approach, in which social viewing and re-reviewing are used to identify patterns in data for which there are no strong orienting hypotheses, predictions or theories; 2. a part- to-whole deductive approach, which involves looking for specific types of events and is appropriate when research is driven by strong questions, hypotheses or theories about those events; and 3. a the manifest content approach, in which interaction focusing on particular pedagogical or subject content is selected out and examined. He provides suggestions about stages of viewing, types of summaries to make at each stage, the importance of time coding, and ways to enhance perception by slowing down or speeding up the tape or watching without sound. These suggestions are very helpful for the beginning or experienced researcher" (p26).

"At the same time one is working with guiding questions, it is important to also remain open to discovering new phenomena. For example, Engle, Conant, and Greeno (2007) were interested in conceptual change, and so they designed a data collection plan that included pre- and post-assessments intended to measure changes in students' conceptual understanding, and they collected video data of classroom conversations that were likely to have generated conceptual growth. But during analysis, some totally unanticipated findings emerged. The researchers addressed the good questions they started with, but ultimately the novel phenomena, they believe, were theoretically more fruitful. Formulating and answering questions does not preclude additional discovery-oriented work with video records. In fact, this is one of the valuable properties of video records – they can be revisited, for continued learning and analysis, at different times with different viewpoints and by different researchers" (p. 25).

SOURCE: Barron, B., and Engle, R.A. (2007). Analyzing data derived from video records. In Sharon J. Derry (Ed.) *Guidelines for Video Research in Education: Recommendations from an Expert Panel*. Chicago, IL: Data

Research and Development Center, NORC at the University of Chicago. Retrieved from http://drdc.uchicago.edu/what/video-research-guidelines.pdf#page=1&view=fitV,0.

6. Generalization

Rationale. It is important to establish what implications single study findings might have beyond the subjects and/or specific context examined.

Standard.[11] A project "Meets 'Appropriate Rigor' Standards" with respect to generalization if the investigator has provided a sound rationale connecting the study to the domain to which generalization is intended (e.g., participants, contexts, measures, classifications, and manipulations).

Considerations in applying the criterion. Although not necessarily recognized as such, generalization may be implied, for example, in some qualitative investigations.[12] More generally, arguments regarding the value of research findings frequently assume (implicitly if not explicitly) some relevance to other samples, conditions, etc. Raters are cautioned, then, to consider the relevance of the generalization standard for all studies, even those which do not explicitly claim generalizable findings.

Summary checklist for rating projects' generalization

This REESE research project …	Meets	Meets, with reservations	Does not meet
Provides a sound rationale connecting the study to the domain to which generalization is intended (e.g., participants, contexts, measures, classifications, and manipulations)			

Exhibit 10: Example A of evidence supporting a rating that a project meets the standard criteria with respect to generalization

For a project to "evaluate the impact of replacement units targeting student learning of advanced middle school mathematics" (p. 1):

"The samples were diverse in terms of campus poverty level, school size, and campus ethnicity. They were also diverse in terms of teachers' gender, ethnicity, years of teaching experience, highest degree obtained, and mathematical knowledge. Comparisons were made to the population in the Texas regions in which the experiments were conducted, as well as to the state of Texas as a whole. For all variables for which we had data at the regional and state levels, the ranges and means were similar among our samples and the middle school mathematics teaching population by region and in the state. Note that the low percentages of African American teachers and students, as well as schools from large urban settings, reflect their small populations in the regions in which the experiments were conducted. Further studies may be needed to examine generalizability of findings to those populations.

Whereas 7 of the 20 geographical regions in Texas participated in the studies, of particular note is the

[11] This standard builds upon the "generalization" standard provided in the AERA reporting standards (American Educational Research Association, 2006: 39).
[12] For a discussion of "Generalization in qualitative inquiry" see Eisenhart (2008).

participation of Region 1 because of its unique demographic and socioeconomic characteristics. Region 1 is in the Rio Grande Valley, adjacent to the Mexican border. It has one of the highest poverty levels in the United States and is predominantly Hispanic. Region 1 participated in the Seventh-Grade Studies; however, because of a shift in local circumstances in the year between recruitment for the two experiments (i.e., the region received a large grant for a major reform in mathematics instruction), the region did not participate in the Eighth-Grade Experiment.

While there was overall attrition in each of the studies, there is evidence that attrition was not differential across experimental groups. In the Seventh-Grade Studies, 140 teachers were accepted into the study, 117 attended the workshop (16% attrition), 95 teachers completed Year 1 (23% attrition), and 67 teachers completed Year 2 (29% attrition). When asked why they dropped from the program, teachers reported reasons that were not related to the project itself (e.g., reassignment or promotion, personal reasons, relocation). In the sample of 95 classrooms that completed the Year 1 experiment, there were no statistically significant differences between groups on any of the student-, teacher-, or school-level variables we examined." (p. 856)

"A possible threat to external validity and generalizability is that our outcome measures were developed within the project to be aligned with the project goals. Methodologically, we chose to develop our own assessment because the Texas state test would not have assessed knowledge of the target M2 content. While there is a danger of overalignment between our intervention and measures, as we described in our assessment development process, we have been explicit in the development of our conceptual assessment framework, vetting the content with experts in the field and collecting several sources of empirical evidence (expert panel review, cognitive thinkalouds, field testing) to support the validity of our assessment argument. While we may not have examined outcomes with other instruments, we have extensive empirical support for the specification of the knowledge, skills, and abilities that were tested.

Finally, we see the consistent replication of our experiments with variations in sample and setting (a wide variety of teachers and schools around the state of Texas), treatments (replacement units), and outcomes (assessments) as sufficient evidence to reject the idea that the findings result from experimental artifacts." (p. 869)

SOURCE: Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J., and Gallagher, L.P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal, 47*(4): 833-878.

Exhibit 11: Example B of evidence supporting a rating that a project meets the standard criteria with respect to generalization

Eisenhart (2009: 62-63) notes that "[t]here are many examples of emergent theoretical generalizations in existing qualitative studies in education research, and some of them appear to be strikingly enduring. Janet Schofield's (1989) study of a desegregating school in the United States provides a case in point. About this study, Schofield writes:

> After I observed extensively in varied areas of the school and interviews a large number of students, it became apparent that the white children perceived blacks as something of a threat to their physical selves. Specifically, they complained about what they perceived as black roughness or aggressiveness… In contrast, the black students perceived whites as a threat to their social selves. They complained about being ignored, avoided, and being treated as inferior by whites, whom they perceived to be stuck-up and prejudiced…. Such findings appear to me to be linked to the black and white students' situation in the larger society and to powerful historical and economic

> forces, not to special aspects of the school [she studied]. The consequences of these rather asymmetrical concerns may well play themselves out differently in different kinds of schools, but the existence of these rather different but deeply held concerns may well be widespread. (p. 221)."

SOURCE: Eisenhart, M. (2009). Generalization from qualitative inquiry. In Kadriye Ercikan and Wolff-Michael Roth (Eds.), *Generalizing from Educational Research: Beyond Qualitative and Quantitative Polarization*. New York, NY: Routledge.

III. Weighing the criteria to arrive at an overall rating

Some flexibility is necessary in applying these standards and criteria to assess the rigor of the wide variety of projects that comprise the REESE portfolio. For example, when considering complex projects with multiple goals and/or methods, it is important that raters weigh elements appropriately to assess constituent parts individually and to arrive at appropriate assessments of such projects as a whole. Nevertheless, all six standards are relevant in assessing the rigor of the major categories of projects found in the portfolio. Raters are cautioned to take all six standards into account in arriving at an overall assessment of the rigor of each project before them, bearing in mind that the ultimate objective of this pilot project is to report on the methodologies employed in the REESE program overall, *not* to advance judgments about specific projects or investigators' work.

Summary checklist for rating projects' methodological rigor

This REESE research project …	Meets	Meets, with reservations	Does not meet
Overall assessment of methodological rigor			

IV. Confidentiality of materials

Raters are reminded that all materials provided for their consideration are confidential, and should be destroyed (hardcopies and electronic files) at the end of the rating process.

V. Value of promoting standards for methodological rigor

All of the panels congratulated NSF for undertaking such an activity that, while difficult, resulted in the standards articulated above. Panelists concurred that it was important to articulate and apply standards that encompassed more than intervention studies. In developing these standards and reviewing their content several panelists suggested that the content had important professional development implications. Specifically they recommended that such standards could be part of solicitations, thus giving investigators a tighter framework for the conduct of their work. The standards could also be used for providing guidelines for annual and final reporting, which would also facilitate comparisons and impact of research findings. Finally, with respect to enhancing capacity, several panelists suggested that such guidelines be used in methodological design courses and offered as professional workshops at science meetings.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6): 33-40.

Boaz, A. and Ashby, D. (2003). "Fit for purpose? Assessing research quality for evidence based policy and practice." *ESRC UK Centre for Evidence Based Policy and Practice*.

Cooper, H.M. (2009). *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*. Thousand Oaks, California: Sage.

Cooper, H.M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research, 52*: 291-302.

Cooper, H.M., Hedges, L.V., and Valentine J. (Eds.). (2009). *The Handbook of Research Synthesis and Meta-Analysis (2nd edition)*. New York: The Russell Sage Foundation.

Crowe, M. and Sheppard, L. (2011). "A review of critical appraisal tools show they lack rigor: Alternative tool structure is proposed." *Journal of Clinical Epidemiology* **64**(1): 79-89.

Eisenhart, M. (2008). Generalization from qualitative inquiry. In K. Ercikan and W-M. Roth (Eds.) *Generalizing from Educational Research: Beyond the Qualitative and Quantitative Polarization*. (pp. 51-66). New York, NY: Routledge Press.

Gersten, R., Fuchs, L.S., et al. (2005). "Quality indicators for group experimental and quasi-experimental research in special education." *Exceptional Children* 71(2): 149-165.

Giaconia, Rose M. and Hedges, L.V. (1982). "Identifying features of effective open education." *Review of Educational Research* 52(4): 579-602.

Greenleaf, C., Hanson, T., Herman, J., Litman, C., Madden, S., Rosen, R., Boscardin, C., Schneider, S., and Silver, D. (2009). Integrating literacy and science instruction in high school biology: impact on teacher practice, student engagement, and student achievement: Final report to the National Science Foundation. Retrieved June 6, 2011 from http://www.wested.org/sli/downloads/nsf-final-report.pdf.

Harden, A., Weston, R., and Oakley, A. (1999). "A review of the effectiveness and appropriateness of peer-delivered health promotion interventions for young people." *EPI Centre London.*

Higgins, J.P.T., and Green, S. (Eds.). (2011). *Cochrane Handbook for Systematic Reviews of Interventions* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org.

Krall, R.M., Straley, J.P., Shafer, S.A., and Osborn, J.L. (2009). Hands-on at a distance: Evaluation of a temperature and heat distance learning course. *J Sci Educ Technol, 18*: 173-186.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Learning Mathematics for Teaching (2006). *A Coding rubric for Measuring the Quality of Mathematics in Instruction* (Technical Report LMT1.06). Ann Arbor, MI: University of Michigan, School of Education. Available via http://isites.harvard.edu/icb/icb.do?keyword=mqi_training.

Mays, N. and Pope, C. (2000). "Assessing Quality in Qualitative Research." *BMJ* 320(7226): 50.

Nathan, M.J., and Koellner, K. (2007). A framework for understanding and cultivating the transition from arithmetic to algebraic reasoning. *Mathematical Thinking and Learning, 9*(3): 179-192.

Powell, A. and Davies, H. (2001). "Reading and assessing qualitative research." *Hospital Medicine* 62(6): 360.

Puma, M.J., Olsen, R.B., Bell, S.H., and Price, C. (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pdf/20090049.pdf.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., Knudsen, J., and Gallagher, L.P. (2010). Integration of technology, curriculum, and professional development for advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal, 47*(4): 833-878.

Schneider, B., M. Carnoy, J. Kilpatrick, W.H. Schmidt, and R.J. Shavelson. (2007). Estimating Causal Effects Using Experimental and Observational Designs. AERA (http://www.aera.net/uploadedFiles/Publications/Books/Estimating_Causal_Effects/Causal_Effects.pdf)

Schochet, P. (2009). *Technical Methods Report: The Estimation of Average Treatment Effects for Clustered RCTs of Education Interventions* (NCEE 2009-0061 rev.). Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pdf/20090061.pdf.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J.R., Porter, J., Smith, J. (2010). Standards for Regression Discontinuity Designs. Retrieved from What Works Clearinghouse website at http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf

Shadish, W.R., Cook, T.D., and Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.

Songer, N.B., Kelcey, B., and Gotwals, A. (2009) When and How Does Complex Reasoning Occur? Empirically Driven Development of a Learning Progression Focused on Complex Reasoning about Biodiversity. *Journal of Research in Science Teaching. (46)*6: 610-631.

Suri, H. and Clarke, D. (2009). "Advancements in Research Synthesis Methods: From a Methodologically Inclusive Perspective." *Review of Educational Research* **79**(1): 395-430.

Valentine, J. C. and Cooper, H.M. (2008). "A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The study design and implementation assessment device (Study DIAD)." *Psychological Methods* **13**(2): 130-149.

Valentine, J., and Cooper, H. (no date). Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes. Available online via http://ies.ed.gov/ncee/wwc/references/iDocViewer/Doc.aspx?docId=1; PDF version available at http://ies.ed.gov/ncee/wwc/pdf/essig.pdf.

What Works Clearinghouse. (2008). *What Works Clearinghouse Procedures and Standards Handbook*, Version 2.0. Washington, D.C.: U.S. Department of Education, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. Retrieved June 2, 2011 from http://ies.ed.gov/ncee/wwc/pdf/wwc_procedures_v2_standards_handbook.pdf.