Technical Report:
Standards of Evidence for Empirical Research, Math and Science Partnership Knowledge Management and Dissemination

by

Daniel J. Heck
Horizon Research, Inc.

Daphne D. Minner
Education Development Center, Inc.

October 2010

Prepared By: Horizon Research, Inc.
326 Cloister Court
Chapel Hill, NC 27514-2296

In order to set out coherent research agendas that aim at cumulating knowledge in particular areas of focus, the field needs to take stock of the current knowledge base at a given point in time, including attention not just to what we know, but also to how well we know it. Education research has been under considerable and mounting scrutiny for the past two decades, leading to a number of calls for improved quality generally, and to proposed standards of evidence for education research and reporting to improve quality (American Educational Research Association, 2006; Burns, 1989; Coalition for Evidence-Based Policy, 2003; Cooper & Hedges, 1994; Eisenhart & Towne, 2003; Feldman, 2003; Howe & Eisenhart, 1990; The Inquiry Synthesis Project, 2006; Lipsey & Wilson, 2001; National Research Council, 2002, 2005; Spencer, Richie, Lewis, & Dillon, 2003; What Works Clearinghouse, 2008). High profile efforts to define quality education research, and to judge the empirical evidence resulting from studies, include the U.S. Department of Education's What Works Clearinghouse (2008), and Great Britain's Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre, 2007). Education scholars have variously responded in support of, and with concerns about, these efforts (e.g., Freeman, de Marrais, Preissle, Roulston, & St. Pierre, 2007; Kelly & Yin, 2007; Slavin, 2008), furthering attention to this issue and providing broader thinking and greater depth of understanding about what constitutes high-quality education research. Somewhat different perspectives persist regarding what knowledge is of most worth. However, there seems to be agreement that taking stock of what we know and how well we know it must form the foundation for cumulating knowledge from education research. Doing so requires standards of evidence that attend to various contributions that research *can* make to the knowledge base, take account of what different methodologies have to offer, and that account for the developmental nature of knowledge growth.

Drawing on the literature cited above and the input and review of an advisory panel of research methodologists, researchers, and reform leaders[1], MSP KMD developed a set of standards of evidence for empirical research. (See Appendix A.) An initial version of the standards of evidence was also reviewed by evaluators of the MSP KMD project from the Consortium for Policy Research in Education, whose feedback led to additional clarifications and explanations regarding how to apply the standards to claims resulting from quantitative, qualitative, and mixed methods approaches.

The resulting standards of evidence, and the process for applying them, result in a careful review of the claims of individual studies to provide ratings based on specific indicators, operationalized for different research methodologies, and narrative justifications for these ratings in six areas:

1. Adequate documentation;
2. Internal validity;
3. Analytic precision;

[1] The Math and Science Partnerships Knowledge Management and Dissemination Project held a meeting of research methodologists, researchers, and reform leaders on March 22–23, 2005 in Washington, DC to provide input on the standards of evidence. Attendees at the meeting later provided feedback on a draft of the standards of evidence. The panel included: Dennis Bartels, Diane Briars, Audrey Champagne, Tom Corcoran, Mel George, Manuel Gomez, Doug Grouws, Frances Lawrenz, Joe Maxwell, Steve Meiring, Andy Porter, Senta Raizen, Karen Seashore Louis, Sharon Senk, Nancy Shapiro, and Carol Weiss.

4. Generalizability/external validity determination;
5. Overall fit; and
6. Warrants for claims.

The ratings identify the strength of contributions of individual studies both efficiently and objectively. Narrative paragraphs for each area provide description, judgment, and justification regarding how the study treated aspects of empirical research relevant to the purpose of the study. (See Appendix B.) These ratings and narratives together portray the nature and strength of the study's contributions to the knowledge base.

To improve knowledge cumulation, the purpose of applying standards of evidence to claims made in research studies is to identify the specific contributions of each study to the field's knowledge base. Contributions are characterized in terms of what is known from the findings based on the intent and substance of the study, and the confidence that can be placed in the findings based on the nature and quality of the study. Applying the standards of evidence is not intended to make "good/bad" or "in/out" judgments on studies, or to suggest that all studies should be strong on every standard. Rather the application of standards of evidence is conducted to aid understanding of the strengths and limitations of each study's contributions to the knowledge base. Applying the standards of evidence to existing research provides a basis for identifying well-supported claims in empirical studies, and to ascertain the specific limitations of claims based on the empirical evidence within particular studies.

References

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

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

Coalition for Evidence-Based Policy. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, and National Center for Education Evaluation and Regional Assistance.

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

Eisenhart, M. & Towne, L. (2003). Contestation and change in national policy on "scientifically based" education research. *Educational Researcher, 32*(7), 31–38.

EPPI-Centre. (2007, March), *EPPI-Centre methods for conducting systematic reviews*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Feldman, A. (2003). Validity and quality in self-study. *Educational Researcher, 32*(3), 26–28.

Freeman, M., de Marrais, K., Preissle, J., Roulston, K., & St. Pierre, E. A. (2007). Standards of evidence in qualitative research: An incitement to discourse. *Educational Researcher, 36*(1), 25–32.

Howe, K. & Eisenhart, M. (1990). Standards for qualitative (and quantitative) research: A prolegomenon. *Educational Researcher, 19*(4), 2–9

The Inquiry Synthesis Project, Center for Science Education, Education Development Center, Inc. (EDC). (2006, April). *Technical report 6: Operationalizing the coding of research rigor, context, and study findings*. Retrieved April 30, 2009, from http://cse.edc.org/work/research/inquirysynth/technicalreport6.pdf.

Kelly, A. E. & Yin, R. K. (2007). Strengthening structured abstracts for education research: The need for claim-based structured abstracts. *Educational Researcher, 36*(3), 133–138.

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

National Research Council. (2002). *Scientific research in education.* Committee on Scientific Principles for Education Research. Shavelson, R.J., & Towne, L., (Eds.) Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2005). *Advancing scientific research in education.* Committee on Research in Education. L. Towne, L. L. Wise, & T. M. Winters, (Eds.) Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5–14.

Spencer, L., Ritchie, J., Lewis, J., & Dillon, L. (2003). *Assessing quality in qualitative evaluation.* London: National Centre for Social Research.

U. S. Department of Education (2008) What works clearinghouse

What Works Clearinghouse. (2008, December). *Procedures and standards handbook (Version 2.0)*. Washington, DC: US Department of Education, Institute for Education Sciences.

APPENDIX A

Codebook for Standards of Evidence for Empirical Research

June 2009

GENERAL GUIDELINES FOR USE OF STANDARDS Purpose—These standards were developed in order to guide the review for methodological integrity and evidentiary utility of existing published research studies. This review will provide information about the state of the knowledge base for a given area of investigation so that rigorous reviews of the knowledge can be generated. The intent for this codebook is to provide a means of systematic review for key indicators which determine a study's level of methodological integrity.

Standards of evidence will be applied to:

1. Studies being conducted within projects funded under the NSF MSP program, including those that report findings directly related to one of the KMD "drill down" topics; and,
2. Studies found in the broader literature that report findings directly related to one of the KMD "drill down" topics.

The KMD project's charge and objective relevant to the standards of evidence work, within focused "drill down" topics, is to summarize what is known from empirical research and to identify the contributions to the knowledge base for mathematics and science educational improvement of what is being learned in the NSF MSP program. The KMD project will summarize and apply standards of evidence to all studies of the first type as they reach the stage of reporting findings. Studies of the second type will be included in the KMD work as a way to situate what is being learned in the MSP program within the field's knowledge base.

The purpose of applying standards of evidence to studies is to identify the contribution(s) of each study to the field's knowledge base. A study's contributions will be characterized in terms of what is known from the findings based on the substance of the study, and the confidence that can be placed in the findings/conclusions/implications based on the nature and quality of the study. Applying the standards of evidence is not intended to cast blanket "good/bad" or "in/out" judgments on studies, nor is it to suggest that all studies should be strong on every standard. Rather the standards of evidence are meant to aid our understanding of the strengths and limitations of each study's contributions to the knowledge base.

Process for using Standards in service of Knowledge Synthesis—A study of either type will first be summarized by KMD staff using the "Article Summary" protocol to record essential parameters of the study (e.g., abstract, findings). When a drill down topic is identified, the article summaries from studies of type 1 and 2 will be examined by the lead KMD staff member who will author the knowledge synthesis for the topic. The lead KMD staff member will confirm the set of studies with findings that are directly related to the drill down topic and pass these studies on to the standards of evidence reviewer(s).

The standards of evidence reviewer(s) will apply standards of evidence to each of the studies, including all findings of the studies that are potentially to be included in any drill down topic. An expedited standards of evidence review (i.e., Overall Fit and Warrants for claims narratives only) will be completed if the reviewer finds any of the identified "substantial limitations" indicated throughout the form. In addition to completing the standards of evidence coding form, the standards of evidence reviewer(s) should revise the article summaries as needed, including inserting any additional findings/conclusions/implications. After the standards of evidence review is complete, the revised summaries and standards of evidence reviews will be returned to the lead KMD staff member for the research review. After the lead staff member has read the results of the standards of evidence review, the lead staff person should consult with the standards of evidence reviewers, as necessary, to discuss the results of the standards of evidence reviews. The purposes of these conversations are to ensure that the lead staff member understands the results of the standards of evidence reviews, and to come to agreement on the contributions of the MSP studies to the existing knowledge base. The contributions could include, for example: (1) new findings, (2) greater confidence in

existing findings through replication or studies that address limitations of the extant literature, or (3) broadened generalizability of findings by their extension to new contexts or populations. Information from the application of standards of evidence will then be combined with the study summary to describe the findings in terms of both their substance and the strength of the evidentiary basis for them.

Language nomenclature used in this codebook:

Conclusion=what the researcher makes of findings, often using a theoretical or conceptual framework for interpreting the findings.

Domain= the bold and lettered items in the codebook.

Domain Narrative=this narrative describes the greatest strengths and weaknesses for a study within a given domain, and provides the rationale/justification for the scale rating that was given. (See also "Overall Fit Narrative" below.)

Expedited review=a truncated review for a study that is identified as having a substantial limitation. Only the Overall Fit and Warrants for Claims ratings and narratives are required for such studies.

Finding=the result of analysis directly in relation to a research question.

Implication=a suggestion or recommendation for what policymakers, practitioners, or other researchers should consider or do as a consequence of the findings/conclusions of the study.

Overall fit narrative=the narrative that summarizes the "take away" messages from the study, including from the other domains any particular strengths in the study, all flaws that have an impact on confidence in or interpretation of the findings, and implications of the scale ratings for the contributions of the study. It also describes whether there are important sources of doubt in the reviewers' judgment regarding any particular finding(s).

Scales=the three point global ratings—

Poor: the study has serious methodological flaws, which undermine the potential contributions of the study

Limited: there are minor flaws methodologically or the reporting is very scant, so the potential contributions of the study should be "interpreted with caution" or "taken with some reservation"

Adequate: this study does a sufficient job methodologically to support its claims; consequently, the study likely makes a sound contribution to the knowledge base

Substantial limitation=If a study includes non-response bias, unfair or unjustified comparisons, selection bias, unfounded instruments, or inappropriate analysis, then it is judged to have a substantial limitation.

Indicators=numbered items in the codebook.

Directions for Coding a Study with the Standards—A reviewer should have all of the relevant articles, documents, and reports for a given study prior to coding and a Standards of Evidence Coding Form

(appended to this Codebook). The indicators listed under each domain should be thought of as a checklist of the things you should consider in terms of possible strengths and weaknesses for this domain. The domain ratings should be made based on consideration of these indicators. There is further description of each of these indicators contained in this codebook. You should also consider the importance and number of items for which there is insufficient information reported in the study.

In the narrative paragraphs for Domains A-D you should include evidence to justify the ratings that you give. These narratives serve as summaries of the key elements/issues within these domains which are then used to assess Overall Fit in Domain E. The *Overall Fit* rating should be based on the overall contribution of the study, taking into account the design and implementation of the study as detailed in Domains A-D. If the rating for Overall Fit is "Poor" then there is no need to proceed further with coding of Domain F, Warrants for Claims.

During the review, each of the substantial limitations should be considered. Indicators to consider for each substantial limitation are marked with an asterisk. A study should be given an expedited Standards of Evidence review if one or more of the substantial limitations are identified. An expedited review requires completing only the rating and narrative for *Overall Fit*. The *Overall Fit* rating should be based on the overall contribution of the study, taking into account any substantial limitation(s) and the design and implementation of the study otherwise. Expedited reviews must still account for the full design and implementation of the study. For all expedited reviews, the *Overall Fit* narrative should include identification of all of the substantial limitation(s), as well as other design and implementation strengths and weaknesses, and an explanation of their impact on the evidentiary basis for the findings reported in the study.

A study with one or more substantial limitations may receive a rating of either "Limited" or "Poor" for *Overall Fit*. A rating of "Limited" should be given for *Overall Fit* if the design and implementation of the study, even with the substantial limitation(s), still result in findings supported by the empirical evidence, particularly when the findings are described with the appropriate limitations. If the *Overall Fit* rating is determined to be "Limited" the expedited review should also include ratings and narratives for *Warrants for Claims*.

Definitions of Substantial Limitations

If any of the following substantial limitations are identified in the study, an expedited review will be sufficient for the study. The expedited review proceeds directly to the Overall Fit rating and narrative. The Overall Fit narrative should identify and describe the evidence for all substantial limitations. It should also describe the implications that the substantial limitations, as well as other strengths/weaknesses of the study have for interpreting the findings. The expedited review should proceed beyond Overall Fit ratings and narrative to Warrant for Claims ratings and narrative. Studies with expedited reviews will still be included in the research review.

1. Unfounded instruments:

a. Data collection procedures are not described in enough detail to make determinations about use of instruments (Indicator A5). In qualitative studies, this limitation may arise because the study lacks description of the data collection activities, particularly if the researcher appears to lack qualifications for collecting data (Indicator B14).
b. Quantitative instruments used in the study have unknown psychometric properties; that is, no validity or reliability information is reported, particularly for instruments that are created specifically for the study (Indicators C15, C16).
c. Qualitative coding, classification, or interpretation is conducted without attention to a systematic, reliable process (Indicators C15, C16).

2. Bias

a. The sample studied is different from the population to which inferences are made in a way that is likely to influence the results of the study substantially (e.g., key demographic differences such as experience level, key contextual differences such as school cultures, volunteers only) (Indicator B6).
b. Bias may not be a substantial limitation if the researchers address its potential effects through analysis (Indicator B11) or discussion (Indicator B12).

3. Unfair comparisons

a. Comparisons are inappropriate for the study, such as when treatment and comparison groups (or other groups that are compared) are not similar on key factors, which may include demographics, contexts, or prior status on outcomes (or this information is unknown, but likely to differ across groups) (Indicator B6).
b. Potentially unfair comparisons may not be a substantial limitation if the researchers address concerns about their possible effects through analysis (Indicator B11) or discussion (Indicator B12).

4. Low response rate

a. A response rate of under 50% will be considered a substantial limitation, whether it occurs due to initial non-response (Indicator B7), later non-response due to attrition from the study (Indicator B8), or dropping of cases in the sample due to missing data (Indicator B9).
b. A low response rate may not be a substantial limitation if the researchers address its potential effects through analysis (Indicator B11) or discussion (Indicator B12).

5. Inappropriate analysis

a. A quantitative study that makes multiple univariate comparisons of the same samples and makes claims about statistically significant findings without accounting for inflated error rates will be considered substantially limited (Indicator C17).
b. A quantitative study that uses an inappropriate unit of analysis for its research questions (Indicator C18) will be considered substantially limited.
c. A qualitative study in which the researcher is likely biased toward certain findings, but has no checks on this bias in place (Indicator B13), will be considered substantially limited.

A. Adequate Documentation of Project Activities: The extent to which there is sufficient and clear description of key elements of the study.

Indicator	Questions and guidance to consider
1. Research question and constructs	**Were the research question(s)/ issues being studied /hypotheses clearly stated?** In more exploratory studies the questions or issues may be less defined, but the reader should still be able to determine the purpose of the study without prior knowledge of the study, or substantial background knowledge of the topic. **Were the factors or constructs being studied clearly described?** Was there sufficient description of how the factors and constructs were operationalized through indicators or illustrative examples? This standard includes both independent and dependent variables, as appropriate to the study. (Some very commonly used or widely understood constructs, such as gender or grade level, may be self evident.)
2. Research site	**Were the research sites sufficiently described?** Were the location(s) where the research took place sufficiently described given the nature of the study? For qualitative studies this typically requires more description than for quantitative studies, because situated description or "thick description" is often critical for qualitative research. Sufficient description includes stating things like the demographics of the community or school setting such as size, socioeconomic status, geographic location, financial resources, etc.

3. Sample demographics	**Were the research samples sufficiently described?** Was relevant information about the sample provided on characteristics likely to relate to the research questions and contexts (such as the ages of the students, years of experience of teachers, SES of the participants, gender, etc.)?
4. Interventions and Implementation	**Were the educational interventions, if applicable, being studied clearly described?** Was there sufficient description of the components, both theoretical and practical, of the intervention? Were the interventions for the treatment groups of interest as well as any comparison groups described? Potentially important information about the implementation of the intervention includes who delivered the intervention, the dosage and duration of the intervention treatment and comparison treatments, and fidelity to original intervention design.
5. Data collection	**Were the data collection strategies articulated?** Was there sufficient information to determine what data collection methods were employed, by whom, and when–relative to the intervention? Were the potential biases of the research team and effects on data collection addressed?

Narrative paragraph for project documentation

This narrative should describe the extent to which there is a sufficient and clear description of the important indicators listed above, as relevant to the nature and purpose of the study. In other words, keep in mind whether the research question asked is exploratory, explanatory, etc. and whether the study is conducted in a quantitative, qualitative or mixed method manner.

The review must discuss whether sufficient information was or was not provided regarding:
-methods of data collection
-sample demographics
-constructs being operationalized in the study
-the intervention and its implementation (if an intervention was being studied)

Poor: there are serious flaws with the project documentation which undermine the potential contributions of the study because:

- there is *little to no information* provided about the methods of data collection; or
- there is *little to no information* provided about the sample demographics; or
- there is *little to no information* provided about the constructs being operationalized in the study; or
- there is *little to no information* provided about the intervention and its implementation.

Limited: there are minor concerns with the project documentation so the potential contributions of the study should be "interpreted with caution" or "taken with some reservation" because:

- there is *some information* provided about the methods of data collection but they are not sufficiently described; or
- there is *some information* provided about the sample's demographics but it is not

sufficiently described; or

- there is *some information* provided about the constructs being operationalized in the study but they are not sufficiently described; or
- there is *some information* provided about the intervention and its implementation but it is not sufficiently described.

Adequate: the project documentation of this study is adequate thus it can make a contribution to the knowledge base because:

- there is sufficient information provided about the methods of data collection; and
- there is sufficient information provided about the sample's demographics; and
- there is sufficient information provided about the constructs being operationalized in the study; and
- there is sufficient information provided about the intervention and its implementation.

B. Internal Validity: "The extent to which the results of a study can be attributed to the treatments rather than to flaws in the research design. In other words, internal validity is the degree to which one can draw valid conclusions about the causal effects of one variable on another. It depends on the extent to which extraneous variables have been controlled by the researcher." (Vogt, 2005). In qualitative research, internal validity refers to the strength of the arguments that link theory, methodology, and interpretation, and the credibility of the supporting evidence. Claims of attribution depend heavily on a study's internal validity.

Indicator	Questions and guidance to consider
6. Sample Bias * Bias	**Was the sampling strategy explained?** Was there sufficient information on how potential participants (schools, students, or teachers) were informed about the study? How they were identified? How they were selected for participation (e.g. random, snowball, volunteers, convenience)? Does the study explain why that particular sampling strategy was used? If this is a comparison study, how were participants assigned to treatment and control groups? Does the study appropriately acknowledge limitations of the selected strategy (e.g. self-selection issues) as they relate to the research questions? For qualitative studies, is there clear logic behind the selection of sites/participants/cases to inform the research question? **Was the sample biased?** Things to consider include how the sample was recruited (e.g. only volunteers, or waitlist, etc.) and whether or not this type of recruitment was likely to create bias. What constitutes bias depends on the nature and purpose of the study. For example, is the study intended to provide a rich description of an intervention, to demonstrate an existence proof that some effect/relationship is possible, or to make a claim that some effect/relationship is likely/common? Sample bias can result from small sample sizes and ceiling effects (treatment group scores high of test of dependent variable prior to any intervention) for certain kinds of quantitative designs.
7. Response bias *Low response rate	**Was there indication of bias in the data due to differences between responders and non-responders on the data collection instruments, or was there an overall low response rate (e.g., survey return rate)?** An example of response bias is: teachers who responded to a survey or agreed to be observed were only those that had the highest content knowledge or years of teaching experience. This kind of bias could be determined only if the researcher provided some basic demographic information on the entire target sample and then explored for systematic bias on these variables between responders and non-responders.

8. Attrition Bias *Low response rate	**Did the overall and differential attrition among participants bias the results?** Attrition refers to the loss of subjects from the sample from the beginning of the study to the end of the study. Was there evidence indicating that either there wasn't differential attrition (even though there could have been in the research design); or that attrition was explored and determined to not be a concern for bias?
9. Missing Data Bias *Low response rate	**Did missing data of the remaining participants bias the results?** Missing data may be an issue if subjects (or researchers) do not complete all of a data collection instrument or subjects do not participate in parts of the data collection used in the study. Was there evidence indicating either that there were no missing data; or, in quantitative studies, that missing data were imputed using standard procedures; or that missingness was explored and determined to not be a concern for bias?
10. Contamination	**Were concerns about contamination between treatment and comparison conditions addressed?** (applicable only in comparative designs) For example, when teachers from the same school are assigned to treatment and comparison conditions, do the researchers provide information about what steps were taken to avoid contamination or why contamination is not a concern?
11. Validity Threats/ Alternative explanations addressed through analysis	**Were viable alternative explanations (threats to validity or credibility) addressed in the analysis strategy?** For example, in quantitative analyses alternative explanations can be explored via covariate analysis. In qualitative studies, alternative explanations can be addressed by follow-up data collection, including "member checking" or "respondent validation" of findings.
12. Validity Threats/ Alternative explanations discussed	**Were viable alternative explanations (threats to validity) addressed in the discussion?** Did the author *discuss* the possible effect of history, maturation, testing, instrumentation, regression artifacts, experimental mortality, or others, as alternative explanations to the claims they made about the findings (e.g., about the effects of treatment, or the relationships among variables)? In qualitative studies, did the researchers acknowledge both limitations that could threaten the quality of the data, and possible alternative interpretations of the data .
13. Investigator Bias/Reflexivity *Inappropriate analysis	**Did the relationship between the researcher(s) and the treatments bias the results?** For example, did the researcher deliver the treatment or have a pre-existing relationship with the research participants? If so, were protections against possible biases employed? In qualitative studies, does the investigator have awareness "of how … interactions in a field site threaten, disrupt, create, or sustain patterns of social interaction [that] might result in a prejudicial account of social behavior in the site? Does the investigator protect against "individual preferences, predispositions, or predilections that prevent neutrality and objectivity"? (Schwandt, 2001). Some bias in terms of predispositions is inevitable, but it is important that the researcher engage in reflexivity, "the process of critical self-reflection on one's biases, theoretical predispositions, and so forth" (Schwandt, 2001).
14. Qualitative descriptive validity * Unfounded Instruments	**Was the descriptive validity of the qualitative data demonstrated?** Descriptive validity (Maxwell, 1992) includes the factual accuracy of the researcher's account of the data. Factors to consider in determining descriptive validity of the data include: 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 the study; the researcher use of memoing, peer debriefing/audit with other researchers, and member checking with subjects. Were the credibility of the researchers and the trustworthiness of the process for collecting data demonstrated?

Narrative paragraph for internal validity

This narrative should discuss the extent to which there is or isn't concern about the trustworthiness of the findings—specifically the relationships between independent and dependent variables that the author postulates. Reviewers should address the extent to which there are extraneous variables, which have or have not been controlled or otherwise accounted for by the researcher, and thus may or may not have influenced the findings of the study.

The review must discuss whether or not:
–bias is present;
–validity threats/alternative explanations are discussed and addressed through analysis;
–qualitative descriptive validity (studies with qualitative data only) is present

Poor: there are serious flaws with the internal validity which undermine the potential contributions of the study because:

- there is so *little information* provided that an adequate determination of validity threats is not possible; or
- the study has a biased sample which, instead of the independent variables, *could* account for the findings; or
- the study has another form of bias which, instead of the independent variables, *could* account for the findings; or
- there are other *compelling* alternative explanations for the findings than the independent variables in the study.

Limited: there are minor concerns with the internal validity so the potential contributions of the study should be "interpreted with caution" or "taken with some reservation" because:

- there is *limited* information to use in determining the level of validity threats; or
- the study has issues related to how the sample was selected which *may* account in part for the findings; or
- the study has another form of bias which may account in part for the findings; or
- there are viable but not compelling alternative explanations for the findings than the independent variables in the study.

Adequate: the internal validity of this study is sufficient thus it can make a sound contribution to the knowledge base because:

- there is *sufficient* information to determine the level of validity threats; *and*
- the sample selection process was satisfactory thus there are few concerns with bias; *and*
- there were not other forms of bias that are likely to account for the findings of the study; *and*
- the most likely viable alternative explanations for the findings other than the independent variables have been addressed in the study through analysis or discussion.

C. Analytic Precision: The extent to which the findings of a study were generated from systematic, transparent, accurate and thorough analyses. In qualitative studies, analytic precision refers to the quality/trustworthiness of the "processes of organizing, reducing, and describing data and continues through the activity of drawing conclusions or interpretations from the data and warranting those interpretations" (Schwandt, 2001).

Indicator	**Questions and guidance to consider**

15. Measurement Validity/Logic of Research Process * Unfounded Instruments	**Was the validity of the measures demonstrated for the constructs being studied?** Was measurement validity appropriately demonstrated for the nature and purposes of the study? Was any type of validation (e.g. content, convergent, discriminant, criterion-related) performed on the measures used in this study? In qualitative studies with naturalistic measures, were the credibility of the researchers and the logic and trustworthiness of the process for analyzing data demonstrated? Evidence of measurement validity should be explicitly mentioned in the study; a single type of validity evidence is generally sufficient.
16. Reliable measures/ Trustworthy techniques * Unfounded Instruments	**Were the data collection measures/techniques determined to be reliable/trustworthy for the groups/conditions under study?** For the purpose(s) of the study, were all the appropriate types of reliability (e.g. test-retest, internal consistency, alternate form, interrater, or agreement among independent coders) determined and results reported? If appropriate, were reported reliability coefficients at a professionally acceptable level, given the uses of the measures (e.g., lower reliability would be acceptable for examining group differences than would be the case for making decisions about individuals)? Were new instruments pilot tested? In qualitative data collection, were the trustworthiness and dependability of the data collection ensured, through strategies such as training for observation/interviewing, systematic adjudication of discrepancies among data replication of accounts by another researcher using transcription or video-taping, interrater checks on coding and classification, and clearly documenting the processes and procedures for recording and analyzing data.
17. Appropriate and Systematic Analysis * Inappropriate analysis	**Were the analysis strategies articulated?** Did the authors provide a description of the steps in their analysis strategy so that a reader can determine if the decisions they made were methodologically sound? This standard applies equally to quantitative and qualitative studies. A description of the overall process and not step-by-step description is sufficient; also, primary interest is in the quality of the analysis done to primary data source(s) that inform the research questions, rather than supplemental data sources that might be used only incidentally to provide substantiating evidence. If mixed methodology was used, how well integrated were the methods? **Were analysis strategies appropriately and systematically used to account for all relevant data?** Were the data for the findings analyzed appropriately? Were all relevant data from multiple measures analyzed to reach the findings (i.e. triangulation was sufficient)? In quantitative comparative studies, were pretest equivalence on covariates and dependent variables determined and handled appropriately in main effect analyses? Were the data for different treatment groups provided? Was significance testing done to determine group differences and results provided? In qualitative studies, were alternative trends and minority opinions included in the analysis?
18. Unit of Analysis issues * Inappropriate analysis	**Was the unit of analysis appropriate to the unit of assignment to the treatment, or to the research question?** For example, if schools were the unit of analysis and schools were assigned to the treatment conditions there is a match. If students were the unit of analysis and classrooms were assigned to the treatment conditions then there is NOT a match. For some research questions, the appropriate unit of analysis may differ from the unit of treatment, such as questions where some interim effect of treatment is the independent variable of interest (e.g., individual teacher knowledge resulting from a PD treatment to which schools were assigned), or where differential experience of the same treatment (e.g., urban/suburban/rural teachers) is of interest.

19. Power	**Was the design of the study conceived with sufficient power, including adequate sample size, to detect differences among participants if they exist?** In quantitative studies, reviewers need to pay particular attention to this issue for studies that have non-significant or no-difference between group findings. Did the design and sample size provide a reasonable opportunity to answer the research question that was posed? Did the researchers provide evidence of power analyses (as appropriate for a given design)? If power analyses were done, was a .80 level reached by the design, anticipated effect, and sample size? For qualitative comparative studies, was attention to the experiences of all groups fair and adequate?
20. Effect size	**Where appropriate, was the effect size of the results reported?** This standard is for quantitative analyses only. There is an increasing expectation that quantitative results should be reported with some indication of effect sizes, not just statistical testing results and significance. If effect sizes are not reported directly, are both means and measures of variability and standard error reported so that effect sizes could be calculated?
21. Multiple instruments/ Sources of evidence	**Were multiple instruments used to assess the dependent variables?** In qualitative studies, were there multiple sources of evidence cited so that the strength and variety of that evidence could be determined? In quantitative studies, were multiple measures used as a way to override possible sources of error or limitations inherent in one instrument or another?
22. Multiple respondents	**Were multiple respondents used to assess the independent and dependent variables?** This standard relates to one kind of triangulation—that is, triangulation of data sources. For example, a study examining something about classroom instruction could be based on information solely from observations by external people (one type of respondent). It could be based on data collected exclusively from teachers, even with multiple instruments, and still be considered one type of respondent. Conversely, a study could use information from observation, measures administered to teachers and measures administered to students and this would constitute multiple respondents.
23. All results	**Were all results reported, including non-significant and/or discrepant findings that do not necessarily support the main findings of the study?** The results of all analyses relevant to a research question should be presented and considered in answering the question.

Narrative paragraph for analytic precision

This narrative should address the extent to which the findings of this study were generated from systematic, transparent, accurate and thorough analyses, referring to those indicators above that are relevant to the study.

The review must discuss the presence or absence of: –measurement validity for all instruments used in the study –measurement reliability for all instruments used in the study –*appropriate* and systematic analysis techniques to support the findings of interest noted on the article summary –sufficient sample size (evenness for treatment groups if applicable)

Poor: there are serious flaws with the analytic precision which undermine the potential contributions of the study because:

- there is so *little information* provided that an adequate determination of analytic precision is not possible; or
- the researchers used measures without accounting for their validity for the purposes of the study; or
- the researchers used measures without accounting for their reliability for the purposes of the study;

or

- the researchers did not use appropriate or systematic analysis techniques; or
- the researchers did not have sufficient sample sizes to detect or document effect/difference if it was
- present.

Limited: there are minor concerns with the analytic precision so the potential contributions of the study should be "interpreted with caution" or "taken with some reservation" because:

- there is limited information to use in determining the level of analytic precision; or
- there are issues with the validity of the measures for the purposes of the study; or
- there are issues with the reliability of the measures for the purposes of the study; or
- the researchers did not use the most appropriate analysis techniques available for the purposes of the study; or
- the sample sizes were marginally acceptable to detect effect/difference if it was present; or
- the sample sizes were drastically uneven between groups (for comparative designs).

Adequate: the analytic precision of this study is sufficient thus it can make a sound contribution to the knowledge base because:

- there is sufficient information to determine the level of analytic precision; and
- valid measures for the purposes of the study were used; and
- reliable measures for the purposes of the study were used; and
- the researchers used appropriate analysis techniques and did so systematically; and
- the sample sizes were acceptable to detect effect/difference if it was present.

D. Generalizability/External Validity Determination: "The extent to which you can come to conclusions about one thing (e.g., population) based on information about another (e.g., sample)." (Vogt, 2005) There are two components to external validity—ecological representativeness and variable representativeness (Kerlinger, 1986). Ecological representativeness means that the study was conducted in such a way that if the social setting in which the research was conducted is changed (e.g., different schools, communities), the relationships found to be significant will remain so in other contexts. Variable representativeness is when the variables in the research consistently mean the same thing in different contexts. For example, does the way that "achievement" was operationalized in one study generalize to other contexts? Internal and external validity are often in tension with each other—it is hard to have the controls you need for good internal validity and have large enough sample sizes for external validity. In qualitative research, generalizability is "the process involved in moving from the specification of patterns, relations, processes, conditions, and meanings discerned in the data generated in the study of some particular event, person, institution, group, and so forth to a more general and abstract understanding of these aspects of human experience. This process is either empirical-statistical or theoretical-analytic, reflecting two different logics of sampling" (Schwandt, 2001).

Generalizability/External Validity

Indicator	Questions and guidance to consider
24. Findings for Whom	**Were analyses performed to establish to whom the findings apply or if the findings apply differentially to subgroups in the study?** Were analyses conducted to determine subgroup differences, dosage effects, and interaction effects relative the research questions of the study?

25. Generalizable to population or theory	**Given the nature of the sample for the study, are the findings generalizable to a larger population?** Was the population from which the sample was drawn adequately described and was the sample chosen sufficiently large and appropriately selected to be representative of the population? For qualitative studies, is there evidence to support analytic generalization? That is, were cases selected in a purposeful manner to support, refute, or refine a theory (analytic generalization)?
26. Generalizable to additional contexts	**Given the context in which the study was carried out, are the findings generalizable to other contexts?** Were the sample and study context adequately described and appropriately selected to provide some confidence that the study would be representative of different contexts?

Narrative paragraph for generalizability

This narrative should address whether or not: –there is something unique about the classrooms/ schools/ communities in which the study was conducted that would limit generalizing to other contexts; or –the way in which the variables in the research were operationalized would be done similarly in different contexts.

The review must discuss: –the extent that the findings are or are not likely to be generalizable beyond the current study sample –differential findings for subjects

Poor: there are serious flaws which undermine the generalizability of the study findings because:

- there is reason to suspect subgroup effects/difference, but analyses to investigate them were not conducted; or
- the population from which the sample was drawn was not adequately described to determine if the sample was appropriate and sufficiently large to be representative of the population; or
- the sample and study context were not adequately described nor appropriately selected to be representative of different contexts to which the findings are generalized.

Limited: there are minor concerns with the generalizability of the study findings so they should be "interpreted with caution" or "taken with some reservation" because:
if there were significant differences/effects among subgroups, they were identified without explanation; or

- the population from which the sample was drawn was not adequately described to determine if the sample was appropriate to be representative but the sample was large enough for generalizability purposes; or
- the sample and study context were marginally described such that they may be representative of different contexts to which the findings are generalized.

Adequate: the generalizability of the study findings is sufficient because:

- if there were significant subgroup differences/effects they were investigated and explained; and
- the population from which the sample was drawn was adequately described and the sample was determined to be representative both in demographic characteristics and sample size; and
- the sample and study context were described and are representative of different contexts to which the findings are generalized.

E. Overall Fit: The extent to which the research questions, study design, data collection procedures, and analysis techniques align. This narrative should focus on the most noteworthy aspects of the Domains A-D as

they relate to the findings of the study. It is not necessary to repeat everything that is mentioned in previous sections. Rather the narrative should focus on the most important issues that affect the interpretation of the study, drawing on previous sections. For a study judged to have one or more substantial limitations, the section A-D narratives will not be written. Therefore this narrative should identify and provide evidence of the substantial limitation, describing implications of the limitation for interpreting the findings of the study.

How well does the research design align with the research questions? Indicate the extent to which the data collection procedures make it possible to answer the questions that are presented.

- Alignment of the research design and research questions relates to **Internal Validity**. Based on this design, how much confidence can we have that any observed effects are a result of the independent variables or treatment, if applicable, rather than something else (such as preexisting or correlated differences that were not accounted for, effects of something other than the treatment of interest, etc.)?
- How well was the research design actually **implemented**? It is possible for a study to identify an appropriate design but fail to properly implement it.
- The selection of **participants** is related to the study's **Generalizability** (i.e., Can you learn about the population to which the research questions relate from the sample that is included in the study?) For example, if a research question refers to practices of science teachers in general, but the sample includes only new science teachers, this would indicate a potential lack of alignment between the participants in the study (given the data collection procedures) and the population that the researchers would like to learn about (given the research questions).

How well does the analysis align with the research design and research questions?

- The alignment of analysis with design and questions relates to **Analytic Precision**. Is the analysis appropriate for addressing the research design? For example, if the research design involves collecting data from both treatment and control (or comparison) participants, administering a pre-test, or collecting data on covariates of interest, is this information used appropriately in the analyses?
- Are the constructs of interest in the research questions measured with **adequate reliability and validity**? For example, in a study where a primary research question refers to teacher quality, how confident are you that what was measured really does capture teacher quality? How confident are you that the measures would provide the same results if implemented by someone else or at a different point in time?

The rating for Overall Fit should not be higher than the rating given for either Internal Validity or Analytic Precision. For studies with a clear primary intention to generalize to a population beyond the sample of participants, the rating for Overall Fit should not be higher than the rating given for Generalizability.

For a study judged to have one or more substantial limitations, the Overall Fit rating cannot be Adequate.

Poor: Studies with poor Internal Validity and/or Analytic Precision. In a study with poor fit, the research design may be inadequate for answering the research question(s). In this case, it would appear impossible to answer the questions posed given the methods described. There may be serious threats to Internal Validity such that it is equally likely that observed effects are due to something other than the independent variables. With all studies, there is a chance that a finding could be attributed to something other than the independent variables; but if the alternative explanation seems to be at least as likely, this is a serious problem. Another example of poor fit would involve an analysis with serious flaws that

undermine the results, such as using a clearly inappropriate statistical technique or using measures with no basis for reliability and validity. If you were presented with a group of several studies with poor fit, you would question the findings even if they seemed to reach the same conclusions.

STOP CODING if you assigned a rating of Poor to Overall Fit

Limited: There is some question about the appropriateness of the research design and/or analysis, but it does not appear to be blatant or serious enough to invalidate the findings. For example, a study with limited fit may use a sample that does not necessarily mirror the population, but is defined well enough that learning something about the sample members would provide information that could at least suggest possible effects for the overall population. A limited study may have some threats to Internal Validity, but there would not be any issues blatant enough to be as likely a cause of an observed effect as the independent variables. The measures used may not have clear support for reliability and validity, but a lack of reliability or validity is not an equally likely explanation for the findings. The analysis strategies may pose some questions as to appropriateness (such as failing to include an important covariate), but the primary assumptions/use of the technique will be appropriate. Any shortcoming identified in the analysis would be unlikely to be the main cause of the findings. Despite some misgivings about a particular study with limited fit, you would feel comfortable if a group of limited studies reached the same conclusions, especially if they had different limitations.

Adequate: Any threats to Internal Validity and Analytic Precision are minor, and potentially negligible. There are always ways that any study can be improved, but if these improvements are not likely to substantially change the findings, then the study is adequate. In a study with adequate fit, there is a reasonable chance of arriving at the same findings and conclusions if the study were replicated in a similar context.

F. Warrants for claims: The extent to which the data interpretation, conclusions, and recommendations are justifiable based on the evidence presented.

Indicator Questions and guidance to consider

27. Limitations	**Were the study's limitations presented?** Does the author describe the shortcomings of the methods used, the sample selected or the interpretability of the findings generated? Is there clear indication that the author is cognizant of major limitations of the study by stating anything that would severely limit the utility/generalizability/trustworthiness of the findings?
28. Decay and Delay of effect	**For studies that have outcomes that may decay over time or take a long time to manifest, was there at least one long-term follow-up at an appropriate interval (e.g., beyond the end of the intervention) on key outcome variables?**
29. Efficacy	For studies that are determining the efficacy of an intervention on outcomes that are measured quantitatively on multiple indicators/instruments, *most or all of the results* for that construct *must be in the positive direction* and at least one must be statistically significant to claim efficacy**. Do this study's results meet this standard?**
30. Conclusions/ implications logically drawn from findings	**Were the conclusions/implications that were drawn logically derived from the findings of the study?** The study should make a logical case/argument that its findings lead to the conclusions and implications that are presented, including acknowledgement/explanation of any discrepant findings.

31. Conclusions/ implications grounded in theory	**Do the conclusions/implications fit within the conceptual or theoretical framework for the study?** Most conclusions and implications will involve some assumptions and inferences beyond the direct findings of the study. Are these conclusions/implications derived logically from both the findings and the conceptual/theoretical frame of the study? If any conclusions/implications call that theory or its appropriateness for the study context into question, is that case made logically from the findings?

Warrants for claims for Conclusions:
Were the conclusions warranted based on the nature and quality of the study and its findings?
A conclusion is what the researcher makes of findings, often using a theoretical or conceptual framework for interpreting the findings. This narrative section addresses the inferential leap from the findings to the conclusions. That is, how much of a stretch is it to go from the findings to the conclusions? Is there evidence for the conclusions, beyond a reasonable doubt? Is theoretical or evidentiary support from other research used appropriately to draw these inferences from the findings? How likely would other researchers in the field be to reach the same conclusions from the study results? If some conclusions are more strongly warranted than others, this should be noted. In terms of limitations being presented, it is important that the limitations be noted in the context of the conclusions, rather than simply stated upfront but ignored in the interpretation of the findings. In terms of the researcher's "subjectivity" or bias, it is important that this is stated and included in the same way that limitations should be addressed.

Warrants for claims for Implications:
Were the implications warranted based on the nature and quality of the study and its conclusions?
An implication is a suggestion or recommendation for what policymakers, practitioners, or other researchers should consider or do as a consequence of the findings/conclusions of the study. This narrative section addresses the inferential leap from the findings/conclusions to the implications. That is, how much of a stretch is it to go from the findings/conclusions to the implications? Is there evidence for the implications, beyond a reasonable doubt? Is theoretical or evidentiary support from other research used appropriately to draw these implications from the study? How likely would other researchers in the field be to derive the same implications from the study results? If some implications are more strongly warranted than others, this should be noted. In terms of limitations being presented, it is important that the limitations be noted in the context of the implications, rather than simply stated upfront but ignored in the interpretation of the findings. In terms of the researcher's "subjectivity" or bias, it is important that this is stated and included in the same way that limitations should be addressed.

Poor: There is little or no basis for reaching the conclusions/implications based on the findings. This could be because the leap from the findings to the conclusions/implications is too vast to be warranted, or because the findings actually present conflicting evidence to what is reported in the conclusions/implications. This rating of poor would indicate that other researchers in the area could just as easily draw different conclusions from the same findings. There may be major limitations in the study that were not acknowledged in the interpretation of the findings. There may also be a major problem with the researchers not accounting for subjectivity or bias from the role that they play in the data collection process.

Limited: There is some pattern in the findings to suggest a trend towards the conclusions, but the evidence may not be adequate or the evidence may be interpreted differently by other researchers. Some study limitations may be explicitly noted at the beginning of the conclusion section, yet some of the conclusions drawn may appear to ignore these limitations. A limited rating may indicate that the findings are more tentative than the conclusions suggest; for example, they may be very likely to decay over time, but perhaps this was not investigated. The study may have found mixed results yet some results were

ignored or inadequately explained in the conclusions.

Adequate: The findings contain reasonable evidence for the conclusions/implications, and the conclusions/implications are consistent with and supported by theory or evidence from other research. There may be some uncertainty about the conclusions/implications, but there are no serious limitations, unwarranted claims, or threats from researcher subjectivity/bias; and the main limitations that exist are either accounted for in the analysis/discussion or are mentioned as caveats to the conclusions that are drawn.

APPENDIX B
MSP-KMD Standards of Evidence Coding Form

Study Authors:
First three words of study title:
Reviewer Name:
Date of Coding:

A. Adequate Documentation of Project Activities

1. Research question and constructs	
2. Research site	
3. Sample demographics	
4. Interventions and implementation	
5. Data collection	

Narrative paragraph for project activity documentation

Poor	Limited	Adequate
1	2	3

B. Internal Validity

6. Sample Bias	
7. Response bias	
8. Attrition Bias	
9. Missing Data Bias	
10. Contamination	
11. Validity threats addressed through analysis	
12. Validity threats discussed	
13. Investigator Bias	
14. Qualitative descriptive validity	

Narrative paragraph for internal validity

Poor	Limited	Adequate
1	2	3

C. Analytic Precision

15. Measurement validity	
16. Reliable measures	
17. Appropriate and Systematic Analysis	
18. Unit of Analysis issues	
19. Power	
20. Effect size	
21. Multiple instruments	
22. Multiple respondents	
23. All results	

Narrative paragraph for analytic precision

Poor	Limited	Adequate
1	2	3

D. Generalizability/External Validity Determination

24. Findings for Whom	
25. Generalizable to population	
26. Generalizable to additional contexts	

Narrative paragraph for Generalizability

Poor	Limited	Adequate
1	2	3

E. Overall Fit: The extent to which the research questions, study design, data collection procedures, and analysis techniques align. Use information from Narratives A-D here for evidence to justify your rating. This is the narrative that the synthesis will use.

Poor	Limited	Adequate
1	2	3

STOP CODING
If selected "1"

F. Warrants for claims

27. Limitations Presented	
28. Decay and Delay of effect	
29. Efficacy	
30. Conclusions/ implications logically drawn from findings	
31. Conclusions/ implications grounded in theory	

Narrative paragraph for Warrants for Claims **for Conclusions**

Narrative paragraph for Warrants for claims for Implications

Poor	Limited	Adequate
1	2	3