STANDARDS OF EVIDENCE REVIEW
REVISED-October 6, 2016

CODEBOOK

As part of the Knowledge Management and Dissemination for the MSPs project (NSF EHR–0445398), HRI developed the original Standards of Evidence review (Heck & Minner, 2010) to evaluate studies and provide detailed feedback on reporting empirical research procedures and evidence to support claims. HRI continues to use the Standards of Evidence review to evaluate the credibility of study claims. The protocol HRI currently uses was revised in 2014 for the evaluation of NSF's Research and Evaluation in Engineering and Science Education (REESE) program.

The Standards of Evidence review is commonly used in the summative phase of the evaluation of research projects. This summative activity focuses on reviewing the project's research findings to judge the extent to which claims based on the research are supported by the data collection, analysis, interpretation, and reporting. As a formative evaluation activity, HRI reviews research products as they are readied for dissemination, providing feedback about a manuscript's strengths and weaknesses, as identified by the Standards of Evidence review.

Another formative component of many evaluations is to provide feedback on the project's research design. The research design review process examines the alignment between the project's research questions and the research plan, guided by a procedure for evaluating the rigor of research designs that HRI developed based on the Standards of Evidence.

Notes about some key nomenclature used in this codebook

- **Results =** the actual output from the analysis (e.g., the F-test and significance values, counts of instances using qualitative coding scheme)
- **Finding** = interpreting the analysis results directly in relation to the research question
- **Conclusion** = what the researcher makes of findings, often using a theoretical or conceptual framework for interpreting the findings
- **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

Note about outside sources

Occasionally research products cite additional work products when describing the study sample or instrumentation. For documentation sections I and II in the form, do not consider what information may be in these additional work products as you complete your review. For sections III and IV, reference to additional work products may be part of the review process, provided the original research product makes clear what information is located in the additional work products and these products are easily accessible to the reader. Reviewers should note whether additional work products were consulted and how they contributed to the rating. In the case where additional work products are cited, but not easily consulted, the reviewer must weigh the importance of the missing or incomplete information in the original product being reviewed in order to determine whether sufficient evidence for the findings is provided. If concerns persist, the lack of information should be reflected in the related indicators and possibly in the final rating.

Codebook for Standards of Evidence for Empirical Research

The initial section asks for record-keeping information (study authors, title, reviewer name, and coding date). To assist readers unfamiliar to the study, the reviewer should copy and paste the study abstract, the findings, and the conclusions/implications in the corresponding box. The grain size of the findings should be consistent with the research questions outlined in the next section. Reporting of the technical information pertaining to statistical tests is not necessary. For greater ease, the reviewer may decide to summarize information instead of copying and pasting directly from the research product.

Section I outlines the criteria for documenting the study. For each criterion, please check the box to confirm that the information outlined is present in the text. Note that checking this box indicates that the information exists in the research product and is not a judgment about whether the information is sufficient or of high quality. On the other side, note that an unchecked box does not necessarily signify a deficiency in the research product. An expectation or a need for every study to report on all the documentation criteria may not exist. The reviewer's responsibility in the documentation sections is to evaluate whether specific information appears in the research product. If you have concerns, note them in the reviewer's comments column. To expedite validation of responses and others researchers' use of the review, please include page numbers.

I. DOCUMENTATION - STUDY

Criterion	Guidance
A. <u>Situating the Research</u>	
1. Intended contribution of this study/research questions	The coordinator will identify the intended contribution(s). 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. If the intended contribution is stated explicitly in the form of a research question, then the research questions will be listed verbatim. If not, or if some, but not all of the intended contributions were stated as research questions, the coordinator will summarize the intended contribution(s). Please confirm the research questions/intended contributions as recorded by the coordinator. If you disagree with the coordinator's decisions, please notify the coordinator. The remainder of the protocol (after this DOCUMENTATION-STUDY section) is conducted per research question/intended contribution. The protocol is intended for one research design per research question. For studies that do not follow this formula, decisions will be made on a case-by-case basis. Alert the coordinator in such situations.

Criterion	Guidance
2. Theoretical background	Was a theoretical frame for the study provided? To satisfy this indicator, the authors need to do more than cite the current literature. However, an explicit statement naming of a theory is not required. The authors need to present an argument behind what they are doing in the study. What is the theory of action for the intervention(s)/construct(s) being studied?
3. Current knowledge	Was there a discussion of the extant literature summarizing what the field knows (and doesn't know) relevant to the current study? Current knowledge is present if, at a minimum, the research product identifies findings from previous studies that the researchers are replicating, extending, or challenging. To distinguish current knowledge from theoretical background, think of theoretical background as the underlying reasoning behind the authors' approach to studying a particular intervention/construct and current knowledge as evidence of the work done in other studies that relate to similar interventions or the same construct.
4. Constructs as they are operationalized	Was there a description of how the key factors and constructs (i.e., the constructs referenced in the research question) were operationalized through indicators or illustrative examples?
5. Researcher disclosure	Was there disclosure of researcher-specific background information which might significantly influence the data collection approach or interpretations of data? Was information disclosed that could allow readers to gain a greater understanding of the lens through which the researchers approached the study? Researcher disclosure is considered particularly important in qualitative research because the researcher is responsible for portions of the data collection and analysis in which their perceptions can substantially influence the findings. However, in both quantitative and qualitative studies in education, researchers often have a relationship to the sample and/or the research event that should be disclosed. For example, if the researcher created the intervention being studied, that information should be stated in the research product. Note that citations of the author's previous work are potential evidence of researcher disclosure. Footnotes and author biographies (occasionally present at the beginning or end of products) should also be examined for evidence of researcher disclosure.

Criterion	Guidance
6. Intended generalizability	Do the researchers indicate the extent of generalizability: either describing the population and context to which the findings apply or acknowledging the lack of generalizability? Explicitly stating the population to which the study generalizes or referring to a population without explicitly labeling the group as the "population" is acceptable. Describing a population of which the sample is representative is sufficient. In basic research, consider generalizations made from a particular instance to a broader theory.
7. Directions for future research	Were extensions of the current study, new lines of inquiry, or new research questions suggested to help guide future research? Suggesting other methodological approaches that would extend or strengthen the claims of the study is acceptable evidence of directions for future research.

Similar to Section I, Section II is concerned with documentation. However, this section needs to be completed for each individual research question/intended contribution. Again, please note any issues about quality in the reviewer's notes column and reserve the checklist as a place to record whether or not information for each criterion was included in the research product.

II. DOCUMENTATION – RESEARCH QUESTION 1

Criterion	Guidance
A. The Units of Study	
1. Research site	Were the location(s) where the research took place described? Note that we are referring to a description, not just a reference to the location. Pertinent information could include the name of the actual school or laboratory, its geographic location, or demographic information on the community in which the research site is situated. This list is not intended to be exhaustive or mandatory, but as a guide to determine whether a description is or is not present. However, information that only describes participants (e.g., undergraduates) is not sufficient for describing the research site.
2. Research participants	Was a relevant description about the sample provided on characteristics likely to relate to the research question and contexts (such as the ages of the students, years of experience of teachers, SES/gender of the participants, etc.)? In order for a description to be provided, at least three relevant indicators describing the sample must be present.
3. Research "event"	Was the phenomenon being investigated (such as the teaching practice, intervention, experiment, lab trial, etc.) described? In some cases, the phenomenon is observed within the data collection process (e.g., lab experiment) and other times it is a separate activity (e.g., teaching practice). It is also possible that the research "event" is a naturally-occurring phenomenon that is described in the section of the product that situates the research.
B. Design	
1. Sampling/assignment strategy	Was there a description of the method used to select participants into the study sample or to assign them to conditions?
2. Design type	Was there a description or designation of the type of study design? For example, is there information indicating that the research was conducted as a case study, a correlational study, an experiment, an ethnography, etc.? Although information that fleshes out the design might be contained in descriptions of the data collection or analysis, this indicator is not satisfied if the design must be inferred from the data collection or analysis. Note also that the design type specified (e.g., "experiment") must be consistent with the design used. If not, this indicator is not satisfied.

Criterion	Guidance
C. <u>Collection of Data and Instrumentation</u>	
1. What data collection methods were used	Was there information to determine what data collection methods were used?
2. What instruments were used	Did the research product name or describe the instruments used? Note that we do not expect authors to provide copies of the instrument, to list the individual items, or give the actual interview questions. However, some descriptive information about the way the instrument is organized and what content it addresses should be present, especially if the instrument is not explicitly named. In cases where multiple instruments are used to target the research question, all of the instruments must meet these guidelines in order for this criterion to be judged sufficient.
3. Where the data were gathered	Was there information to determine where the data collection took place? This indicator refers to a more specific location of data collection than the research site (II.A.1), such as in a school, in students' homes, in a university lab, or online. Naming a geographic place, such as Cleveland, is not sufficient.
4. When the data were gathered (relative to the research event)	Was there pertinent information on when the data were gathered relative to the research event? The information needs to be specific when it is likely to matter for making judgments in the validity section. For example, to say "post-test" may not be sufficient because the passage of time between the intervention and the assessment may have ramifications for judging the validity of this study design to adequately measure the outcome of the intervention.
5. How the data were processed	Was there information on how the data were processed (e.g., information on the scoring of an assessment, the application of a rubric, the development of a coding scheme), with "processed" meaning how information gathered through data collection activities were translated into data used in analysis?
D. <u>Analysis</u>	
1. Analysis strategy	Were the analysis strategies articulated? Were the steps in the analysis strategy described so that a reader can determine if the decisions made were methodologically sound? This standard applies equally to quantitative and qualitative studies.
2. Results	Were results relevant to this research question reported?
E. <u>Findings</u>	
1. Empirical support for findings	Were the findings explicitly connected to the empirical results?
2. Limitations of findings	Was there a description of the limitations of the methods used, the sample selected or the interpretability of the findings generated?

Section III requires a more sophisticated level of analysis. The reviewer is moving from recording the presence of information about criteria to making a judgment about the quality or sufficiency, as reported, of how the criterion was addressed. For each criterion, the reviewer must assess whether a validity threat should have been addressed and, if so, whether important concerns about the integrity of the study remain due to that validity threat. In this situation the reviewer should check "Concerns Remain" and support that judgment with a brief description of how the criterion was not sufficiently addressed. If the research is not vulnerable to a particular validity threat or resolves the major concerns related to a criterion, the reviewer should select "Satisfactory or NA". Each identified issue in a study should be recorded under the criterion for which the concern is the most prevalent. An identified issue should be mentioned under another criterion only if it raises a different concern.

III. VALIDITY– RESEARCH QUESTION 1

Criterion	Guidance
A. <u>Avoiding Bias in Design</u>	
1. Protections against sample bias*	Considerations include how the sample was recruited (e.g., only volunteers, or waitlist, a typical instantiation of the phenomenon, 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. Studies intending to generalize to a broader population must ensure that the sample is representative of the population. In order to determine if potential bias is a concern, the reviewer must evaluate whether the population to which the claims are generalized is adequately represented in the sample. Remember, many studies do not intend to make broad generalizations. Note that a recruitment/selection strategy in and of itself (e.g., snowball, purposeful) neither automatically addresses nor fails to address potential bias. A rationale for the method and discussion of the potential impacts of the sample selection method on findings and how analysis will address those possible impacts should be present. If bias is likely, does the research product appropriately acknowledge limitations of the sample selection strategy (e.g., self-selection issues from using a volunteer sample) as they relate to the research question? For qualitative studies, is there clear logic behind the selection of sites/participants/cases to inform the research question?
2. Protections against unfair comparisons	If this is a comparison study, how were participants assigned to treatment and control groups? Were steps taken to increase the reader's confidence that these groups were initially equivalent? Were those steps (e.g., RCT, matching techniques, using covariates) performed satisfactorily?

Criterion	Guidance
B. Avoiding Bias in Conducting Research	
1. Protections against non-response bias*	In order to select "Concerns Remain" for this criterion, the research product must contain explicit information about the participant recruitment. Consider issues surroundings samples where participants choose to opt-in to the study under the sample bias criterion (III.A.1). 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 potential response bias is: teachers who responded to a survey or agreed to be observed were primarily 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. Alternatively, if bias is suspected, additional information may be provided that reduces the likelihood of bias. For example, there would be less concern about low response to a school-wide teacher survey on teacher beliefs if the characteristics of responding teachers are similar to those of all teachers in the school (e.g., in average years of experience, student test scores).
2. Protections against attrition bias*	Attrition refers to the loss of participants from the sample from the beginning of the study to the end of the study. Was the overall attrition rate explored? If the rate was high, were the implications for the results addressed or discussed? Could differential attrition among participant groups have biased the results? If there was differential attrition, were the implications for the results addressed or discussed? If attrition information is not reported explicitly, the reviewer may have other sources of information available for evaluating whether attrition was likely a potential source of bias such as the number of cases in the sample compared to the number of cases in the analysis.

Criterion	Guidance
3. Protections against missing data bias*	Missing data may be an issue if participants (or researchers) do not complete all of a data collection instrument, participants do not participate in parts of the data collection, or if the researcher excludes or ignores some part of the data. Was there information in the research product indicating either that there were no missing data; or, in quantitative studies, that missing data were imputed using standard procedures; or that extent of missing data was explored and determined to not be a concern for bias? If information about the extent of missing data is not reported explicitly, the reviewer may have other sources of information available for evaluating whether missing data was likely a potential source of bias such as the number of cases in the sample compared to the number of cases retained in the analysis.
4. Protections against contamination	If a comparison study, were concerns about contamination between treatment and comparison conditions addressed? For example, when teachers from the same school are assigned to treatment and comparison conditions, was information provided about what steps were taken to avoid contamination or why contamination is not a concern? In its classic sense, contamination relates to diffusion of the treatment, like a spillover effect of the treatment into control conditions. You might also think of contamination in terms of altered perceptions of the participants because of knowledge about the existence of various experimental conditions (e.g., if a participant in the control group believed the treatment condition was more desirable, it may affect the measured outcome). Or you may think of a situation where simply the presence of an observer likely results in atypical behavior. Contamination arises when the participants themselves or the event being investigated could be affected in unintended ways by the research.

Criterion	Guidance
5. Protections against investigator bias	Could the relationship between the researcher(s) and the treatments or participants have biased the interpretation of the findings? For example, did the researcher deliver the treatment or have a pre-existing relationship with the research participants? If so, were appropriate 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). Unlike contamination which is related to participant behavior, investigator bias is related to researcher behavior. An evaluation of protections against investigator bias requires an examination of the extent to which the researcher's stance during the data collection, analysis, or interpretation stages may bias the findings.
C. <u>Appropriate Data Collection Methods</u>	
1. The data collection methods were justified for addressing the research question/purpose	If a case is made for the data collection strategy selected, is the choice justified given the research question? If a case is not made, is the choice of data collection strategy logical given the research question? The focus of this criterion is on the data collection method more broadly, and should not move to a critique of any actual instruments used. Note concerns about the design not covered under other criteria under this criterion (e.g., concerns about the timing of the data collection).

Criterion	Guidance
2. The instruments were justified for addressing the research question/purpose	To evaluate this criterion consider the following for each instrument: Do the authors make a case for the instrument's appropriateness given the research questions? Was an appropriate case made for the validity, reliability, and/or credibility of the instrument for addressing the purposes of the study? For quantitative studies, was any type of validation (e.g., content, convergent, discriminant, criterion-related) reported on the measure used in this study? In data collection for quantitative measures, were appropriate types of reliability information (e.g., test-retest, internal consistency, alternate form, interrater, or agreement among independent coders) reported given the purposes of the study? Standard instruments in the field (e.g., WISC IQ test) satisfy this indicator so long as the instrument is normed appropriately for the study sample. For qualitative studies, was the logic of the process for gathering and analyzing data reported? The protocol or process for obtaining and recording data must be appropriate to the research question and should provide an opportunity to collect evidence to confirm or disconfirm the researcher's hypotheses or assumptions. In qualitative data collection, were the trustworthiness and dependability of the data collection documented through strategies such as training for observing/interviewing, systematic adjudication of discrepancies in coding, replication of accounts by another researcher using transcription or video-taping, or inter-rater checks on coding and classification? In addition, to meet this criterion the authors must provide a justification for the instrument. Whether we as reviewers have doubts on the quality of the instrument should not play a part in the rating. (If a reviewer has expertise in the content area and is troubled by the instrument, please note this in the reviewer comments box. The rating, however, should be based solely on whether a justification for the instrument was made by the authors along with appropriate evidence supporting the justification.)

Criterion	Guidance
3. Triangulation across multiple data collection methods, researchers, instruments, or respondents was appropriate	Evaluation of this criterion moves beyond the merits of an individual instrument and instead looks across instruments, data collection, and data processing. Did the authors use multiple methods, instruments, researchers, or respondents to assess key variables? If so, was the triangulation appropriate given the research question? If not, did the lack of triangulation leave the study vulnerable to bias? 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?
D. <u>Appropriate and Systematic Analysis</u>	
1. Appropriate unit of 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. However, it is problematic if the analysis is not tied to or does not pertain to the phenomenon of interest.
2. Appropriate/logical method of analysis	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 results? In quantitative comparative studies, was pretest equivalence on covariates and dependent variables determined and handled appropriately in main effect analyses? Was significance testing done to determine group differences and results provided? In contrast, was there evidence of statistical "fishing"? If multiple comparisons were made, did the researcher make appropriate adjustments? In qualitative studies, were alternative trends, discrepant evidence, and minority opinions included in the analysis?

Criterion	Guidance
3. Characteristics and size of sample suitable for planned analyses	Was the study designed with sufficient ability to detect differences, including a sample with enough variation and adequate size to address the research question, e.g., to detect differences among groups if they exist? Focus on whether the sample met the needs for completing the planned analysis and **NOT** on whether the sample was appropriate for making generalizations. 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 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? If a power analysis was not described, based on the information provided, did the design and sample size provide a reasonable opportunity to answer the research question that was posed? (Note: If you find the authors found a statistically significant result, the sample size was adequate.) For qualitative comparative studies, was attention to the experiences of all groups fair and adequate? Does the sample adequately reflect the variation necessary for addressing the research question?
E. <u>Appropriate Reporting of Results</u>	
1. All results, including null results and discrepant evidence	Were all results relevant to this research question reported, including non-significant and/or discrepant results? If discrepant results are not presented, is there some evidence that they looked for it and didn't find it?
2. Size of effect	For quantitative analyses, was the size of the effect stated? Was the magnitude and practical importance of the size of the effect conveyed?

Criterion	Guidance
3. Qualitative descriptive validity	For qualitative methodologies, 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's use of memoing, peer debriefing/audit with other researchers, and member checking with participants. When making your rating, be sure that you are focusing on the description itself. One sign of concern for this indicator would be if the product referred to a lot of anecdotal evidence that was not dependent on systematic data collection and analysis. This criterion should capture any doubts the reviewer has about the trustworthiness of the research methods used that have not been noted elsewhere in the validity section of the protocol.

Criterion	Guidance
F. <u>Considering Alternative Explanations</u>	
1. Alternative explanations considered through either the design, analytic strategy, discussion, or in recommendations for future research	Here are some questions to consider as this indicator is rated. Note that not all of these questions must be addressed for a product to receive a "Satisfactory (or NA)" rating. Were design decisions made to rule out plausible alternative explanations for the findings? For example, did the study include an additional comparison group to rule out historical threats to validity? Or, were all participants interviewed in comparable locations (e.g., their own homes) to rule out the effects a different environment might have on participant responses? Were viable alternative explanations (threats to validity or credibility) addressed in the analysis strategy, either explicitly or implicitly? 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. Were viable alternative explanations (threats to validity) addressed in the discussion? Did the author discuss the possible effects 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? Were alternative explanations discussed in terms of guiding future research? For example, did the author point out other plausible explanations for the findings and suggest testing these explanations in future studies?
2. Consistency of results with findings	Were the results consistent with the findings? If discrepant results or alternative explanations were presented, were they accounted for in the findings?

*Use the following table to assist in classifying threats to validity A1, B1-B3.

Study Phase	**Ask Yourself:**	**If yes, there is potential for…**
Deciding whom to recruit	Is it likely that the process of inviting individuals to participate into the study introduced bias?	Sample bias
Recruiting participants	Given that invited individuals can decide not to participate, is it likely that the participants are qualitatively different than people who decided not to participate?	Non-response bias
Executing the study	Did a significant proportion of participants exit the study before it was finished? Are they likely to be different from those who completed the study? If the study involves comparisons of different groups of participants, did a large majority of participants who exited belong to one group?	Attrition bias
Processing/analyzing the data	Considering only those participants who completed the study, were pieces of data missing for some participants?	Missing data bias

Section IV requires reviewers to synthesize all of the information from the protocol. Keeping in mind this information, reviewers are asked to rate (by research question) the extent to which the research product meets standards. If the design was not consistent with the research problem or if the analysis was not consistent with the design, it is not possible for the findings to be empirically supported, rendering a rating of "1" which means the product does not meet standards. If the design is consistent with the research problem and the analysis aligns with the design, a product can meet standards. However, the extent of confidence the reader should place in a research product's findings is determined by how well the research product's evidence supports the findings and rules out alternative explanations. Because all studies include some weaknesses, the reviewer must carefully consider the study's approach and the vulnerabilities inherent in the approach as s/he weighs the evidence supporting the claims against competing doubts. When revisiting the indicator ratings to assess a study's validity, reviewers should distinguish between knowing there is a problem and speculating that there might have been a problem. If significant reservations about the findings persist that outweigh evidence in support of the findings, a rating of "2" should be assigned. If all of the evidence that supports the findings outweighs reasonable reservations, a rating of "3" should be assigned. A rating of "3" should be assigned even if concerns remain about one or more criteria when there is sound evidence for the findings. All ratings should be supported with a narrative of 1-2 paragraphs which explains the key factors for selecting the rating.

Research products should receive a rating of "1" if the author could not possibly answer the research question due to serious concerns about the sample or analysis strategy or some significant bias present in the design. 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. If you were presented with a group of several studies with research questions rated a "1" you would question the findings even if they seemed to reach the same conclusions.

Research products should receive a rating of "2" if the findings shed some light on the research question, but important reservations remain. Concerns do not appear to be blatant or serious enough to invalidate the findings, but the holes or gaps in some parts of the study leave plausible alternative explanations unaddressed. Despite some misgivings about a particular study with a rating of "2" you would feel comfortable if a group of studies with "2" ratings reached the same conclusions, especially if they had different limitations.

Research products should receive a rating of "3" if the findings are supported by the evidence and the researchers have addressed alternative explanations by: (1) designing ways to test them or rule them out, (2) analyzing data to control for or account for them, or (3) discussing their likelihood. No study is perfect; the reviewer is to weigh the extent to which the authors made a case for the approach to answering the research question and whether the evidenced produced was sufficient for shedding light on the research question. If addressing any remaining concerns would more than likely have strengthened the case for the findings, then a research product should be rated a "3." Ratings of "3" apply when there is a reasonable chance of arriving at the same findings and conclusions if the study were replicated in a similar context.

IV. OVERALL RATING: QUALITY OF THE RESEARCH PRODUCT—
RESEARCH QUESTION 1

Level 1 – Does not meet standards Research design does not align with the stated problem. (Check if applicable ☐) and/or The analysis conducted does not align with the design. (Check if applicable ☐) and/or Findings are not supported by the evidence. (Check if applicable ☐) and/or Documentation lacks elements necessary to determine if findings are supported by the evidence. (Check if applicable☐)
Level 2 – Meets standards with reservations Limitations of the described design and/or analysis raise important concerns about the extent to which the findings are supported by the evidence.
Level 3 – Meets standards Strengths of the described design and analysis outweigh the limitations, indicating that the findings are generally supported by the evidence.

Overall rating: _____ (drop down)

Please explain your rating:

If the research product was rated a "1" for all research questions your coding process for this product is complete. Otherwise, respond to Sections V for the research product as a whole. Section V requires the reviewer to evaluate the connection between the findings, conclusions, generalizations, and implications in addressing the research question (see page 2 in this codebook for definitions of these terms). Most studies provide conclusions. Some studies provide conclusions, but do not extend beyond the study sample in making generalizations or giving implications for the research. Please select NA for the appropriate criteria for research products that do not include conclusions, generalizations, and/or implications.

V. GENERALIZATIONS, CONCLUSIONS, IMPLICATIONS OF THE STUDY

	Guidance
1. Conclusions aligned with the study's findings	Were the conclusions that were drawn logically derived from the findings of the study? The research product should make a logical case/argument that its findings lead to the conclusions that are presented, including acknowledgement/explanation of any discrepant findings.

	Guidance
2. Generalizations of the findings stated with appropriate caveats or bounds	Were the generalizations of the findings sensitive to the constraints of using a particular sample or context for the study? Was the sample sufficiently large and appropriately selected to be representative of the population and any sub-populations to which generalizations are made? In situations where the research was conducted in one or more contexts and generalizations are made regarding other contexts, were the studied contexts adequately described and appropriately selected to provide some confidence that they would be representative of other contexts? For qualitative studies in particular, is there evidence to support analytic generalization? That is, were cases/contexts selected in a purposeful manner and adequately described to support, refute, or refine a theory or framework? If generalizations are included, are the caveats and bounds of generalizing the study's findings stated? (A statement of the caveats or bounds of generalizability must be present at least once in the research product to meet this criterion.)
3. Implications aligned with findings and sensitive to the study's limitations	Most implications (e.g., decisions/actions that might be taken in response to a study's findings) will involve some assumptions and inferences beyond the direct findings of the study. The study may or may not have acknowledged/reported all of the study's limitations, but implications should still be sensitive to important limitations. Are these implications derived logically from the findings given the study's limitations?