The Strength of Evidence Pyramid: One Approach for Characterizing the Strength of Evidence of Geoscience Education Research (GER) Community Claims

Kristen St. John[1,a] and Karen S. McNeal[2]

ABSTRACT

During the past two decades, the Geoscience Education Research (GER) community has been increasingly recognized as an evidence-based research subdiscipline in the geoscience and in the larger discipline-based education research (DBER) field. Most recently, the GER community has begun to address the current state of the field and discuss the best course forward so that it can have the greatest collective impact on advancing teaching and learning in the geosciences. The community has formally recognized that practice should be evidence based and that the strengths and limitations of community-level research claims should be transparent. As such, this commentary article describes a conceptual model—the Strength of Evidence Pyramid—as a pathway to organize the strength of evidence in the GER community of generalizable claims generated by both geo–Scholarship of Teaching and Learning and geo-DBER efforts. Its design is informed by a rubric and the outcomes of a DBER synthesis, as well as by parallels we see in the concept of evidence-based medicine in the health sciences. The proposed GER Strength of Evidence Pyramid uses five levels to categorize GER-community claims: (1) practitioner wisdom/expert opinion; original qualitative and quantitative studies, including (2) case studies and (3) cohort studies; and analyzed published literature in the form of (4) meta-analyses and (5) systematic reviews. The goal of the Pyramid is to assist geoscience-education researchers and geoscience educators to visualize, organize their thinking, and evaluate the quality of the evidence of GER-community claims. The potential applications and limitations of the model for use in the GER community are described. [DOI: 10.5408/17-264.1]

Key words: strength of evidence, Strength of Evidence Pyramid, generalizability of findings

INTRODUCTION AND BACKGROUND

How Do We Define GER?

A primary goal of geoscience education research (GER) is to improve teaching and learning in the geosciences through scholarly work and research. There are two interrelated, scholarly fields that support that goal. One is the scholarship of geoscience teaching and learning, which we refer to here as "Geo–Scholarship of Teaching and Learning (geo-SoTL)." Geo-SoTL involves the development, application, and evaluation of new geoscience teaching innovations and curricula. The other is geoscience discipline-based education research (DBER), which we refer to here as "geo-DBER." Geo-DBER involves the development and testing of questions and hypotheses in GER, which often (but not always) are motivated by the goal of improving geoscience teaching and learning. It is our observation that the GER community has generally embraced both geo-SoTL and geo-DBER as being of equal value and complimentary and reinforcing in nature. This is not necessarily the case for other science, technology, engineering, and math (STEM) education research fields (Shipley et al., 2017, this issue). Nevertheless, like all DBER,[3] geo-DBER is interdisciplinary and typically involves the use of social science methods to develop and test hypotheses (Singer et al., 2012; Lukes et al., 2015; Dolan et al., 2017; Fig. 1).

Although scholarly work on teaching and learning in the geosciences have been published for decades (e.g., *Journal of Geoscience Education* was first published in 1951), arguably, it was the publication of the Wingspread Report (Manduca et al., 2003) that first synthesized community thinking on GER and helped establish GER as an important research field that has value to both the geosciences and the social sciences. More recently, the growth and interest in GER is evident from the increase in the quality and frequency of GER articles (Pilburn, et al., 2011), the 2014 establishment of the National Association of Geoscience Teachers (NAGT) GER Division[4] (Lukes et al., 2015), an increase in the number of GER-focused graduate programs (Libarkin, 2015), the establishment of an online "home" for GER,[5] and an increase in tenure-track faculty positions at U.S. colleges and universities that support GER (St. John, 2015).

Within this GER landscape, there have been a series of recent workshops (St. John et al., 2015, 2016, 2017; Macdonald, 2016) aimed at bringing members of the GER community together to take stock of the current state of the field and to discuss the best course forward so that it can have the greatest collective impact on advancing teaching and learning in the geosciences. This *Journal of Geoscience Education* theme issue is one outcome of the 2015 workshop on Synthesizing Geoscience Education Research: Where Are We? What Is the Path Forward? This collection of articles, in

Received 20 March 2017; Revised 20 July 2017; Accepted 13 August 2017; published online 16 November 2017.

[1]Department of Geology and Environmental Science, James Madison University, Memorial Hall 7335, MSC 6903, 395 South High Street, Harrisonburg, Virginia 22807, USA

[2]Department of Geosciences, College of Sciences and Mathematics, Auburn University, 2081 Bear Eaves Coliseum, Auburn, Alabama 36849-5305, USA

[a]Author to whom correspondence should be addressed. Electronic mail: stjohnke@jmu.edu. Tel.: 540-568-6130. Fax: 540-568-8058.

[3] At present, the term *DBER* has only been applied to education research in STEM disciplines (Henderson et al., 2017), but the possibility of DBER in non-STEM disciplines also exists (Singer et al 2012).

[4] NAGT GER Division Web site: http://nagt.org/nagt/divisions/geoed/index.html.

[5] http://nagt.org/nagt/geoedresearch/index.html.

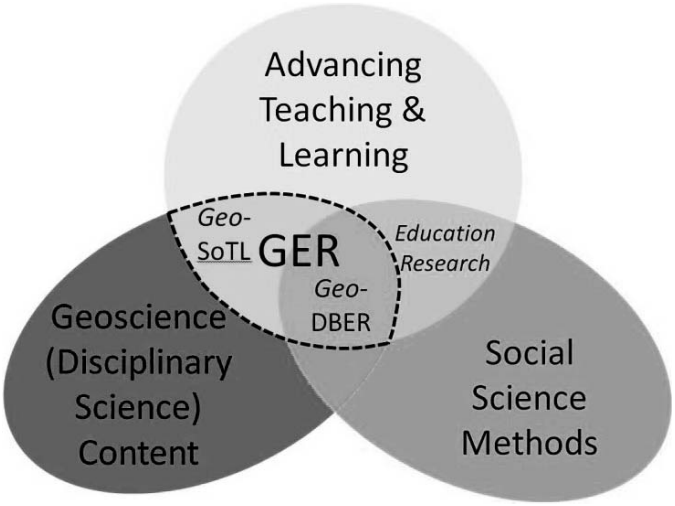

FIGURE 1: GER Venn diagram. Modified from Lukes et al. (2015).

particular the literature reviews, helps identify and articulate the current state of collective research. Another important outcome was an agreement among participants that conclusions and recommendations that emerge from the GER community and translate to practice should be evidence-based and the strengths and limitations of those community-level claims should be transparent (Macdonald, 2016). By community-level claims, we mean claims that assert to be generalizable across multiple contexts. There are two underlying assumptions to that tenet: (1) that decisions that affect teaching and student learning in the geosciences are best made when informed by evidence (including both empirical and theoretic), and (2) that there is a hierarchy to the types of GER evidence that exists, which may not be obvious to stakeholders (e.g., geoscience educators [practitioners], administrators, funding agencies).

Although the strengths and limitations of claims made in individual studies are often addressed through author attention to journal standards and peer-review feedback, GER community-level claims currently do not have a framework in which to evaluate their strength of evidence. The purpose of this commentary is to address that need by proposing a conceptual model, derived from DBER synthesis findings (Singer et al., 2012) and from a model for evidence-based medicine from the health sciences community (Glover et al., 2008), as one approach for characterizing the strength of evidence of GER community claims.

The Concept of Strength of Evidence in DBER

Our examination of the literature suggests that, although there are published rubrics (e.g., Perkins, 2004) to facilitate basic literature reviews, there are no conceptual models to serve as broader frameworks in which to situate the different types of studies (e.g., case studies, literature reviews) according to their relative strength of evidence and generalizability for the discipline-specific, education-research fields. Why this situation exists is unclear to us. This may be due to the emerging nature of DBER itself (Singer et al., 2012); perhaps such organizational exercises are undertaken only when a new research field has grown (matured and expanded) to a point at which the community members

Table I: DBER report rubric to characterize the strength of report conclusions by connecting to the evidence base. Modified from Singer et al., 2012, Box 1-1, p. 18.

DBER Report Level of Evidence	Characterized by
Limited	Few peer-reviewed studies of limited scope · with some converged of findings *or*
	· Converge with nonpeer-reviewed literature *or*
	· Convergence with practitioner wisdom.
Moderate	A well-designed study of appropriate scope that has been replicated by at least one other similar study and often including both quantitative and qualitative data *or*
	A few large-scale studies (e.g., across multiple courses, departments, or institutions) with similar results *or*
	A moderate number of small-scale studies (e.g., in single course or section) with general convergence but possibly with contradictory results. If the results are contradictory, more weight must be given to studies that reflect methodological advances or a more-current understanding of teaching and learning or those that are conducted in more-modern learning environments.
Strong	Numerous, well-designed, qualitative and/or quantitative studies, with a high convergence of findings.

"feel" the limitations of how things are being done (e.g., small-scale case study research) and seek to ask questions that require research designs that support larger-scale operation for broader generalizability of findings. Or, perhaps the need for that research is externally driven; it may depend on users or funders (i.e., educators and/or funding agencies) asking for more clarity from researchers on the strength of evidence behind the body of research conclusions and resulting recommendations.

Although rubrics for literature reviews are of a different scope and scale than conceptual models for framing the strength of evidence of community claims, we can look at the rubric and findings from the DBER report (Singer et al., 2012) for elements that may inform development of a conceptual model for GER community claims. The synthesis nature of the DBER report (Singer et al., 2012) on the status, contributions, and future direction of DBER in physics, biosciences, geosciences, chemistry, astronomy, and engineering necessitated the development of an evaluation model to qualify the conclusions and recommendations of the synthesis. That development was in the form of a three-tiered rubric of levels of evidence and offered an organizational structure for the study authors to articulate their confidence in their conclusions drawn from reviewing the DBER literature. A modified version of that rubric is shown in Table I. Important characteristics of the rubric are that the number of studies, the scope of the studies (e.g., single course versus multicourse; single institution versus multi-institution), and the convergence of findings were all deemed important factors to consider when evaluating the

Table II: Connecting outcomes from the DBER report (Singer et al., 2012) to issues important to the design and use of a conceptual model of the strength of evidence for GER-community claims.

From DBER Report (Singer et al., 2012, p 54): Challenges to DBER	Considerations for Evaluating Strength of Evidence of GER Studies and GER Community Claims
Many DBER findings are from studies in which the authors/researchers are the developers or implementers of the curriculum/instructional method/instruments. Therefore, there is the potential for bias, and it is uncommon to find independently reproduced research findings of most studies.	Are the potential biases transparent in the studies? What measures have been taken to reduce researchers' biases?
Most DBER studies are on a single course (low scale), and assessments are very course specific, making it difficult to generalize findings broadly.	Are the studies focused more on only specific courses or are they of a broader nature? Do the instruments have broader use within the GER community?
Few DBER studies focus on different subpopulations of students: (a) social/economic/ethnic diversity (b) majors versus nonmajors in introductory courses (c) structural differences among introductory courses, service courses for majors in other disciplines, and courses for majors	Are the studies addressing different subpopulations of one or more of the following: (a) social/economic/ethnic diversity (b) majors versus nonmajors in introductory courses (c) structural differences among introductory courses, service courses for majors in other disciplines, and courses for majors, (d) rural/urban differences, *and* (e) gender differences?

strength of evidence for community claims. Importantly, the breadth of research study design is also recognized in this rubric with quantitative and qualitative studies both being included and presumably having equal weight. In addition, practitioner wisdom, which draws on the knowledge and experience of classroom educators, is also a source of evidence, albeit limited according to this rubric. For all of these reasons, a simple, three-tiered rubric is an appealing starting point for a GER conceptual model for strength of evidence of community claims. However, it also contains vague terms (e.g., "well-designed," "appropriate scope") that may not be uniformly applied. In addition, the hierarchy organization (from limited to strong) and the corresponding relative differences in number (e.g., few to numerous) and scope (e.g., limited to large scale) of studies may be better served with a different visualization, one in which a conceptual diagram, as opposed to a table, could help represent the hierarchical differences more effectively.

Applying this rubric (Table II), Singer et al. (2012) were able to draw broad conclusions about the state of DBER findings (in addition to the original study, see the Mogk [n.d.] Web site summary and the Kastens and Mogk [2016] Webinar for a summary of DBER findings in the context of geosciences), based on their synthesis of results from commissioned reports (see Pilburn et al. [2011] for geoscience) and their review of the literature. The review and synthesis process also enabled Singer et al. (2012) to identify broad challenges for DBER studies going forward. We think this is important to note in the context of this commentary because the challenges not only help identify directions of future research, they highlight issues that researchers and educators need to be aware of when considering strength and applicability of research community claims. These include recognizing issues of researcher/author bias, the limits that the use of custom-designed instruments and surveys pose to generalizability of findings, and the limits that nondiverse study populations pose to the generalizability of findings. These are summarized in Table II and are paired with questions that we think should be considered in the design and use a conceptual model of strength of evidence for GER community claims.

Thinking Outside the Box: The Concept of Strength of Evidence in the Health Sciences

Another way of thinking about strategies for developing a conceptual model for strength of evidence in GER is to look at examples outside the science education research community; in particular, turning to a model for evidence-based medicine (EBM). We think EBM is a relevant model to consider because there is a large body of practitioners (in medicine and in education) who were not involved in doing the research and yet who can benefit from the findings if provided with appropriate guidance as to what findings are more trustable, based on their strength of evidence. In addition, both fields have additional stakeholders (patients in medicine and students in education) that depend on practitioners making well-informed decisions in how to apply research findings to their context. The goal of EBM is to identify and integrate the best research evidence with clinical expertise and patient values to make health-related decisions (Sackett et al., 2000). EMB emerged in early 1990s and stressed the examination of evidence from clinical research (Guyatt, 1991; Evidence-based Medicine Working Group, 1992; Guyatt et al., 1992; Montori et al., 2008). It required physicians to become more familiar with the medical literature, but also provided a model for helping physicians evaluate strength of evidence. EBM models have taken the form of tables, and more recently, as pyramids. One widely used and adapted model is by Dartmouth College and Yale University (Glover et al., 2006). Versions of this EBM model are used in the nursing (Ackley et al., 2008; Melnyk and Fineout-Overholt, 2011), primary care,[6] occupational therapy (Arbesman et al., 2008), mental health,[7] and medical library science.[8] In the Dartmouth and Yale model (Glover, 2006; Fig. 2), the width of the diagram

[6] See http://www.phcris.org.au/guides/about_research.php [PHCRIS, 2017].
[7] See http://www.dartmouth.edu/~biomed/resources.htmld/guides/ebm_psych_resources.html.
[8] See http://libguides.gwumc.edu/ebm/studytypes.

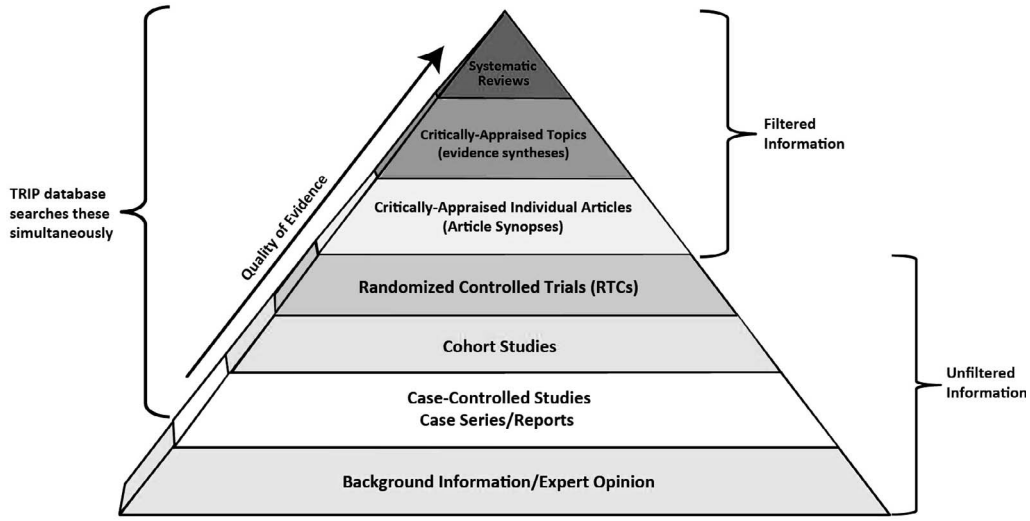

FIGURE 2: A modified version of the evidence-based medicine pyramid (originally produced by Glover et al, 2006; Dartmouth College and Yale University). Color for this figure is available in the online journal.

narrows, moving upward, to visually represent the greater number of studies that are of lower evidence (wider) to the lower number of studies that are of higher evidence (narrower). In addition, a distinction is made between filtered (analyzed published literature) and unfiltered (primary literature/original studies) information. Lastly, the EMB conceptual model also serves as the organizational starting point for launching a database for Translating Research into Practice (TRIP), to search for studies at each of the levels in the pyramid. In this way, the practitioners are immediately aware of the strength of evidence for a claim. The purpose of the conceptual diagram is to help ensure that medical practitioners can situate medical findings and provide advice in terms of its strength of evidence.

Although medical research may seem far-a-field from geoscience-education research, we see strong benefits to adopting their evidence-based research approach to designing a conceptual model of strength of evidence for GER (and other disciplines of educational research). Similar to the health sciences, research results should inform practitioners' decisions on how to best help the people they serve (medical patients and geoscience students, respectively). We also see value in organizing the model in a visual hierarchy that differentiates among different types of studies, so that the practitioners can make informed decisions based on the best evidence. In addition, we recognize that practitioners are not separate from this process; their expert opinions have value, they are an important part of the knowledge base and are where the outcomes of research need to be translated into practice. The model is not fully adaptable however, because the types of medical research do not consistently have parallels to education research. Case studies and cohort studies are common in disciplinary education research, including GER. However, based on the definition of randomized clinical trials (RCTs) by the Coalition for Evidence-Based Policy (2003), we see RCTs as relatively uncommon in GER (and DBER in general), in large part, because the control and intervention groups are typically known to the researchers, and those groups are not randomly selected; they are students in particular classes or participants in workshops, among other factors (AEA 267, 2017). This type of experimental approach is often called "quasi-experimental" in the educational arena because truly random approaches are rare. However, one setting in which some randomized experiments have been conducted is in the massive open online course (MOOC) platform, which gives an opportunity for different segments of the class population to be randomly assigned different instructional materials or prompts, and learning outcomes compared (e.g., Raffaghelli et al., 2015; Reich, 2015). We propose that a consequence of the dearth of RCTs in DBER is the increased potential for bias (an extension of the bias noted in Table II). In addition, although some articles and reports in DBER achieve critical appraisal via the publication venue and/or number of citations and are topically relevant across disciplines (e.g., Freeman, 2014; Manduca et al., 2017; Gentile et al., 2017), the volume of studies for particular topics in GER (or in DBER) is simply not great enough to develop a mechanism for, or a distinction between, synopses and syntheses, as opposed to reviews, as is done in medical research. Finally, although well-established in medical research, systematic reviews are a comparatively recent development in education (Bennett et al., 2006). Systematic reviews involve explicit criteria for selecting studies for the review, thorough coverage of studies published on the chosen review topic, and transparent measures of quality assurance by the researcher involved in the review (Petrosino et al., 2000; Bennett et al., 2006; Higgins and Green, 2011).

A PROPOSED CONCEPTUAL MODEL FOR THE GER STRENGTH OF EVIDENCE PYRAMID

The evaluation rubric used in, and outcomes from, the DBER report (Singer et al., 2012) and the Glover et al. (2006) model for evidence-based medicine from the health sciences community are the primary building block we used to design a conceptual model to characterize the strength of evidence of GER-community claims. A preliminary version of this model was presented at the 2015 GER workshop on

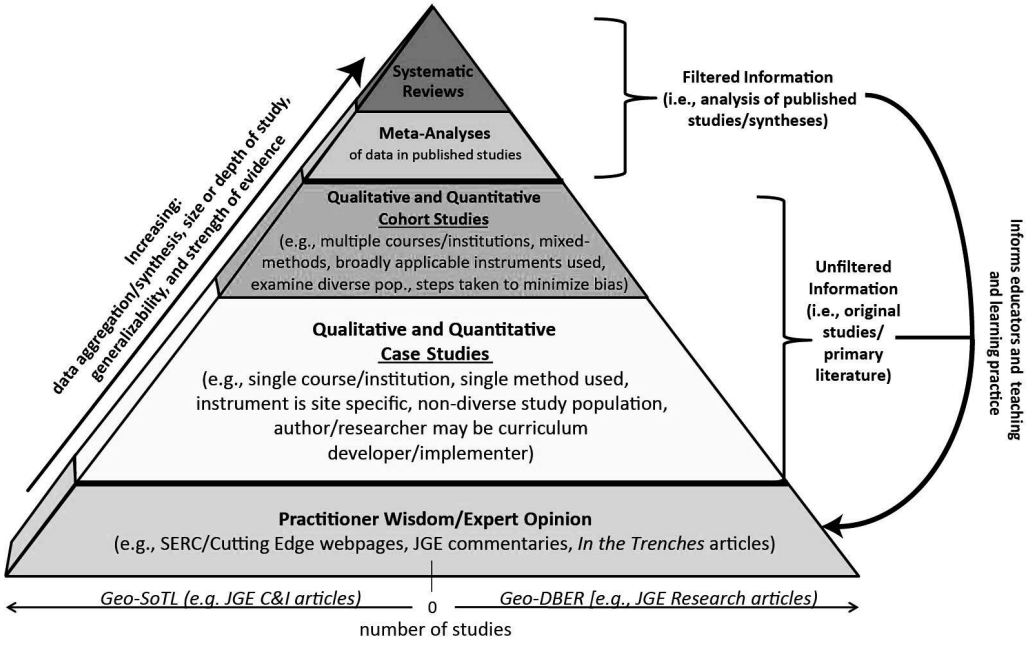

FIGURE 3: Proposed model for strength of evidence of GER-community claims. Color for this figure is available in the online journal.

Synthesizing Geoscience Education Research: Where Are We? What Is the Path Forward?[9] Feedback from workshop participants was collected and used to revise the model.

Does Shape Matter?

Like the Glover et al. (2006) EBM model, we chose the shape of the model to be a pyramid, with all labels on the front face of the pyramid (Fig. 3). We recognize that additional approaches could be taken, however, to truly make it a diagram that has a three-dimensional perspective to better illustrate the different types of studies in GER. For example, we could instead view the pyramid from the pinnacle and place research methods (e.g., qualitative, quantitative, experimental) on different sides of the pyramid. Such a design modification was proposed (Tomlin and Borgetto, 2011) for the EBM model as applied to occupational therapy. However, the traditional model (i.e., that of Glover et al., 2006) continues to dominate that profession (Podvey et al., 2013). Alternatively, the sides of the pyramid could represent different research themes (e.g., studies on students' conceptual understanding of the solid Earth; access and success in the geosciences) within GER. However, this poses problems because we would be limited to only three themes were we to retain a pyramid shape. In the end, we recognize that all models are imperfect (Box, 1979; Hayes, 2007) and chose simplicity over multidimensionality for the diagram design. We think the simple approach can convey sufficient visual structure to support understanding and community discussions on GER levels of evidence.

The width of the pyramid represents the relative number of studies (or other available resources, such as Web pages on geoscience course design[10] or GER,[11] in the case of sharing practitioner wisdom) at each level of the pyramid. Because we believe that both geo-SoTL and geo-DBER are important for improving the geoscience teaching practice, those are both included as valid lines of evidence in GER. Both are also areas for peer-review publication; for example, in the *Journal of Geoscience Education*, geo-SoTL typically results in curriculum and instruction articles, and geo-DBER typically results in research articles. The proportion of geo-SoTL or geo-DBER articles are not expected to be equal at any level of the pyramid, and it is likely that geo-DBER studies would dominate the upper levels (i.e., meta-analyses and reviews) because of the nature of the questions that would drive the analyses. However, we can envision review articles that would examine and synthesize geo-SoTL primary studies as well, for example, on how a particular learning goal is addresses and measured across the geoscience curriculum.

Moving Through the Pyramid Levels

There are five levels to this GER-community claims pyramid. The first level (green) is Practitioner Wisdom/ Expert Opinion. This knowledge base of "what we know" about GER is also the interface in which results of GER directly connect to teaching practice, and it may be a starting point for reflective practitioners to move into GER. Practitioner Wisdom recognizes that practitioners are uniquely positioned to contribute pedagogic content knowledge (PCK) to the research process (Shulman, 1986). PCK is an integration of what practitioners know about *how* to teach and *what* they teach (Cochran, 1997). We (and others[12]) see these as essential parts of the teaching process that should inform research by highlighting potential challenges, prom-

[9] 2015 GER Workshop information: http://serc.carleton.edu/earth_rendezvous/2015/morning_workshops/w3/index.html

[10] https://serc.carleton.edu/NAGTWorkshops/coursedesign/index.html.

[11] http://nagt.org/nagt/geoedresearch/toolbox/index.html.

[12] This idea was introduced by Kim Kastens at the 2017 Earth Educators Rendezvous Geoscience Education Research and Practice Forum: https://serc.carleton.edu/earth_rendezvous/2017/program/ger/index.html.

ising practices, and puzzling questions to address. In turn, it was the consensus at the 2015 GER workshop, with which we concur, that practitioner wisdom and expert knowledge are informed by GER studies at higher levels of the pyramid (depicted by the feedback loop arrow in Fig. 3), as well as personal practitioner experience (including through their own action research). Practitioner Wisdom/Expert Opinion is not, however, a level that directly involves typical scholarly peer-review publication or original research and analysis. At this level, wisdom about teaching and learning is often shared via professional develop workshops and through dissemination outlets, such as the Science Education Resource Center (SERC[13]), NAGT Teach the Earth resources,[14] and On the Cutting Edge workshops (Manduca et al., 2010) and Web pages,[15] as well as articles in the NAGT practitioner magazine *In the Trenches* and commentaries in the *Journal of Geoscience Education* (which are peer-reviewed).

The next two levels of the GER pyramid represent original qualitative and quantitative studies that are peer-reviewed and published as primary literature (and are, therefore, "unfiltered"). Most common are case studies (yellow) that focus on a single course or institution that is taught by the researcher using curriculum or instructional methods that they developed and are testing in their classes. The methods of analysis often rely on a single instrument appropriate to that site (e.g., course or institution). The population depends on the location and scope of the study but may be of limited diversity. These examples are not intended to be interdependent; that is, a case study of a single course does not require that a researcher only use a single instrument in his or her research design. The intent of the examples within the model is to convey that the scope of the case study is generally small, or the design is limited, thereby limiting the strength of evidence and generalizability of findings. Certainly, within the case study level, there are differences in how robust one study design is compared with another, and thus, an argument could be made for adding sublevels to this conceptual model to accommodate smaller-scale differences in levels of evidence (e.g., as was done in Bitting et al., 2017, this issue; and McConnell et al., 2017, this issue).

Less common, but important for determining generalizability of study findings, are cohort studies (orange). These may address some of the same research questions as case studies, but they investigate a broader cross section of courses, institutions, and/or populations. The instruments, therefore, must be broadly applicable as well, and the research design typically uses a mixed-methods approach. By increasing the depth and/or size of the study, researchers also take steps to reduce potential bias (e.g., they are not instructors of all of the courses that are testing an intervention).

The upper most levels of the GER-community claims pyramid, meta-analyses and systematic reviews, are the least common, in part, because they depend on access to data, methods, and findings from previously published research. The goal is to provide a more comprehensive description and analysis of a topic or question than could be addressed by smaller-scale case or cohort studies. These types of studies result in increased data aggregation, syntheses, and generalizability and are powerful for elevating confidence in GER-community claims of what we know and how well we know it. Both types of studies need to be done carefully to minimize bias and avoid inclusion of data from poor quality studies (e.g., weak methods used in previously published work) in their analyses (Palermo, 2012). Meta-analyses (light blue) involve application of statistical methods to look at a broad suite of existing quantitative or qualitative data. Interestingly, meta-analyses were not included in the EMB model (Glover et al., 2006), but recent arguments have been made for putting meta-analysis near the top of the EMB pyramid (Berlin and Golub, 2014).

We envision systematic reviews (dark blue) in GER as using systematic and transparent methods to identify, select, and evaluate relevant published literature on a particular topic or question (Higgins and Green, 2011). Reviews should encompass a significant time frame of study for a topical area of research. They should include the most up-to-date, as well as historical, research to provide an overview of the research evolution on that topic. Reviews of this nature can help identify patterns, trends, and gaps in GER and thereby help identify important questions and areas of needed future GER. Systematic reviews may use meta-analyses when appropriate and available.

A Snapshot or a Trajectory?

This conceptual model was initially conceived as a snapshot of the state of GER. Realistically we will always have more geoscience educators than geoscience researchers, so the wide base of Practitioner Wisdom/Expert Opinion will always be large. Small-scale, single-course or single-institution case studies are more manageable and may require less funding than multi-institutional ones, so we will likely and persistently see more case studies than cohort studies. Even if motivation and funding for multi-institutional studies increase, many researchers can implement case studies with little to no funds, so that level of the pyramid will probably stay larger than multi-institutional. The meta-analyses and reviews are fewest because they depend on a wide base of case and cohort studies. It is here, however, that we also see how the model is a trajectory—from practitioner wisdom (e.g., see Kastens and Krumhansl, 2017, this issue) and unfiltered information (case and cohort studies) to the filtered (meta-analyses and reviews)—with results ultimately feeding back into the broad base to support educators and increase expert knowledge.

APPLICATIONS AND LIMITATIONS OF THE GER STRENGTH OF EVIDENCE MODEL

Our goal in proposing the conceptual model on GER strength of evidence was to help geoscience education researchers and geoscience educators visualize, organize their thinking, and evaluate the quality of evidence within GER-community claims, which assert to be generalizable across multiple contexts in the geosciences. We envision several ways in which the model may support conversations

[13] https://serc.carleton.edu/index.html.
[14] https://serc.carleton.edu/teachearth/index.html.
[15] https://serc.carleton.edu/NAGTWorkshops/index.html.

and activities to support stronger evidence-based decisions for improving geoscience teaching and learning. We will address those below, after a cautionary note on what we see as inappropriate uses of the GER Strength of Evidence Pyramid.

Unintended and Inappropriate Uses

In considering the GER Strength of Evidence model, we must also consider what it is not well designed to address. All models have limitations (Box, 1979; Hayes, 2007) and have the potential to be misused. Although our goal was to create a model that is inclusive for all GER, that may not be possible. The model may be best matched to quantitative and mixed-method studies on geoscience teaching and learning. Studies that use a strictly qualitative research design may be harder to fit within the tiered structure because some qualitative studies are necessarily small (e.g., focus groups) yet powerful in design. We do not mean to diminish their value in the structure of this model.

The model also is best matched to GER studies that are considered applied research and use-inspired basic research, rather than being strictly basic research. Applied and use-inspired basic research is motivated by the need to solve problems, whereas pure basic research is a quest for fundamental understanding, which may be only theoretic or abstract (Stokes, 1997). Although increasing the knowledge in a field certainly has fundamental value, the feedback into practice (indicated by the arrows on the right of the pyramid in Figure 3) is a critical piece of the proposed model. That said, even pure, basic research may be able to fit into this model indirectly, with the link to practice being perhaps one or two steps removed from the original study. Basic research studies may provide a theoretic foundation for case or cohort studies, which, in turn, have direct bearing on practice. Studies of instrument development or research methodology would fit this situation. This may be analogous to the concept of the broader impact in science proposals, where the primary benefit of the research is to gain new knowledge that will be most beneficial to other researchers, but those finding have a broader reach as well, which can support new advances in teaching and learning. Additionally, some commentaries may fit only awkwardly into the model; they are a form of Practitioner Wisdom (i.e., opinions and viewpoints) but may be written about issues or problems in GER that only indirectly speak to teaching and learning. This commentary, for example, may not fit well in the pyramid because it is not directly applied to geoscience teaching and learning.

In addition, because this model was designed as a framework for considering the collective nature of geoscience education results, it should not be used to evaluate the particular effect of any specific original study in the primary literature. It cannot be used to judge whether a particular study is "good" versus "bad." It should not be used as a tiered model for evaluating the quality of work in a researcher's promotion and tenure dossier. Each of these situations have their own sets of metrics (e.g., citations, downloads, adaptation for classroom use) or rubrics (e.g., promotion and tenure institutional evaluation guidelines) that address accomplishment, value, and success. The model should not be used to judge the value of qualitative versus quantitative research methods in individual studies. Qualitative, quantitative, or mixed-method studies have no hierarchical placement in terms of strength of evidence; these are all different ways of knowing (or discovering), and their use depends on the specific questions or hypothesis being tested and the research design of a particular study. Furthermore, the GER strength of evidence model does not prioritize the use of one theoretic framework over another (e.g., constructivism, social learning theory); the learning theories that are employed to situate or construct GER questions are completely open.

Applications and Potential Next Steps

We see the most important application of the GER Strength of Evidence Pyramid as facilitating discussions on the generalizability of findings and on the strengths and limitations of claims that influence geoscience education practitioner knowledge. It can also help the community to identify gaps in GER. For example, in the initial conversations around the GER strength of evidence model at the 2015 GER workshop, it became strikingly evident that literature reviews in GER were almost absent. The need for literature reviews of GER topical areas thus became a deciding motivation to initiate this theme issue on *Synthesizing Results and Defining Future Directions in Geoscience Education Research* and spurred the development of literature-review manuscript-submission guidelines for the *Journal of Geoscience Education*.[16] A gap analysis can be particularly useful in directing research energies (and research funding) to areas of greatest need and to levels of the hierarchy (e.g., meta-analyses) that are not well represented.

Another application of the model may be to situate or contextualize GER results from different types of studies. For example, in the introduction of this theme issue the authors (McNeal et al., 2017, this issue) place each article in context of the model to visualize the nature of research collection for this theme. Similarly, contextualizing results could be done for studies included in literature reviews on a particular topic or meta-analyses to visualize where the supporting studies are situated within GER. For example, authors of literature review papers could use the model to characterize the types of studies their analysis draws from (e.g., whether they are largely case studies or cohort studies), as was done in Bitting et al. (2017, this issue) in their literature review of teaching assistant training. Bitting et al. (2017, this issue), in fact, modified the model to better serve their needs in the literature review process, finding it useful to add sublevels to categories to make finer distinctions among different types of case studies and different types of cohort studies. McConnell et al., (2017 this issue) followed a similarly modified model for their literature review of active learning strategies.

We also see a potential for the model to highlight research support needs. It may spur researchers and funders to take actions to increase support in equitable ways in order to acquire stronger evidence to make GER recommendations and to increase collective impact of the research. For example, there currently are few to no meta-analyses of GER data. Because meta-analyses address research questions that require analysis of a broad suite of existing

[16] JGE literature review guidelines, http://nagtjge.org/userimages/ContentEditor/1447705357878/JGE_Lit_Rev_Guidelines_Nov13_2015.pdf

quantitative or qualitative data, they fundamentally depend on access to data from previously published articles and reports. Aggregate data are more valuable than individual data. However, this highlights a research challenge, not only for GER but also for DBER in general: how to support the archiving and sharing of data? Other research fields have constructed ways of doing this; for example, in the geosciences, there are society or journal databases, such as the GSA Data Repository,[17] and there are government-run topical databases, such as the World Data Service for Paleoclimatology for ice core data,[18] and social science data repositories also exist (e.g., Databrary[19]) that have found functional ways to manage data in ways that are acceptable to the Human Subject Institutional Review Boards. To address the meta-analysis gap in GER requires either the creation of a data repository (for GER or for all DBER) or the development of a relationship with an existing social science data repository (Kastens and Shipley, 2016). If such a database were to be developed, it would also be useful to design it in the spirit of the health sciences' "Translating Research Into Practice (TRIP)" database (Fig. 2), enabling searches at each level of the pyramid.

Finding better ways for translating geoscience education into practice is another critical need for the GER and educator communities. Although the proposed model can facilitate conversations among researchers, we must also consider how it can be made most useful to educators because they, like medical practitioners, are at the interface of research and practice. Making the strength of evidence more clear when the GER community makes recommendations for practice is one way to support more informed decisions by educators. Sharing the GER Strength of Evidence Pyramid with geoscience educators may help address that need, but we should be open to considering other ways to make the concepts more clear and accessible. We may need to follow the medical research field's approach and develop a user guide to geoscience education literature, organized around the levels of the GER Strength of Evidence Pyramid. Doing so might involve developing a grading system for recommendations similar to that in EBM[20] (Shekelle et al., 1999) in which a "recommendation grade" is anchored in a level of evidence. In addition, in our exploration of the EBM and the Glover et al. (2006) EBM Pyramid, we were struck by how many medical library Web sites[21,22] used the pyramid as both an organizing framework and an entry point to studies for each level of the pyramid. Perhaps the GER Strength of Evidence Pyramid could be used in a similar way if embedded in online resource sites that geoscience educators and researchers already use, such as the SERC site,[23] which contains both teaching resources (e.g., On the Cutting Edge and Teach the Earth) and researcher resources (e.g., GER Toolbox). In that case, the GER Pyramid could be a front door for a collection of well-supported, generalizable community claims on "what we know" about geoscience teaching and learning, with links to recommended readings (e.g., literature reviews in particular theme areas of GER).

Lastly, several of the issues discussed above for which we see the GER Strength of Evidence Pyramid being potentially useful, are also issues for research in other discipline based fields (e.g., physics education research, chemistry education research). There is actually little in the model that is necessarily restricted to GER (perhaps only the examples of journal or SERC resources); and the origins of the model are in DBER and evidence-based medicine. Therefore, although our interest here was in developing a model for our GER community, the GER Strength of Evidence Pyramid may also serve as starting point for a model for other STEM educational research fields and for DBER in general. In this way, the GER Strength of Evidence Pyramid could be another resource that GER can contribute to advance cross-STEM DBER connections (Shipley et al., 2017, this issue).

[17] GSA Repository https://www.geosociety.org/datarepository/.
[18] World Data Service for Paleoclimatology for ice core data https://www.ncdc.noaa.gov/data-access/paleoclimatology-data/datasets/ice-core.
[19] Databrary https://nyu.databrary.org/.
[20] Levels of evidence and grades of recommendations, https://hsl.lib.umn.edu/biomed/help/levels-evidence-andgrades-recommendations
[21] University of Pittsburgh Health Sciences Library System, https://hsls.pitt.edu/resources/ebm.
[22] Walden University Library evidence-based practice research http://academicguides.waldenu.edu/healthevidence/evidencepyramid#s-lg-box-8700027,
[23] https://serc.carleton.edu/index.html.

Acknowledgments

This work has benefitted from thoughtful discussions with Heather Macdonald, Kim Kastens, Tony Feig, Laura Lukes, Eric Riggs, and participants in the 2015 GER workshop at the Earth Educators' Rendezvous. We also thank the Editor and reviewers for their constrictive comments, which improved the description and discussion of the proposed GER Strength of Evidence Pyramid. The initial idea for this conceptual model stemmed from planning activities for the 2015 GER workshop, which was funded by the National Science Foundation under grant DUE-1513519 (principal investigator, Heather Macdonald), *Shaping the Future of Geoscience Education Research: Synthesizing Results and Articulating Future Directions*.

REFERENCES

Ackley, B.J., Swan, B.A., Ladwig, G., and Tucker, S. 2008. Evidence-based nursing care guidelines: Medical-surgical interventions. St. Louis, MO: Mosby Elsevier. p. 7. Available at http://libguides.winona.edu/c.php?g=11614&p=61584#s-lg-box-179330 (accessed 20 March 2017).

AEA 267 (n.d). Area Education Agency 267. Available at https://www.aea267.k12.ia.us/system/assets/uploads/files/1469/quasi-experimental_design_2.pdf. (accessed 20 March 2017).

Arbesman, M., Scheer, J., and Lieberman, D. 2008. Using AOTA's critically appraised topic (CAT) and critically appraised paper (CAP) series to link evidence to practice. *OT Practice,* 13(5):18–22.

Bennett, J., Lubben, F., and Hogarth, S. 2006. Bringing science to life: A synthesis of the research evidence on the effects of context-based and STS approaches to science teaching. *Science Education,* 91(3):347–370. doi:10.1002/sce.20186.

Berlin, J.A., and Golub, R.M. 2014. Meta-analysis as evidence—building a better pyramid. *JAMA,* 312(96):603–605.

Bitting, K., Tasdale, R., Ryker, K. 2017. Applying the GER strength of evidence pyramid: Developing a rubric to characterize existing geoscience teaching assistant training studies. *Journal of Geoscience Education*. 65(4):519–530.

Box, G.E. 1979. Robustness in the strategy of scientific model building. *In:* Launer, R.L., and Wilkinson, G.N., eds. Robustness in statistics. New York: Academic Press.

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 National Center for Education Evaluation and Regional Assistance, NCEE EB2003. 19 p. Available at https://ies.ed.gov/ncee/pubs/evidence_based/evidence_based.asp (accessed 2 September 2017).

Cochran, K.F. 1997. Pedagogical content knowledge: Teachers' integration of subject matter, pedagogy, students, and learning environments. *Research Matters—To the Science Teacher,* No. 9702 Reston, VA: National Association for Research in Science Teaching.

Dolan, E., Elliot, S., Henderson, C., Curran-Everett, St. John, K., and Ortiz, P. 2017. Evaluating discipline-based education research for promotion and tenure. *Innovative Higher Education*, pp. 1–7. doi:10.1007/s10755-017-9406-y.

Evidence-Based Medicine Working Group. 1992. Evidence-based medicine: a new approach to teaching the practice of medicine. *Journal of the American Medical Association,* 268(17):2420–2425.

Freeman, S., Eddy, S., McDonough, M., Smith, M.K., Okoroafor, N., Jordt, H., and Wedderoth, M.P. 2014. Active learning increases student performance in science, engineering, and mathematics. *PNAS,* 111(23):8410–8415. doi:10.1073/pnas.1319030111.

Gentile, J., Brenner, K., and Stephens, A.; National Academies of Sciences, Engineering, and Medicine, eds. 2017. Undergraduate research experiences for STEM students: Successes, challenges, and opportunities. Washington, DC: The National Academies Press. doi: 10.17226/24622.

Glover, J., Izzo, D., Odato, K., and Wang, L. 2006. EBM Pyramid and EBM Page Generator, Trustees of Dartmouth College and Yale University. Available at https://www.dartmouth.edu/~biomed/resources.htmld/guides/ebm_resources.shtml (accessed 2 September 2017).

Guyatt G. Evidence-based medicine [editorial]. 1991. *ACP J Club,* 114:A16. doi:10.7326/ACPJC-1991-114-2-A16.

Guyatt, G., Cairns, J., and Churchill, D. 1992. Evidence-based medicine: A new approach to teaching the practice of medicine. *JAMA,* 268(17):2420–2425. doi:10.1001/jama.1992.03490170092032.

Hayes, R.B. 2007. Of studies, syntheses, synopses, summaries, and systems: the "5S" evolution of information services for evidence-based healthcare decisions. *Evidence-based Nursing,* 10:6–7.

Henderson, C., Connolly, M., Dolan, E., Finkelstein, N., Franklin, S., Malcolm, S., Rasmussen, C., Redd, K., and St. John, K. 2017. Towards the STEM DBER alliance: Why we need a discipline-based STEM education research community. *Journal of Engineering Education,* 106(3):349–355.

Higgins, J.P.T., and Green, S., eds., 2011. Cochrane handbook for systematic reviews of interventions (version 5.1). The Cochrane Collaboration. Available at http://training.cochrane.org/handbook (accessed 2 September 2017)

Kastens, K. 2017. A Community of practice for GER. Available at http://nagt.org/nagt/geoedresearch/toolbox/basics/CoP.html (accessed 7 March 2017).

Kastens, K., and Krumhansl, R., 2017. Identifying curriculum design patterns as a strategy for focusing geoscience education research: A proof of a concept based on teaching and learning with geoscience data. *Journal of Geoscience Education*, 65(4):373–392.

Kastens, K., and Mogk, D., 2016. Discipline-based education research (DBER) and geoscience [Webinar]. Available at http://nagt.org/nagt/profdev/workshops/geoed_research/dber_webinar.html (accessed 7 March 2017).

Kastens, K., and Shipley, T., 2016. Exploring options for archiving, disseminating, and sharing data in geoscience education research, presentation at the geoscience education research community planning workshop; Madison, WI. Available at http://serc.carleton.edu/earth_rendezvous/2016/program/morning_workshops/w3/program.html. (accessed 20 March 2017).

Libarkin, J. 2015. Alphabetical list of graduate programs in geocognition and geoscience education research. Available at https://geocognitionresearchlaboratory.wordpress.com/graduate-study/geocognition-geoscience-education-research-programs/ (accessed 7 March 2017).

Lukes, L., LaDue, N., Cheek, K., Ryker, K., and St. John, K. 2015. Creating a community of practice around geoscience education research: NAGT-GER [Abstract]. *Journal of Geoscience Education,* 63(1):1–6. doi:10.5408/1089-9995-63.1.1.

Macdonald, H. 2016. Meeting report—First steps toward synthesizing geoscience education research: Notes from workshop discussions. Available at https://serc.carleton.edu/earth_rendezvous/2015/morning_workshops/w3/program.html (accessed 2 September 2017).

Manduca, C.A., Iverson, E.R., Luxenberg, M., Macdonald, R.H., McConnell, D.A., Mogk, D.W., and Tewksbury, B.J., 2017. Improving undergraduate STEM education: The efficacy of discipline-based professional development. *Science Advances,* 3(2):e1600193.

Manduca, C.A., Mogk, D.W., Stillings, N. 2003. Bringing research on learning to the geosciences: Report from a workshop sponsored by the National Science Foundation and the Johnson Foundation. Cincinnati, OH: Johnson Foundation. p. 36.

Manduca, C.A., Mogk, D.W., Tewksbury, B., Macdonald, R.H., Fox, S.P., Iverson, E.R., Kirk, K., McDaris, J., Ormand, C., and Bruckner, M. 2010. On the cutting edge: Teaching help for geoscience faculty. *Science,* 327(5969):1095–1096.

McConnell, D.A., Chapman, L., Czajka, C.D., Jones, J.P., Ryker, K.D., Wiggens, J. 2017. Instructional utility and learning efficacy of common active learning strategies. *Journal of Geoscience Education,* 65(4):604–625.

McNeal, K.S., and Petcovic, H.L. 2017. parking conversations about graduate programs in geoscience education research. *Journal of Geoscience Education,* 65(4):399–406.

Melnyk, B.M., and Fineout-Overholt, E. 2011. Evidence-based practice in nursing and healthcare: A guide to best practice. Philadelphia, PA: Lippincott, Williams & Wilkins. Available at http://guides.lib.umich.edu/c.php?g=282802&p=1888246 (accessed 2 September 2017).

Mogk, D. [n.d.]. Discipline-based education research (DBER) understanding and improving learning in undergraduate science and engineering—contributions and opportunities for the geosciences. Available at http://serc.carleton.edu/NAGTWorkshops/DBER.html (accessed 7 March 2017).

Montori, V.M., and Guyatt, G.H. 2008. Progress in evidence-based medicine. *JAMA,* 300(15):1814–1816. doi:10.1001/jama.300.15.1814.

Palermo, T. 2013. New guidelines for publishing review articles in JPP: Systematic reviews and topical reviews. *Journal of Pediatric Psychology,* 38(1):5–9. doi:10.1093/jpepsy/jss124.

Perkins, D. 2004. Scholarship of teaching and learning, assessment, and the Journal of Geoscience Education. *Journal of Geoscience Education,* 52(2):113–114.

Petrosino, A., Roruch, R., Rounding, C., McDonald, S., and Chalmers, I. 2000. The Campbell collaboration social, psychological, educational and criminological trials register (*C2-SPECTR*) to facilitate the preparation and maintenance of systematic reviews of social and educational interventions. *Evaluation and Research in Education,* 14(3–4):206–219.

Pilburn, M.D., van der Hoeven Kraft, K., and Pacheco, H. 2011. A new century for geoscience education research—A commissioned study to inform the 2012 DBER Report. Washington, DC: Board on Science Education, The National Academies of Sciences, Engineering, and Medicine. p. 24.

Podvey, M.C., Hoover, K.D., and Henderson, K.S. 2013. Evaluating

evidence in occupational therapy and applied behavior science. Extended abstract/white paper from a poster at the 93rd Annual AOTA Conference and Expo. Available at http://files.abstractsonline.com/CTRL/E7/4/80A/F68/A8F/434/BAA/7BE/FC5/E66/620/C9/a901_1.pdf (accessed 2 September 2017).

Primary Health Care Research & Information Service. 2017. PHCRIS getting started guides: Introduction to... Different research models. Available at http://www.phcris.org.au/guides/about_research.php (accessed 2 September 2017).

Raffaghelli, J.E., Cucchiara, S., and Persico, D. 2015. Methodological approaches in MOOC research: Retracing the myth of Proteus. *British Journal of Education Technology,* 46(3);488–509.

Reich, J. 2015. Rebooting MOOC research: Improve assessment, data sharing, and experimental design. *Science,* 347(6217):34–35.

Sackett D.L., Straus S.E., Richardson W.S., Rosenberg, W., Haynes, R.B. 2000. Evidence-based medicine: How to practice and teach EBM. 2nd ed. Edinburgh, UK: Churchill Livingstone.

Shekelle, P.G., Woolf, S.H., Eccles, M., and Grimshaw, J. 1999. Developing clinical guidelines. *Western Journal of Medicine,* 170(6);348–351.

Shipley, T.F., McConnell, D., McNeal., K.S., Petcovic, H.L., and St. John, K.E. 2017. Transdisciplinary science education research and practice: Opportunities for GER in a developing STEM discipline based education research alliance (DBER-A). *Journal of Geoscience Education,* 65(4):355–362.

Shulman, L.S. 1986. Those who understand: Knowledge growth in teaching. *Educational Researcher,* 15:4–14.

Singer, S., Nielsen, N., and Schweingruber, H., eds. 2012. Discipline-based education research: Understanding and improving learning in undergraduate science and engineering. committee on the status, contributions, and future directions of discipline-based education research; board on science education; division of behavioral and social sciences and education. Washington, DC: National Academies Press, p. 282.

St. John, K. 2015. Is there a better model for promotion and tenure preparation and evaluation of geoscience education researchers in geoscience departments? *Journal of Geoscience Education,* 63(4):265–267. http://dx.doi.org/10.5408/1089-9995-63.4.265.

St. John, K., Cervato, C., Kastens, K., Macdonald, H., McDaris, J., McNeal., K., Petcovic, H., Pyle, E., Riggs, E., Ryker, K., Semken, S., and Teasdale, R. 2017. Identifying and prioritizing geoscience education research grand challenges: Draft plans for a community research agenda. *Geological Society of America Abstracts With Programs,* 49(6):259-5.

St. John, K., Kastens, K., Macdonald, H., McNeal, K., McDaris, J. 2016. Emerging priorities and new online resources to support geoscience education researchers. *Geological Society of America Abstracts with Programs*, 48(7). doi: 10.1130/abs/2016AM-283820.

St. John, K., Macdonald, H., Feig, A., LaDue, N., Lukes, L., McNeal, K., Riggs, E., and McDaris, J. 2015. Shaping the future of geoscience education research: a community effort [abstract 68-10]. *Geological Society of America Abstracts with Programs,* 47(7):255.

Stokes, D.E. 1997. Pasteur's quadrant—Basic science and technology innovation. Brookings Institution Press, pp. 196.

Tomlin, G., and Borgetto, B. 2011. Research pyramid: A new evidence-based practice model for occupational therapy. *American Journal of Occupational Therapy,* 65:189–196. doi:10.5014/ajot.2011.000828.