Defining a Literature

by Mary M. Kennedy

As scholars and their audiences pursue standards of evidence, standards for literature reviews have also become salient. Many authors advocate "systematic" reviews and articulate standards for these. This article compares the bodies of literature derived from systematic and other types of review, which the author labels conceptual, and examines problems associated with different approaches to defining a body of literature. These problems include (a) defining the boundaries of the literature, (b) distinguishing studies from citations, (c) distinguishing literature from lore, (d) deciding which reporting venues to include, and (e) weeding out anomalous studies. The article demonstrates that although systematic reviews may remove some biases through their inclusion rules, they may introduce other biases through their exclusion decisions and may thwart conceptual advances in a field.

Keywords: literature review; meta-analysis; synthesis; teacher qualifications

Although the literature review is a widely recognized genre of scholarly writing, there is no clear understanding of what constitutes a body of literature. Each reviewer must decide which specific studies to include or exclude from a review and why. And each such decision alters the character of the set as a whole and could also therefore alter the net conclusions drawn from the set. In this article, I examine a number of examples of inclusion and exclusion decisions and illustrate how they affect the resulting bodies of literature. For purposes of illustration, I draw on literature examining the relationship between teachers' qualifications and the quality of their teaching practice.

Questions about inclusion have become more salient with recent advocacy for a particular type of literature review, often called "systematic." A systematic review typically focuses on a very specific empirical question, often posed in a cause-and-effect form, such as "To what extent does A contribute to B?" The term *systematic* (Centre for Reviews and Dissemination, 2001; Cooper, 1984; Evidence for Policy and Practice Information Centre, 2005) means that the authors have defined the research question as clearly and specifically as possible and have made a concerted effort to ensure that they have found all evidence relevant to that question. An example of a systematic review of literature on qualifications would be Kennedy, Ahn, and Choi's (in press) review of the effects of teachers' college course work on their current students' achievement in mathematics. Both the question and the rules of inclusion are laid out in detail, and literature is sought not just from journals but also from dissertations, conference presentations, and independent reports.

Advocates of systematic reviews tend to label all other reviews as *nonsystematic*, a term that implies deficiency. But there are many other approaches to literature reviews, and each makes its own contribution to the field. The American Educational Research Association (2006) lists the following as eligible for publication in the *Review of Educational Research*: integrative reviews, theoretical reviews, methodological reviews, and historical reviews. To simplify my discussion, I group all of these under a general heading of "conceptual reviews," meaning that these approaches share an interest in gaining new insights into an issue. For example, when Ball, Lubienski, and Mewborn (2001) reviewed literature on the role of teachers' mathematical knowledge in teaching, they did not ask what we know, empirically, about the problem but asked instead why we don't know more, how people have thought about the problem in the past, and what other issues are intertwined with this one. Similarly, in the policy arena, when Goldhaber and Anthony (2003) reviewed literature on teacher qualifications and student achievement, they did not ask which qualification had the greatest impact but instead tried to think aloud about the nature of the question. They examined problems with research methods, problems with how teachers' qualifications are distributed among student populations, and also research findings about whether these qualifications matter. This article helps educate readers about a variety of interwoven and complicated problems of definition, the hiring and allocation of resources, and the research methods used to sort all this out.

One reason people worry about the purposes and rigor of reviews is that reviews are sometimes conducted in the service of arguments. Given the value-laden nature of educational decision making, arguments can be a legitimate and important form of discourse, but they can also introduce problems for researchers when they are presented as if they are either conceptual or systematic reviews. In fact, arguments have characteristics borrowed from both other forms. An argument may use literature selectively, as a conceptual review might, but it might also make summary claims of the sort found in systematic reviews. Some arguments, especially commissioned reports such as that by the National Commission on Teaching and America's Future (1996), are presented explicitly as arguments, but others appear in scholarly journals, in which their argumentative intention may be less obvious.

Still, even systematic reviews require nontrivial judgments as researchers try to define the boundaries of their research questions

Educational Researcher, Vol. 36, No. 3, pp. 139–147
DOI: 10.3102/0013189X07299197

and their standards for acceptable literature. Often, when readers approach a systematic review, they may be unaware that micro-level decisions have influenced the composition of the literature as a whole. This article reviews a collection of problems associated with all forms of review: (a) defining the boundaries of the literature, (b) distinguishing studies from citations, (c) distinguishing literature from lore, (d) deciding which reporting venues to include, and (e) weeding out anomalous studies.

Defining the Boundaries of a Body of Literature

The literature I use for this examination comes from a literature database called the Teacher Qualifications and the Quality of Teaching (TQQT) database. Because my colleagues and I anticipated publishing systematic reviews of this literature, we tried to be thorough in our search for relevant studies, to define explicit rules for searching, and to define rules for what would be included or excluded from our body of literature. Here are the rules we developed:

1. Each study must include at least one teacher qualification and at least one indicator of the quality of teaching, and it must demonstrate a link between the two.
2. Each study must take place in the context of K–12 schools in the United States and focus on the teachers of record. (This rule meant that we excluded studies of student teachers, preschool teachers, teachers of college students and adults, and teachers working outside the United States.)
3. Each study must have been published no earlier than 1960.

Literature was obtained by searching the Education Resources Information Center (ERIC), PsycINFO, *Dissertation Abstracts International*, and EconLit. Search terms included those commonly used to refer to qualifications, such as *assessment*, *certification*, *teacher education*, *teacher effectiveness*, and so forth. In addition, we searched the bibliographies of numerous other literature reviews and policy analyses in this area and searched entire journals whose domains encompassed this general topic. Studies were screened to ensure that they met our rules. As of this writing, the database included 465 records. Although these procedures and rules seem straightforward, many complications arose during the search, most having to do with what constitutes a qualification or an indicator of quality.

What Is a Qualification?

Our original search criteria defined qualifications to include mainly things having to do with educational background and credentials: things teachers earned prior to seeking teaching positions. We excluded assessments of beliefs, values, personality traits, or other personal variables, with an eye toward concentrating on variables that states or districts are most likely to incorporate into their policies. We included not only obvious qualifications, such as credentials, test scores, and degrees, but also such things as the particular courses teachers took, their grade point averages, the status of the institutions they attended, and related indicators of their educational backgrounds that might be relevant to local hiring decisions. However, during our search, we discovered that a number of other qualifications were of interest to school districts that hire teachers. One is National Board for Professional Teaching Standards certification, which is not something earned prior to teaching but is nonetheless relevant when hiring experienced teachers. In addition, we recognized that years of teaching experience or the possession of an advanced degree is relevant to such hiring decisions. Consequently, we expanded our list of qualifications to include things that are typically acquired after teachers obtain full-time teaching positions but would be relevant to selecting experienced teachers. Then we discovered that many school districts use commercial screening systems, such as the Teacher Perceiver Interview or the Star Teacher Interview, as part of their selection process. These screens are intended to help districts select teachers with the most desirable beliefs, attitudes, and values, and they are not required by state policies. These systems reflect interest in a type of qualification that had not been considered in any other reviews of literature on teacher qualifications, yet thousands of districts had subscribed to such selection systems, so we added studies of them to our database.

What Is an Indicator of the Quality of Teaching?

We began with the notion that teaching quality could be inferred from classroom observations, student test scores, principals' ratings, and artifacts from instruction, such as assignments. However, a number of decisions had to be made to clarify these rules. There were studies, for instance, that assessed teachers' practices outside their regular classrooms, in artificial settings, or that assessed them using artificial tasks (e.g., Popham, 1971). We decided to exclude these and to concentrate on studies that examined the quality of the practice that occurs during regular classroom teaching, as teachers do their assigned work. We also found some studies that assessed student gains over a 1- or 2-week period. We decided to eliminate these studies as well, on the grounds that they do not necessarily generalize to the task of teaching for an entire academic year.

Can a Single Measure Be a Qualification as Well as an Indicator of Quality?

One of the most complicated problems arising from this literature search was the distinction between "TQs" (teacher qualifications) and "QT" (the quality of teaching). In the abstract, the difference between a qualification and an indicator of quality seems straightforward. Qualifications such as credentials and test scores are granted to teachers outside the classroom, whereas indicators of quality emanate from classroom practice itself. But many studies examine relationships between educational backgrounds and test scores and speak of test scores as if these are indicators of quality. Could a test score be a qualification as well as an indicator of quality? We decided not and so excluded these studies on the grounds that they are actually TQ-TQ studies, not TQ-QT studies. A similar complication occurred on the other side, because we found numerous studies that looked at the relationship between observed teaching practice and student achievement, both of which we considered to be indicators of the quality of teaching practice. Can observed practice be considered both a qualification and an indicator of quality? Again, we decided not. Hence, these studies were considered QT-QT studies and were eliminated from our literature database.

These decisions matter. It is essential in systematic reviews that reviewers define what is relevant to their questions and then ensure that all relevant studies and no irrelevant studies are incorporated into the reviews. We defined a relevant study as one that included at least one qualification and at least one indicator of quality. But

the sequence of events leading to teaching actually looks like this: Teachers first get educated, then earn test scores, then engage in teaching practices, and then influence students' achievement. So a study that looks at any pair of sequentially linked events is generally relevant. Viewed in this way, hundreds of additional studies could be construed as relevant to our question and would be of interest to many people. Our inclusion rules defined the issue in a particular way, but not in the only way it could have been defined.

Distinguishing Citations From Studies

When examining a literature database, it would be a mistake to assume that citations and studies are coterminous. We found some cases in which one citation described multiple studies and some cases in which multiple citations described the same study. In the first case, we found nine citations that described 2 studies each, one that described 3 studies, and one that described 4 studies. These citations created 15 additional studies that were not reflected in our citation count, but each required a separate record in our database so that each could be uniquely characterized.

On the other side, we have multiple citations describing the same study. Because our search included dissertations, conference presentations, and reports as well as journal articles, we might expect some redundancy. But in an empirical summary of literature, one wants to ensure that each finding is counted only once. This ensures that the summary reflects the volume of evidence on the issue rather than the productivity of authors, and it also ensures that the effects we examine are statistically independent of one another.

In our literature database, we found 19 pairs of citations referring to the same studies. We also found five instances in which 3 citations described the same study, four instances in which 4 citations described the same study, and one in which 5 citations described the same study. So deriving the number of studies requires us to start with our 450 citations, add the 15 additional studies that were reported by authors who described more than 1 study (for a total of 465 records), and then subtract 45 studies that were presented in redundant citations, leaving a true count of 420 discrete studies.

Distinguishing Literature from Lore

What we consider to be "knowledge" in a field usually extends well beyond the formal literature, though most of it probably derives in some way from that literature. We may think of knowledge in a given field as consisting of three layers. First, there are the primary studies that researchers conduct and publish. Next, there are reviews of those studies, whether systematic or conceptual, that provide summaries and new interpretations built from but often extending beyond the original literature. Finally, there are the perceptions, conclusions, and interpretations that people share in informal hallway conversations that become part of the lore of the field. This third layer is the one most scholars actually believe to be true, but it can have a relatively loose relationship to the primary studies and even to the literature reviews.

Each of these layers offers a unique portrait of "the literature." The first is the most complete and thorough rendition but also the most splintered and incoherent. The second and third render the first more coherent by skipping over some studies, emphasizing others, and reinterpreting still others. Each field of study has its own lore about what "the literature" shows, and because most of us are too busy to carefully examine entire bodies of literature, we accept secondhand summaries, both in print and in the hallway, a practice that allows each field to generate a particular lore that may or may not match the full scope of primary literature that presumably underlies it. In the lore, some studies get magnified over time and others recede from view.

One way lore is created is through secondhand citations. A particular study may be cited by one author and then re-cited by other authors who never read the original piece but instead are basing their citations on the first author's description of the study. We found evidence of this phenomenon in our search. We found cases in which particular articles were miscited in the same way by multiple authors, a pattern suggesting that one author's mistake was copied by several others.

When reviewers are not systematic in their coverage, the gap between the primary studies and the lore can grow larger. We found evidence of this phenomenon when, at the time we were engaging in our search, another researcher (Walsh, 2001) published a literature review examining the same topic. But Walsh did not set out to conduct a systematic review. Instead, her aim was to examine the literature that had become lore: the literature that was used by advocates in this arena to justify their arguments about the merits of teacher education or certification. She was concerned about the quality of the literature that had been used to make these arguments. That she has gathered this literature together enables us to ask whether "the literature" that contributes to the lore is different from "the literature" that arises from a systematic search of primary studies. Table 1 shows the number of studies unique to each body of literature and the number that appeared in both.

The first thing Table 1 reveals is that the TQQT database is substantially larger than Walsh's (2001) database, with 465 citations compared with 200. This difference suggests that systematic searches are indeed capable of finding many more studies than are normally cited by authors who rely on less systematic procedures. However, Table 1 also suggests that there are substantial differences between the two lists. Only 83 studies appeared in both lists. Of Walsh's 200 citations, 117 were not included in the TQQT database, and of the 465 records in the TQQT database, 382 were missing from Walsh's database.

Understanding the reasons for these differences is important, for the differences are likely to influence any summary judgments made about "the literature" as a whole. On one side, when authors claim to be reviewing a literature but have not done so systematically, they can increase the gap between the lore of the field and its actual corpus of primary studies. On the other side, if Walsh's (2001) authors reviewed things that our systematic search missed, we may wonder about the adequacy of even "systematic" review procedures. To see what might have accounted for these differences, I examined the two off-diagonals more closely.

The TQQT literature database includes 382 studies that did not appear in Walsh's (2001) database. Table 2 summarizes the characteristics of these studies, focusing on characteristics that might account for the fact that they have apparently been omitted from the lore as it is represented in policy arguments. The first row of Table 2 contains studies published since 2000, the year before Walsh's review was published. These are in the TQQT database because that database is still growing but are not in Walsh's because her compilation is finished.

Table 1
Overlap Between Walsh's (2001) Literature and TQQT Literature

	In TQQT Database		
In Walsh's Database	**Yes**	**No**	**Sum**
Yes	83	117	200
No	382	0	382
Sum	465	117	582

Note. TQQT = Teacher Qualifications and the Quality of Teaching.

Table 2
Characteristics of Studies in the TQQT Database but Not in Walsh's (2001) Database

Number of Studies	Possible Reason for Omission From Walsh's Review
82	Studies published in 2000 or later
32	Studies of the Teacher Perceiver Interview
84	Salary variables
165	Dissertations
19	Qualitative studies
382	

Note. TQQT = Teacher Qualifications and the Quality of Teaching.

Table 3
Characteristics of Studies in Walsh's (2001) Database but Not in the TQQT Database

Number of Studies	Reason for Omission From the TQQT Database
56	Rejected as literature reviews
26	Rejected, no link explicated
4	Rejected, no acceptable teacher qualifications
4	Rejected, no acceptable indicators of the quality of teaching
8	Rejected, studies of preservice study
3	Rejected, publication dates were too early
1	Rejected, study of an inservice program
1	Rejected, study of early childhood teachers
2	Surfaced (i.e., citations were found) but could not be located
11	Did not surface, were not published, and were not listed in the Education Resources Information Center database
1	Did not surface, yet was a published article[a]
117	

Note. TQQT = Teacher Qualifications and the Quality of Teaching.
a. On discovering this article in Walsh's database, we looked it up and determined that it was a policy analysis, not an empirical study. Hence, it was rejected from the TQQT database.

The remaining rows reveal the importance of inclusion rules in the definition of a literature. For example, the second row refers to 32 studies published before 2000 that evaluated the predictive validity of commercial hiring interviews (for a review of this literature, see Metzger & Wu, 2003). The Teacher Perceiver Interview is rarely considered a "qualification" in the lore of the literature, even though it is widely used by school districts as part of their hiring procedures. The third row represents qualifications that are acquired after initial hiring decisions, such as years of experience, the possession of an advanced degree, or National Board for Professional Teaching Standards certification. These are variables we had originally planned to exclude and later included because of widespread interest in them. Because the literature Walsh (2001) examined had mostly to do with initial certification, it should not be surprising that it did not include these salary-relevant variables.

The remaining rows in Table 2 may reveal potential biases in the lore of the literature. These rows have less to do with which kinds of teacher qualifications to include and more to do with which kinds of studies to include. These rows suggest that the lore of the literature has largely ignored dissertations and has also overlooked almost all of the qualitative research that has been done in this area. To the extent that these studies yield different findings from other studies, the lore of the literature will be biased and will not reflect the full scope of primary studies in this field.

Now consider the other side: Why did the TQQT search miss so many of the citations included in Walsh's (2001) review? Our systematic review was built not only from searches of extant databases such as ERIC and PsycINFO but also from the reference lists of other literature reviews, so it should have found everything Walsh found. Yet there were 117 studies included in Walsh's synthesis that were not in the TQQT literature database. Why?

In fact, virtually all of these studies had been found using the TQQT search procedures but were later rejected. Table 3 shows the number of studies found and rejected for different reasons, and, in the bottom two rows, studies that had not been found.

The first two rows reveal one important feature of "literatures" that needs more attention. In both of these rows are studies that were cited by policy advocates but rejected from the TQQT literature database on the grounds that they did not provide direct evidence on the question. The first row reveals one big difference between what is allowable in arguments or conceptual reviews compared with systematic reviews: In a systematic review, no study can be counted more than once (Dunkin, 1996). Therefore, other literature reviews are generally eliminated from systematic reviews. But arguments and conceptual review may freely include such reviews and discuss their arguments and interpretations. Table 3 shows that 56 articles that appeared in Walsh's (2001) list, but not in the TQQT list, were literature reviews that had been captured and then rejected from the TQQT database.

The second row reveals another difference between systematic reviews and others. Because a systematic review aims to summarize empirical outcomes, it requires that all studies provide explicit evidence linking one variable to another. In our case, we asked that each study include at least one qualification, at least one indicator of quality, and a link between the two. This link could be established either by group comparison, by correlational techniques, or qualitatively. We excluded many studies that described the distribution of teachers' qualifications across different student populations because they did not show how these qualifications were linked to

any indicator of quality. Yet clearly, the question of how teachers with various qualifications are distributed is relevant. So it should not be a surprise that 26 articles were present in Walsh's (2001) literature but eliminated from the TQQT literature because they did not provide explicit links between qualifications and indicators of quality, nor that another eight citations in Walsh's database were eliminated from the TQQT database because they lacked either a TQ or a QT measure. These differences do not necessarily mean that one set of inclusion rules is better than another but rather that literatures gathered for different purposes can be quite different. Because the authors whose references Walsh gathered were engaged in conceptual reviews and in advocacy arguments, they could freely draw on related literature that was relevant to their broader conceptual aims but that would not be relevant to a systematic review that focused on a clearly defined, but narrower, question.

Remaining differences also have to do with inclusion rules. For instance, one of the studies in Walsh's (2001) database (Lutz & Hutton, 1989), used teachers' job satisfaction as an outcome measure. The study was rejected from the TQQT database because we did not consider job satisfaction to be an indicator of teacher quality. Another article (Sanders & Horn, 1998) has been widely cited as demonstrating that teachers vary substantially in their effectiveness, and it was apparently cited by at least one of the authors surveyed by Walsh. However, Sanders and Horn's study did not address the question of which teacher qualifications are related to differences in effectiveness. The authors merely established that there are differences among teachers in student outcomes. So the study was eliminated from the TQQT database for want of a valid TQ, but it is certainly relevant to the general issue of teacher quality and thus reasonable to cite by one of the reviewers surveyed by Walsh.

Perhaps the most important difference revealed in Table 3 is that 11 of the studies referred to by Walsh's (2001) authors were not published and were also not included in available literature databases such as ERIC or EconLit. They may have been presented at conferences 15 years ago, and their authors still had copies. But their presence in these reference lists suggests that advocates may be building their arguments on evidence that is not accessible, even to an earnest reader who makes an effort to further pursue the issue. Because readers have no access to these primary studies, they must either accept their relevance and quality as part of the lore of the literature or discount them because of their inaccessibility. That such studies can become part of our lore when other, more accessible studies such as dissertations are not, illustrates the potential for bias in the lore.

If Walsh's (2001) database captures the lore of the literature, whereas the TQQT database captures the full panoply of primary studies, this comparison suggests that there may indeed be important differences between the sum of primary research available on an issue and the lore about that issue. In this case, the lore does not reflect dissertations, which are difficult to search through and expensive to obtain, but does reflect sources that are not, in fact, part of a publicly available original literature. This is an important feature of lore.

Deciding Which Reporting Venues to Include

Some reviewers, in an effort to be more systematic, define the boundaries of their searches to include only articles that have been published in peer-reviewed journals (e.g., Wilson, Floden, & Ferrini-Mundy, 2001). The assumption underlying this inclusion rule is that because these articles have been reviewed by peers, they are more likely to meet minimum standards of quality. However, it is not clear that peer-reviewed journals necessarily provide higher quality evidence, especially in the field of education, which differs from other fields of scholarship in two important ways. First, because education is a cultural enterprise, a wide spectrum of the public may participate not just in the debates but also in the production of knowledge. Second, because the field of academic education (i.e., university and college departments) is so large, a large number of journals have been created to accommodate the demand for publication outlets. These factors increase the likelihood that non-peer-reviewed studies may nevertheless be of high quality and that even peer-reviewed journal articles may have quite variable quality.

Of course, study quality itself is also an arguable phenomenon; researchers argue over the merits of virtually every aspect of study methodology.[1] Methodology is an especially important (and contentious) issue in the literature on teacher qualifications because teachers are not randomly assigned either to their qualifications or to their jobs, and students are not randomly assigned to teachers except in rare cases. In fact, studies of how teachers are allocated to students suggest that students and teachers are roughly matched according to their qualifications, with relatively less prepared students receiving relatively less prepared teachers and vice versa (Kain & Singleton, 1996; Lankford, Loeb, & Wykoff, 2002). To the extent that teachers and students are systematically matched, one could expect to see a correlation between teachers' qualifications and their students' achievement, even if teachers' qualifications had no causal effect on students' achievement. In this situation, study quality is an important factor in evaluating findings.

To see if study quality in journals is superior to that of studies published elsewhere, I developed two quality criteria. One is that studies should use pretests as a way to eliminate prior differences in achievement and to control for possible allocation biases. This is a very rough criterion and does not stipulate how the pretests are used. Pretest scores could be used to form groups of students, to create gain scores, or to contribute to regression equations. Another minimum criterion for study quality is that studies examine teachers as individual units rather than relying on school or district averages. The argument here is that institutional averages mask important differences among teachers within institutions and confound teachers' qualifications with other differences among institutions, thus making interpretations of findings more difficult. These two criteria offer very minimal standards for study quality and leave open the possibility of myriad other important differences in quality. But they provide a way of asking whether journals publish higher quality studies than other reporting venues.

Table 4 shows the percentage of citations appearing in different reporting venues that used student achievement as their outcome measures and that met either of these two criteria for study quality. The reporting venues appear in order of the percentage of citations that met these minimal quality criteria. Thus, the lowest percentages of studies meeting these criteria appear in the leftmost column, which encompasses conference presentations, and the largest percentages are in the rightmost column, which encompasses independent reports and electronic sources. This latter group consists

Table 4
Proportion of Studies Using Student Achievement That Also Met Specific Study Quality Standards

Criterion for Quality of Study	Reporting Venue				
	Conference Presentation	Book or Book Chapter	Journal Article	Dissertation	Report or Electronic Source
Percentage with pretest	26	35	45	49	55
Percentage with teacher as unit	52	54	57	75	88
Percentage meeting both criteria	21	35	36	43	51

mainly of research reports published by independent organizations that take an interest in education and engage in educational research.

Notice that the column for journals is in the center of this distribution. Journal articles are virtually indistinguishable from books and book chapters on these two criteria. They are more likely to meet these standards than are conference presentations but less likely than are dissertations, independent reports, and electronic publications. This pattern suggests that the widespread reliance on journal articles as an inclusion criterion may reflect convenience more than study quality.

Even if journal articles were generally better at meeting these quality criteria, a review that is restricted to journal articles may introduce other biases into a review. For example, journals may publish only those studies that find statistically significant relationships, thus leaving the field ignorant of studies that do not find such relationships. Moreover, the peers who review manuscripts submitted to journals tend to be members of the academy and therefore do not represent the full range of people who seek to participate in educational debates. This is an important issue in the field of education because education is inherently a cultural and political enterprise, of concern to many people outside the academy who may subscribe to different cultural and political views than university scholars. Many people outside the academy publish reports through their institutions and websites yet are capable of producing studies that meet these standards of quality. Yet among the studies examined here, only 12% of journal article authors were affiliated with noneducational institutions, in contrast to 53% of authors of independent research reports. If these nonacademy researchers are not part of the communities of scholars who maintain the journal peer-review system, journal articles may reflect only the points of view of university-based scholars, thus rendering subtle cultural, political, or value biases into "the literature" that would be corrected if reviewers expanded their searches to include more sources. This issue is especially salient today, as critics of the academy are growing more vocal about their perception of political bias in the academy. In preparation for this article, I conducted a Google search for "liberal bias in universities" and found more than 10 million sites. Whether the concerns of these critics are warranted or not, they certainly are salient and worth considering when inclusion rules for literature reviews are being developed.

So far, the advantages of systematic reviews have outweighed those of more limited reviews. But systematic reviews are not free of their own problems, and the next sections make these problems more clear.

Weeding Out Anomalous Literature

The unique contribution of a systematic review is that it can provide a definitive summary of what is known about a clearly articulated research question. One particular form of systematic review, meta-analysis, raises the bar a bit further by quantitatively synthesizing findings, yielding a new summary data point. Meta-analysis invites questions about which studies are sufficiently similar that their findings can be aggregated in this way. The literature on meta-analysis is filled with arguments about the "apples and oranges" problem (Glass, 1977; Hedges, 1986; Slavin, 1984). That is, when can we say that a group of studies actually do address the same research question, and when can we say that the studies address different questions and hence should not be lumped together? Most of this discussion has to do with the problem of combining studies that are too dissimilar, but there are also problems associated with partitioning studies in a way that eliminates literature that may be relevant. There are many ways this second possibility can arise.

One problem, articulated by Pressley, Duke, and Boling (2004), is that systematic reviews are usually conducted after a particular hypothesis, or type of treatment, has been under consideration long enough for an empirical body of work to accumulate. If a field has examined one type of treatment for the past 15 years, for instance, but has become interested in a new type of treatment only within the past 5 years, a meta-analysis is more likely to examine evidence regarding the first treatment than the second. In this respect, these authors argued, systematic reviews that rely on meta-analysis are necessarily backward looking, unable to attend to the most recent ideas and less examined innovations. Technically, of course, this does not have to be the case; a definitive statement of what is known can summarize 1 or 2 studies as easily as 50, but these authors may be right that reviewers prefer to summarize larger bodies of literature.

A similar problem occurs with respect to the evolution of research methods over time. Just as research questions change over time, so do research methods, and more recent studies rely on more recently introduced methods. An examination of the TQQT literature database, for instance, shows that earlier studies are more likely to rely on relatively less complex statistics, such as Pearson product–moment correlations, t tests, or analyses of variance, whereas later studies are more likely to rely on multiple regression and, even more recently, hierarchical linear models. In our literature database, 43% of studies published in the first 10 years (1960–1969) presented Pearson product–moment correlation coefficients, whereas only 23% of those in the most recent 10-year periods reported these. And of course, none of the early studies

presented hierarchical linear models, whereas 10% of those in the most recent period presented such analyses.

These more advanced statistical techniques are especially important in research on teachers' qualifications because virtually all studies in this area are susceptible to confounding variables. Teachers decide for themselves which qualifications to seek, so their qualifications are automatically confounded with their personal interests and values. They choose the districts and schools in which to seek employment, again confounding their personal aspirations with their teaching assignments and with the types of students they teach. Moreover, districts and schools have their own methods for recruiting and retaining teachers and for assigning them to schools and to students. So teachers arrive in their positions after complicated interactions between hiring and job-seeking practices. These processes, taken together, suggest that the positions teachers eventually take are likely to be *affinity assignments*, in which their social backgrounds and qualifications match the social backgrounds and qualifications of their students.

In such an environment, it makes no sense to compute a Pearson product–moment correlation coefficient when we know that this statistic lacks any causal message. Analyses such as multiple regression and hierarchical linear modeling are not panaceas, of course, but they are more able to take these confounding variables into account. But meta-analysts are still working to develop strategies for synthesizing slopes (see Becker & Wu, 2006) and have not even begun to think about how to handle the statistics yielded by hierarchical linear models. Because the technical capacity of meta-analysis necessarily lags behind the technical capacity of the primary researchers, it is possible for meta-analysts to summarize technically weaker studies simply because they are able to do so and to exclude studies that are more complicated because the analysts do not know how to treat them.

Meta-analyses also frequently need to eliminate studies that provide unusual and innovative approaches to study design, because they are not sufficiently comparable with other studies. Our database includes some highly innovative approaches to answering questions about the contribution of teachers' qualifications. For example, Grissmer, Flanagan, Kavata, and Williamson (2000) mapped national historical trends in teacher qualifications against concomitant trends in student achievement, measured by repeated National Assessment of Educational Progress assessments. Their unit of analysis was the entire nation, and they wanted to see whether broad societal changes in the level of teachers' qualifications might be related to broad societal changes in students' achievement. The study offered a unique approach but was so unique that it cannot readily fit into a meta-analysis with other studies. In another example, Rowan, Correnti, and Miller (2002) presented a series of analyses designed to put the question of qualifications into a larger context. They began with a variance decomposition, which suggested that about 15%–20% of the variance in student achievement lies among classrooms within schools, an amount that converts to a effect size (d) of about .42–.50. They then partialed out prior student achievement and student socioeconomic status and found that this classroom effect was reduced to .35. Recognizing that classroom effects are not the same as teacher effects, they then looked to see how stable these classroom effects were from year to year within teachers, reasoning that this analysis would provide a clue as to how much of the variation among classrooms was associated with teachers as opposed to other classroom variables. Through this step-by-step approach, these authors established a context that allowed them to finally ask about the contribution of teachers' qualifications to student achievement. This provocative article reads more like an essay than a quantitative analysis, and many details about sample sizes, standard deviations, and so forth, were not reported. Hence, the article cannot be included in a meta-analysis, even though its argument is relevant, and even though its approach to the problem is conceptually important. Idiosyncratic and original studies such as these are likely to be excluded from a meta-analysis not because they are of low quality but because they are methodological anomalies. Yet such anomalies might be high quality, timely, and theoretically important.

This is one of the thorniest and least recognized problems associated with systematic reviews and with meta-analysis in particular. Among meta-analysts, this problem has been given the relatively innocuous label "missing data" (e.g., Hedges, 1986). When Hedges described the problem of missing data, he referred both to studies whose designs were too complicated to use and to studies that failed to provide needed information such as sample sizes or standard deviations. Hedges noted that missing data from either of these causes could result in a bias in the sample of studies summarized, because both sparse reporting and elaborate designs are likely to be correlated with the studies' outcomes. Depending on circumstances, this problem can be far more than merely a missing data problem. It can introduce a *method bias* into a review by eliminating studies that do not provide the kind of statistical information that meta-analysts need or are able to analyze.

These problems are especially important in a domain that is steeped in confounding variables and in which researchers use a wide range of methods to try to control for them. In such a domain, homogeneous literatures (i.e., literatures amenable to meta-analysis) can be produced only by defining small and more narrow subsets of literature within the larger and more amorphous collection. In the case of the TQQT project, subsets have been formed that are homogeneous not only by the type of qualification they address but also by their research methods. For example, a meta-analysis of research on alternative certifications focused mainly on group comparisons. But it eliminated a study by Raymond and Fletcher (2002a, 2002b; Raymond, Fletcher, & Luque, 2001) because the study did not provide enough information about samples to allow the meta-analysts to compute an effect size. It also eliminated the only true experiment in our entire database (Decker, Mayer, & Glazerman, 2004) because the study did not report on exactly the same comparison as other studies examined. Studies such as these are eliminated not because they are of low quality but because their analyses did not exactly match other analyses or did not provide specific needed statistics. Yet they are highly relevant to the issue in general and would likely not have been eliminated from a nonquantitative, but still systematic, review.

There is another problem here: The very act of being systematic may actually hinder conceptual progress by eliminating the very studies that seek to redefine our questions or by grouping studies according to their technical features (those that provide correlation coefficients or t tests compared with those that provide regressions or hierarchical linear models) rather than by the issues they address. For instance, conceptually, a review should include all studies that

tell us about, say, secondary mathematics teachers' content knowledge. But it is difficult to aggregate studies that measure content knowledge with test scores and studies that measure it using college majors or grade point averages. Equally legitimate estimates of teachers' content knowledge may be derived from measures with different properties—say, categorical versus continuous—and may be tested using different statistics, yet we may not be able to aggregate across their varying procedures. So when research syntheses are done, literature may be parsed into bundles that are computationally coherent even if their conceptual ties are lost.

The question motivating the construction of the TQQT literature database is far too broad to succumb to a single empirical summary. It asks if any kind of teacher qualifications are related to any indicators of the quality of practice for any group of teachers. Within the database are studies that inquired about a wide range of qualifications, a wide range of indicators of teaching quality, and a wide range of teaching populations. To make sense of it, smaller and more coherent subsets of related literature need to be defined. Oswald and McCloy (2003) called these "local" analyses. To date, the TQQT project staff has produced a handful of research syntheses, each of which focuses on relatively small but coherent subsets taken from this larger literature. Most focus on a particular type of qualification, such as alternative routes (Qu & Becker, 2003), college coursework (Kennedy et al., in press), or the Teacher Perceiver Interview (Metzger & Wu, 2003); but some focus on particular types of teachers, such as those teaching mathematics (Choi & Ahn, 2003), and one focuses exclusively on a particular research method, qualitative research, because this methodology is so different from others that the meaning of these findings is difficult to merge with the meaning of other findings. Even though the full TQQT literature database includes 465 studies, the number shrinks rapidly when we ask about specific qualifications or specific groups of teachers. Consequently, individual syntheses often involve one to two dozen studies.

Discussion

Problems associated with defining a literature range from relatively minor nuisance complications, such as multiple redundant citations, to relatively substantial and conceptually difficult complications, such as how inclusion and exclusion rules are defined and how differences in study quality are addressed. Recent interest in systematic reviews, reflected in groups such as the What Works Clearinghouse and the Campbell Collaboration, groups that explicitly aim to generate definitive findings with their systematic procedures and to encourage policy applications of these findings, have raised the visibility of these issues.

The literature database described here was intended for use in systematic reviews. It was formed from thorough and systematic searches with explicit inclusion and exclusion rules. Comparisons of this literature database with Walsh's (2001) database suggest that differences in search and inclusion strategies can have a substantial impact on the resulting body of literature. These comparisons suggest the importance of thorough search procedures, but they also remind us of the role of lore in a field. The literature cited by Walsh's reviewers included not only original studies but also numerous literature reviews and essays, as well as articles that are not accessible to others, so that readers of these reviews are forced to take the reviewer's word for each citation's unique contribution and significance to the overall argument. There are also substantial omissions from those reviews, suggesting that the lore of the field is not based on a representative compilation of primary studies. Lore is an essential part of every field, for few scholars have time to read the entire corpus of literature on any one topic. But the lore is also likely to include secondhand interpretations, miscited studies, and misrepresented studies. Moreover, the lore that Walsh identified in the area of teacher qualifications misses a large fraction of studies and appears to be systematically biased against dissertations and qualitative research.

The problem is further complicated when reviewers try to improve study quality by restricting their reviews to journal articles. Within the TQQT literature at least, journal articles demonstrated no advantage in terms of study design quality. Furthermore, because their authors and peer reviewers represent mainly the university community, these articles may be biased toward the interests and prejudices of that community, at the expense of scholars in other institutions whose scholarship is of equal quality but whose interests and views differ.

As scholars have worked to develop rigorous procedures for conducting meta-analyses and systematic reviews, new complications arise. We have no clear rules for settling on which collections are homogeneous and which are heterogeneous. We have no clear rules for settling on which primary sources are of relatively higher quality than others. And the more a reviewer strives to make a literature both large and homogeneous, the more likely he or she is to also make it less informative, for such reviews are likely to eliminate studies that introduce new ideas, use new methodologies, or use unique methodologies. Such reviews could thwart conceptual and theoretical advances because their questions are necessarily narrow. At some point, the balance of benefit between systematic and conceptual reviews tips back toward conceptual reviews, for in their lack of "system," they have the flexibility to address the complexity of the substantive issues we care about.

NOTES

I acknowledge and appreciate financial support from the U.S. Department of Education's Office of Educational Research and Improvement, now the Institute for Educational Sciences; from the National Science Foundation's Program on Research, Evaluation and Communication; and from the National Science Foundation's Program on Math and Science Partnerships. None of these agencies is responsible for the contents, opinions, or conclusions presented here, of course. I also acknowledge and appreciate very helpful comments from my colleague Betsy Becker and from anonymous reviewers.

[1]Some advocates for rigor argue that only true experiments should be included in systematic reviews. If that criterion were applied to the Teacher Qualifications and the Quality of Teaching (TQQT) database, it would be reduced from 465 records to 1 record. And even this 1 study did not randomly assign teachers to their qualifications. Rather, it randomly assigned students to teachers who had already self-selected their own qualifications.

REFERENCES

American Educational Research Association. (2006). Review of Educational Research: *Submissions*. Retrieved March 6, 2006, from http://www.aera.net/publications/?id=505

Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical

knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433–456). Washington, DC: American Educational Research Association.

Becker, B. J., & Wu, M. J. (2006). *The synthesis of regression slopes in meta-analysis.* Manuscript submitted for publication.

Centre for Reviews and Dissemination. (2001). *Undertaking systematic reviews of research effectiveness.* Retrieved March 2006 from http://www.york.ac.uk/inst/crd/report4.htm

Choi, J., & Ahn, S. (2003, April). *Measuring teachers' subject-matter knowledge as a predictor of the quality of teaching.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Cooper, H. (1984). *The integrative research review: A systematic approach.* Beverly Hills, CA: Sage.

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). *The effects of Teach for America on students: Findings from a national evaluation.* Princeton, NJ: Mathematica Policy Research.

Dunkin, M. J. (1996). Types of errors in synthesizing research in education. *Review of Educational Research, 66*(2), 87–97.

Evidence for Policy and Practice Information Centre. (2005). *What is a systematic review?* Retrieved February 2, 2006, from http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?&page=/reel/about_reviews.htm

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5,* 351–379.

Goldhaber, D. D., & Anthony, E. (2003). *Teacher quality and student achievement.* New York: ERIC Clearinghouse on Urban Education.

Grissmer, D., Flanagan, A., Kavata, J., & Williamson, S. (2000). *Improving student achievement: What NAEP state test scores tell us.* Washington, DC: RAND.

Hedges, L. V. (1986). Issues in meta-analysis. *Review of Research in Education, 13,* 353–398.

Kain, J. F., & Singleton, K. (1996, May-June). Equality of educational opportunity revisited. *New England Economic Review,* 87–111.

Kennedy, M. M., Ahn, S., & Choi, J. (in press). The value added by teacher education. In M. Cochran-Smith, S. Feiman-Nemser, & J. McIntyre (Eds.), *Handbook of research on teacher education.* New York: Macmillan.

Lankford, H., Loeb, S., & Wykoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*(1), 37–62.

Lutz, F. W., & Hutton, J. B. (1989). Alternative teacher certification: Its policy implications for classroom and personnel practice. *Educational Evaluation and Policy Analysis, 11*(3), 237–254.

Metzger, S. A., & Wu, M.-J. (2003, April). *Commercial teacher interviews and their problematic role as a teacher qualification.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future.* New York: Author.

Oswald, F. L., & McCloy, R. A. (2003). Meta-analyses and the art of the average. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 311–336). Mahwah, NJ: Lawrence Erlbaum.

Popham, W. J. (1971). Performance tests of teaching proficiency: Rationale, development, and validation. *American Educational Research Journal, 8*(1), 105–117.

Pressley, M., Duke, N. K., & Boling, E. C. (2004). The educational science and scientifically-based instruction we need: Lessons from reading research and policymaking. *Harvard Educational Review, 74*(1), 30–61.

Qu, Y., & Becker, B. J. (2003, April). *Does traditional teacher certification imply quality? A meta-analysis.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Raymond, M., & Fletcher, S. (2002a). Education Next *summary of CREDO's evaluation of Teach for America.* Available at http://media.hoover.org/documents/ednext20021unabridged_raymond.pdf

Raymond, M., & Fletcher, S. (2002b). The Teach for America evaluation. *Education Next, 1,* 62–68.

Raymond, M., Fletcher, S. H., & Luque, J. (2001). *Teach for America: An evaluation of teacher differences and student outcomes in Houston, Texas.* Houston, TX: CREDO.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects Study of elementary schools. *Teachers College Record, 104*(8), 1525–1567.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) data base: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247–256.

Slavin, R. E. (1984). Meta-analysis in education: How has it been used. *Educational Researcher, 13*(7), 6–15.

Walsh, K. (2001). *Teacher education reconsidered: Stumbling for quality.* Baltimore: Abell Foundation.

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2001). *Teacher preparation research: Current knowledge, gaps, and recommendations.* East Lansing: Michigan State University.

AUTHOR

MARY M. KENNEDY is a professor in the Department of Teacher Education, Michigan State University, East Lansing, MI 48824; *mkennedy@msu.edu.* Her research focuses on teacher education and teacher knowledge and on how these contribute to teaching quality.

Manuscript received March 8, 2006
Revision received July 18, 2006
Accepted July 26, 2006