Chapter 2. Methods

In this chapter, we document the procedures that the RTI-UNC EPC used to develop this comprehensive evidence report on health literacy. To set the framework for the review, we first present the key questions and their underlying analytic framework. We then describe our strategy for identifying articles relevant to our key questions, our inclusion/exclusion criteria, and the process we used to abstract relevant information from the eligible articles and generate our evidence tables. We also discuss our criteria for grading the quality of individual articles and the strength of the evidence as a whole. Last, we explain the peer review process.

Key Questions and Analytic Framework

Based on the growing appreciation of the relationship between literacy and health, the complexity that can be involved in obtaining medical care, and health outcomes, we pose two key questions in this report, both of which have four parts. The AMA and AHRQ initially offered these questions, and we put them into final form with input from the TEAG:

- **Key Question 1**: Are literacy skills related to:
  a. Use of health care services?
  b. Health outcomes?
  c. Costs of health care?
  d. Disparities in health outcomes or health care service use according to race, ethnicity, culture, or age?
- **Key Question 2:** For individuals with low literacy skills, what are effective interventions to:
  a. Improve use of health care services?
  b. Improve health outcomes?
  c. Affect the costs of health care?
  d. Improve health outcomes and/or health care service use among different racial, ethnic, cultural, or age groups?

In the analytic framework for these key questions (Figure 1), the exposure of interest (the characteristic that is the focus of the study) is the literacy level of an individual. The literacy level may be related to the effectiveness of interventions to improve the use of health care services or the actual health of the patient. Literacy may affect the cost of health care by interacting with the level and/or effectiveness of health care services used and the cost of interventions. Patient characteristics including race, ethnicity, sex, and age and cross-cultural communication barriers may confound these relationships. Provider characteristics may influence the relationships as well. This analytic framework is merely a lattice for understanding our approach to this issue. The relationship between literacy and health-related outcomes may, in reality, have many subtle aspects that cannot be adequately represented on such a figure.

Note: Appendixes and Evidence Tables cited in this report are provided electronically at http://www.ahrq.gov/clinic/epcindex.htm

For Key Questions (KQ) 1a or 2a, we considered any process of care as a health service, including clinic and hospital visits and use of preventive health care and screening. For KQ 1b or 2b, the phrase "health outcomes" can take various meanings. We included knowledge and comprehension as either a health service or a health outcome, depending on context. Knowledge and comprehension and other categories of health outcomes are described below:

- *Knowledge.* Because level of literacy constitutes the exposure of interest in the analytic framework, one may consider health knowledge as a proximal outcome. However, because much of the research on literacy and health has focused on understanding health information, not to consider these as a health outcome would eliminate a substantial portion of research. A common assumption is that knowledge improves health outcomes, but this relationship has not been proven definitively and most likely depends on the type of knowledge.
- *Biochemical or biometric health outcomes.* Although patients often cannot directly feel them, biochemical or biometric measures such as blood pressure or glycosylated hemoglobin (HbA1c) can be important intermediate markers of more tangible health outcomes.
- *Measures of disease incidence, prevalence, morbidity, and mortality.* This category includes such outcomes as stage of cancer presentation, arthritis disease severity, and diabetes control.
- *General health status.* This outcome includes general measures of health status, usually assessed by self-report questionnaires, that have been shown to predict health outcomes.

For KQ 1c measuring the cost of health care, we included any study that measured the monetary cost of health care services. For KQ 2c, we also included studies measuring the cost of the intervention. Finally, to address KQ 1d and 2d concerning disparities in health outcomes and use of health care services, we looked for studies that reported the interaction between literacy and race, ethnicity, culture, or age with respect to health outcomes.

Literature Review Methods

Inclusion and Exclusion Criteria

Based on the final key questions specified above, we generated a list of inclusion and exclusion criteria (Table 3). We limited studies to those with outcomes related to health and health services. To ensure that the literature reviewed was relevant to current practice in the United States, we decided in agreement with our TEAG to restrict our searches to more current literature (1980 publication to the present, May 2003) and to studies conducted in developed countries, including the United States, Canada, the United Kingdom, Australia, New Zealand, and Europe. Therefore, we excluded the body of population-based studies concerning the role of poor literacy on public health outcomes in the developing world. Study participants included individuals of all ages and caregivers concerned with the outcomes of children.

As described in Table 3, we excluded studies for several reasons, including lack of a health-related outcome or results limited to the readability of materials. We also excluded studies that focused on literacy as an outcome rather than an exposure, as is seen in studies of physician office-based programs designed to improve children's literacy. We also excluded studies that used cognitive impairment or dementia as an outcome of interest because we would not be able to determine whether literacy was causing or being affected by the condition. Studies measuring only subjects' ability to interpret numerical information, without a clear health outcome, were excluded as well.

Literature Search and Retrieval Process

Databases and Search Terms. To identify the relevant literature for our review, we searched a variety of databases and employed different search strategies depending on the database (Table 4). In MEDLINE, our primary database, we had to rely on key word searches because no MeSH headings specifically identify literacy-related articles. Similarly, the terms "literacy" or "health literacy" were searched in different databases with the choice based on the scope of the database. For example, in health and biomedical databases such as MEDLINE, the Cumulative Index to Nursing and Allied Health (CINAHL), and the Cochrane Library, we searched on "literacy" because the health orientation was expected in those databases. In databases such as PSYCINFO, the Educational Resources Information Center (ERIC) or Public Affairs Information Service (PAIS), which include articles concerning a variety of literacy issues, we used "health literacy" to narrow the search to articles of interest. We also searched the Industrial and Labor Relations Review (ILRR) database to determine if any employer health literacy initiatives were discussed in the labor relations literature.

In addition, the searches in MEDLINE and CINAHL included the term "numeracy." In MEDLINE only, we searched for additional articles using the name or accepted acronym for standardized tests of literacy related to health outcomes including WRAT (Wide Range Achievement Test), REALM (Rapid Estimate of Adult Literacy in Medicine), and TOFHLA (Test of Functional Health Literacy in Adults). We reviewed the Web-based bibliographies produced by the Department of Society, Human Development, and Health of the Harvard School of Public Health[18] and the National Library of Medicine's bibliography concerning Health Literacy from their Current Bibliographies in Medicine series.[19] Finally, we also asked the TEAG and our external peer reviewers for titles of articles that we may have missed.

Table 4 presents the yield and results from our search. We conducted our initial search in late 2002 and updated it in May 2003. Beginning with a yield of 3,015 articles, we retained 73 articles that we determined were relevant to address our key questions and met our inclusion/exclusion criteria.

Article Selection Process. Once we had identified articles through the electronic database search, review articles, and bibliographies, we examined abstracts of articles to determine whether studies did, in fact, meet our criteria. One reviewer performed an initial evaluation of the abstracts for inclusion or exclusion. If one abstractor concluded that the article should be included in the review, it was retained in the analysis. Abstracts initially excluded from the study by one reviewer received a second review. The group included three physician health services researchers—Michael Pignone, MD, MPH (Scientific Director), Darren DeWalt, MD

(Co-Investigator), and Stacey Sheridan, MD, MPH (Co-Investigator)—and one health policy and health services researcher—Nancy Berkman, PhD, MLIR (Study Director).

Approximately 700 articles required review of the full article because of missing or uninformative abstracts. For the full article review, one reviewer read each article and decided whether it met our inclusion criteria. Those articles the reviewer determined did not meet our eligibility criteria, as presented in Table 3, were assigned a reason for exclusion. A second reviewer re-reviewed all initially excluded articles, and the decision to include any once-excluded articles was made as a group by the four senior staff members of the project. A list of articles excluded at full article review is provided at the end of this report, along with the reason for their exclusion.

Literature Synthesis

Development of Evidence Tables and Data Abstraction Process

The four senior staff members for this systematic review jointly developed the evidence tables. We created two sets of evidence tables, one for KQ 1 and one for KQ 2. They were designed to provide sufficient information to enable readers to understand the study and to determine quality; we gave particular emphasis to essential information on our key questions. The format of the tables, which was based on successful designs used for prior systematic reviews, varied slightly by key questions; the tables for KQ 2 include a column that describes the intervention.

For this work, the RTI-UNC EPC team decided to abstract data from included articles directly into evidence tables, in part because three of the senior staff members had prior experience conducting evidence-based systematic reviews for AHRQ. This decision meant that we bypassed the use of data abstraction forms. Following this approach created efficiencies in production and did not result in any major changes in the type of information included in the evidence tables as the project progressed.

The abstractors trained themselves on entering data into the tables by abstracting several articles and then reconvening as a group to discuss the utility of the table design. This process was repeated through several iterations until they decided that the tables included the appropriate categories for gathering the information contained in the articles. The design was then reviewed by the TEAG through a teleconference.

The first reviewer (Dr. Pignone, Dr. DeWalt, or Dr. Sheridan) initially entered data from an article into the evidence table, and the second reviewer (Dr. Berkman) also reviewed the article and edited all initial table entries for accuracy, completeness, and consistency. All disagreements concerning the information reported in the evidence tables were reconciled by the two abstractors. The full research team met regularly throughout the period of article abstraction and discussed global issues related to the data abstraction process.

The final evidence tables are presented in their entirety in Appendix C. Entries for both tables are listed alphabetically. A list of abbreviations used in the tables appears at the beginning of the appendix.

Quality and Strength of Evidence Evaluation

Rating the Quality of Individual Articles. The RTI-UNC EPC's approach to assessing the quality of individual articles was developed based on the domains and elements recommended in the evidence report by West and colleagues, *Systems to Rate the Strength of Scientific Evidence*.[20] We developed one form for reviewing all studies, which is presented at the end of this report and in Appendix B. However, because we included both intervention and observational studies in our review, several questions were relevant only to certain studies. In cases in which the item was not relevant, the quality rating was "not applicable" (NA). The categories reviewed included the following:

1. *Study population* (whether it was adequately described and appropriate for drawing relevant conclusions). Both concerns were combined to form one score.
2. *Intervention* (whether it was clearly described). This category was only relevant and answered in relation to KQ 2. For KQ 1, the response was "NA."
3. *Comparability of subjects.* This item judged the quality of the methods used for creating the sample population, including the sampling strategy, the inclusion/exclusion criteria, and the approach to randomization or allocation. It also concerned the comparability of experimental and comparison groups.
4. *Literacy measurement* (whether the instrument used was valid, reliable, and clearly defined). This measure was important for our studies because it determined how the investigators evaluated the literacy of participants. For KQ 2, interventions in populations previously characterized by literacy measurement were included, but if participants' literacy was not directly evaluated, we graded the study as "poor" for this item.
5. *Maintenance of comparable groups.* This item captured the integrity of the samples among those studies that were conducted at more than one point in time. If the study included only one contact with participants, the grade was "NA."
6. *Outcome measurement* (whether the outcome was clearly defined and whether the method of assessment was reliable). This item also rated (in studies where it was appropriate) whether the study included blinding of participants or outcome assessors.
7. *Statistical analysis.* This factor included whether the tests used were conducted in an appropriate manner and whether the effect of multiple comparisons was taken into account.
8. *Appropriate control of confounding.* This item rated the study's use of multivariate statistical techniques and/or participant restriction, stratification, or randomization to control for confounding.
9. *Funding source.* Studies were recorded as being funded by government or private foundation or by private corporate sponsorship or as not stating their funding source.

The two article abstractors independently rated each article on each of the first eight categories as "good," "fair," or "poor." We then created a composite rating in which we gave

each item equal weight. Specifically, we converted ratings for each item into numeric values in which 0 = poor, 1 = fair, and 2 = good. We averaged the ratings of the two evaluators for each item. The total score was the average of all these scores. Because one or more items may be rated as "NA" and excluded as evaluation criteria for a particular study, the number of ratings being averaged varied across studies. We included in this final rating only those items that had been rated individually (i.e., given scores of good, fair, or poor); we excluded items judged "NA." The only items reconciled between the two abstractors were those in which one rater provided a score for the item and the second said the item was not applicable. Corresponding to our individual item ratings, we concluded that, overall, an article should be considered poor with a rating of < 1.0, fair with a rating of = 1.0 and < 1.5, and good with a rating of = 1.5.

We did not integrate our evaluation of funding source into the numeric quality score for each article because of a lack of comparability between the scores. Many articles did not list their funding source (24 in total), and it was not clear what the relative score should be for a study that provided no information. Therefore, we reported these data separately and descriptively only. We include overall article ratings, individual item ratings, and funding source in the evidence table entry for each article.

Grading the Strength of Available Evidence. We developed a scheme for grading the quality or strength of our body of evidence as a whole. Using the West et al.[20] report that compared various schemes for grading bodies of evidence, we based our evaluation on criteria developed by Greer et al.[21] that we deemed most applicable to the study designs included in our literature. That system included three domains: quality of the research, quantity of studies (including number of studies and adequacy of the sample size), and consistency of findings. Grades were developed by consensus of the four senior staff members.

We graded the body of literature applicable to each of the four components of the two key questions separately. The possible grades in our scheme are as follows:

I. The evidence is from studies of strong design; results are both clinically important and consistent with minor exceptions at most; results are free from serious doubts about generalizability, bias, or flaws in research design. Studies with negative results have sufficiently large samples to have adequate statistical power.

II. The evidence is from studies of strong design, but some uncertainty remains because of inconsistencies or concern about generalizability, bias, research design flaws, or adequate sample size. Alternatively, the evidence is consistent but derives from studies of weaker design.

III. The evidence is from a limited number of studies of weaker design. Studies with strong design either have not been done or are inconclusive.

IV. No published literature.

Peer Review Process

Among the more important activities involved in producing a credible evidence report is conducting an unbiased and broadly based review of the draft report. External reviewers are

clinicians, researchers, representatives of professional societies, and potential users of the report, including TEAG members (see Appendix D). We asked peer reviewers to provide comments on the content, structure, and format of the evidence report and to complete a peer review checklist. We revised the report, as appropriate, based on comments from peer reviewers.