Chapter 2. Methodology

Literature Search and Selection Methods

Sources

To identify relevant literature we searched five databases: MEDLINE (with HealthSTAR), CINAHL, PsycINFO, EBSCO, and the Campbell Collaboration. PsycINFO indexes dissertations as well as published articles in a wide range of journals not included in MEDLINE. Within EBSCO we searched Health Source: Nursing/Academic Edition, Academic Search Elite, Business Source Premier, and MasterFILE Premier. The Campbell Collaboration is an international effort modeled on the Cochrane Collaboration whose mission is to prepare, maintain, and disseminate systematic reviews of the effectiveness of social and educational policies and practices. Its Social, Psychological, Educational and Criminological Trials Register (C2-SPECTR, http://128.91.198.137/) is a registry of studies in education, social work and welfare, and criminal justice.[29] C2-SPECTR contains over 10,000 reports. It was constructed by searching three bibliographic databases that included Educational Research Information Clearinghouse (ERIC), Sociological Abstracts, and Criminal Justice Abstracts as well as hand searching 48 journals in sociology, psychology, education, criminology and other fields and other specialized reference lists, bibliographies, collections of individuals, and other sources. Only randomized and possibly randomized trials are included in the C2-SPECTR database.

We searched MEDLINE (back to 1980) and CINAHL (back to 1982) to capture the healthcare literature, and we searched PsycINFO (back to 1984) and EBSCO (back to 1980) to capture all available industry literature. We searched MEDLINE, CINAHL, and PsycINFO again during the course of the project, with the final search in August 2002. Following the searches of literature databases, additional published studies were identified through hand searches of reference lists and selected tables of contents. Unpublished studies and government reports were identified through discussions with content experts and electronic mail lists.

Search Methods

Each search was organized by a lead investigator and was based on the set of six key questions; each of the five working conditions (workforce staffing, workflow design, personal/social, physical environment, and organizational factors) was searched separately. We defined the timeframe of the searches to be 1980 to the present, because most current quality management and accreditation systems have been implemented since 1980. Some of the databases did not go back as far as 1980, so these were searched in their entirety. Searches were limited to human studies and the English language but included foreign articles with an English abstract. Additionally, MEDLINE and CINAHL searches included a set of search terms that addressed medical errors, safety, and quality, and were constrained to articles pertaining to health care. Each search strategy was complemented by terms specific for one of the working condition categories. PsycINFO searches were not constrained to a specific industry and addressed performance measures and errors. As mentioned above, each search strategy was complemented by terms specific for one of the working condition categories. The search strategy for EBSCO was also not constrained to a specific industry. The search terms were working conditions and

productivity, human performance, and employee productivity and performance. The search strings for the other databases (MEDLINE, CINAHL, PsycINFO, and the Campbell Collaboration) are provided in Appendix B. We entered retrieved titles/abstracts into an EndNote® database, except for the EBSCO and Campbell Collaboration databases, where only included studies were entered.

Selection Processes

The searches resulted in a total of 23,179 citations. The lead investigator for each of the working condition categories reviewed the titles/abstracts for the citations that fell into their working condition category. Table 1 lists the eligibility criteria that were applied during the title/abstract review. For the non-healthcare industries, studies which met the criteria in Table 1 were considered for inclusion if they reported any measure of work results or productivity.

Table 1. Inclusion/exclusion criteria for judging titles/abstracts and full text papers

Code #	Justification for codes
	Inclusions
1	IN: For this category of working conditions
2	IN: For *another* category of working conditions
	Indicate which one:
	A. Physical Environment
	B. Workflow Design
	C. Workforce Staffing
	D. Organizational Factors
	E. Personal/Social
3	IN: Good review or background article
	Exclusions
4	OUT: Does not address any key question
5	OUT: Does not report original data
6	OUT: Wrong population (animal study, etc.)
7	OUT: <specific to topic> Write justification below

Due to budgetary constraints, citations that required inter-library loans or foreign language translation were assessed again by each lead investigator to determine if they were key citations to obtain. The full texts of citations that met inclusion criteria from the title/abstract review were then obtained and assessed again for inclusion/exclusion using the same criteria (Table 1). The search and selection of citations is depicted in Appendix C. The EBSCO and Campbell Collaboration databases were not searched independently for each of the five working condition topics and are indicated as such in Appendix C.

Prior to the review of citations for inclusion, we conducted a process to ensure consistency in application of the review criteria. A dual review by the lead investigator for each category paired randomly with one of the other investigators was conducted on a random sample of the citations for each of the topics. Kappa values were calculated to assess inter-rater reliability.[30] The kappa values ranged from 0.10 to 0.56. The investigators met to review the disagreements and the inclusion/exclusion criteria. Consensus was achieved on the points that studies of human performance, staff turnover rates, and quality be included for the appropriate working condition so long as there were original data. Studies focusing on patient satisfaction and communication (except for technology) were excluded, in that they did not address any key question. Studies of providers' perceptions of health care were included.

A second dual review was initiated on a separate random sample to reassess inter-rater reliability. This review was conducted after the lead investigators had completed their reviews of approximately half of the search result sets. The kappa values were similar among four of the topic areas and ranged from 0.41 to 0.48. Based on this outcome, a single review of the titles/abstracts by lead investigator was deemed acceptable for these four categories.

The kappa value for the second dual review of one topic (organizational factors) was lower at 0.14. The source of disagreement between the two reviewers for this working condition category was related to the rates of studies judged relevant. The lead investigator rated 5 percent of the studies relevant, while the second reviewer rated 15 percent as relevant. For the other four working condition categories, the rates of relevance judgments ranged from 11 percent to 20 percent for all reviewers. A third investigator then reviewed a random sample of 200 of the citations in this category. The three investigators then met to review and revise the definition of organizational culture. Disagreements among the three investigators were found to be due to differences in how they defined the scope of organizational culture. The investigators gained consensus that the definition should be broadened (as reflected in the definition included in Chapter 1). The lead investigator then re-reviewed the entire set of citations.

Data Abstraction and Synthesis

Data Abstraction

All studies rated as relevant on the basis of review of titles and abstracts were retrieved, photocopied, and reviewed by the lead investigator for each working condition category. Studies judged to have evidence about a key question were then abstracted, with the details of methods and results recorded on data abstraction tables. The ratings of quality of study design and execution were assigned at this time (see further details in next section). To provide

guidelines for extraction of information into the data abstraction tables, the investigative team developed data abstraction guidelines (Appendix D).

Assessment of Study Quality

Our system for rating the quality of individual studies was based on previously published methods. Several approaches to evaluating quality were examined. We reviewed concepts from the U.S. Preventive Services Task Force, the *Guide to Community Preventive Services*, the Oxford Centre for Evidence-based Medicine, and AHRQ Evidence Report/Technology Assessment Number 47 *Systems to Rate the Strength of Scientific Evidence*.[31-34]

The approach that was best suited to the types of studies included in this report is based on the system used by the Task Force on Community Preventive Services.[32] This approach differs from that used in a previous evidence report on patient safety practices.[14] Each study was assessed for quality using two ratings. The first, suitability of study design, was a three-tier approach (greatest, moderate, and least) relating the strength of the study design to threats of internal validity. We developed a modified version of design suitability for our topic. We rated studies greatest, moderate, or least based on comparison group status and measures of other factors affecting outcomes (Table 2). The second measure of quality was the quality of study execution. Here, six areas of threats to validity have been described,[32] with the ratings of good, fair, or limited corresponding to 0-1, 2-4, or 5 or more limitations. We used a similar approach and rated study execution as good, fair, or poor based on internal and external validity. Internal validity was assessed by considering such factors as comparability of groups, differential loss to followup, measurement/instrumentation issues, maturation/pre-testing effects, and whether there was a clear description of interventions. External validity was assessed by considering such factors as a sensitized or pre-tested population, specialized/atypical population, selection biases (non-random subject selection), reactive effects of experimental settings, and multiple interventions.

Table 2. Design suitability

Greatest	Concurrent comparison groups and sufficient measures for other factors affecting outcomes
Moderate	Non-concurrent comparison groups or insufficient measures for other factors affecting outcomes
Least	Non-concurrent comparison or no comparison groups and insufficient measures for other factors affecting outcomes

The overall strength of the evidence per topic area was assessed based on the criteria outlined by the Task Force on Community Preventive Services.[32] The quality and quantity of the studies and size and consistency of the results were used to grade the overall strength of the

evidence. Where evidence was available, the body of evidence for a topic area was rated as strong, sufficient, or insufficient according to the parameters outlined in Table 3.

Table 3. Assessment of strength of evidence	
Strong	At least two studies having greatest design suitability, good execution, and consistent results; or at least five studies having greatest design suitability, good or fair execution, and consistent results
Sufficient	At least one study with greatest design suitability and good execution; or at least three studies having moderate or better design suitability, fair or better execution, and consistent results
Insufficient	Too few studies to meet definition of sufficient evidence; or inconsistent results among multiple studies having some design or execution flaws

Methodologic Limitations

In this report we have adapted the methods of the systematic review[35] to collect, evaluate, and synthesize the best available evidence that addresses the key questions. However, these methods were not developed for the domain of inquiry in which we are applying them, and this has implications for the conclusions that can be reached and the degree of certainty with which they can be stated. The methods and assumptions of the systematic review of health interventions, as practiced by organizations such as the AHRQ Evidence-based Practice Centers and the Cochrane Collaboration[36] are not entirely applicable to a broad and diverse domain of inquiry such as that addressed in this report. Differences include those relating to a) underlying research traditions and assumptions; b) the search for relevant literature; c) selection of evidence for inclusion; d) appraisal of the relative validity and generalizability of studies; and e) combination or synthesis of evidence.

For healthcare interventions, the research tradition and assumptions of clinical epidemiology serve as the foundation both for original studies and syntheses of evidence. The underlying conditions, the interventions used to identify or treat them, and the outcomes of interest are generally well defined, with established and agreed upon means of identifying patients for inclusion and measuring outcomes of interest. The research designs used to investigate these interventions and conditions are well established and agreed upon, with defined limitations and biases and a familiar hierarchy of levels of evidence.[33] As a result, studies of a particular intervention for a particular condition can usually be compared, results of selected studies can be combined, and the result often be expressed in terms of a single numeric estimate of effect, with an appropriate and precise estimate of precision.

In contrast, the evidence for this report is drawn from both healthcare and non-healthcare fields and includes diverse domains of inquiry, with different research traditions and assumptions, including cognitive science, sociology, industrial and human factors engineering, and others. In many cases there are no uniform definitions of underlying conditions, interventions, or outcome states, and no single commonly accepted means of identifying or

measuring them. Furthermore, these domains of inquiry do not share a common, agreed upon hierarchical framework of research designs as in clinical epidemiology. As a result, it is often difficult to compare the results of studies, and it may be wholly inappropriate to attempt to combine them.

Controversy continues about whether research from the paradigm of clinical epidemiology should be the sole basis for recommendations for action, in particular regarding behavioral, organizational, or information interventions.[37-41] In view of this ongoing controversy, and the limitations of the application of the systematic review methods to the evidence examined herein, we have taken a hybrid approach, by summarizing evidence (where similarity of definitions and methods permits this) and separately describing individual studies (where summarization is not possible).

Synthesis of Evidence

The lead investigator for each category of working condition reviewed all studies included in the data abstraction tables and selected studies to be included in evidence tables. When multiple studies on a similar topic were available, studies that had the lowest design and execution ratings were not included in evidence tables. All studies included in evidence tables were reviewed by a second investigator to verify the accuracy of the summary information and quality ratings in the evidence tables. All evidence tables were then compiled by the principal investigator, and some studies were moved among tables if their results fit better with a different group of studies.

A technical advisory panel was identified, based on nominations from members of the expert panel. The technical advisory panel included six members with diverse backgrounds and broad knowledge of the field of patient safety (Appendix A). The final set of evidence tables and a draft of the narrative of the results chapter was submitted to the technical advisory panel for review. The panel then discussed the tables by a telephone conference call. The purpose of this review was to identify gaps in coverage of evidence from the entire domain of working conditions and to initiate supplementary literature searches based on the identified gaps.