Review of Educational Research
Spring 2005, Vol. 75, No. 1, pp. 27–61

Effects of Problem-Based Learning: A Meta-Analysis From the Angle of Assessment

David Gijbels
University of Antwerp
Filip Dochy
University of Leuven
Piet Van den Bossche
University of Maastricht
Mien Segers
University of Leiden and University of Maastricht

This meta-analysis investigated the influence of assessment on the reported effects of problem-based learning (PBL) by applying Sugrue's (1995) model of cognitive components of problem solving. Three levels of the knowledge structure that can be targeted by assessment of problem solving are used as the main independent variables: (a) understanding of concepts, (b) understanding of the principles that link concepts, and (c) linking of concepts and principles to conditions and procedures for application. PBL had the most positive effects when the focal constructs being assessed were at the level of understanding principles that link concepts. The results suggest that the implications of assessment must be considered in examining the effects of problem-based learning and probably in all comparative education research.

KEYWORDS: assessment, meta-analysis, problem-based learning.

Problem-based learning (PBL) represents a major development in higher education practice that continues to have a large impact across subjects and disciplines around the world. As indicated by many authors (Engel, 1997; Gagné, Yekovich, & Yekovich, 1993; Poikela & Poikela, 1997; Segers, 1997), today's society requires that graduates be able to solve complex problems efficiently. The claims made for PBL promise an important improvement in outcomes for higher education. The results of studies examining the effects of PBL are conclusive regarding the problem-solving ability of students. However, the results are inconclusive regarding the effects on the acquisition of knowledge (Albanese & Mitchell, 1993; Dochy, Segers, Van den Bossche, & Gijbels, 2003; Vernon & Blake, 1993). This inconclusiveness seems to contradict the vast amount of research indicating that knowledge acquisition is a prerequisite for successful problem solving (Bransford, Vye, Adams, & Perfetto, 1989; Glaser, 1992; Schoenfeld, 1985; Segers, Dochy, & De Corte, 1999; Smith, 1991; Spilich, Vesonder, Chiesi, & Voss, 1979). However, in other contexts, research has shown that assessment methods influence the findings of studies. Dochy, Segers, and Buehl (1999) reviewed 183 studies related to prior knowledge and concluded that, whereas prior knowledge generally had positive

effects on students' performance, the effects varied as a function of the method used for assessment. More specifically, prior knowledge was very likely to have negative or no effects on performance when "flawed" assessment measures were used, such as methods measuring aspects other than the possession of prior knowledge (e.g., familiarity with the topic, as indicated by the participants in the study, or participants' perceptions of the possession of prior knowledge). This means that the findings on the effect of prior knowledge on students' performances were influenced by the assessment method and the feature it was measuring.

Prior reviews (e.g., Albanese & Mitchell, 1993; Vernon & Blake, 1993) gave an overview of the effects of the implementation of PBL in comparison with more conventional education methods. A recent meta-analysis by Dochy et al. (2003, p. 550) included the method of assessment as a moderator variable and indicated that the larger an instrument's efficacy to evaluate the student's knowledge application, the larger the ascertained effect of PBL. In the present study, we want to go one step further and investigate the influence of assessment as the main independent variable. However, whether different methods of assessment actually tap different aspects of a student's knowledge base remains unclear (Bennett, 1993). Messick (1993) suggests a separation of variation in assessment method from variance relevant to the focal constructs being measured in the assessment. The latter is taken as a unit of analysis in this study. As a consequence, a theoretical framework describing the underlying focal constructs to be measured in higher education is needed to further investigate the influence of assessment on the reported effects of PBL as compared with conventional education settings. Before describing a theoretical framework on the components of problem solving, we first elaborate on problem-based learning and assessment in problem-based learning.

Problem-Based Learning

Although new in some aspects, PBL is based on ideas that have a long history and have been nurtured by many researchers (e.g., Ausubel, Novak, & Hanesian, 1978; Bruner, 1959, 1961; Dewey, 1910, 1944; Piaget, 1954; Rogers, 1969). The idea that learning is fostered when students have the opportunity to formulate and achieve their own learning goals is mentioned clearly in the work of Dewey (1910, 1944) and can also be found in Piaget and in Bruner (1959, 1961). Other aspects go back much further. The view that learning should take place in concrete situations that have a relationship with students' prior knowledge and experiences goes back to ancient Greece:

> In "The Meno" (370 B.C.), Plato presents a famous passage where Menon pushes Socrates on the issue of how one is able to leap ahead of what is known in the search for new understanding. Understanding depends on prior learning, Menon argues. When new knowledge is incompatible with this learning, one lacks a base on which to build. (Prawat, 1999, p. 48)

PBL, as it is known today, originated in Canada in the 1950s and 1960s in response to dissatisfaction with common practices in medical education there (Barrows, 1996). Although originally developed for medical training at McMaster, the McMaster version of PBL has been applied globally in many disciplines not necessarily related to the study of medicine (Gijselaers, 1995). For instance, it has been applied to the study of architecture (Donaldson, 1989; Maitland, 1991); busi-

ness administration (Merchant, 1995); economics (Garland, 1995); engineering studies (Cawley, 1989); geology (Smith & Hoersch, 1995); law (Kurtz, Wylie & Gold, 1990; Pletinckx & Segers, 2001); nursing (Higgings, 1994); social work (Heycox & Bolzan, 1991); psychology (Reynolds, 1997); and other domains of postsecondary education (Boud, 1987).

Definition

In the literature, PBL has been defined and described in various ways. PBL is used to refer to many contextualized approaches to instruction that anchor much of learning and teaching in concrete problems (Evenson & Hmelo, 2000). This focus on concrete problems as initiating the learning process is central in most definitions of PBL. For example, Barrows and Tamblyn (1980, p. 18) defined the concept of PBL as "the learning that results from the process of working toward the understanding or resolution of a problem. The problem is encountered first in the learning process and serves as a focus or stimulus for the application of problem solving or reasoning skills, as well as for the search for or study of information or knowledge needed to understand the mechanisms responsible for the problem and how it might be resolved." Boud (1987, p. 13) states that "the starting point for learning should be a problem, a query or a puzzle that the learner wishes to solve." A much-quoted definition is the one given by Albanese and Mitchell (1993, p. 53): "Problem-based learning at its most fundamental level is an instructional method characterized by the use of patient problems as a context for students to learn problem-solving skills and acquire knowledge about the basic and clinical sciences." Vernon and Blake (1993, p. 550) defined PBL by its instructional design components, students' cognitive processes, and teachers' role: "a method of learning (or teaching) that emphasizes (1) the study of clinical cases, either real or hypothetical, (2) small discussion groups, (3) collaborative independent study, (4) hypothetico-deductive reasoning, and (5) a style of faculty direction that concentrates on group progress rather than imparting information." Other authors, such as Boud and Feletti (1997, p. 15), have related PBL to a way of approaching a curriculum: "Problem based learning is an approach to structuring the curriculum which involves confronting students with problems from practice which provide a stimulus for learning."

This range of definitions illustrates how difficult it is to come to one universal definition (Chen, Cowdroy, Kingsland, & Ostwald, 1995). PBL can adopt various forms, depending on the nature of the domain and the specific goals of the programs it is part of (Barrows, 1986; Boud, 1987). Savin-Baden (2000) argues that there simply are no narrowly defined characteristics of PBL, only people working in various contexts using various PBL-approaches. However, despite the many variations of PBL that aim to match it with specific educational or disciplinary contexts, for comparative research a core model or basic definition is needed to serve as a basis of comparison with other education methods. Barrows (1996) developed a core model based on the original method from McMaster University. The McMaster approach that originated in the context of medical education has served as a robust basis for many other contexts (Boud & Feletti, 1997). Barrows's (1996) core model describes six core characteristics of PBL:

1. Learning is student-centered.
2. Learning occurs in small student groups.

3. A tutor is present as a facilitator or guide.
4. Authentic problems are presented at the beginning of the learning sequence, before any preparation or study has occurred.
5. The problems encountered are used as tools to achieve the required knowledge and the problem-solving skills necessary to eventually solve the problems.
6. New information is acquired through self-directed learning.

It should be noted that, just as the definition of PBL is ambiguous, so too is the definition of what constitutes a conventional lecture-based program. For the most part, conventional instruction is marked by large group lectures and instructor-provided learning objectives and assignments (Albanese & Mitchell, 1993).

Main Goals of PBL

Problem-based learning environments in higher education are intended to guide students to become experts in a field of study, capable of identifying the problems of the discipline and analyzing and contributing to the solutions. The findings of cognitive psychological research, especially the results of expert-versus-novice studies, have contributed to insights on the nature of expertise. Two general characteristics of expert performance can be identified (Feltovich, Spiro, & Coulson, 1993; Gagné et al., 1993; Glaser, 1990):

1. *Experts possess coherent knowledge.* They have command of a well-structured network of concepts and principles in the domain that accurately represents key phenomena and their relationships. In contrast, beginners' knowledge is patchy, consisting of isolated definitions without an understanding of underlying principles and patterns.
2. *Experts know how to use the relevant elements of knowledge in a flexible way to describe and solve novel problems.* As Glaser (1990, p. 447) noted, "experts and novices may be equally competent at recalling specific items of information, but the more experienced relate these items to the goals of problem solution and conditions for action." Novices often know facts, concepts, and principles without knowing the conditions under which they apply and how they can be used most effectively.

These distinctions provide a basis for unraveling the general goal of PBL, which is to develop successful problem solving in two dimensions: the acquisition of knowledge and the application of knowledge.

Effects of PBL

If one ponders the implementation of PBL, a major question is: Do students who use PBL reach the goal in a more effective way than students who receive conventional instruction? Albanese and Mitchell (1993, p. 56) posed the question this way: "Stated bluntly, if problem-based learning is simply another route to achieving the same product, why bother with the expense and effort of undertaking a painful curriculum revision?" To date, the interest in this question has produced six systematic reviews on the effects of problem-based learning. Three were published in the same year and the same journal (Albanese & Mitchell; Berkson, 1993; Vernon & Blake, 1993). More recently, Colliver (2000), Smits, Verbeek, and Buisonjé (2002), and Dochy et al. (2003) undertook systematic reviews, each from a different point of view.

The review by Albanese and Mitchell (1993) is probably the best known. The core question in that review—What are the effects of problem-based learning?—was investigated by means of five subquestions:

- What are the costs compared with those of lecture-based instruction?
- Do PBL students develop the cognitive scaffolding necessary to easily assimilate new basic science information?
- To what extent are PBL students exposed to an adequate range of content?
- Do PBL students become overly dependent on a small group environment?
- Do faculty dislike PBL because of the concentrated time commitment required?

The study categorizes and lists the qualitative results of studies in medical education from 1972 to 1993. The results are presented in a review that reports effect sizes and *p* values with institutions as the unit of analysis. The main results of this review were that (a) students found PBL to be more nurturing and enjoyable than conventional instruction, and (b) PBL graduates performed as well, and sometimes better, on clinical examinations and faculty evaluations than did students who had received conventional instruction. However, PBL students scored lower on basic science examinations and viewed themselves as less well prepared in the basic sciences in comparison with their conventionally trained counterparts. Furthermore, PBL graduates tended to engage in backward reasoning rather than the forward reasoning that experts engage in. Finally, the costs of PBL were high when class sizes exceeded 100.

In the same year (1993), Vernon and Blake synthesized all available research from 1970 through 1992, comparing PBL with more conventional methods of medical education. They performed five statistical meta-analyses, with the following main results: PBL was found to be significantly superior with respect to students' attitudes and opinions about their programs and with respect to measures of students' clinical performance. Contrary to the previous review findings, the scores of PBL students on miscellaneous tests of factual and clinical knowledge were not significantly different from the scores of conventionally taught students. However, the conventionally taught students performed significantly better than their PBL counterparts on the National Board of Medical Examiners (NBME) Step 1 (see method section for a description of this test).

Berkson (1993) also searched for evidence of the effectiveness of PBL in the medical PBL literature through 1992. Six topics on the effectiveness of PBL as compared with conventional curricula underlie this narrative meta-analysis in the medical domain: problem solving, imparting knowledge, motivation to learn medical science, promoting self-directed learning skills, student and faculty satisfaction, and financial costs. The results showed no distinction between graduates of PBL and conventional instruction, but PBL was stressful for both students and faculty and appeared to be unreasonably expensive.

More recently, Colliver (2000) questioned the educational superiority of PBL relative to standard approaches. Colliver focused on the credibility of claims about ties between PBL, education outcomes, and magnitude of effects. To examine these claims he conducted a review of medical education literature starting with three reviews published in 1993 (discussed above) and moving on to research published from 1992 through 1998 by the primary sources for research in medical education.

For each study, a summary is written, which includes the study design, outcomes measures, effect sizes, and further information relevant to the research conclusion. Colliver concludes that there is no convincing evidence that PBL improves students' knowledge base or clinical performance, at least not of the magnitude that would be expected given the resources required for a PBL curriculum. Nevertheless, PBL may provide a more challenging, motivating, and enjoyable approach to medical education.

One of the more recent reviews by Smits et al. (2002) is limited to the effectiveness of PBL in continuing medical education. This review includes only controlled evaluation studies in continuing medical education from 1974 to 2000. Smits et al. found limited evidence that PBL increases participants' knowledge and performance and patients' health. They found only moderate evidence that doctors were more satisfied with PBL.

The most recent review by Dochy et al. (2003) is the first to search for studies beyond the domain of medical education. The main questions are very similar but more focused than those posed in the other reviews: What are the main effects of PBL on students' knowledge and knowledge application, and what are the potential moderators of the effect of PBL? The results of their meta-analysis suggest that problem-based learning has statistically and practically significant positive effects on the students' knowledge application. The effect of problem-based learning on the knowledge base of students tends to be negative. However, the effect is found to be strongly influenced by outliers, and the moderator analysis suggests that students in a problem-based learning environment can rely on a more structured knowledge base.

Assessment in Problem-Based Learning

Widely varied methods have been used to assess students learning in PBL, from traditional multiple-choice exams and essay exams to new assessment techniques such as case-based assessment, self- and peer assessment, performance-based assessment, and portfolio assessment. Since the early 1990s, many educators and researchers have advocated new modes of assessment to be congruent with the education goals and instructional principles of PBL (Segers, Dochy, & Cascallar, 2003). It is now generally recognized that a seventh characteristic should be added to the six characteristics in Barrows's (1996) core model of PBL: That is, it is essential to PBL that students to learn by analyzing and solving representative problems. Consequently, a valid assessment system would evaluate students' problem-solving competencies in an assessment environment that is congruent with the PBL environment. This means that assessment in PBL should take into account both the organization of the knowledge base and the students' problem-solving skills (Segers et al., 2003). In addition, congruency with the learning environment implies the following:

1. Students' problem-solving skills are evaluated in an authentic assessment environment, i.e., using authentic assessment tasks or problems (Baxter & Shavelson, 1994; Shavelson, Gao, & Baxter 1996).
2. The authentic problems are novel to the students, asking them to transfer knowledge and skills acquired previously and to demonstrate understanding of the influence of contextual factors on problem analysis as well as on problem solving (Birenbaum & Dochy, 1996).

3. The problem-analysis assessment tasks ask students to argue for their ideas on the basis of various relevant perspectives (Segers, 1997).
4. The test items ask for more than the knowledge of separate concepts: Integrative knowledge, requiring the integration of relevant ideas and concepts, is stressed. Because real-life problems are mostly multidimensional and, as such, integrate various disciplines within one field of study, assessment focuses on problems with this integrative characteristic (Segers, 1997).
5. Assessment of the application of knowledge in problem solving is the heart of the matter.

The last item above deserves elaboration. Test items require examinees to apply their knowledge to commonly occurring and important problem-solving situations (Segers et al., 1999; Swanson, Case, & van der Vleuten, 1991). Because a sufficient level of domain-specific knowledge is a determinant of productive problem solving, items measuring the coherence of students' knowledge base serve at least a feedback function. For feedback reasons, the use of knowledge profiles rather than unidimensional scores is preferable (e.g., see Dochy, 1992). Dochy defined knowledge profiles as "a plotting as a graph of raw or standardized scores of a group or individual on certain parameters" (p. 143). The knowledge profiles indicate strengths and weaknesses in students' knowledge base. Research has shown that such profiles can be seen as basic determinants of academic achievement and can accurately identify specific deficits that contribute significantly to low achievement (Dochy; Letteri, 1980; Letteri & Kuntz, 1982). In the current situation, those findings imply that items assessing knowledge have to indicate the weaknesses in students' knowledge base. For example, are the students able to define or describe the central concepts of the domain studied, and do they understand the interrelations among the concepts? This kind of information enhances future learning by students in the direction of the knowledge base necessary to tackle problems.

Theoretical Framework on Problem Solving

The literature on problem solving is characterized by a wide variety of theoretical frameworks (e.g., De Corte, 1996; Glaser, Raghavan, & Baxter, 1992; O'Neil & Schacter, 1997; Schoenfeld, 1985; Smith, 1991). Despite their differences in details and terminology, all models agree that an organized domain-specific knowledge base—and metacognitive functions that operate on that knowledge—are essential parts of successful problem solving. There is also a fairly broad consensus that differences in motivation and beliefs account for problem-solving styles.

Starting from a review of three comprehensive models of the components of problem solving (Glaser et al., 1992; Schoenfeld, 1985; Smith, 1991), Sugrue (1993, 1995) presents an integrated theory-based model of the cognitive components of problem solving. The great advantage, for our purposes, of this model over the Glaser et al. model, the Schoenfeld model, and the Smith model is that the Sugrue model is translated into specifications for the assessment of the main cognitive components of problem solving. Sugrue assumes that successful problem solving in a given domain results from the interaction of knowledge structure, metacognitive functions, and motivation. For each of the three categories of cognitive components, Sugrue describes a limited set of variables that should be targeted by assessment. Because it would be impracticable to measure all of the variables that relate to the three cognitive components of problem solving, two criteria guided

TABLE 1
Cognitive components of problem solving to be assessed

Knowledge structure	Metacognitive functions	Motivation
(1) Concepts (2) Principles (3) Links from concepts and principles to conditions and procedures for application	(1) Planning (2) Monitoring	(1) Perceived self-efficacy (2) Perceived demands of the task (3) Perceived attraction of the task

Note. Adapted from "A Theory-Based Framework for Assessing Domain-Specific Problem Solving Ability," by B. Sugrue, 1995, *Educational Measurement: Issues and Practice, 14*(3), p. 31. Copyright 1995 by the National Council on Measurement in Education. Reprinted with permission.

the selection of variables in each category. They were either (a) shown to be critical by research, or (b) open to instructional intervention. (The second criterion was suggested by Snow, 1990.) These two criteria led to a model of the cognitive components of problem solving that should be targeted by assessment as presented in Table 1.

In line with the aforementioned main goals of PBL, we will now focus on the influence of the assessment of students' knowledge and knowledge application on the reported effects in studies comparing PBL with more conventional learning environments. Both the acquisition and the application of knowledge can be situated in the knowledge structure component of the problem-solving model as will be outlined below. Sugrue (1995) argued that good problem solvers draw on a store of automated, task-specific procedures. Assessment should permit identification of the nature and extent of a student's knowledge of principles and procedures in the domain of interest. In addition, because principles are rules that involve relationships among concepts, then the student's knowledge of the individual concepts should also be measured. It may be that a student has knowledge of individual concepts but has little or no knowledge of the general rules (principles) governing the relationships among the concepts. Finally, one should be able to identify students who have knowledge of principles but whose knowledge of specific procedures is limited (pp. 29–30).

The preceding paraphrase of Sugrue's understanding of problem solving supports a distinction between three aspects of the knowledge structure that can be targeted by assessment of problem solving. First, the understanding of concepts, which can be defined as "a category of objects, events, people, symbols or ideas that share common defining attributes or properties and are identified by the same name" (Sugrue, 1993, p. 9). It belongs to the category of what cognitive psychologists have called declarative knowledge. Next, understanding of the principles that link concepts should be distinguished. Sugrue (p. 9) defined a principle as "a rule, law, formula, or if-then statement that characterizes the relationship (often causal) between two or more concepts. Principles can be used to interpret problems, to guide actions, to troubleshoot systems, to explain why something happened, or to predict the effect a change in some concept(s) will have on other

concepts." If-then production rules or sequences of steps have often been called procedural knowledge (Anderson, Reynolds, Schallert, & Goetz, 1977). Finally, the linking of concepts and principles to conditions and procedures for application should also be targeted by assessment. A "procedure" defined as "a set of steps that can be carried out either to classify an instance of a concept or to change the state of a concept to effect a change in another" (Sugrue, p. 22) and "conditions" defined as "aspects of the environment that indicate the existence of an instance of a concept, and/or that a principle is operating or can be applied and/or that a particular procedure is appropriate" (Sugrue, p. 22) can be placed in the category of conditional knowledge (Paris, Lipson, & Wixson, 1983). In this final aspect of the knowledge structure—linking of concepts and principles to conditions and procedures for application—declarative knowledge becomes encapsulated in procedural knowledge. To facilitate problem solving, concepts and principles are linked to conditions and procedures to facilitate their use in unfamiliar situations (Gagné et al., 1993).

Sugrue translated her model into specifications for the assessment of the main cognitive components of problem solving (1993, 1995). Various assessment methods for measuring each of the three levels of the knowledge structure can be identified (see Table 2). Whether the assessment method that is used to assess the knowledge structure has a multiple choice, open-ended, or hands-on format, the focus should be on the level of assessment: the extent to which the student's knowledge structure is organized around key concepts and principles that are linked to

TABLE 2
Construct-by-format matrix for measuring constructs related to the knowledge structure

Elements of the knowledge structure	Method		
	Selection (MC)	Generation (open ended)	Explanation (hands on)
Concepts	Select examples.	Generate examples.	Explain why examples reflect concept attributes. Select live examples.
Principles	Select similar problems. Select best prediction. Select best explanation.	Generate predictions or solutions. Explain an event.	Explain predictions or solutions.
Application conditions and procedures	Select correct procedure for identifying instances. Select most appropriate procedure to change the state of a concept by manipulating another concept.	Perform task-specific procedures. Generate (describe) a procedure.	Explain how to perform a procedure.

Note. Adapted from "A Theory-Based Framework for Assessing Domain-Specific Problem Solving Ability," by B. Sugrue, 1995, *Educational Measurement: Issues and Practice, 14*(3), p. 32. Copyright 1995 by the National Council on Measurement in Education. Reprinted with permission.

conditions and procedures for application. As described above, three levels can be distinguished in the knowledge structure. In the first level, the assessment of the understanding of concepts is the core issue. For example, voltage and resistance are physical concepts. At the second level, understanding of the principles that link concepts is the subject of assessment. In physics, for example, the law of Ohm is a principle that prescribes current as a function of voltage and resistance in an electrical circuit. In the third and final level, the concepts and principles are linked to conditions and procedures for application. At this level the organized knowledge is applied under the appropriate circumstances, for instance, to connect an electrical circuit with bulbs and batteries in such a way that a certain level of current flows through it (Sugrue, 1995).

Research Questions

Prior reviews have given an overview of the effect of the implementation of PBL as compared with more conventional education methods. A recent meta-analysis by Dochy et al. (2003) included the method of assessment as a moderator variable, suggesting that the more an instrument is capable of evaluating the students' competence in knowledge application, the larger the ascertained effect of PBL. In this study, we want to go a step further and investigate the influence of assessment as the main independent variable. The goal of this study is to describe these effects of PBL from the angle of the underlying focal constructs being measured with the assessment. Using Sugrue's model (1993, 1995) as a frame of reference, the research questions can be formulated as follows: What are the effects of PBL when the assessment of its main goals focuses respectively on (a) the understanding of concepts, (b) the understanding of the principles that link concepts, and (c) the linking of concepts and principles to conditions and procedures for application?

On the basis of the described main goals of PBL and the suggestion that follows the moderator analysis in the review by Dochy et al. (2003, p. 550), it is expected that the effect of PBL, as compared with that of conventional education methods, should increase with each level of the knowledge structure.

Method

Criteria for Inclusion

Before searching the literature for work pertaining to the effects of PBL, we determined the criteria for inclusion in our analysis. First, each study had to be empirical, meaning that some data collection on students had to be included. Although more non-empirical literature and literature reviews were selected as sources of relevant research, this literature was not included in the analysis. Second, the characteristics of the problem-based learning environment had to fit the previously described core model of PBL (Barrows, 1996). Third, each study had to include some course or curriculum comparison. Specifically, it had to compare students in a PBL environment with students in a more conventional educational setting. The dependent variables used in the studies had to be operationalized aspects of the main goals of PBL (i.e., knowledge acquisition and knowledge application). Fourth, the subjects of study had to be students in higher education (including college and university students in all possible domains of interest). Finally, to maximize ecological validity, each study had to be conducted in a real-

life classroom or programmatic setting rather than under more controlled laboratory conditions.

Literature Search

The review and integration of research literature begins with the identification of the literature. Locating studies is the stage at which the most serious form of bias enters a meta-analysis (Glass, McGaw, & Smith, 1981). As Glass (1976, p. 6) stated, "How one searches determines what one finds; and what one finds is the basis of the conclusions of one's integration." The best protection against this source of bias is a thorough description of the procedure used to locate the studies.

We started a literature search in 1997 that included both published and unpublished studies. A wide variety of computerized databases were used, including the Educational Resources Information Center (ERIC) catalogue, PsycLIT, ADION, and LIBIS, as well as the Current Contents (for Social Sciences). The following keywords were used: *problem-solving, learning, problem-based learning, higher education, college(s), research,* and *review.* The literature was selected on the basis of the abstracts. This reading resulted in the selection of 14 publications that met the aforementioned criteria. Next, we employed the "snowball method" and reviewed the references in the selected articles for additional works. We also gathered review articles and theoretical overviews to check their references. This method yielded 17 new studies.

A second literature search that began in 1999 followed the same procedure. In addition, we contacted several researchers active in the field of PBL and asked them to provide relevant studies or to identify additional sources of studies. The second search yielded 9 additional studies.

Although our search for literature was not limited to a single domain of interest, almost all studies that met the criteria for inclusion were conducted in the domain of medical education. Only one study (Son & VanSickle, 2000) was situated outside the medical domain, in the field of economics. The strategies that we used to search for literature were meant to uncover both published and unpublished studies to prevent publication bias. We found a great many papers, but further reading revealed that all of them had eventually been published, either as articles in peer-reviewed journals (e.g., Schmidt et al., 1996) or as chapters in edited books (e.g., Boshuizen, Schmidt, & Wassamer, 1990).

Coding Study Characteristics

Using our research question as a guide, we defined the characteristics central to our review and analyzed the articles that we had selected on the basis of these characteristics. Specifically, the following information was recorded in tables:

- First author and year of publication;
- Number of subjects;
- Dependent variable (i.e., method of assessment);
- Level of assessment;
- Principal outcomes of the research; and
- Statistical values.

In coding this information and constructing overview tables, we used the following coding guidelines: With respect to the *dependent variable,* only the outcomes

related to the main goals of PBL were coded. The studies that were included in our review assessed the effects of PBL in very different ways. Some studies were more broadly based and examined other effects of PBL (e.g., satisfaction), but we included only the main goals of PBL (i.e., knowledge acquisition and knowledge application) in our results. To classify the outcomes at one of the three levels in the model by Sugrue (1995), we added an extended description of the assessment and constructs being assessed to the study characteristics coding table. We searched for additional information when the original data were too limited or unclear (e.g., Donner & Bickley, 1990; Kaufman, Mennin, Waterman, Duban, Hansbarger, Silverblatt, et al., 1989). Some assessment methods were always classified at the same level. Other methods, such as the use of essay questions, did not always measure at the same conceptual level and were classified at different levels depending on the particular study. For the main methods of assessment, we summarize in the next section the classification that resulted from the rating at the three levels by three independent raters. When there was disagreement among the raters, we discussed the classification until a clear consensus was reached. A complete overview can be found in the Appendix.

National Board of Medical Examiners: United States Medical Licensing Examination

The United States Medical Licensing Examination (USMLE) is the examination that doctors must pass to be allowed to practice medicine in the United States. The examination consists of three parts: Step 1 (at the end of Year 2 of the student's medical schooling), Step 2 (at the end of Year 3 of the student's medical schooling), and Step 3 (at the end of the study). The focus of each part is different. Step 1 stresses concepts of basic science that are important to the practice of medicine, with special emphasis on principles and mechanisms underlying health, disease, and methods of therapy. Step 2 assesses whether the candidate can apply the medical knowledge and understanding of clinical science that are essential for the provision of patient care under supervision; this step includes emphasis on health promotion and disease prevention. Step 3 assesses whether the candidate can apply the medical knowledge and understanding of biomedical and clinical science that are essential for the unsupervised practice of medicine, with emphasis on patient management in ambulatory settings (Federation of State Medical Boards & NBME, n.d.). The FLEX examination (Jones, Bieber, Echt, Scheifley, & Ways, 1984) is a similar examination that is no longer administered (S. Case, personal communication, December 7, 1999). The "NBME medicine shelf test," used in the study by Richards, Ober, Cariaga-Lo, Camp, Philip, McFarlane, et al. (1996), is a test from the NBME with items from the Step 2 examination (S. Case, personal communication, December 7, 1999).

On the basis of this information, in every study using the NBME or USML examinations we coded Step 1 as assessing the first level of the knowledge structure, Step 2 as assessing the second level, and Step 3 as assessing the third level.

Modified Essay Questions (MEQ)

The Modified Essay Questions test (MEQ) is a standardized series of open questions about a problem. The information on the case is ordered sequentially: The student receives new information only after answering a certain question (Verwijnen, Imbos, Snellen, Stalenhoef, Sprooten, & Van der Vleuten, 1982). The

student must relate theoretical knowledge to the particular situation of the case (Knox, 1989). Because the context of the particular situation of the case plays an important role in these questions, all MEQ questions were classified as assessing the third level of the knowledge structure.

Progress Test

The progress test is a written test consisting of about 250 true/false items sampling the full domain of knowledge that a graduate should be able to recall. The progress test is designed by a progress test review committee on the basis of a predefined blueprint of content domains to provide a longitudinal assessment of progress toward the final curricular objectives (Verwijnen et al., 1982). The test is constructed to also assess "rooted" knowledge, not just details (Mehrens & Lehmann, 1991). We classified all progress tests as assessing the first level of the knowledge structure.

Free Recall

This task makes a strong appeal to the students' retrieval strategies. Students are asked to write down everything that they can remember about a certain subject. Free recall was used in the study by Tans, Schmidt, Schade-Hoogeveen, and Gijselaers (1986) as a retention test; in their study the test calls, to a relatively large extent, upon the organization of the knowledge base (Patel & Groen, 1986). Thus, in the study by Tans et al., we classified the free recall test as assessing the second level of the knowledge structure. In the study by Moore, Block, Style, and Mitchell (1994), students had to recall in their 4th year the material learned in two courses in Years 1 and 2; there was no attention to the structure of the material. In the study by Moore et al., therefore, we classified the free recall test as assessing the first level of the knowledge structure.

Standardized Patient Simulations

The standardized patient simulation tests are developed by the OMERAD Institute of the University of Michigan. A patient case is simulated, and students' knowledge and clinical skills are assessed on the basis of their answers to specific questions (Jones, Bieber, Echt, Scheifley, & Ways, 1984). In the studies by Barrows and Tamblyn (1976) and Distlehorst and Robbs (1998), we classified standardized patient simulations as assessing the third level of the knowledge structure.

Essay Questions

Essay questions require an elaborated written answer (Mehrens & Lehmann, 1991). The classification of essay questions was dependent on the kind of response that was required. In the study by Aaron et al. (1998) students had to use their knowledge in a new context; thus we classified the essay questions as measuring the third level of the knowledge structure. In the study by Martenson, Eriksson, and Ingelman-Sundberg (1985), the focus was on understanding and representing the second level of the knowledge structure.

Short-Answer Questions

In comparison with the answer to an essay question, the length of the desired answer to a "short-answer" question is restricted. But, as with the classification of essay questions, the kind of response to be given is determinant. The study by

Martenson et al. (1985) focused on understanding (second level); and in both of the studies by Antepohl and Herzig (1997, 1999), the questions were used to assess factual knowledge, representing the first level of the knowledge structure.

Multiple-Choice Questions

Multiple-choice questions can be used to assess all three levels in the knowledge structure, as is indicated in Table 2. However, in all the studies using the multiple-choice format in this review, the focus was on reproduction. As a consequence, all multiple-choice questions were classified as assessing the first level of the knowledge structure.

Oral Examinations

The classification of oral examinations was dependent on the kind of response that was required. In the study by Goodman et al. (1991), we classified the oral examination as assessing the first level, but in the same study it also assessed the second level of the knowledge structure.

Performance-Based Testing: Rating

Standardized rating scales are used to evaluate the performance of the students (performance assessment; Shavelson et al., 1996). They can be used to evaluate all three levels in the knowledge structure. Ratings are used to assess the amount of factual knowledge (first level) and also to assess the organization of information (second level), as in the study by Richards et al. (1996). And in the study by Santos-Gomez, Kalishman, Rezler, Skipper, and Mennin (1990), they are used to assess the third level by rating, for example, the student's communication with patients and teamwork.

Case-Based Examinations

In case-based examinations, students have to answer questions about authentic cases. We classified case-based examinations at Level 2 of the knowledge structure if students were asked to explain predictions or solutions. This criterion applied to most of the case-based examinations (e.g., Schmidt 1996; Hmelo, 1998). When students were asked to select and explain how to perform a procedure (e.g., Schuwirth et al., 1999), we classified the cases as assessing Level 3 of the knowledge structure.

Synthesizing Research

Literature reviews can take any of three approaches: narrative, quantitative method, and statistical meta-analysis. In a narrative review, the author tries to make sense of the literature in a systematic and creative way (Van IJzendoorn, 1997). Quantitative methods use elementary mathematical procedures for synthesizing research studies (e.g., counting frequencies to produce box scores). A quantitative approach is more objective than a narrative review but also has less depth (Dochy et al., 1999). Glass (1976) systematized the approach of quantitative procedures and introduced the term *meta-analysis:* the analysis of analyses, i.e., the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings (Kulik & Kulik, 1989). Two important advantages of meta-analyses are that (a) large numbers of studies that vary substantially can

be integrated, and (b) the integration is not greatly influenced by the interpretation or use of the findings by the reviewers.

The oldest procedure for integrating studies is the narrative review. In the narrative or qualitative review, results from each study are considered at "face value," and one tries to integrate the findings in an umbrella theory (Hunter & Schmidt, 1990). This takes place in a systematic and, at the same time, creative way. However, the integration often arises from taking only a small number of studies into account and classifying all other studies as deficient (Glass, 1976).

Van IJzendoorn (1997) points out that the narrative reviewer and the quantitative reviewer set about the formulation of hypotheses and the systematic gathering of relevant studies in the same way; it is at the stage of data analysis that their methods diverge. Thus the interpretation of the more statistical, quantitative meta-analysis presupposes the narrative reviewer's strengths: creativity and intuition. With Van Ijzendoorn (1997, p. 4), we conclude that "a narrative component should always be integrated in the meta-analytic approach."

For our purposes, we conducted a statistical meta-analysis, using the MetaStat 1.5 software. We supplemented this analysis with more inclusive vote counts and the associated sign test. The simplest and most conservative methods for combining results of independent comparisons are the vote-counting methods. To do a vote count of directional results, the reviewer must count the number of comparisons that report significant results in the positive direction and compare this to the number of comparisons reporting significant results in the negative direction (Cooper, 1989). After the count is complete, a sign test is performed to discover if the cumulative results suggest that one direction occurs more frequently than chance would suggest (Cooper, 1989; Hunter & Schmidt, 1990). In performing this procedure, one assumes that under the null hypothesis of no relation in the population, the frequencies of significant positive results and negative results are equal (Hedges & Olkin, 1980).

In performing the vote count, we counted the number of experiments with significant positive and negative findings. If one study contained multiple experiments, they were all counted. To perform this procedure, only limited information is needed. In this context, Cooper (1989, p. 94) suggested that "vote counts should always be described in quantitative reviews, but . . . should always be supplemented with more sensitive procedures." In our review, the vote counts allow us to include studies that reported insufficient exact statistical data to be included in the more sensitive procedure that we used: the statistical meta-analysis. Hunter and Schmidt (1990) define the statistical meta-analysis as the quantitative accumulation and analysis of effect sizes and other descriptive statistics across studies.

Metric for Expressing Effect Sizes

A statistical meta-analysis integrates statistically empirical studies investigating the same phenomenon. The findings of all of the studies have to be expressed in a common form—the effect size—to make any comparison possible. Regarding the nature of the studies included in our statistical meta-analysis, we used the standardized mean difference effect size (Glass's delta). This metric is appropriate when the means of two groups are being compared (Cooper, 1989; Glass et al., 1981). Glass's delta expresses the distance between the two group means in terms of their common standard deviation. The common standard deviation is calculated

by using the standard deviation of the control group because it is not affected by the treatment.

In the present study, calculation of the effect size was sometimes difficult or even impossible because the research reports and articles varied in the completeness of their reporting of research results (e.g., Mennin, Friedman, Skipper, Kalishman, & Snyder, 1993). Several research reports failed to report the mean and standard deviation of the control group. In that case, the effect size was calculated as much as possible by means of the transformation of *t,* chi-square, and *F* statistics (Cooper, 1989; Glass et al., 1981). In some cases, the effect size was calculated starting from data on the significance level (e.g., Santos-Gomez et al., 1990). When no exact *p* values were reported, the *p* value corresponding to the highest value in the probability range reported by the researchers was used. (e.g., when in the original work "$p < .05$" was mentioned, the value on the basis of which the effect size was calculated became $p = .05$; e.g., Schuwirth et al., 1999). As a consequence, the effect sizes calculated and used in the analysis represent a conservative approach and tend to underestimate the real values. Effect sizes were not deduced from unreliable data in the reported findings (e.g., from graphs). In such cases, only the sign of the difference between the two conditions was reported.

Identifying Independent Hypothesis Tests

Sometimes, a single study may contain multiple tests of the same hypothesis. Because one of the assumptions underlying meta-analysis is that effects are independent from one another, a decision must be made about what will be considered as an independent estimate of effect when a single study reports multiple outcomes. Several strategies can be suggested regarding how to decide on the unit of analysis when calculating average effect sizes. In this study, we used the shifting units method from Cooper (1989). Each effect size resulting from hypothesis tests with independent samples was initially coded as if it were an independent event. However, study outcomes resulting from the same sample are also aggregated within the separate categories of the influencing variable (see aggregated effect sizes in bold in the Appendix at the end of this article). This strategy is a compromise that allows studies to retain their maximum information value while keeping any violation of the assumption of independence of hypothesis tests to a minimum.

Combining Effect Sizes Across Studies

Once an effect size had been calculated for each study or comparison, the effects testing the same hypothesis were averaged. Unweighted and weighted procedures were used. In the unweighted procedure, each effect size was weighted equally in calculating the average effect. In the weighted procedure, more weight was given to effect sizes with larger samples (factor w = inverse of the variance), on the assumption that the larger samples more closely approximate actual effects (Cooper, 1989; Hedges & Olkin, 1985). These weighted combined effect sizes were tested for statistical significance by calculating the 95% confidence interval (Cooper, 1989).

Analyzing Variance in Effect Sizes Across Studies

The last step was to examine the variability of the effect sizes by means of a homogeneity analysis (Cooper, 1989; Hedges & Olkin, 1985; Hunter & Schmidt,

1990). First, a Q statistic is calculated for each subgroup of comparisons. The value of these statistics is added up to obtain a value Q_w (within-group chi-square). Then, this value is subtracted from Q_t (chi-square distribution, $N - 1$ degrees of freedom) to obtain Q_b (between-group chi-square, $Q_b = Q_t - Q_w$). The statistic Q_b is used to test the homogeneity of the *mean* effect of grouping. If Q_b reaches a significant level, the grouping factor provides a significant contribution to the variance in the set of effect sizes. Q_w is comparable to Q_t, meaning that if this statistic reaches a significant level, there is a need for further grouping of the data (Hunter, Schmidt, & Jackson, 1982; Springer, Stanne, & Donovan, 1999).

Results

Forty studies met the inclusion criteria for the meta-analysis; of these, 31 were published in peer-reviewed journals and 9 were published in edited books. Of the 40 studies, 31 (77%) presented data on knowledge-concepts effects, 17 (42%) presented data on knowledge-principles effects, and 8 (20%) presented data on effects concerning the application of knowledge (conditions and procedures). The percentages add up to more than 100 because several studies presented outcomes of more than one category (see Appendix).

When the effect sizes are plotted by study, three studies are seen to be serious outliers (Eisenstaedt, Barry, & Glanz, 1990; Mennin et al. 1993; Tans et al., 1986). When these three studies (all situated at the concept level of the knowledge structure) are left aside, the main effects of PBL on the three levels of the knowledge structure measured appear to be different. The results of the analysis are summarized in Table 3.

In general, the results of the vote count were statistically significant, except for the assessment of the first level of the knowledge structure. These results suggest that students in PBL perform better at the second and third levels of the knowledge structure. None of the studies reported significant negative findings at the

TABLE 3
Main effects of problem-based learning

				Average effect sizes			
Outcome	Signif. +	Signif. –	N	Unweighted	Weighted (*CI* 95%)	Q_b	Q_w
Concepts	3	5	21	–0.042	0.068 (+/– 0.864)^{ns}	18.998**	113.563**
Principles	17	1*	15	+0.748	+ 0.795 (+/– 0.782)		82.196**
Application	6	0*	13	+0.401	+0.339 (+/– 0.662)^{ns}		23.356**

Note. Unless noted with ^{ns}, all weighted effect sizes are statistically significant (the 95% confidence intervals do not include zero). Unweighted effect sizes were not tested for significance. Signif. + = number of studies with a significant (at 5% level) positive finding. Signif. – = number of studies with a significant (at 5% level) negative finding. N = number of independent outcomes measured. *CI* = confidence interval. Q_b = between-group chi-square. Q_w = within-group chi-square. ^{ns} = not significant.

*Two-sided sign test is significant at the 5% level. **$p < .05$.

third level of the knowledge structure. Only one study reported negative findings at the second level of the knowledge structure. At the first level—understanding conceptions—the vote count shows a negative tendency, with 5 studies yielding a significant negative effect and only 3 studies yielding a significant positive effect. However, this difference is not significant at the 5% level. If we look to the weighted average effect sizes (*ES*), this negative effect is close to zero but positive (weighted average $ES = 0.068$) based on 21 studies. On the basis of 15 studies, students studying in PBL classes demonstrated better understanding of the principles that link concepts (weighted average $ES = 0.795$) than did students who were exposed to conventional instruction. On the basis of 13 studies, students in PBL were better at the third level of the knowledge structure (weighted average $ES = 0.339$) than were students in conventional classes. It is important to note that the weighted average *ES* of 0.795, belonging to the second level of the knowledge structure, was the only statistically significant result.

As can be seen from the statistically significant Q_b statistics reported in Table 3, the grouping into three levels of assessment provides better insight into the effects of PBL. However, the results of the homogeneity analysis suggest that further grouping of the data is necessary for a full understanding of the moderators of the effects of PBL. As indicated by the statistically significant Q_w statistics, one or more factors other than chance or sampling error account for the heterogeneous distribution of effect sizes.

Conclusion and Discussion

The purpose of this review was to examine the effects of PBL from the angle of the underlying focal constructs being measured with the assessment. We were interested in empirical and quasi-experimental studies with clear descriptions of the conditions and measures used to assess the effects of PBL. As a result, our search for literature examining the effects of PBL yielded 40 studies, which we reviewed by means of a statistical meta-analysis, supplemented by the more inclusive vote counts and the associated sign test. When appropriate, we have made some narrative comments. We used Sugrue's (1995) model on the cognitive components of problem solving to classify the methods of assessment used in various studies into three categories of analysis. These three categories were the three levels of the knowledge structure that can be targeted by assessment of problem solving: understanding of concepts, understanding of the principles that link concepts, and linking of concepts and principles to conditions and procedures for application.

Before discussing the conclusions that can be drawn from the analysis, it is important to consider the limitations of this review. The selection of studies for a review of any type is subject to selection bias. Bias, for example, can be caused by selecting only published works, which tend to report only statistically significant results (Glass, 1976). We attempted to avoid this form of bias by also searching for unpublished works. However, all of the studies involved turned out to be published eventually: 31 studies were published in a peer-reviewed journal, and 9 studies were published in an edited book. Another remarkable fact is that, although our literature search was broad, all but one of the studies that met our criteria for inclusion were situated in the domain of medical education (the exception, Son & VanSickle, 2000, was in the field of economics).

Although PBL originated in medical education, it has been applied globally for many years in several disciplines. Nevertheless, claims about the effects of PBL seem to rely almost exclusively on literature in medical education. Generalizations should therefore be made with special caution. It is also known that selection bias problems are sometimes inherent in "between institution" or "elective track" studies. Another criticism of meta-analysis, the "garbage in, garbage out" critique (Hunt, 1997), which refers to the mixing good and bad studies, may also apply to our review. A clear description of our criteria for inclusion and the use of a weighting technique that takes into account the sample sizes of studies—based on the assumption that larger samples more closely approximate actual effects (Cooper, 1989; Hedges & Olkin, 1985)—should overcome these critiques to a certain extent.

A final limitation of this study is related to the theoretical framework used to concretize the research question and to interpret and code the studies involved. The model presented by Sugrue (1993, 1995) is only one possible framework for the components of problem solving. The translation of the model into specifications for the assessment of problem solving made the model useful for our purpose. The strength of the model is at the same time the greatest weakness of the model. The classification according to the three levels of the knowledge structure can be done relatively easily in most domains, such as mathematics, science, economics, and medical education. The extraction of unambiguous principles governing relationships among concepts might be more difficult to use in other domains, such as history (Sugrue, 1995).

Despite these limitations, we feel that several useful conclusions may be drawn from the analysis of studies in this review. In general, the effect of PBL differs according to the levels of the knowledge structure being measured. PBL had the most positive effects when the focal constructs being assessed were at the level of understanding the principles that link concepts, the second level of the knowledge structure. Only one study presented significant negative findings (Martenson et al., 1985). No negative findings were found at the third level of the knowledge structure. These findings seems to be in line with the tentative conclusion of Dochy et al. (2003), suggesting that the better the capacity of an instrument for evaluating the application of knowledge by the student, the larger the ascertained effect of PBL. This conclusion is confirmed when one looks at the first level of the knowledge structure: More studies report negative effects of PBL when assessing the understanding of concepts. However, when the weighted average effect sizes are taken into account, a different picture emerges. PBL has a small positive effect size (weighted average $ES = 0.068$), meaning that when the understanding of concepts is the subject of the assessment, students in PBL perform at least as well as students in conventional learning environments. However, the effect size is not statistically significant. In line with the conclusion of Dochy et al., the effect of PBL is more positive when understanding of the principles that link concepts is at the heart of the assessment (weighted average $ES = 0.795$). Contrary to the suggestion that the effects of PBL should be larger when more complex levels of the knowledge structure are being assessed, the effect size belonging to the third level of the knowledge structure, although still positive, is smaller (weighted average $ES = 0.339$) but not statistically significant.

Linking these results to the main goals of PBL and the expert–novice studies, it could be concluded that students' path toward expertise has been accelerated

(Glaser, 1990). First, students in PBL seem to posses a highly structured network of concepts and principles (Level 2). Second, students in PBL are equally competent at recalling specific items of information, as compared with students in more conventional learning environments. Although students in PBL are better in relating their knowledge base to the goals of problem solution and conditions for action, the magnitude of the effect belonging to Level 2 knowledge is not very large and not statistically significant. The question is to what extent students' year of study is a moderating variable. It might be expected that when the students' assessment is close to graduation, the results will show higher positive effect sizes than in research settings measuring at an earlier stage of the study. The expertise level of the students was one of the moderating variables in the meta-analysis by Dochy et al. (2003). Their results suggest that the advantage of the conventional education method in knowledge acquisition disappears after the second year. The effects of PBL on the application of knowledge differentiated for the expertise level of students show as positive.

PBL aims to educate students who are able to solve complex problems. To be congruent with its education goals and resulting instructional principles and practices, the assessment of the application of knowledge in solving problems is at the heart of the matter in PBL. Therefore, one could expect students in PBL to perform better at this level of the knowledge structure. The effect of PBL is larger when assessment appeals to the understanding of principles that link concepts. In only 8 of the 40 studies did the assessment focus on the third level of the knowledge structure. Most of the studies ($N = 31$) reported assessment at the level of understanding of concepts. These results present an implicit challenge for PBL and comparative research on PBL: It is important to pay more attention to this third level of the knowledge structure. If PBL aims to educate better problem solvers, more attention should be paid to the third level, both during the learning activities that take place and during students' assessment in PBL. The aim of educating graduates who can solve complex problems in an efficient way is a general goal in higher education. This concern accounts for all learning environments in the context of higher education and probably beyond.

The evaluation of the practical significance of an effect size is a subject of discussion between researchers in education and other fields (Springer et al., 1999). As Glass et al. (1981, p. 104) note, "There is no wisdom whatsoever in attempting to associate regions of the effect-size metric with descriptive adjectives such as 'small,' 'moderate,' 'large,' and the like." Cohen (1988) and Kirk (1996) hesitantly suggested general guidelines ($ES = 0.20$, small effect; $ES = 0.50$, moderate effect; and $ES = 0.80$, large effect), stating that "there is a certain risk inherent in offering conventional operational definitions for those terms for use in power analysis in as diverse a field of inquiry as behavioral science" (Cohen, 1988, p. 25). In general in the field of education research, an effect size of 0.33 is seen as the minimum necessary to establish practical significance (Gall, Borg, & Gall, 1996). Considered in this light, the results of the present meta-analysis are of practical significance ($ES = 0.339$) for the assessment of the organization of the knowledge structure, and they certainly are of practical significance when assessment addresses the linking of concepts and principles to application conditions and procedures ($ES = 0.795$). The latter effect size comes closer to the desired level ($ES = 1.00$) for major curricular interventions as assumed in the critical overview of Colliver (2000).

From the homogeneity analysis, it is clear that the method of assessment has an important influence in the reported effects of PBL, as stated above, but also that other moderators of PBL play a substantial role when the effects of PBL are examined. Study design, scope of implementation, and year of study have been shown to be possible moderating variables in the reporting of the effects of PBL (Dochy et al., 2003). However, the scope of this review was to investigate the assessment of the three levels of the knowledge structure as main independent variable. In connection with the influence of assessment as a moderator variable, it would also be interesting for future research to take into account the context of assessment (for licensure and grading purposes, or for research purposes alone). Nevertheless, it became clear from this study that the implications of assessment and the levels in the knowledge structure being measured must be considered when one examines the effects of problem-based learning, and probably should be considered in all comparative education research.

Note

The authors are grateful to Dan Hickey and Rachel Lewis, University of Georgia; Eduardo Cascallar, Assessment Group International; the late Alicia Schmid, of the Educational Testing Service; and Neville Bennett, University of Exeter (UK), for their comments on earlier draft of this article.

References

References marked with an asterisk indicate studies included in the meta-analysis (see Appendix A).

*Aaron, S., Crocket, J., Morrish, D., Basualdo, C., Kovithavongs, T., Mielke, B., et al. (1998). Assessment of exam performance after change to problem-based learning: Differential effects by question type. *Teaching and Learning in Medicine, 10*(2), 86–91.

Albanese, M. A., & Mitchell, S. (1993). Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine, 68,* 52–81.

*Albano, M. G., et al. (1996). An international comparison of knowledge levels of medical students: The Maastricht progress test. *Medical Education, 30,* 239–245.

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. (1977). Frameworks for comprehending. *American Educational Research Journal, 14,* 367–381.

*Antepohl, W., & Herzig, S. (1997). Problem-based learning supplementing in the course of basic pharmacology: Results and perspectives from two medical schools. *Naunyn-Schmiedeberg's Archives of Pharmacology, 355,* R18.

*Antepohl, W., & Herzig, S. (1999). Problem-based learning versus lecture-based learning in a course of basic pharmacology: A controlled, randomized study. *Medical Education, 33*(2), 106–113.

Ausubel, D., Novak, J., & Hanesian, H. (1978). *Educational psychology: A cognitive view* (2nd ed.). New York: Holt, Rinehart & Winston.

*Baca, E., Mennin, S. P., Kaufman, A., & Moore-West, M. (1990). Comparison between a problem-based, community-oriented track and a traditional track within one medical school. In Z. M. Nooman, H. G. Schmidt, & E. S. Ezzat (Eds.), *Innovation in medical education: An evaluation of its present status* (pp. 9–26). New York: Springer.

Barrows, H. S. (1986). A taxonomy of problem-based learning methods. *Medical Education, 20,* 481–486.

Barrows, H. S. (1996). Problem-based learning in medicine and beyond. In L. Wilkerson & W. H. Gijselaers (Eds.), *New directions for teaching and learning: Vol. 68. Bringing*

problem-based learning to higher education: Theory and practice (pp. 3–13). San Francisco: Jossey-Bass.
*Barrows, H. S., & Tamblyn, R. M. (1976). An evaluation of problem-based learning in small groups utilizing simulated patient. *Journal of Medical Education, 51,* 52–56.
Barrows, H. S., & Tamblyn, R. M. (1980). *Problem-based learning: An approach to medical education.* New York: Springer.
Baxter, G. P., & Shavelson, R. J. (1994). Science performance assessments: Benchmarks and surrogates. *International Journal of Educational Research 21*(3), 279–299.
Bennett, R. E. (1993). On the meanings of constructed response. In R. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–28). Hillsdale, NJ: Lawrence Erlbaum.
Berkson, L. (1993). Problem-based learning: Have the expectations been met? *Academic Medicine, 68*(10), S79–S88.
*Bickley, H., Donner, R. S., Walker, A. N., & Tift, J. P. (1990). Pathology education in a problem-based medical curriculum. *Teaching and Learning in Medicine, 2*(1), 38–41.
Birenbaum, M., & Dochy, F. (Eds.). (1996). *Alternatives in assessment of achievements, learning processes and prior knowledge.* Boston: Kluwer Academic Publishers.
*Block, S. D., & Moore, G. T. (1994). Project evaluation. In D. C. Tosteson, S. J. Adelstein, & S. T. Carver (Eds.), *New pathways to medical education: Learning to learn at Harvard Medical School.* Cambridge, MA: Harvard University Press.
*Boshuizen, H. P. A., Schmidt, H. G., & Wassamer, L. (1993). Curriculum style and the integration of biomedical and clinical knowledge. In P. A. J. Bouhuys, H. G. Schmidt, & H. J. M. van Berkel (Eds.), *Problem-based learning as an educational strategy* (pp. 33–41). Maastricht: Network Publications.
Boud, D. (1987). Problem-based learning in perspective. In D. Boud (Ed.), *Problem-based learning in education for the professions* (pp. 13–18). Sydney: Higher Education Research and Development Society of Australia.
Boud, D., & Feletti, G. (1997). Changing problem-based learning [Introduction]. In D. Boud & G. Feletti (Eds.), *The challenge of problem-based learning* (2nd ed.; pp. 1–14). London: Kogan Page.
Bransford, J. D., Vye, N. J., Adams, L. T., & Perfetto, G. A. (1989). Learning skills and the acquisition of knowledge. In A. Lesgold & R. Glaser (Eds.), *Foundations for a psychology of education* (pp. 199–249). Hillsdale, NJ: Lawrence Erlbaum.
Bruner, J. S. (1959). Learning and thinking. *Harvard Educational Review, 29,* 184–192.
Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31,* 21–32.
Cawley, P. (1989). The introduction of a problem-based option into a conventional engineering degree course. *Studies in Higher Education, 14,* 83–95.
Chen, S. E., Cowdroy, R. M., Kingsland, A. J., & Ostwald, M. J. (Eds.). (1995). *Reflections on problem-based learning.* Campbelltown, New South Wales, Australia: Australian Problem-based Learning Network.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
Colliver, J. A. (2000). Effectiveness of problem-based learning curricula: Research and theory. *Academic Medicine, 75*(3), 259–266.
Cooper, H. M. (1989). *Integrating research: A guide for literature reviews* (2nd ed.). Newbury Park, CA: Sage.
De Corte, E. (1996). Instructional psychology: Overview. In E. De Corte & F. E. Weinert (Eds.), *International encyclopedia of developmental and instructional psychology* (pp. 33–43). Oxford, UK: Elsevier Science.
Dewey, J. (1910). *How we think.* Boston: D. C. Heath & Co.
Dewey, J. (1944). *Democracy and education.* New York: Macmillan.

*Distlehorst, L. H., & Robbs, R. S. (1998). A comparison of problem-based learning and standard curriculum students: Three years of retrospective data. *Teaching and Learning in Medicine, 10*(3), 131–137.

Dochy, F. (1992). *Assessment of prior knowledge as a determinant for future learning.* Utrecht/London: Lemma B. V./Jessica Kingsley Publishers.

Dochy, F., Segers, M., & Buehl, M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research, 69*(2), 145–186.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction, 13,* 533–568.

Donaldson, R. (1989). A good start in architecture. In B. Wallis (Ed.), *Problem-based learning: The Newcastle Workshop* (pp. 41–53). Newcastle, Australia: University of Newcastle.

*Donner, R. S., & Bickley, H. (1990). Problem-based learning: An assessment of its feasibility and cost. *Human Pathology, 21,* 881–885.

*Doucet, M. D., Purdy, R. A., Kaufman, D. M., & Langille, D. B. (1998). Comparison of problem-based learning and lecture format in continuing medical education on headache diagnosis and management. *Medical Education, 32*(6), 590–596.

*Eisenstaedt, R. S., Barry, W. E., & Glanz, K. (1990). Problem-based learning: Cognitive retention and cohort traits of randomly selected participants and decliners. *Academic Medicine, 65*(9), 11–12.

Engel, C. E. (1997). Not just a method but a way of learning. In D. Boud & G. Feletti (Eds.), *The challenge of problem-based learning* (2nd ed.; pp. 17–27). London: Kogan Page.

Evenson, D. H., & Hmelo, C. E. (Eds.). (2000). *Problem-based learning: A research perspective on learning interactions.* Mahwah, NJ: Lawrence Erlbaum.

*Farquhar, L. J., Haf, J., & Kotabe, K. (1986). Effect of two preclinical curricula on NMBE Part I examination performance. *Journal of Medical Education, 61,* 368–373.

Federation of State Medical Boards & National Board of Medical Examiners (n.d.). *The United States Medical Licensing Examination.* Retrieved February 6, 2003, from *http://www.usmle.org*

Feltovich, P. J., Spiro, R. J., & Coulson, R. L. (1993). Learning, teaching, and testing for complex conceptual understanding. In N. Frederksen, R. J. Mislevey, & I. I. Bejar (Eds.), *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum.

*Finch, P. M. (1999). The effect of problem-based learning on the academic performance of students studying pediatric medicine in Ontario. *Medical Education, 33*(6), 411–417.

Gagné, E. D., Yekovich, C. W., & Yekovich, F. R. (1993). *The cognitive psychology of school learning* (2nd ed.). New York: HarperCollins College Publishers.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research* (6th ed.). White Plains, NY: Longman.

Garland, N. J. (1995). Peer group support in economics: Innovations in problem-based learning. In W. Gijselaers, D. Tempelaar, P. Keizer, E. Bernard, & H. Kasper (Eds.), *Educational innovation in economics and business administration: The case of problem-based learning* (pp. 331–337). Dordrecht: Kluwer.

Gijselaers, W. (1995). Perspectives on problem-based learning. In W. Gijselaers, D. Tempelaar, P. Keizer, J. Blommaert, E. Bernard, & H. Kasper (Eds.), *Educational innovation in economics and business administration: The case of problem-based learning* (pp. 39–52). Norwell, MA: Kluwer.

Glaser, R. (1990). Toward new models for assessment. *International Journal of Educational Research, 14,* 475–483.

Glaser, R. (1992). Expert knowledge and processes of thinking. In D. F. Halpern (Ed.), *Enhancing thinking skills in the sciences and mathematics* (pp. 63–75). Hillsdale, NJ: Lawrence Erlbaum.

Glaser, R., Raghavan, K., & Baxter, G. P. (1992). *Cognitive theory as the basis for design of innovative assessment: Design characteristics of science assessments* (CSE Tech. Rep. No. 349). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Glass, G. V. (1976). Primary, secondary and meta-analysis. *Educational Researcher, 5,* 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* London: Sage.

*Goodman, L. J., Brueschke, E. E., Bone, R. C., Rose, W. H., Williams, E. J., & Paul, H. A. (1991). An experiment in medical education: A critical analysis using traditional criteria. *Journal of the American Medical Association, 265,* 2373–2376.

Hedges, L. V., & Olkin, I. (1980). Vote counting methods in research synthesis. *Psychological Bulletin, 88,* 359–369.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Heycox, K., & Bolzan, N. (1991). Applying problem-based learning in first-year social work. In D. Boud & G. Feletti (Eds.), *The challenge of problem-based learning* (pp. 186–193). New York: St. Martin's Press.

Higgins, L. (1994). Integrating background nursing experience and study at the postgraduate level: An application of problem based learning. *Higher Education Research and Development, 13,* 23–33.

*Hmelo, C. E. (1998). Problem-based learning: Effects on the early acquisition of cognitive skill in medicine. *The Journal of the Learning Sciences, 7,* 173–236.

*Hmelo, C. E., Gotterer, G. S., & Bransford, J. D. (1997). A theory-driven approach to assessing the cognitive effects of PBL. *Instructional Science, 25,* 387–408.

Hunt, M. (1997). *How science takes stock.* New York: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverley Hills, CA: Sage.

*Imbos, T., Drukker, J., van Mameren, H., & Verwijnen, M. (1984). The growth in knowledge of anatomy in a problem-based curriculum. In H. G. Schmidt & M. L. de Volder (Eds.), *Tutorials in problem-based learning: New directions in training for the health professions* (pp. 106–115). Assen, The Netherlands: Van Gorcum.

*Imbos, T., & Verwijnen, G. M. (1982). Voortgangstoetsing aan de medische faculteit Maastricht [Progress testing at the Maastricht faculty of medecine]. In H. G. Schmidt (Ed.), *Probleemgestuurd Onderwijs: Bijdragen tot Onderwijsresearchdagen 1981* (pp. 45–56). Harlingen, The Netherlands: Stichting voor Onderzoek van het Onderwijs, Flevodruk Harlingen b.v.

*Jones, J. W., Bieber, L. L., Echt, R., Scheifley, V., & Ways, P. O. (1984). A problem-based curriculum: Ten years of experience. In H. G. Schmidt & M. L. de Volder (Eds.), *Tutorials in problem-based learning: New direction in training for the health professions* (pp. 181–198). Assen, The Netherlands: Van Gorcum.

*Kaufman, A., Mennin, S., Waterman, R., Duban, S., Hansbarger, C., Silverblatt, H., et al. (1989). The New Mexico experiment: Educational innovation and institutional change. *Academic Medicine, 64,* 285–294.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

Knox, J. D. E. (1989). What is . . . a modified essay question? *Medical Teacher, 11*(1), 51–55.

Kulik, J. A., & Kulik, C. L. (1989). The concept of meta-analysis. *International Journal of Educational Research, 13*(3), 227–234.

Kurtz, S., Wylie, M., Gold, N. (1990). Problem-based learning: An alternative approach to legal education. *Dalhousie Law Journal, 13,* 787–816.

Letteri, C. A. (1980). Cognitive profile: Basic determinant of academic achievement. *Journal of Educational Research, 5,* 195–198.

Letteri, C. A., & Kuntz, S. W. (1982, March). *Cognitive profiles: Examining self-planned learning and thinking styles.* Paper presented at the annual meeting of the American Educational Research Association, New York City.

*Lewis, K. E., & Tamblyn, R. M. (1987). The problem-based learning approach in baccalaureate nursing education: How effective is it? *Nursing Papers, 19*(2), 17–26.

Maitland, B. (1991). Problem-based learning for an architecture degree. In D. Boud & G. Feletti (Eds.), *The challenge of problem-based learning* (pp. 203–210). New York: St. Martin's Press.

*Martenson, D., Eriksson, H., & Ingelman-Sundberg, M. (1985). Medical chemistry: Evaluation of active and problem-oriented teaching methods. *Medical Education, 19,* 34–42.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology.* New York: Holt, Rinehart and Winston.

*Mennin, S. P., Friedman, M., Skipper, B., Kalishman, S., & Snyder, J. (1993). Performances on the NMBE I, II, III by medical students in the problem-based learning and conventional tracks at the University of New Mexico. *Academic Medicine, 68,* 616–624.

Merchand, J. E. (1995). Problem-based learning in the business curriculum: An alternative to traditional approaches. In W. Gijselaers, D. Tempelaar, P. Keizer, E. Bernard, & H. Kasper (Eds.), *Educational innovation in economics and business administration: The case of problem-based learning* (pp. 261–267). Dordrecht, The Netherlands: Kluwer.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. Bennett & W. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 61–74). Hillsdale, NJ: Lawrence Erlbaum.

*Moore G. T., Block S. D., Style C. B., & Mitchell, R. (1994). The influence of the New Pathway curriculum on Harvard medical students. *Academic Medicine, 69,* 983–989.

*Morgan, H. R. (1977). A problem-oriented independent studies programme in basic medical sciences. *Medical Education, 11,* 394–398.

*Neufeld, V., & Sibley, J. (1989). Evaluation of health sciences education programs: Program and student assessment at McMaster University. In H. G. Schmidt, M. Lipkin Jr., M. W. de Vries, & J. M. Greep (Eds.), *New directions for medical education: Problem-based learning and community-oriented medical education* (pp. 165–179). New York: Springer Verlag.

O'Neil, H. F., & Schacter, J. (1997). *Test specifications for problem-solving assessment* (CSE Tech. Rep. No. 463). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology, 8,* 293–316.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science, 10,* 91–116.

*Patel, V. L., Groen, G. J, & Norman, G. R. (1991). Effects of conventional and problem-based medical curricula on problem-solving. *Academic Medicine, 66,* 380–389.

Piaget, J. (1954). *The construction of reality in the child.* New York: Basic Books.

Pletinckx, J., & Segers, M. (2001). Programme evaluation as an instrument for quality assurance in a student-oriented educational system. *Studies in Educational Evaluation, 27,* 355–372.

Poikela, E., & Poikela, S. (1997). Conceptions of learning and knowledge: Impacts on the implementation of problem-based learning. *Zeitschrift fur Hochschuldidactic, 21*(1), 8–21.

Prawat, R. S. (1999). Dewey, Peirce, and the learning paradox. *American Educational Research Journal, 36*(1), 47–76.

Reynolds, F. (1997). Studying psychology at degree level: Would problem-based learning enhance students' experiences? *Studies in Higher Education, 22*(3), 263–275.

*Richards, B. F., Ober, P., Cariaga-Lo, L., Camp, M. G., Philp, J., McFarlane, M., et al. (1996). Rating of students' performances in a third-year internal medicine clerkship: A comparison between problem-based and lecture-based curricula. *Academic Medicine, 71*(2), 187–189.

Rogers, C. R. (1969). *Freedom to learn.* Colombus, OH: Charles E. Merill.

*Santos-Gomez, L., Kalishman, S., Rezler, A., Skipper, B., & Mennin, S. P. (1990). Residency performance of graduates from a problem-based and a conventional curriculum. *Medical Education, 24,* 366–377.

*Saunders, N. A., McIntosh, J., McPherson, J., & Engel, C. E. (1990). A comparison between University of Newcastle and University of Sydney final-year students: Knowledge and competence. In Z. H. Nooman, H. G. Schmidt, & E. S. Ezzat (Eds.), *Innovation in medical education: An evaluation of its present status* (pp. 50–54). New York: Springer.

Savin-Baden, M. (2000). *Problem-based learning in higher education: Untold stories.* Buckingham, UK: Society for Research in Higher Education and Open University Press.

*Schmidt, H. G., Machiels-Bongaerts, M., Hermans, H., ten Cate T. J., Venekamp, R., & Boshuizen, H. P. A. (1996). The development of diagnostic competence: Comparison of a problem-based, an integrated, and a conventional medical curriculum. *Academic Medicine, 71,* 658–664.

Schoenfeld, A. H. (1985). *Mathematical problem solving.* San Diego, CA: Academic Press.

*Schuwirth, L. W. T., Verhoeven, B. H., Scherpbier, A. J. J. A., Mom, E. M. A, Cohen-Schotanus, J., van Rossum, H. J. M., et al. (1999). An inter- and intra-university comparison with short case-based testing. *Advances in Health Sciences Education, 4,* 233–244.

*Schwartz, R. W., Burgett, J. E., Blue, A. V., Donnelly, M. B., & Sloan, D. A. (1997). Problem-based learning and performance-based testing: Effective alternatives for undergraduate surgical education and assessment of student performance. *Medical Teacher, 19,* 19–23.

Segers, M. (1997). An alternative for assessing problem-solving skills: The overall test. *Studies in Educational Evaluation, 23*(4), 373–398.

Segers, M., Dochy, F., & Cascallar, E. (2003). *Optimizing new modes of assessment: In search of qualities and standards.* Boston/Dordrecht: Kluwer Academic.

Segers, M., Dochy, F., & De Corte, E. (1999). Assessment practices and students' knowledge profiles in a problem-based curriculum. *Learning Environments Research: An International Journal, 2,* 191–213.

Shavelson, R. J., Gao, X., & Baxter, G. P. (1996). On the content validity of performance assessments: Centrality of domain specification. In M. Birenbaum & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior learning* (pp. 131–143). Boston: Kluwer Academic Press.

Smith, D., & Hoersch, A. L. (1995). Problem-based learning in the undergraduate geology classroom. *Journal of Geological Education, 43,* 149–152.

Smith, M. U. (1991). *Toward a unified theory of problem-solving: Views from the content domains.* Hillsdale, NJ: Lawrence Erlbaum.

Smits, P. B. A., Verbeek, J. H. A. M., & De Buisonje, C. D. (2002). Problem based learning in continuing medical education: A review of controlled evaluation studies. *British Medical Journal, 321,* 153–156.

Snow, R. E. (1990). New approaches to cognitive and conative structures in education. *International Journal of Educational Research, 14*(5), 455–473.

*Son, B., & VanSickle, R. L. (2000). Problem-solving instruction and students' acquisition, retention, and structuring of economics knowledge. *Journal of Research and Development in Education, 33*(2), 95–105.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behaviors, 18,* 275–290.

Springer, L., Stanne, M. E., & Donovan, S. S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research, 6*(1), 21–51.

Sugrue, B. (1993). *Specifications for the design of problem-solving assessments in science: Project 2.1 designs for assessing individual and group problem-solving.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Sugrue, B. (1995). A theory-based framework for assessing domain-specific problem solving ability. *Educational Measurement: Issues and Practice, 14*(3), 29–36.

Swanson, D. B., Case, S. M., & van der Vleuten, C. M. P. (1991). Strategies for student assessment. In D. Boud & G. Feletti (Eds.), *The challenge of problem-based learning* (pp. 260–273). London: Kogan Page.

*Tans, R. W., Schmidt, H. G., Schade-Hoogeveen, B. E. J., & Gijselaers, W. H. (1986). Sturing van het onderwijsleerproces door middel van problemen: een veld-experiment [Guiding the learning process by means of problems: A field experiment]. *Tijdschrift voor Onderwijsresearch, 11,* 35–46.

*Van Hessen, P. A. W., & Verwijnen, G. M. (1990). Does problem-based learning provide other knowledge? In W. Bender, R. J. Hiemstra, A. J. J. A. Scherpbier, & R. P. Zwierstra (Eds.), *Teaching and assessing clinical competence* (pp. 446–451). Groningen, The Netherlands: Boekwerk Publications.

Van IJzendoorn, M. H. (1997). Meta-analysis in early childhood education: Progress and problems. In B. Spodek, A. D. Pellegrini, and O. N. Saracho (Eds.), *Issues in early childhood education: Yearbook in early childhood education.* New York: Teachers College Press.

*Verhoeven, B. H., Verwijnen, G. M., Scherpbier, A. J. J. A., Holdrinet, R. S. G., Oeseburg, B., Bulte, J. A., et al. (1998). An analysis of progress test results of PBL and non-PBL students. *Medical Teacher, 20*(4), 310–316.

Vernon, D. T. A., & Blake, R. L. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine, 68,* 550–563.

Verwijnen, M., Imbos, T., Snellen, H., Stalenhoef, B., Sprooten, Y., & Van der Vleuten, C. (1982). The evaluation system at the Medical School of Maastricht. In H. G. Schmidt, M. Vries, & J. M. Greep (Eds.), *New directions for medical education: Problem-based learning and community-oriented medical education* (pp. 165–179). New York: Springer.

*Verwijnen, M., Van der Vleuten, C., & Imbos, T. (1990). A comparison of an innovative medical school with traditional schools: An analysis in the cognitive domain. In Z. H. Nooman, H. G. Schmidt, & E. S. Ezzat (Eds.), *Innovation in medical education: An evaluation of its present status* (pp. 41–49). New York: Springer.

Authors

DAVID GIJBELS is a Researcher at the Centre of Excellence in Higher Education (ECHO), University of Antwerp, Lange Nieuwstraat 55, 2000 Antwerpen, Belgium;

e-mail *david.gijbels@ua.ac.be.* His research and development interests focus on the effectiveness of teacher training programs and on various aspects of new learning environments, such as problem-based learning and its assessment and evaluation. Correspondence should be addressed to him.

FILIP DOCHY is a Professor at the Research Center for Teaching and Training Methodology, University of Leuven, Dekenstraat 2, 3000 Leuven, Belgium; e-mail *filip.dochy@ped.kuleuven.ac.be.* He is also a Professor in the Department of Educational Innovation and Information Technology, University of Maastricht. He is currently president of the European Association for Research on Learning and Instruction (*www.EARLI.org*). His research interests are teacher training, new modes of assessment, learning arrangements, assessment engineering, and edumetrics.

PIET VAN DEN BOSSCHE is a doctoral student in the Department of Educational Development and Research, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands; e-mail *piet.vandenbossche@educ.unimaas.nl.* His research focuses on knowledge sharing and development in teams. As part of this broader interest he also conducts research on student-centered learning environments, such as problem-based learning.

MIEN SEGERS is a Professor of Educational Sciences in the Department of Educational Sciences, University of Leiden, Wassenaarseweg 52, Postbus 9555, 2300 RB Leiden, The Netherlands; e-mail *segers@fsw.leidenuniv.nl.* She is also a Professor in the Department of Educational Development and Research, University of Maastricht. She is currently coordinator of the Special Interest Group on Higher Education at the European Association for Research on Learning and Instruction. Her major research interests are the evaluation and optimization of student learning in learner-centered learning environments and the qualities of new modes of assessment in those environments.

APPENDIX

Studies measuring the knowledge structure

Study	Subjects PBL/conv	Method of assessment	Level of assessment	Result	
				ES	*p* value
Aaron et al., 1998	113/121	MCQ	1	**−0.440**	$p < .05$
	17/12	MCQ	1	**−0.769**	$p > .05$
	17/12	Essay questions	3	**0.0**	$p > .95$
Albano et al., 1996		Progress test	1	=	
Antepohl & Herzig 1997	110/110	Short answer	1	**+0.603**	$p = .0013$
Antepohl & Herzig 1999	55/57	MCQ	1	–0.125	$p = .4$
		Short answer questions	1	+0.424	$p = .07$
		Total	1	**+0.167**	$p = .43$
Baca et al., 1990	37/41	NBME Step 1	1	**−0.919**	$p < .0001$
Barrows & Tamblyn, 1976	10/10	Standardized patient simulation + MCQ	3	**+1.409**	$p < .005$
Bickley et al., 1990	23/	NBME Step 1	1	−	
	24/	Pathology		+	
	20/			+	
	24/			−	
	20/			+	
Block & Moore, 1994	62/63	NBME Part 1	1	=	n.s.
		Behavioral science subtest		+	sign.
	62/63	NBME Part 2	2	/	n.s.
		Public health		+	sign.
Boshuizen et al., 1993	4/4	1 case	2	**+2.268**	$p = .024$
Distlehorst & Robbs, 1998	47/154	USMLE Step 1	1	**+0.18**	$p = .6528$
	47/154	USMLE Step 2	2	+0.390	$p = .0518$
		Performance rating	2	+0.5	$p = .0028$
				+0.445	
		Standardized patient simulations:			
		Overall case score	3	+0.3	$p = .0596$
		Post station encounter	3	+0.33	$p = .0742$
		Patient checklist ratings	3	+0.14	$p = .1669$
				+0.26	

(continued)

APPENDIX *(Continued)*

Study	Subjects PBL/conv	Method of assessment	Level of assessment	Result	
				ES	*p* value
Donner et al., 1990		First try pass rate			
		NBME Part 1	1	=	
Doucet et al., 1998	34/29	MCQ (40 items)	1	**+0.434**	$p = .05$
	21/26	Key feature problem examination (28 cases)	2	**+1.293**	$p = .001$
Eisenstaedt et al., 1990	32/58	MCQ (traditional exam)	1	–8.291	$p < .001$
					$p < .001$
		MCQ	1	–2.211	
				–5.251	
Farquhar et al., 1986	40/40	NBME Part 1	1	–	n.s.
		Anatomy		+	n.s.
		Physiology		+	n.s.
		Biochemistry		+	n.s.
		Pathology		–	n.s.
		Microbiology		–	$p < .05$
		Pharmacology		–	n.s.
		Behavioral		+	n.s.
Finch, 1999	21/26	MCQ (60 items)	1		$p > .05$
	21/26	Essay questions	2	**+1.904**	$p < .0005$
Goodman et al., 1991	72/501	NBME Part 1	1	**–0.044**	$p = .40$
		Pathology	1	–0.242	$p = .01$
	12/12	Oral in 1985	1	**0.0**	$p = .97$
	15/13	Oral in 1987	1	**–0.667**	$p = .06$
	36/297	NBME Part 2	2	**–0.133**	$p = .73$
		Oral (problem solving)	2		
	12/12	In 1985	2	**–0.071**	$p = .89$
	15/13	In 1987	2	**+0.769**	$p = .04$
Hmelo et al., 1997	20/20	1 case	**2**	**+0.7305**	
		Length reasoning		+0.883	$p < .01$
		Use scientific concepts		+0.578	$p < .1$
Hmelo, 1998	39/37	6 cases	**2**	**+0.768**	
		Accuracy		+0.521	$p < .05$
		Length reasoning		+0.762	$p < .005$
		Number of findings		+0.547	$p < .05$
		Use scientific concepts		+1.241	$p < .001$
Imbos & Verwijnen, 1982		Progress test	1	=	

APPENDIX *(Continued)*

Study	Subjects PBL/conv	Method of assessment	Level of assessment	Result	
				ES	*p* value
Imbos et al., 1984		Anatomy (progress test)	1	Years 1 to 4:	=
				Years 5 and 6:	+
Jones et al., 1984	63/138	NBME Part 1	1		
		Overall		+	n.s.
		Anatomy		−	n.s.
		Physiology		+	$p < .05$
		Biochemistry		+	$p < .05$
		Pathology		−	n.s.
		Microbiology		−	n.s.
		Pharmacology		+	n.s.
		Behavioral		+	$p < .004$
	170/331	Subject matter part of "clerkships exams" (pretest)	1		
		Obstetrics		−	n.s.
		Pediatrics		−	$p < .01$
		Surgery		−	n.s.
		Medicine		−	$p < .05$
	60/142	FLEX weighted average	1	+	n.s.
Kaufman et al., 1989		NBME Part 1	1	−	$p < .0001$
		in 1983		−	$p < .05$
		1984		−	$p < .01$
		1985		−	$p < .1$
		1986		−	n.s.
		1987		−	$p < .1$
		1988		−	n.s.
		1989		−	$p < .05$
		1990		−	$p < .01$
		NBME Part 2	2	+	$p < .01$
		in 1983		−	n.s.
		1984		−	n.s.
		1985		+	$p < .1$
		1986		+	$p < .1$
		1987		+	n.s.
		1988		+	n.s.
		Overall 3rd-year Clerkship	3	/	n.s.
		in 1983		−	n.s.
		1984		−	n.s.

(continued)

APPENDIX *(Continued)*

Study	Subjects PBL/conv	Method of assessment	Level of assessment	Result	
				ES	*p* value
		1985		–	n.s.
		1986		+	n.s.
		1987		+	n.s.
		1988		+	n.s.
		1989		+	n.s.
		Clinical subscores of clinical rotations		+	signif.
		in 1983		–	n.s.
		1984		+	n.s.
		1985		/	n.s.
		1986		+	n.s.
		1987		+	$p < .1$
		1988		+	n.s.
		1989		+	n.s.
Lewis & Tamblyn, 1987	22/20	MCQ (100 items)	1	**+0.24**	$p = .479$
	22/20	Clinical performance	3	**+0.234**	$p = .2265$
Martenson et al., 1985	1651/818	Short answer	2	+	$p < .001$
		Short answer	2	–0.15	$p < .00003$
	1651/818	Essay questions	2	+0.15	$p < .00003$
				0.00	
Mennin et al., 1993	167/508	NMBE Part 1	1	**−7.908**	$p < .0001$
	144/447	NBME Part 2	2	**+0.046**	$p = .29$
	103/313	NBME Part 3	3	**+0.307**	$p < .001$
Moore et al., 1994	60/61	NBME Part 1	1		
		Anatomy		–0.257	$p = .16$
		Behavioral		+0.455	$p = .01$
		Biochemistry		–0.138	$p = .50$
		Microbiology		+0.323	$p = .10$
		Pathology		+0.029	$p = .89$
		Pharmacology		–0.037	$p = .71$
		Physiology		–0.159	$p = .43$
		Total		**−0.01**	$p = .96$
		Free recall (preventive medicine and biochemistry)	1	=	
	60/61	Diagnostic and clinical tasks in 1989	1	No difference	
		1990			

APPENDIX *(Continued)*

Study	Subjects PBL/conv	Method of assessment	Level of assessment	Result *ES*	Result *p* value
Morgan, 1977		Subjects 2nd year	1		
	15/82	NBME Part 1		+	
	15/76	NBME Part 1		+	
	16/81	NBME Part 1		+	
Neufeld & Sibley, 1989		First-try pass rate Exam Medical Council of Canada	1	–	
		Canadian specialty board examinations	1	+	
Patel et al., 1991	12/12	1 case	2	–	/
	12/12			–	/
	12/12			–	/
Richards et al., 1996	88/364	Clinical rating scale: (1) Amount of factual knowledge	1	**+0.5**	$p = .0001$
	88/364	Clinical rating scale:		+0.426	
		(1) Take history and perform physical;	2	+0.425	$p = .002$
		(2) Derive differential diagnosis;	2	+0.462	$p = .0005$
		(3) Organize and express information.	2	+0.390	$p = .004$
		NBME Medicine Shelf Test (Part 2)	2	**+0.3375** +0.073	$p = .80$
Santos-Gomez et al., 1990	41/78	Rating (supervisors)	3	+0.257	$p = .21$
	39/71	Rating (nurses)	3	–0.446	$p = .09$
	43/70	Rating (self)	3	+0.525	$p = .05$
				+0.112	
Saunders et al., 1990	45/243	MCQ	1	–0.716	$p < .001$
	47/242	MCQ		–0.476	$p < .01$
				−0.596	
	45/240	MEQ1	3		
	44/243	MEQ2		–0.066	n.s.
				+1.017	$p < .001$
				+0.4755	

(continued)

APPENDIX *(Continued)*

Study	Subjects PBL/conv	Method of assessment	Level of assessment	Result	
				ES	*p* value
Schmidt et al., 1996	Total = 612	30 cases	2	**+0.310**	/
Schuwirth et al., 1999	30/32	60 short cases	3	**+0.06**	
	30/30			**+0.25**	
	30/30			**−0.114**	
	29/30			**+0.238**	
	27/30			**+0.732**	$p < .01$
	32/25			**+1.254**	$p < .001$
Schwartz et al., 1997		MCQ Test 1	1	–	
		MCQ Test 2		/	
		MCQ Test 3		–	
		MCQ Final exam		/	
		Standardized patient simulation	3	+	/
		MEQ	3	+	/
		NBME II	2	/	n.s.
		Surgery subsection		+	sign.
Son & Van Sickle, 2000	72/68	Knowledge acquisition Instrument: 16 MCQ 8 correct/ incorrect 1 short answer	1	**+0.381**	$p = .05$
	72/80	Knowledge structure instrument: modified order tree technique	2	**+0.384**	$p = .05$
Tans et al., 1986	74/45	MCQ (60 items)	1	**−2.583**	$p = .0013$
	6/5	Free-recall test	2	**+2.171**	$p < .005$
Van Hessen & Verwijnen, 1990	/179	Progress test	1	–	$p < .05$
Verhoeven et al., 1998	190/124	Progress tests	1	**−0.203**	n.s.
	146/104			**+0.211**	n.s.
	135/87			**+0.288**	n.s.
	188/151			**+0.193**	n.s.
	144/140			**−0.037**	n.s.
	135/122			**−0.385**	$p < .01$

APPENDIX *(Continued)*

Study	Subjects PBL/conv	Method of assessment	Level of assessment	Result	
				ES	*p* value
Verwijnen et al., 1990	266/1253	MCQ (64 questions)	1	−	
	471/894	MCQ (70 questions)		−	
	565/1234	MCQ (64 questions)		/	
	565/167	MCQ (264 questions)			

Note. Effect sizes in bold refer to results from hypothesis tests with independent samples that are coded as if they were independent events. Study outcomes resulting from the same sample are aggregated within the separate categories of the influencing variable (level of assessment) and are also marked in bold. Subjects PBL/conv = number of subjects in the problem-based learning condition / number of subjects in the conventional instruction condition; *ES* = effect size; MCQ = multiple-choice questions; NBME = National Board of Medical Examiners; n.s. = not significant; sign. = significant; USMLE = United Nations Medical Licensing Examination; FLEX = Federated Licensing Examination; MEQ = modified essay question. The sign of the *ES* indicates whether the PBL result is greater (+) or smaller (−) than the result for conventional instruction conditions. If it was not possible to compute an *ES*, then only the sign of the results is given. A slash (/) indicates that no effect was found.