Problem-Based Learning: Outcomes Evidence From the Health Professions

Mark A. Albanese
University of Wisconsin - Madison

Laura Dast
National Conference of Bar Examiners

Over the past 30 years, problem-based learning (PBL) has become a major force in health professions education and even in the broader educational world. This article focuses on the outcomes that have been found from using PBL in the health professions based on at least 20 reviews done since 1990. The outcomes identified in these reviews are described as well as the strength of the evidence used in their support. These review results are augmented with results from selected articles that elaborate on how PBL can produce the identified outcomes.

The Outcomes of Problem-Based Learning

Background

The question that has dogged problem-based learning (PBL) since its inception is this: “To what degree does PBL produce the types of changes in learners that it was designed to produce?” These changes include whether they become self-directed learners who have a deeper knowledge of their discipline and who are better prepared to apply the science of medicine to patient care. Demonstrating these changes, however, has been a daunting challenge. Over the past 40 years, there have been hundreds of studies designed to test these differences and at least 20 major reviews since 1990 (Albanese & Mitchell, 1993; Berkson, 1993; Colliver,
Attempts to do systematic reviews of PBL’s effectiveness have either relied on a very small set of controlled studies (see Colliver, 2000; Newman, 2003; Vernon and Blake, 1993) or adopted a thematic approach (see Berkson, 1993), or what might be called a “best evidence approach” that used effect sizes when possible and a thematic approach when not (see Albanese and Mitchell, 1993). For those who are certain that PBL is the best instructional choice, these reviews have been disappointing. Newman (2003) summarized the earlier reviews as follows:

Vernon and Blake (1993) concluded that “results generally support the superiority of the PBL approach over more traditional academic methods.” Albanese and Mitchell (1993), while acknowledging the weaknesses of the research literature, concluded that PBL was more nurturing and enjoyable and that PBL graduates performed as well and sometimes better on clinical examinations and preceptor evaluations. However, the authors also concluded that PBL graduates showed potentially important gaps in their cognitive knowledge base, did not demonstrate expert reasoning patterns, and that PBL was very costly for institutions to begin and conduct. Berkson (1993) was unequivocal in her conclusion that “the graduate of PBL is not distinguishable from his or her traditional counterpart.” She further argued that the experience of PBL can be stressful for the student and faculty, and implementation may be unrealistically costly. (pp. 12-13)

Van den Bossche and colleagues (2000) determined that PBL had a positive, robust effect on the skills of students but a negative, non-robust effect on their knowledge. The review by Smits and colleagues (2002) concluded that “there was no consistent evidence that PBL is superior to other educational strategies in improving doctors’ knowledge and performance” (p. 13). Newman’s review was no more encouraging. Only for the outcome “accumulation of knowledge” were there more than three
studies that met the inclusion criteria. For this outcome, of 39 effect sizes computed, 16 favored the PBL group and 23 the control. Generally, the state of the literature was not sufficient for Newman to draw any firm conclusions. Dochy and colleagues (2003) performed a meta-analysis of 43 studies, concluding that PBL had a negative effect on the knowledge base of students (effect size = -0.776) but a positive effect on their application of that knowledge (effect size = +0.658). Gijbels and colleagues (2005) reported another meta-analysis of 40 studies in which they analyzed the effects of PBL as a function of the type of cognitive skill assessed in the outcome: concepts, principles and application. They found a slight negative effect size for concepts (-0.042) and positive effect sizes for principles and application (0.748 and 0.401, respectively).

Recent Reviews

More recent reviews of PBL have expanded the skills examined to include the performance of graduates in practice as well as expanding the health professions examined beyond medicine to others such as dentistry and nursing. Moreover, the reviews have gone global, with studies coming from Asia, South America, and the Middle East as well as established sources from North America and Europe.

Koh and colleagues (2008) performed a qualitative analysis of a literature search for evidence that PBL in medical school leads to greater physician competencies after graduation. Of the 2,675 articles identified in their search, only 13 met the full inclusion criteria. The researchers concluded that the strongest support for PBL was in its being superior to lecture for the competencies of teamwork skills, appreciation of social and emotional aspects of health care, appreciation of legal and ethical aspects of health care, and appropriate attitudes toward personal health and well-being. Moderate support for PBL was found for the competencies of communication and interpersonal skills, continuity of care, coping with uncertainty, the use of computers or information resources, and understanding evidence-based medicine.

Yuan, Williams, and Fan (2008) reviewed the literature to determine whether there is evidence that PBL can be used to promote critical thinking in nursing students. Computerized literature searches from 1990-2006 yielded only 10 studies that met the inclusion criteria. The authors concluded that the available evidence did not show PBL to be significantly better than conventional methods in developing nursing students’ critical-thinking skills. However, the most compelling study included in that review was a randomized clinical trial conducted by Tiwari and colleagues (2006.
First-year undergraduate nursing students at a university in Hong Kong were randomly assigned to parallel courses delivered using either PBL \((n = 40)\) or lecture \((n = 39)\) over one academic year. The primary outcome measure was students’ critical-thinking disposition as measured by the California Critical Thinking Disposition Inventory (CCTDI). Data were collected at four time points spanning three years. The two groups were not statistically different at the first time point, but from the first to the second time point, the PBL group showed significantly greater improvement in overall CCTDI \((P = 0.0048)\), and the subscores of truthseeking \((P = 0.0008)\), analyticity \((P = 0.0368)\) and critical thinking self-confidence \((P = 0.0342)\). From the first to the third time point, the significant scores were the CCTDI \((P = 0.0083)\) and the subscale scores truthseeking \((P = 0.0090)\) and analyticity \((P = 0.0354)\). From the first to the fourth time point, the significant subscale scores were truthseeking \((P = 0.0173)\) and systematicity \((P = 0.0440)\). The overall CCTDI score reached a maximum effect size of 0.56, a moderate value using Cohen’s criteria. Tiwari and colleagues (2006) concluded that there were significant differences in the development of students’ critical thinking between those who undertook the PBL and lecture courses, respectively, but it was not enough to convince Yuan and colleagues (2008) of the effect of PBL in their overall review.

Strobel and van Barneveld (2009) used a qualitative meta-synthesis approach to compare and contrast the assumptions and findings of the meta-analytical research on the effectiveness of PBL. They specifically examined how differences in the definition of and measurement of learning contributed to the inconclusiveness of the different meta-analyses. Considering these definitional complexities, they sought to determine what generalizable value statements about the effectiveness of PBL can be made and supported by the majority of the meta-analyses. What separates a meta-synthesis from a meta-analysis is that the analysis consists of a “translation” of different conceptualizations and comparisons instead of a quantitative analysis of treatment differences. Strobel and van Barneveld analyzed eight meta-analyses meeting study entry criteria. After conducting a qualitative summary, they grouped studies according to four levels of outcomes (non-performance or skill-oriented, knowledge assessment, performance or skill-oriented, and mixed knowledge and skill). They then subdivided each of the outcomes into the data source and/or method of assessment. For the main analysis, they formed a matrix indicating whether PBL was better (+), worse (-), or no different from the comparison group as a function of method of assessment, data source, and the four levels of outcomes for each of the eight meta-analyses.

One of the most interesting aspects of Strobel and van Barneveld’s
(2009) analysis is that they linked study outcomes to the task demands of the assessment. Measures requiring a recognition task (multiple choice items and variations) of short-term knowledge acquisition and retention, as well as those assessing understanding and application of important concepts of the sciences basic to the practice of medicine (for example, the United States Medical Licensing Examination, Step 1) tended to favor traditional learning approaches. Measures requiring a production task (for example, recall), particularly if assessing long term retention, tended to favor PBL. Performance ratings, such as those done by clinical supervisors, also favored PBL. Outcomes that required both knowledge and skill for performance such as oral examinations also favored PBL.

The researchers concluded that “PBL is significantly more effective than traditional instruction to train competent and skilled practitioners and to promote long-term retention of knowledge and skills acquired during the learning experience or training session” (p. 55). They further recommend that the focus of research should shift from determining PBL’s effectiveness to studying the differences in the effectiveness of support structures to find optimal scaffolding, coaching, and modeling strategies for successful facilitation of PBL.

It would be reassuring to conclude that by using a meta-synthesis approach, Strobel and van Barneveld (2009) have achieved the big picture missed by all of the component meta-analyses; however, such an assumption overlooks that all of these meta-analyses are incestuous, building upon earlier ones. Meta-syntheses were originally developed for qualitative studies; the outcomes examined are all idiosyncratic to the study context, and a plus/minus summary can help cut through the fog. In the case of Strobel and van Barneveld, each meta-analysis basically keeps reproducing what came before augmented by a relatively small number of new studies. In contrast to the temerity expressed by each meta-analysis in their consideration of individual studies, the consistent, but incestuous, findings across the meta-analyses looks like an unmistakable trend. So, while Strobel and van Barneveld’s recommendations to do more research on the support structures and coaching and modeling strategies for PBL are reasonable, the extent it is based upon their conclusion that PBL is significantly more effective than traditional instruction needs to be regarded with some skepticism.

Neville (2009) conducted a narrative review of the literature, focusing prominently on a synthesis of the various reviews that have occurred. Given that Neville was one of the original investigators from McMaster University when the use of PBL began, his review was both very deep and quite insightful, if not somewhat biased (he is quite willing to throw out a
study or two as outliers to support a point). He concluded that despite the heterogeneity of the literature, there is sufficient cognitive psychological evidence to validate the PBL approach to learning and a significant amount of empirical evidence in support of PBL having effective learner outcomes.

Hartling and colleagues (2010) conducted a systematic review under the auspices of the Best Evidence Medical Education group. Of 6,521 articles identified, 25 curriculum-wide studies and five course-based studies survived the study entry criteria. Owing to the failure of heterogeneity tests, the authors used a qualitative synthesis. They concluded that the existing literature provided inconsistent findings regarding the effectiveness of PBL compared to more traditional methods in undergraduate medical education. They also vigorously criticized the quality of the literature, a common practice of all reviews during this period.

Polyzois and colleagues (2010) performed a systematic review of PBL focusing on comparative studies within undergraduate education of health care professionals. They specifically integrated studies from outside medical education in their review. Of 1,233 articles initially found, 39 studies survived their selection criteria. These studies were grouped into three classifications as to whether their outcomes favored PBL, favored conventional instruction, or showed no difference. It is not clear exactly how the grouping process was accomplished or how reliably it was done. Regardless, in curriculum-wide studies, PBL was identified as more favorable in 16 studies, as no different in 15 studies, and as worse in two studies. Of seven single course studies identified, PBL was favored in six studies, with the remaining study showing no difference. Polyzois and colleagues concluded that PBL showed better results in developing critical reasoning, problem-solving abilities, and creativity. Further, PBL is described by students in two studies cited as a more enjoyable way of learning. This review was probably the most positive toward PBL of those conducted in the last 20 years, even though the large majority of the studies assessed were the same. The inclusion of studies of non-medical health professionals may be one factor. Unfortunately, effect sizes were not reported, nor were the methods used for grouping studies described in sufficient detail to determine if this was a factor.

In what was ultimately a follow-up to the review by Yuan and colleagues (2008), Oja (2011) reviewed the literature from 2006 forward to determine whether more recent research showed that PBL had effects on the critical-thinking skills of nursing students. An additional four studies were identified and the combination evaluated. The totality of five studies that reported effect sizes had values that ranged from 0.07 to 0.76. Applying Cohen’s criteria, three of the studies had small effect sizes, and
Oja concluded that the studies reviewed indicate a positive relationship between problem-based learning and improved critical thinking in nursing students. A critical point in this review was that the studies were international. The studies that showed the largest positive effect sizes came from Hong Kong and Shanghai. The smallest effect size was from a study conducted in Turkey. Given these differences and factoring in that adopting PBL curricula is often discussed as requiring a shift in institutional culture, there may be cultural differences that influence the extent to which PBL is successful. This possibility has not been explored.

An analysis of PBL as an instructional strategy must make a distinction between those studies conducted before 1993 and those conducted since. The three reviews that came out in 1993 apparently moved curricula toward what might be called a hybrid format—a combination of structured activities directed at giving students the disciplinary conceptual frameworks (usually via lecture) combined with substantial time devoted to doing PBL. At the risk of “cherry-picking” only the studies that have shown positive results, we will highlight what have been some of the findings of interest in the post-1993 time period.

One of the more compelling studies comes from 10 years of experience with PBL at the University of Missouri - Columbia. Hoffman and colleagues (2006) present results for the USMLE Steps 1 and 2 and residency director perceptions of their PBL graduates versus all other graduates for the period of one year pre-implementation of PBL (the 1996 graduating year) and post-implementation (1997-2006 graduating years). The performance of the classes on the MCAT was also presented to demonstrate to what extent USMLE performance diverged from the pattern of MCAT scores. Hoffman et al. presented their results as effect sizes with national means and standard deviations serving as comparison values for those from Missouri. The results were presented in graph form. We estimated values from the graphs to arrive at effect sizes for the year before PBL was implemented, the transition year (the first year of PBL), and the mean effect size over the 8-9 years during which PBL was in use. The results are shown in Table 1.

Using Medical College Admissions Test (MCAT) effect sizes as an index for students’ entering academic ability, one would expect the Step 1 and Step 2 effect sizes to be comparable unless the curriculum somehow affected students’ ability in a disproportionate way. In the pre-PBL year, the Step 1 effect size was almost identical to the MCAT effect size, which was approximately one quarter of a standard deviation below the national mean. The Step 2 effect size was even more negative, approximately one
half standard deviation below the national mean. There was a relatively
dramatic improvement in the transitional year that continued for the 8-9
years after PBL was introduced into the curriculum. In that period, the
MCAT effect sizes rose from -0.23 to an average of -0.06, a +0.29 rise. But
the changes in Step 1 and Step 2 effect sizes have been stunning. Step 1
increased from a -0.25 to an average of +0.30, a 0.55 increase. Step 2 in-
creased from -0.50 to an average of 0.38, a 0.88 increase. Thus, during the
PBL period, students’ Step 1 and Step 2 performance exceeded that which
would have been predicted from MCAT scores by over one quarter and
one half standard deviation, respectively. Even more impressive is that the
performance improvement has been sustained. The type of PBL used in
the curriculum at Missouri would probably be termed hybrid. There are
about 10 hours of lecture concurrent with equal time spent in PBL. This
exceeds the limit for lecture time that Barrows (1985) recommended (1.5
hours per day/7.5 hours per week). One other point is that concurrent
with implementing PBL, Missouri reduced entering class sizes from 112
to 96 students. While the smaller class size may have made the change
cost neutral (Albanese & Mitchell, 1993), it seems unlikely to account for
such a large and sustained difference in objective licensing test scores.

Schmidt and colleagues (2006) obtained self-ratings of professional
competence from a survey of 820/2081 (39%) of graduates of a PBL school
and 621/3268 (19%) of graduates of a traditional school in the Netherlands.
For interpersonal competencies such as working in a team, interpersonal
skills, and skills required for running meetings, the PBL graduates rated
themselves more skilled by an effect size of 1.30. For PBL-related competen-
cies such as self-directed learning, problem solving, and information
gathering, the PBL graduates rated themselves as more highly skilled by an
effect size of 0.78. For general academic competencies and task-supporting
competencies, the differences were 0.14 and 0.31, respectively—small but
still more positive for the PBL graduates.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pre-PBL</th>
<th>Transition</th>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCAT</td>
<td>-0.23</td>
<td>-0.18</td>
<td>-0.06</td>
<td>-0.17 to 0.07</td>
</tr>
<tr>
<td>Step 1</td>
<td>-0.25</td>
<td>-0.10</td>
<td>0.30</td>
<td>0.10 to 0.60</td>
</tr>
<tr>
<td>Step 2</td>
<td>-0.50</td>
<td>0.00</td>
<td>0.38</td>
<td>0.25 to 0.50</td>
</tr>
</tbody>
</table>
Schafer and colleagues (2006) reported a randomized trial of PBL versus traditional curricula regarding basic science and clinical knowledge. Students who had applied for the PBL track but due to limits were randomly assigned to the PBL track ($N = 122$) or to the traditional track ($N = 129$) and the remaining students in the traditional track ($N = 617$) were compared at three time points (the beginning of the first semester, the third semester, and the fifth semester) using a 200-item progress test (on third basic science, two thirds clinical). The results showed comparable gains by all groups on the basic science portion of the exam, but by the third administration, performance on the clinical section by the PBL students exceeded that of the other two groups by effect sizes greater than 1.17.

Albanese and colleagues (2006) examined the relationship of undergraduate science grade point averages (SGPA) to Step 1 failure rates for students in a PBL track and traditional track at one medical school and those at three other traditional medical schools. The relationship between the total MCAT score and Step 1 failure is shown in Figure 1.

While the overall failure rates for the three different groups were not different by a great margin (2%), the pattern by which the three different groups reached the overall failure rate was strikingly different for the PBL track. The traditional track and the three traditional medical schools had a relatively linear relationship between SGPA and Step 1 failure rate. Students in the PBL track, however, had almost no failures among those with SGPA values below 3.0. For SGPA values between 3.0 and 3.4, the PBL students had a much higher rate of failure that merged with the other two groups for SGPA values beyond 3.4 (of a maximum 4.0).

These results raise the concern that PBL may be better for some students, particularly those who have had relatively poor grades in their prior coursework (SGPA < 3.0), and worse for students whose grades were relatively good, but not stellar (SGPA = 3.0-3.4). These results need to be confirmed with other schools, but it may explain why the results from studies of PBL have been so variable. If some students do better and others do worse, effects will be difficult to find because they will cancel each other out. Another factor that may affect studies of PBL is that Schmidt et al. (2009) found the attrition rates of PBL schools in the Netherlands to be substantially below that of the conventional schools. Assuming academic problems are the main reason for student attrition, the higher rate of elimination of poorer-performing students in the conventional schools may inflate their outcomes, masking the overall poorer performance of students in the conventional curricula (Albanese, 2006).
Figure 1
Percent Below Cutpoint Who Failed: Science GPA (SGPA)
Summary

PBL use began when McMaster University’s medical school opened its doors in the late 1960s. It has since spread far beyond North America to almost every corner of the globe and to other health professions and many other disciplines. With their long history of PBL use, the health professions have perhaps the greatest wealth of research to draw upon to explore the effectiveness of PBL. While studies of the efficacy of PBL are not conclusive, there is some evidence of improved clinical knowledge and skills in PBL-trained students. It has been consistently found that students and faculty enjoy PBL, potentially reducing student attrition in some cases. However, the evidence for PBL is not as great as would be expected from its rapid spread, particularly since practically every review conducted has lamented the poor condition of the research literature.

The problems with the PBL literature cited include variable and often incomplete descriptions of the nature and extent of the PBL employed, voluntary assignment of learners to research treatments, poor and non-standard outcome measures employed, incomplete or inadequate presentation of results, potential confounding with cooperative assessment (a potent force in its own right), and the like. These weaknesses have led several investigators to forego plans to conduct meta-analyses in favor of approaches that did not employ combining effect sizes. These weaknesses also led us to highlight individual studies whose strengths or results set them apart and that perhaps deserve greater attention. Characteristics that set some studies apart for us are random assignment, use of standard instruments for assessment, analysis over multiple years or institutions, or results that seem to explain in greater detail a particular finding. These are the components that future PBL research needs to adopt.

There are, however, insights from the PBL health professions research that may benefit the larger education community. The most salient discovery was the greater satisfaction of students and faculty with PBL. In the Netherlands, PBL schools have had substantially lower attrition rates than traditional schools. This finding could be of special relevance to those making the push to improve retention in the science, technology, engineering and math (STEM) disciplines. In the Netherlands, selection for medical education is done by lottery for all students who obtain a minimum score on an entry examination. Furthermore, medical education begins after most students complete their basic education, or what is called high school in the U.S. Thus, it may be that using PBL in the STEM disciplines with students who are moving from basic to applied knowledge could reduce the high attrition rates.
Another take-home point is that cooperative learning assessment methods have been shown to be fairly potent in their own right. Incorporating them into PBL is likely to be value added and magnify the advantages intrinsic to PBL, regardless of assessment method. Also, in early studies of PBL, there were consistent concerns about incomplete knowledge structures for novice learners. These concerns dissipated after hybrid curricula were initiated that included more lecture early in the curriculum to establish better organization for assimilating the clinical material. There seems to be a tipping point in using organized versus student-determined activities, although it is not clear exactly where that point is, which dictates the need for further research.

Finally, one of the reasons it has been difficult to document the effects of PBL is that the effects may not be the same for all students. If our study of the PBL track holds up, it appears that PBL may benefit the lowest-performing students (science GPA < 3.0) at the expense of those who are in the next tier up (science GPA = 3.0-3.4). Perhaps PBL schools are directing disproportionately greater resources to students with science GPAs <3.0, compared with those directed at students in the 3.0-3.4 range. It could also be that students with science GPAs in the 3.0-3.4 range are in greater need of the structure provided by the conventional curriculum. If this is what is happening, the consequences are unclear. More research is needed in this area.

In closing, beginning a PBL curriculum is not for the faint-hearted. There is much infrastructure that needs to be put into place, and there may be increased costs. While the documentation of PBL’s effectiveness appears to be gaining ground, and we are gaining a better understanding about how to do it well, there is still much we need to learn.

References


ditional versus problem-based curriculum. Proceedings of the 2006 annual meeting of the Association for Medical Education in Europe, Cotone Congressi, Genoa, Italy (Abstract 10H2, p. 208).

**Mark Albanese**, Ph.D., is emeritus professor of educational psychology and population health at the Schools of Education and Medicine and Public Health at the University of Wisconsin – Madison and director of testing and research for the National Conference of Bar Examiners. Among other awards, he received the 1998 John P. Hubbard Award from the National Board of Medical Examiners for “significant contributions to the pursuit of excellence in the field of evaluation in medicine.” Albanese has authored over 125 articles in peer-reviewed publications. **Laura C. Dast** has been the curriculum development manager for University of Wisconsin School of Medicine and Public Health’s undergraduate medical education office for the past 13 years, and in that role, promoted curriculum mapping, integration of educational technology, and active learning. She has recently become manager of testing operations at the National Conference of Bar Examiners, where her interest in education and assessment continues.