Problem-based learning: why curricula are likely to show little effect on knowledge and clinical skills

Mark Albanese

Objectives A recent review of problem-based learning's effect on knowledge and clinical skills updated findings reported in 1993. The author argues that effect sizes (ES) seen with PBL have not lived up to expectations (0.8–1.0) and the theoretical basis for PBL, contextual learning theory, is weak. The purposes of this study were to analyse what constitutes reasonable ES in terms of the impacts on individuals and published reports, and to elaborate upon various theories pertaining to PBL.

Design Normal theory is used to demonstrate what various ESs would mean for individual change and a large meta-analysis of over 10 000 studies is referred to in identifying typical ESs. Additional theories bearing upon PBL are presented.

Results Effect sizes of 0.8–1.0 would require some students to move from the bottom quartile to the top half of the class or more. The average ES reported in the literature was 0.50 and many commonly used and accepted medical procedures and therapies are based upon studies with ESs below 0.50.

Conclusions Effect sizes of 0.8–1.0 are an unreasonable expectation from PBL because, firstly, the degree of changes that would be required of individuals would be excessive, secondly, leading up to medical school, students are groomed and selected for success in a traditional curriculum, expecting them to do better in a PBL curriculum than a traditional curriculum is an unreasonable expectation, and, thirdly, the average study reported in the literature and many commonly used and accepted medical procedures and therapies are based upon studies having lesser ESs. Information-processing theory, Cooperative learning, Self-determination theory and Control theory are suggested as providing better theoretical support for PBL than Contextual learning theory. Even if knowledge acquisition and clinical skills are not improved by PBL, the enhanced work environment for students and faculty that has been consistently found with PBL is a worthwhile goal.

Keywords Clinical competence, *standards; *curriculum; education, medical, undergraduate, *organization; KAP; problem-based learning, *organization; United States.

Medical Education 2000;34:729–738

Introduction

In a recent article, Jerry Colliver[1] reviewed the research on problem-based learning. My analysis of issues raised by Colliver will be divided into the following six sections:

1 What effect size should we expect?
2 The active ingredient of problem-based learning (PBL)
3 Theory from the near and far side
4 The new United States Medical Licensure Examination (USMLE)
5 What if PBL has no effect on knowledge acquisition and clinical skills?
6 Summary/conclusions.

1 What effect size should we expect?

An effect size refers to the difference between two (or more) treatment means (usually experimental group vs. control group) divided by an estimate of the standard deviation (usually combined within-group estimate). Essentially, it is the difference between group means expressed in standard deviation units. Values have the potential to range from negative to positive infinity, but rarely exceed 1·0 in absolute value. Cohen[2] labelled

University of Wisconsin Medical School, Madison, Wisconsin, USA
Based on a presentation at the American Educational Research Association, April 24–29 2000, New Orleans, Louisiana, USA
Correspondence: Dr Mark Albanese, Office of Medical Education, Research and Development, University of Wisconsin Medical School, 1300 University Avenue, Madison, Wisconsin 53706-1532, USA

values as 0·2 = small, 0·5 = medium, 0·80+ = large, but did not give any rationale for the labels. Colliver[1] cites Bloom[3] in his assessment of expectations for effect sizes with PBL. He argues that an effect size of half of the 2 standard deviations that Bloom identified with one-on-one tutoring should be reasonable (0·8–1·0). This recommendation would discount any effect size less than what Cohen[2] labelled as large. It raises the issue of whether effect sizes of less than 0·80 have been found to be useful.

Cohen's text was created primarily as a means of computing statistical power for planning sample sizes for research studies. The concept of effect size was useful because it allowed one to express expectations for group differences that were independent of the sample size or the units of the measure being used in the study. One advantage of using the large effect size as the minimum to have practical value would be that the sample sizes needed for research studies would be greatly reduced. For instance, for a two-group study to have a statistical power of 0·80 with a type 1 error rate of 0·05 and an effect size of 0·80, 26 subjects per group would be needed. This compares with 64 subjects if a medium effect size of 0·50 were adopted and 400 if a small effect size of 0·20 were anticipated. Clearly, adopting the Colliver[1] recommendation would substantially reduce the number of subjects needed to conduct research studies. However, what would we miss if such a stringent effect size were adopted? To address this question, it would be helpful to analyse what it takes in the way of change to obtain effect sizes of various values and to identify typical effect sizes reported in the literature.

In gaining an understanding of what an effect size of a given value means, it helps to examine how distributions of scores and individuals are influenced. For a small effect size (0·2) in normally distributed data, 85·3% of the control and experimental distributions would overlap. This declines to 67% for a medium effect size (0·5) and 52·6% for a large effect size (0·80).[2] As far as individuals are concerned, again under normal assumptions, an effect size of 1·0 would move the average student from the 50th percentile in the control condition to the 84th percentile. Table 1 shows how students at various points in the score distribution of a control group would be influenced by various effect sizes.

Table 1 Terminal percentiles* for selected effect sizes and initial percentiles

	Effect size				
Initial percentile	0·20	0·30	0·50	0·80	1·00
5th	7	9	13	20	26
10th	14	16	22	32	39
15th	20	23	30	41	48
25th	32	35	43	54	62
50th	58	62	69	79	84

* Terminal percentiles were computed by identifying the standard normal score which corresponded to the point which cut off the initial percentile of the normal distribution, adding the effect size to the standard normal score and then identifying the percentage of the normal distribution that fell below the new standard normal score. For instance, the standard normal score that corresponds to the initial 25th percentile was −0·685. To arrive at the terminal percentile for an effect size of 0·80, 0·80 was added to −0·685, to give +0·115. A standard normal score of 0·115 cuts off 54% of the normal distribution.

To meet Colliver's criteria, some subjects would need to make some very impressive gains. A subject initially at the 10th percentile would need to almost make it to the middle third. A subject at the 25th percentile would have to make it into the third quartile. These are clearly gains that would be impressive. However, even an effect size of 0·50 requires gains on the part of some subjects that would be impressive. For example, a subject at the median would have to move into the top third of the distribution. The issue becomes even more daunting if one considers the fact that ceiling effects often limit the ability of the highest performing subjects to achieve commensurate gains upon treatment exposure. This is especially a problem for locally made measures and competency-based assessments. Ceiling effects refer to the common occurrence where the top performing subjects receive the highest, or close to the highest, score on the measuring scale on the initial measurement. The potential for them to respond to treatment conditions is limited by the fact that they cannot go much higher. If ceiling effects operate, the lower-performing subjects must 'pick up the slack' created by the inability of the high performers to demonstrate their gains. If one thinks about an effect size as moving a population a given distance, even a small effect size is equivalent to 'moving a mountain'.

As to typical effect sizes, Lipsey & Wilson[4] examined 302 meta-analyses involving a total of 14 000 studies. The mean effect size they found was 0·5 – a medium effect size by Cohen's criteria. Lipsey & Wilson also examined the effect sizes associated with a number of medical interventions. As a few examples, antiarthritic drugs had effect sizes ranging from 0·45 to 0·77; anti-hypertensive drugs had effects on quality of life that ranged from 0·11 to 0·28, and by-pass surgery had an effect on angina of 0·80. Effect sizes associated with reduction in mortality were much smaller, from 0·08

Table 2 Effect sizes from Colliver[1]

Study	Measure	Results	Comments
Mennin *et al.* (1993)[5] New Mexico	NBME I NBME II NBME III	−0·85 −0·16 −0·33	Non-randomized had NBME III effect size = 0·33
Moore *et al.* (1994)[6] Harvard	NBME I	−0·01	Diagnostic reasoning and clinical problem-solving tasks showed no differences, and other comparisons were confounded by non-participation
Schmidt *et al.* (1996)[7] Three Dutch medical schools	Diagnostic ability scores	1–2% variance	Maximum effect size = +0·50
Richards *et al.* (1996)[8] Wakeforest	Med. Shelf Clinical ratings	+0·07 +0·39 to 0·50	
Hmelo (1998)[9] Rush	Pathophysiological explanation tasks	+0·8 to 2·36	Colliver questions whether this was teaching to the task, potentially easily remedial to traditionally taught students
Distlehorst & Robbs (1998)[10] SIU	Step 1 Step 2 Post clerkship SP exam Clerkship ratings	+0·18 +0·39 +0·30 +0·50	PBL students were older and had higher MCATs (effect size = 0·46)
Kaufman & [11] Mann (1998) Dalhousie	MCC Part I	+0·12 to 0·29	
Ripkey *et al.* (1998)[12]	Step 1 (compared 4 types of curricula adjusting for MCAT)	means: 209·1 to 210·8	PBL combined with other NBME non-standard curricula; did not separate out schools that changed over study period

NBME, National Board of Medical Examiners; PBL, problem-based learning; MCAT, Medical College Admissions test; SIU, Southern Illinois University; SP, Standardized Patients; MCC, Medical Council of Canada Qualifying Examination Part 1.

(streptokinase for myocardial infarction; chemotherapy for breast cancer) to 0·47 (AZT for AIDS). Thus, a medium effect size of 0·50 is well within the range of the effect of drug therapy for non-lethal outcomes. If only effect sizes of 0·80 or greater were considered to be of practical value, over half the psychological, educational and behavioural treatment literature and a number of drug therapies in common use (chemotherapy for breast cancer) would have been dismissed.

At this point, it would be useful to return to the results that Colliver[1] reported. Table 2 shows a summary of the findings from that study.

Of the eight studies, four showed effect sizes of 0·5 or more for at least one comparison. The results reported by Hmelo[9] actually exceeded Colliver's criteria. Colliver[1], however, discounted most of these by saying they used measures that were directly tied to the PBL intervention and with specific instruction, the control students would be easily able to achieve such gains. This certainly raises the bar to a level that is beyond what innovations have generally been required to hurdle.

Clearly, if we were to ignore gains that do not meet Colliver's criteria, we will be demanding much from any innovation. To determine if such expectations are reasonable from another vantage point, it is sometimes helpful to analyse the systems which intersect to produce a curriculum innovation such as PBL. The systems I will consider are: student selection, curriculum delivery and quality control.

Student selection system

Students admitted to medical school generally have demonstrated superlative achievement in lecture-based competitively graded courses. Those who support PBL often consider the traditional teaching methods to be outmoded relics of the past, dinosaurs if you will. If that is the case, medical students represent the tyrannosaurus rex (T. rex) of that Jurassic curriculum. They have not only survived a brutal 'Darwinian' selection process, but thrived. Expecting students who are selected through a process which ensures survival in a traditional curriculum to perform even better in a PBL curriculum seems like transporting a T. rex from the Jurassic period to modern times and expecting it to thrive in a petting zoo. After a few 'kiddie' meals, it should be clear that simply relocating a lean, mean killing machine to a more docile environment will not change its eating habits. Raising wild animals in captivity sometimes enables them to be adaptable to a more civilized environment. In other cases, it has required breeding generations of animals to have characteristics that make them more suitable for domestication. Similarly, medical students are a product of an entire educational system from grade school onwards, which has been largely like the traditional curriculum, lecture-based and competitively graded. Expecting students who have risen to the top after 15+ years of being cultivated and culled by a traditional educational approach, in what might be likened to an educational genetic engineering process, to suddenly excel in a different type of educational milieu seems to be overly optimistic. They can probably be domesticated to some degree by modifying the medical school admission process, but it will probably take a change in the entire educational process leading up to medical school before PBL (or any similar innovation) will be likely to reach its true potential.

Curriculum delivery system

The Jurassic curriculum has a number of delivery features which make it extremely resource-efficient. First, lectures maximize the efficient use of faculty resources. The faculty member who lectures has contact with all students (at least those in attendance) for the period that he/she is teaching. It may not be the kind of contact that all faculty would agree is optimal for learning, but contact it is. Faculty can also prepare for their lecture in an optimally efficient manner. They can organize their slides on the plane, train, bus or boat. With laptop computers, the options are limitless. Second, the amount of content delivered per unit of instruction time can be maximized if the lecturer engages in careful planning. With the continual exponential growth in the amount of biomedical knowledge, this is a very appealing feature of the lecture. Third, the curriculum governance system is aligned with the medical school governance system. It is generally left to departments to govern the content and delivery of the traditional curriculum. Since departments are the basic element of governance in most medical schools, the medical curriculum fits nicely in this larger governance structure.

A PBL curriculum, on the other hand, is not designed to achieve efficiency. Colliver[1] only accepted studies that dealt with curriculum-wide implementations of PBL, dismissing studies which implemented PBL within a course or in less broad-spectrum approaches. While I do believe PBL can be successfully implemented in a less than curriculum-wide mode, there are elements of PBL which make a broad implementation more likely to succeed. The most critical element that makes single-course implementations of PBL problematic is competition for student attention from courses operated in Jurassic mode. Faced with an imminent test or assignment deadline in a Jurassic course, students will often forgo activities in a PBL course.

If PBL is implemented curriculum-wide, it faces many challenges. First, the curriculum delivery system tends to be centrally governed due to its interdisciplinary nature. Since many medical schools are primarily governed by departments, a centrally governed curriculum often conflicts with the governance of the medical school. Resolving that conflict is no small problem. One of the many conflicts that this disconnect creates comes in the faculty reward system, with faculty often feeling inadequately rewarded for efforts devoted to small group instruction. Generally speaking, a curriculum governance system that is not aligned with the medical school governance system will require additional resources devoted to overcoming incompatibilities.

Studies devoted to estimating the faculty costs of PBL versus a traditional curriculum suggest that PBL may be as cost-effective or more so for class sizes up to 100.[13] This does not, however, include infrastructure costs of small group rooms, increased library utilization, availability of information technologies, etc. Restructuring the curriculum support systems for PBL offers many challenges.

The point of discussing curriculum delivery issues in considering why outcomes of PBL may not be as great as anticipated is that as long as the delivery system is more compatible with the traditional curriculum, there will be continual challenges to delivering a PBL curriculum in its optimal form. Developing compensatory methods to meet these challenges may make implementation take longer than it would if the delivery systems were in alignment. Thus, a PBL curriculum may take 3 or 4 years before evaluations of its effects can adequately reflect its success.

Quality control system

Changes in curriculum are often evaluated using performance measures designed for the previous curriculum. This places the new curriculum in the unenviable position of having to do what the old curriculum did, only better, in addition to meeting whatever added objectives exist for the new curriculum. This is a relatively tall order. In addition to the measures that are created by the institution, there are the licensure examinations. The United States Medical Licensure Examination (USMLE) is arguably the most influential quality control mechanism in the United States. The sensitivity of such an 'omnibus' examination to curriculum changes has been found to be relatively limited.[12] Thus, quality control mechanisms often place innovative curricula at a disadvantage in demonstrating their effectiveness.

2 The active ingredient of PBL

It may seem odd that although our review of 1993[13] generally concurred with the findings from Colliver[1] that I should be discussing the active ingredient in PBL. After all, something has to do something in order to be active. From Colliver,[1] one might conclude that there is no evidence of PBL doing anything. However, there is evidence that PBL does something. Perhaps the most compelling evidence of PBL doing something has been the rapid spread of PBL within and beyond the health professions.

The rise of PBL

In the years since we published our review,[13] PBL has undergone extraordinary dissemination. The percentage of medical schools reporting self-instruction (a surrogate for PBL, although it can also reflect other types of instruction) as an innovation increased from 79% in 1994–95 to 94% in 1998–99.[14] PBL has also expanded far from the health professions. The Center for Problem-Based Learning was established at Southern Illinois University. It maintains a list-serve which facilitates communication between individuals who are interested in PBL, primarily in the public schools. Beyond North America, the Problem-Based Learning Assessment and Research Centre (PBLARC) has existed at the University of Newcastle, Australia since 1987. It maintains a database listing publications from over 30 different disciplines. Finally, there are the rabid supporters of PBL. It induces an intensity that is analogous to a religious fervour. One need only read the enthusiastic interchanges on the PBL list-serve to get a sense of the excitement that PBL has generated.

Theoretical basis for PBL

One of the arguments used to support the superiority of PBL is the concept of contextual learning. The basic premise is that when we learn material in the context of how it will be used, it promotes learning and the ability to use the information. In PBL, the problem is usually portrayed in the real-life context of a patient coming to visit a doctor, or some variation. Colliver[1] criticizes the contextual learning argument on the grounds that it was drawn from a weak research finding and that almost all of clinical education occurs in the contextually relevant process of patient care. I do not disagree with Colliver's criticism of contextual learning theory as an argument for PBL and considered it to be one of the least compelling theories in support of PBL. There are other theories, however, that provide better support.

Information-processing theory

Schmidt (1983)[15] argues that information-processing theory underlies PBL. This theory involves three major elements: prior knowledge activation, encoding specificity and elaboration of knowledge. Prior knowledge activation refers to students using knowledge they already possess to understand and structure new information.

Encoding specificity refers to the fact that the more closely a situation in which something is learned resembles the situation in which it will be applied, the more likely it is that transfer of learning will occur. This aspect of information-processing theory resembles the contextual learning theory used to support PBL. However, in this case it is only part of a larger set of elements. The third element, elaboration of knowledge, refers to the fact that information will be better understood and remembered if there is opportunity for elaboration (discussion, answering questions, etc.). These three elements are commonly a part of PBL. They also have relatively strong documentation from the wider education and psychology literature. The fact that encoding specificity incorporates most of the salient features of contextual learning theory, suggests that information-processing theory provides a more comprehensive and parsimonious basis of theoretical support for PBL.

Cooperative learning

The use of cooperative learning (CL) is another concept that can be used to support PBL. CL situations are those where individuals perceive that they can reach their goals if and only if the other group members also do so. The small group format used in PBL often fits this definition. Qin *et al.*[8] conducted a meta-analysis of studies assessing the effect of cooperative vs. competitive learning on problem solving. Cooperation was operationally defined by the presence of joint goals, mutual rewards, shared resources, and complementary roles among members of a group. Competitive learning situations were those where individuals perceived that they could reach their goals if and only if the other participants could not attain their goals. Competition was operationally defined by the presence of a goal or reward that only one or a few group members could achieve by outperforming the others. Problem solving was defined by situations which required participants to form a cognitive representation of a task, plan a procedure for solving it, execute the procedure and check the results.

Their study spanned the twentieth century. Out of 63 findings, 52 occurred since 1970.

The authors concluded that overall

> *… cooperation resulted in higher-quality problem solving than did competition (effect size = 0·55). No differences were found between studies that focused on children or adults. The average person (at the 50th percentile) in the cooperation condition solved problems better than 72·5% of the participants in the competitive condition.*

If CL is the active ingredient that 'works', what is it about cooperative learning that enables it to work? One possible reason is that CL enables material to better mesh with students' level of cognitive development. Qin *et al.*[16] noted that in cooperative efforts, learners exchanged ideas and corrected each other's errors more frequently and effectively than did individuals competing with each other. It is possible that students who are struggling to understand the material are more likely to be able to identify the sources of other students' misunderstandings than is the expert instructor.

PBL systems often use CL as part of their process. If CL is one of the active ingredients of PBL, it will be helpful to determine whether CL is a part of the process in evaluating the effects of various PBL interventions. Perhaps one of the reasons for the ambiguous results from evaluations of PBL is the presence/absence of CL.

3 Theories from the near and far side

Self-determination theory

If information processing theory and cooperative learning theory provide support for various elements of the PBL process, there are other theories which address issues of motivation and behaviour that are also relevant. Williams *et al.*[17] argue that self-determination theory (ST), a theory which has been demonstrated to be effective in counselling and pharmacological treatments for patients who smoke or who have hypertension or coronary artery disease, has the potential to greatly improve education. ST distinguishes between two types of motivating conditions: controlled and autonomous. Controlled motivators are termed maladaptive and include external demands and contingencies as well as 'introjected regulation' which are internalized contingencies about what one 'should' do. These are all accompanied by either explicit or implicit rewards or punishments, or in the case of the introjected regulation, what are termed 'intrapsychic', or internal representations of rewards and punishments (self-aggrandizement and self-derogation). Under controlled forms of motivation, people act with a sense of pressure and anxiety. In educational situations, this takes the form of learning which is rote, short-lived, and poorly integrated into students' long-term values and skills. It does not take much of a leap to infer that many traditional curricula are steeped in controlled forms of motivation.

Autonomous motivators are those which are personally endorsed by the learner and reflect what the individual finds interesting and important. They include learners engaging in an activity simply because it is interesting and enjoyable as well as situations in which the learner has identified with its value for functioning as a physician. In contrast to the external element of the rewards and punishments administered in controlled motivating conditions, autonomous motivation involves behaving with a sense of volition, agency, and choice. Williams *et al.*[17] cite research which has demonstrated that, relative to controlled motivation, autonomous motivation for learning promotes greater conceptual understanding, better academic performance, higher academic achievement, stronger feelings of competence, enhanced creativity, a preference for optimal challenge over easy success, more positive feelings while learning, a tendency to cope more positively with failures and setbacks, greater persistence and better psychological adjustment.

A learning climate which promotes autonomous motivators includes one in which educators take the perspectives of students into account, provide relevant information and opportunities for choice, and encourage students to accept more responsibility for their own learning and behaviour. It also includes teachers' being meaningfully involved in students' learning through dialogue, listening, asking students what they want, providing satisfying rather than superficial replies to student-generated questions, providing factual information and advice and suspending judgement when soliciting the opinions and reactions of students. Such an environment minimizes pressure and control while encouraging a high level of performance. Clearly, autonomous motivators can be employed in either a PBL or traditional curriculum. However, autonomous motivators would seem to be especially compatible with collaborative learning environments. Further, PBL would seem to be an easier fit with autonomous motivators than would the traditional curriculum.

Control theory

If ST represents theory from the near side, control theory (Glasser[18]) might be considered theory from the far side. Control theory (CT) posits that all behaviours are based upon satisfying one or more of five basic needs. It further posits that one cannot make someone do something, especially learn, unless to do so satisfies some need of the person doing it. CT argues that people keep mental images in their head and attach them to which needs they satisfy. People feel pleasure when a need is satisfied and pain when a need is frustrated.

The five basic needs proposed by CT are: survive and reproduce; belong and love; gain power; be free, and have fun.

Survive and reproduce. Survival and reproduction are the genetic imperative. Educationally, it is hard to concentrate when one is worried about starving to death. Fortunately, this is a rare occurrence in medical education in the United States and other developed countries. However, some of the survival instincts ingrained in us all can manifest themselves in students who are on the verge of flunking out.

Belong and love. Children who have all their physical needs met but who lack love and attention fail to thrive, and can die of a peculiar starvation called marasmus. Without long-term parental care, no mammalian babies will survive. The social nature of the human experience has contributed to our dominance of the planet.

Gain power. Power can be used for good or evil, but CT considers it to be just one more genetic need. Examples of power needs include winning, pride, integrity, possessions, influence, appearance, etc. Most people have trouble going very long without complaining about someone who is frustrating their need for power.

Be free. People will go to great lengths to exercise freedom, even to their own disadvantage. Many academics value their freedom above remuneration. More generally, we desire the other four needs until they encroach too much on our need to be free.

Have fun. CT argues that learning is innately fun: just watch a child learning something new. A work environment where good-natured ribbing and joking is going on generally is more productive and satisfying. The immediate fun of learning is what keeps us going day by day.

Control theory in operation

CT argues that everything we do is initiated by a satisfying picture of that activity which we store in our heads as a pleasant memory. We choose our behaviours, including emotions. For instance, rather than saying one is depressed, one should say 'I am choosing to depress'. Behaviours are a person's best attempt to gain control of their lives, 'to reduce the difference between what we want at the time and what we see is available in the real world'. CT argues that one must meet one or more needs in order to influence another person. The reason instruction fails is because it fails to meet the needs of the learner. It would seem that small-group PBL satisfies the following needs:

- Freedom: less structured time, students can choose when to meet and what they are to do next.
- Power: students have the power to determine their learning needs.
- Love and belonging: small groups and facilitators become much more personally involved with one another than in lectures.
- Fun: the most dominant finding is that both students and faculty enjoy PBL.
- Survive and reproduce: students helping students promotes survival.

To summarize, the theoretical development of PBL has been relatively weak as noted by Colliver[1]. This does not mean that there is not a wealth of theory that applies. This section described four theories that seem to have ready application to PBL and offer opportunities for research on PBL to determine what 'makes it tick'. At least two of the theories offer some insights as to why PBL has not shown uniformly similar results (collaborative learning vs. competitive learning; controlled vs. autonomous motivation). There are likely to be other theories that are also relevant. The challenge for research is to delineate which theories seem to define effective PBL and, perhaps, education more generally.

4 The new United States Medical Licensure Examination (USMLE)

With the introduction of the USMLE in 1993, the nature of the licensure examination has changed. The questions have increasingly taken on a clinical context. While the clinical context would previously have been considered 'window dressing', it is now considered to be an important part of assessing competence in biomedical education. To determine whether the USMLE is sensitive to the kinds of effects promoted by PBL and other innovative curricula, Ripkey *et al.*[13] studied USMLE performance according to the type of curriculum at medical schools in the US and Canada. Their results suggested that Step 1 performance is fairly insensitive to curricular differences. However, PBL curricula were grouped with other types of innovations, weakening any attempt to generalize the results to PBL. Further, the period studied was a time when many schools were changing to include PBL in their curricula. The 1995–96 AAMC curriculum directory served to classify schools, but the sample cohorts entered medical school during the period 1992–95. Any school which had changed its curriculum in the 1992–95 period would have at least some of its students misclassified, further limiting the ability to interpret the results.

While the Ripkey *et al.* results still leave questions as to the sensitivity of the USMLE to differences brought about by PBL, Blake *et al.*[19] reported on one of the first studies to use the USMLE as a criterion for evaluating the effectiveness of a new PBL curriculum at the University of Missouri. They examined student performances on Steps 1 and 2 of the USMLE for two classes of medical students who completed their traditional curriculum (1995–96 graduating classes) in comparison with four classes who experienced their recent PBL curriculum (1997–2000). The authors report mean scores for each class as well as those for all exam takers in the US and Canada. The mean Step 1 scores for the traditional curriculum were both below the mean for North America and had an average effect size of −0·19. After the implementation of PBL, three out of the four means exceeded the North American mean, producing an average effect size of +0·20. The only negative effect size occurred for the first year in which the new curriculum was implemented (−0·05). For the last 2 years, the effect sizes exceeded +0·30. A similar result was obtained on the Step 2 exam. The mean performances of the classes experiencing the traditional curriculum were below the North American mean, with effect sizes which averaged −0·23. The mean performance of classes experiencing the PBL curriculum all exceeded the North American means and had an average effect size of +0·29.

To determine whether the results could be attributed to changes in the composition of the classes, data on the undergraduate GPAs (Grade Point Averages) and Medical College Admissions Test (MCAT) were reported. The MCAT scores were all at or below the US mean while GPAs were above the national mean. While the GPAs leave some residual doubt, the below average MCATs suggest that the change in the results for the USMLE exams with the implementation of PBL is unlikely to be due to changes in the class composition. One potential explanation for these findings could be Hawthorne effects, where subjects improve their performance simply because they are participating in an experiment. While it cannot be totally ruled out, given the multiyear nature of the data reported and the continued improvements, even in the last year of the study, it is unlikely that Hawthorne effects can be considered the cause. Further, a meta-analysis of 86 studies in which Hawthorne effects were controlled concluded that 'there was no evidence of an overall Hawthorne effect' (Adair *et al.*).[20] Based on anecdotal reports I have heard from other institutions, and trends in data (i.e. Way *et al.*[21] show a difference in Step 1 scores in 1995 which suggest that PBL and independent study pathway students scored higher than students in the lecture-discussion pathway), it is possible that there will be additional studies showing results similar to those reported by Blake *et al.*[19]

5 What if PBL has no effect on knowledge acquisition and clinical skills?

For a moment, let us assume that PBL has no effect on either knowledge acquisition or clinical skills: what then? There is reasonably strong evidence that PBL has an effect on the learning environment. Woodward *et al.*[22] used records from the Canadian health care system to follow graduates of McMaster, a PBL medical school, in comparison with graduates of traditional medical schools as they entered medical practice. The authors analysed differences in how and where they practised medicine as well as their efforts at furthering their medical education. The results showed that McMaster graduates were more likely to: spend time in direct patient care, bill for more psychotherapy services per month, have an academic appointment, enter family medicine and be in group practice. The common thread in these findings seems to be a greater desire to affiliate. This is consistent with the finding that both students and faculty in PBL schools enjoy the educational process more than those in traditional schools. This latter finding was found in all three reviews published in 1993. Perhaps it is a more humane learning environment that promotes collegial interactions which is the ultimate outcome of PBL. If that is the case and we can show no reduction in learning, is that so bad? With the exception of the few masochists in the group, it would seem likely that students who enjoy their medical education experience and their interactions with their peers and instructors would be more likely to engage in lifelong learning. As long as the costs do not overtax the medical school resources, it would seem that investing in a more positive workplace environment for both faculty and students would be a worthwhile goal in and of itself.

6 Summary/conclusions

It is hard to argue with the numbers reported by Colliver.[1] They are consistent with the results obtained in three reviews published in 1993[13,23,24] and they do not add much to what is already known about the effects of PBL on knowledge acquisition and clinical skills. Where we diverge is in the interpretation of the results and their implications.

While Colliver[1] argues that the effect sizes seen with PBL have not lived up to reasonable expectations (0·8–1·0), are those expectations reasonable? Effect sizes of that magnitude would require some students to move from the bottom quartile to the top half of the class. Even greater movement would be required if the measure used is subject to ceiling effects. Further, data are provided which suggest the average effect size reported in the literature is 0·50, and many very commonly used and accepted medical procedures and therapies are based upon studies that had effect sizes even below 0·50. There are also systems issues that are likely to attenuate effect sizes, including the fact that students are groomed for success in a traditional curriculum. Why should they be expected to do better in a PBL curriculum than in the one in which they have been 'genetically engineered' to excel? PBL also has to overcome hurdles because it generally requires a centralized organizational structure instead of the decentralized organizational structure common to medical school governance. In addition, quality control measures used to measure curricular success have generally been designed for the curriculum being replaced, placing a demand on the PBL curriculum to not only do better than the former curriculum in meeting the old goals, but to meet the new curriculum goals as well. Given these hurdles, it is not surprising that PBL has not generated the kinds of effect sizes that Colliver and others had expected.

Colliver[1] also argues that the theoretical basis for PBL, contextual learning theory, is weak. I concur with his observation, but offer four other theories that offer promise for better explanation and prediction of what elements of PBL are effective, including information-processing theory, cooperative learning, self-determination theory, and control theory. Much research needs to be done to really understand what is happening with PBL.

However, change may be in the offing. The results of a recent study (Blake *et al.*[19]), one of the first using the USMLE as an outcome measure, appear to demonstrate fairly compelling gains in scores. As more results are reported using the USMLE as the outcome, we may see some of the long hoped-for results demonstrating positive gains in test performance with PBL curricula.

Finally, results of a compelling study utilizing the database from the Canadian health care system suggest that PBL graduates are more likely to seek affiliation; bolstering the consensus finding from three 1993 reviews of PBL that students and faculty enjoy PBL more than traditional teaching methods. I conclude by arguing that even if knowledge acquisition and clinical skills are not improved by PBL, enhancing the work environment for students and faculty is a worthwhile goal in and of itself.

References

1 Colliver J. Effectiveness of problem based learning curricula. *Acad Med* 2000;**75**:259–66.
2 Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* Rev. edn. Englewood Cliffs, New Jersey: Erlbaum; 1977.
3 Bloom BS. The 2-sigma problem: the search for methods of group instruction as effective as one-on-one tutoring. *Educ Res* 1984;**13** (6):4–16.
4 Lipsey MW, Wilson DB. The efficacy of psychological, educational and behavioral treatment. confirmation from meta-analysis. *Amer Psychol* 1993;**12**:1181–209.
5 Mennin S, Friedman M, Skipper B, Kalishman S, Snyder J. Performances on the NBME I, II, and III by medical students in the problem-based learning and conventional tracks at the University of New Mexico. *Acad Med* 1993;**68**:616–24.
6 Moore GT, Black SD, Style CB, Mitchell R. The influence of the New Pathway Curriculum on Harvard medical students. *Acad Med* 1994;**69**:983–9.
7 Schmidt HG, Machiels-Bongaerts M, Hermans H, ten Cate TJ, Venekamp R, Boshuizen HPA. The development of diagnostic competence: comparison of a problem-based, an integrated, and a conventional medical curriculum. *Acad Med* 1996;**71**:658–64.
8 Richards BF, Ober P, Cariaga-Lo L, Camp MG, Philip J, McFarlane M, Rupp R, Zaccaro DJ. Ratings of students' performances in a third-year internal medicine clerkship: a comparison between problem-based and lecture-based curricula. *Acad. Med.* 1996;**71**:187–9.
9 Hmelo CE. Cognitive consequences of problem-based learning for the early development of medical expertise. *Teaching Learning Med* 1998;**10**:92–100.
10 Distlehorst LH, Robbs RS. A comparison of problem-based learning and standard curriculum students. three

years of retrospective data. *Teaching Learning Med* 1998;**10**:131–7.

11 Kaufman DM, Mann KV. Comparing achievement on the Medical Council of Canada Qualifying Examination Part I of students in conventional and problem-based learning curricula. *Acad Med* 1998;**73**:1211–3.

12 Ripkey DR, Swanson DB, Case SM. School-to-school differences in Step 1 performance as a function of curriculum type and use of Step 1 in promotion/graduation requirements. *Acad Med* 1998;**73** Suppl 10:S16–S18.

13 Albanese MA, Mitchell S. Problem-based learning: a review of literature on its outcomes and implementation issues. *Acad Med* 1993;**68**:52–81.

14 Association of American Medical Colleges. Curriculum Directory 1998–99. Washington, DC: AAMC; 1999.

15 Schmidt HG. Problem-based learning: rationale and description. *Med Educ* 1983;**17**:11–16.

16 Qin Z, Johnson DW, Johnson RT. Cooperative versus competitive efforts and problem solving. *Rev Educ Res* 1995;**65** (2):129–43.

17 Williams GC, Saizow RB, Ryan RM. The importance of self-determination theory for medical education. *Acad Med* 1999;**74** (9):992–5.

18 Glasser W. *Control Theory in the Classroom*. New York: Harper and Row; 1986.

19 Blake RL, Hosokawa MC, Riley SL. Student performances on Step 1 and Step 2 of the United States Medical Licensing Examination following implementation of a problem-based learning curriculum. *Acad Med* 2000;**75** (1):66–70.

20 Adair JG, Sharpe D, Huynh C-L. Hawthorne control procedures in educational experiments: a reconsideration of their use and effectiveness. *Rev Educ Res* 1989;**59**:215–28.

21 Way DP, Giagi B, Clausen K, Hudson A. The effects of basic science pathway on USMLE Step 1 scores. *Acad Med* 1999;**74** Suppl 10:S7–S9.

22 Woodward CA, Ferrier BM, Cohen M, Goldsmith C. A comparison of the practice patterns of general practitioners and family physicians graduating from McMaster and other Ontario medical schools. *Teaching Learning Med* 1990;**2** (2):79–88.

23 Berkson L. Problem-based learning. Have the expectations been met? *Acad Med* 1993;**68** Suppl 10:S79–S88.

24 Vernon DTA, Blake RL. Does problem-based learning work? A meta-analysis of evaluative research. *Acad Med* 1993;**68**:550–63.

Received 19 *June* 2000; *accepted for publication* 19 *June* 2000