Is Project Based Learning More Effective than Direct Instruction in School Science Classrooms?

An Analysis of the Empirical Research Evidence

by

Clifford Dann

A Thesis Submitted to the Faculty of Graduate Studies

In Partial Fulfillment of the Requirements for the Degree of

MASTER OF EDUCATION

Department of Curriculum, Teaching and Learning

University of Manitoba

Winnipeg, Manitoba, Canada

Abstract

An increasingly loud call by parents, school administrators, teachers, and even business leaders for “authentic learning”, emphasizing both group-work and problem solving, has led to growing enthusiasm for inquiry-based learning over the past decade. Although “inquiry” can be defined in many ways, a curriculum called “project-based learning” has recently emerged as the inquiry practice-of-choice with roots in the educational constructivism that emerged in the mid-twentieth century.

Often, project-based learning is framed as an alternative instructional strategy to direct instruction for maximizing student content knowledge. This study investigates the empirical evidence for such a comparison while also evaluating the overall quality of the available studies in the light of accepted standards for educational research. Specifically, this thesis investigates what the body of quantitative research says about the efficacy of project-based learning vs. direct instruction when considering student acquisition of content knowledge in science classrooms. Further, existing limitations of the research pertaining to project based learning and secondary school education are explored. The thesis concludes with a discussion of where and how we should focus our empirical efforts in the future.

The research revealed that the available empirical research contains flaws in both design and instrumentation. In particular, randomization is poor amongst all the studies considered. The empirical evidence indicates that project-based learning curricula improved student content knowledge but that, while the results were statistically significant, increases in raw test scores were marginal.

Acknowledgements

This project was not a singular effort. Many, many people made significant sacrifices to see this thesis through to its completion. I don't think that it is possible for one to imagine the amount of coordination, effort, and especially discipline that is required to embark upon and finish a graduate degree…and that's from the people supporting the student! I had no idea how much work it would be and I never would have finished without the help of my family, friends, and mentors.

The first person I owe thanks to is my advisor Dr. Don Metz who has been patient and supportive throughout this (seemingly never-ending) process. He has been practical and pragmatic in helping me get to the end. Dr. Barbara McMillan, Dr. Rod Clifton and Dr. John Long have provided unlimited support. And I mean unlimited. Endless revisions are necessary in the writing process to achieve a satisfactory final product and they must have read, re-read, marked-up and patiently explained thousands of words worth of work. They are, all three, master teachers.

My wife Sarah and my son Eddy have held down the fort while I have been away, locked in an office for months. Sarah has been a single parent this whole time balancing the needs of a toddler with absolutely every other domestic duty in addition to her own, demanding professional work. She earned this degree as much as I did. My mother-in-law, Deloris Long, has cooked me about a thousand meals and given up her home to me so that I could have a quiet place to work. Her own work ethic is inspiring. Finally, I am grateful to my parents, Cheryl and Terry Dann, who have never backed down from a challenge, ever, and taught their kids to do the same.

TABLE OF CONTENTS

TABLES

Table

Chapter 1

Introduction

Background and Significance of the Study

I have taught general science, physics, and chemistry in Winnipeg, Canada since 2004 and, since 2005, have taught at a large, urban high school in south Winnipeg. With over 1200 students and 100 staff, our school is the largest in its division. A major shift in my school's culture occurred in 2010 when we adopted a model of teaching and learning advocated by The Partnership for 21st Century Skills (hereafter P21) (2011). The school's new direction was formalized within the school plan and it was made clear to teaching staff that a coherent vision of learning between P21 and the school was the new operational norm. Accompanying this new allegiance was a tacit approval of constructivist ideology and teaching practices, primarily inquiry-based learning and, specifically, project-based learning (hereafter PBL), associated with that movement.

Founded in 2002, P21 (2002) was formed "as a coalition bringing together the business community, education leaders, and policymakers to position 21st century readiness at the center of (United States) K-12 education, and to kick-start a national conversation on the importance of 21st century skills for all students" (Partnership for 21st Century Skills, 2011). The Partnership for 21st Century Skills advocates school curricula focused on reading, writing, and arithmetic framed within the '21st century skills' of critical thinking and problem solving, communication, collaboration, and creativity and innovation collectively known as the 'four Cs'. The P21 focus is rooted in a belief that, among other deficiencies, U.S. public education is not consistently producing graduates

capable of working effectively in globally competitive economies leaving the economy of the United States vulnerable to international economic pressure. In fact, the report says:

> Every child in America needs to be ready for today's and tomorrow's world. A profound gap exists between the knowledge and skills most students learn in school and the knowledge and skills they need for success in their communities and workplaces. To successfully face rigorous higher education coursework, career challenges and a globally competitive workforce, U.S. schools must align classroom environments with real world environments by fusing the three Rs and four Cs. (P21 Common Core Toolkit, 2011)

Although a US based organization, the 21st century learning movement has a strong following in Canada. In 2012 Canadians for 21st Century Learning and Innovation, an organization strikingly similar to P21, issued a report, *Shifting Minds: A 21st Century Vision for Public Education in Canada* (Canadians for 21st Century Learning and Innovation, 2012) which outlines a vision of public education in Canada similar to P21's for the American public system. Taking aim at elementary through secondary education, these organizations frame the problem with the public education system as two-fold: not only are current curricula not capable of producing graduates who can compete successfully on an international stage, but interventions commonly used in classrooms, namely Direct Instruction (hereafter DI), are insufficient to effectively teach the new, 21st century, competencies in the reformed curricula. Tim Magner, former executive director of P21 articulated the organization's position in a recent interview:

> The idea that you can get by in this world with just knowledge I think misses the point entirely. In a pre-Google world I think you could make a case that acquisition of information had its own reward. [Today] there is a foundation of content knowledge that students need to have but there is also a much richer cognitive process that students need to have to be able to analyze, synthesize, and to apply what they know in the real world. From our perspective that is much more rigorous educational articulation. To us it's not about what students know but what students can do with that knowledge. (DeWitt, 2012)

It is this 'richer cognitive process' that the P21 camp has identified as an essential component of the 21st century classroom, a classroom that must provide students with the opportunity to learn not only content as prescribed by government curricula (knowledge) but the more nebulous (as P21 argues) no less important, 'four Cs' (skills). P21 views PBL as fitting the bill in this regard and has teamed with experts in 21st century learning such as the Buck Institute for Education (hereafter BIE) who remain dedicated to advancing the use of PBL in classrooms. These experts and organizations argue vigorously in favour of project based teaching practice and paint PBL as a necessity for producing graduates capable of meaningfully applying knowledge. In the words of the BIE, "…forty years of accumulated evidence that the instructional strategies and procedures that make up standards-focused PBL are effective in building deep content understanding, raising academic achievement, and encouraging student motivation to learn" (Buck Institute for Education, 2009).

Origin and Development of Project-Based Learning: As if the Space Race Mattered

Although PBL has more recently found its way into mainstream classrooms, in a real and practical sense, we can find evidence that it had become a compelling idea to leaders in the 1960s. For instance, in a speech in 1962, President J. F. Kennedy said:

> But if I were to say, my fellow citizens, that we shall send to the moon, 240,000 miles away from the control station in Houston, a giant rocket more than 300 feet tall, the length of this football field, made of new metal alloys, some of which have not yet been invented, capable of standing heat and stresses several times more than have ever been experienced, fitted together with a precision better than the finest watch, carrying all the equipment needed for propulsion, guidance, control, communications, food and survival, on an untried mission, to an unknown celestial body, and then return it safely to Earth, , re-entering the atmosphere at speeds of over 25,000 miles-per-hour, causing heat about half that of the temperature of the sun…and do all this, and do it right, and do it first before this decade is out, then we must be bold (Kennedy, 1962).

John Kennedy was a gifted orator. He needed to be. He was president of the United States during a fractious time in that country's history. A time that required the unification of disparate groups split along religious, racial, and ideological lines. Although slavery had been outlawed almost one hundred years earlier, extreme tensions existed in many states between the African American population and whites. Many

schools and other public places were still segregated. Although the allies' triumph in World War II had been total, America had in effect fought two wars, one in Europe and one in the Pacific. The emotional, physical, and economic toll had been high: hundreds of thousands of soldiers had been killed and billions of dollars borrowed to pay for victory. And then, discouragingly, a new enemy, communism, had appeared even as Marshall executed a bold plan for Europe's reconstruction and MacArthur wiped the slate clean on 2,000 years of militaristic culture in Japan.

However, World War II and the communist threat were not all bad for America. Kennedy understood how to use these two events, one receding into history the other in the present, to bring Americans together. He grasped that, for America to keep moving forward, to stay a world leader in science and technology, to maintain global economic dominance, he would have to push the nation's scientists, engineers, and business leaders towards audacious challenges that would force existing abilities to be stretched and new skills to be developed and tested. Kennedy wrapped this belief around the communist threat, clearly outlining his philosophy of American's challenge during the first ever presidential candidates' debate with Richard Nixon on September 26, 1960. Referring to the global rise of communism, Kennedy challenged the citizens of the United States:

> In the election of 1960, and with the world around us, the question is whether the world will exist half-slave or half-free, whether it will move in the direction of freedom, in the direction of the road that we are taking, or whether it will move in the direction of slavery. I think it will depend in great measure upon what we do here in the United States, on the kind of

> society that we build, on the kind of strength that we maintain (Kennedy, 1960).

The nascent space race provided the perfect theatre for Kennedy's battle between slavery and freedom. At the time, Americans were terrified by the Soviet Union's early dominance of space. In 1957 the Soviets launched 'Sputnik' the first human-made satellite and then, in 1961, Yuri Gagarin became the first person to orbit the Earth. By contrast, the first American satellite was not launched until 1958, and although Alan Shepard also went into space in 1961 it was after Gagarin and was a parabolic flight without any orbits of the Earth. America lagged behind the Soviets and her people knew it.

Throughout history, grand, national projects, whether dedicated to the gods (Giza's pyramids, Europe's cathedrals) or engineering works to serve the public good (Roman aqueducts and roads), have often been used by leaders as unifying events meant to bind their peoples and give them a sense of common purpose. Kennedy's Apollo program was no different and was, in part, designed to tap into the capitalist traits of ingeniousness and entrepreneurial acumen; two national qualities of which Americans are still intensely proud 60 years later. Further, he understood that although stories of individuals overcoming adversity are seductive, they do not resonate with the same power as tales of people who come together to find a common aim, working towards the solution to a problem that resonates not just with the individual but with the citizenry as a whole.

As a member of ‘the greatest generation’ he had experienced and participated in the ‘project’ of defeating the Nazis and Imperial Japan, where a group of republics (both soviet and democratic), kingdoms, and constitutional monarchies put aside their ideological differences to defeat a common enemy. Further, he understood that there were rewards to undertaking these projects that stretched beyond victory. In his Rice University speech, Kennedy makes reference to the ancillary benefits America could hope to enjoy “the growth of our science and education will be enriched by new knowledge of our universe and environment, by new techniques of learning and mapping and observation, by new tools and computers for industry, medicine, the home as well as the school” (Speech at Rice University, September 12, 1962)

The space race, for both the soviets and America was different than the great projects of the past. The Egyptians had thousands of years of practice building pyramids before their masterpieces at Giza were constructed. Likewise, the cathedral builders of medieval Europe sometimes took over a century to complete just one building. While these accomplishments are certainly impressive and required many unique solutions to unexpected problems, consider that only 25 years passed between the first V2 launches (the forerunner of the Saturn V) and Neil Armstrong’s footsteps on the moon. This is a breathtaking pace for such a complex achievement.

The President’s remarks about ‘new ways of learning’ were prescient. The speed with which America met Kennedy’s challenge in only seven years was impressive. However, it was the way in which those working on the *project learned things that they did not know they needed to know as they moved towards a solution* that makes this historical narrative compelling to schools and pedagogy. Although the theory that one

could acquire an underlying body of knowledge (a curriculum) while engaged in the hands-on solution of a problem was not new in the sixties, the attractive qualities of such a theory were likely not diminished by Neil Armstrong's successful landing of the Eagle in the Sea of Tranquility. The enormous amount of *learning* that occurred during the Apollo project could not help but pique the interest of contemporary educators to wonder if engaging in projects was and is a good way for their own students to acquire a curriculum.

The Apollo scientists were, in effect, *constructing* a new reality as they engaged with the problem of landing a human on the moon. They had to combine and apply their previous experiences, both in school (formal learning) and with nature (informal learning) to construct *the idea* that a person could leave Earth and travel a great distance to a place no person had ever been before.

Pinning down a complete, all-encompassing definition of constructivism is a lengthy and difficult task (Matthews, 2003) as it is a subset of multiple lines of post-modernist thought including radical constructivism, social constructivism, and even deconstructivism. However, for the purposes of this study, it is enough to note that constructivism has had a major impact on teaching methods in North American schools primarily because it is easy to agree with a central aspect of the constructivists claim that learning must be an active process involving inquiry and problem solving.

Thus, in North American schools there has developed a strong emphasis among "progressive" pedagogues of project-based learning, discovery learning, inquiry learning, cooperative learning, and hands-on learning – all of which are presumed to limit, even reduce, the amount of time and attention given to direct instruction. As a result, according

to some analysts of this development, "advocates of…progressivism currently hold sway in the educational establishment and in many schools of education where teachers are trained….Even though there is some debate in schools of education, the assumptions and ideas of the progressives have considerable influence throughout North America" (Zwaagstra, Clifton, and Long, 2010, p.6). Constructivism and this so called progressive education is the North American parentage of PBL, along with the buoyant optimism of Kennedyesque public policy for North American schools.

But What Shall Teachers Do on Monday Morning in Science Class?

Effective teaching often requires deftly managing the balance between the curriculum and the day-to-day school and teaching practices (Crosthwaite, Cameron, Lant, Litster, 2006; Stronge, 1997). Although it can be designed on many different models, generally, 'curriculum' means *what* the students are supposed to learn (Lunenburg, 2011). 'Practice' is what the teacher *does* (and, by extension, also what the students do) to best help students achieve the goals set by the curriculum. Teaching is a complex endeavor in part because so many theories of practice exist. There is currently no consensus regarding the notion of 'best practice'. Studying and critiquing, in a rigorously objective manner, the myriad ways teachers might organize their students' interaction with the curriculum and assess their subsequent learning is arguably the most important practical and intellectual challenge facing modern educational practitioners and theorists.

Whereas many teachers have significant control over which subject they teach and which practices they use to teach it, usually only a few have input into the selection of the curricular model or outcomes. As a result, teachers must often focus on *effective practice* as a way of encouraging learning and match what they believe to be the *best practices* to use with an existing curriculum. This is a deceptively complex task made more complex given the wide array of student learning characteristics encountered in a typical classroom. The interplay between the student, curriculum, practice, and other factors such as school climate, funding, and resources (to name a few) can give rise to a system of seemingly infinite complexity where, conceivably, many interventions will have positive effects for some students and negative effects for others. As such, the term

‘best practice’ – though complex – has taken on an increasingly important role in the modern teachers’ lexicon, forcing many practitioners to pay more than lip service to the question, ‘What can I do to best teach most of my students, most of the time?’

As if the task of selecting the best practices was not hard enough, the teaching profession has historically been subjected to a seemingly endless number of trends and fads often based on anecdotal evidence, scientifically unsupported dogma and ideology, and poorly constructed studies (Barrow, 2006). What’s more, some argue that teacher training programs have historically done a poor job of teaching teachers how to interpret the existing body of scientific research on teaching practices and their relative effectiveness in the classroom (Chaddock, 1998; Slavin, 1989; Stanovich & Stanovich, 2003).

Perhaps as a result of the arguably clouded relationship between practice and research, teachers are often eager to critique their practice by conducting periodic surveys of their students’ comprehension and overall progress (Bakula, 2010). These practice “checks” often occur at two levels: First, judging the efficacy of a practice to precipitate learning and, secondly, comparing the effect of two different practices on the learning of the same curricular outcomes. One might make the analogy to a doctor having the choice of prescribing a number of different drugs to treat the same disease knowing that their effectiveness will be nuanced from patient to patient. However, an important difference between the environments of doctors and teachers is that the physician can prescribe a treatment with the understanding that the medicine’s effectiveness has been the subject of a comprehensive and independent assessment and a rigorous approval process (Health

Canada, 2001). However, teachers have no such assurances and usually work in a much "softer" assessment environment.

Vigilant self-assessment of practice by classroom-teachers is important to ensure quality instruction (Bruce & Ross, 2007). However, like medical researchers, social scientists have known for decades that well-constructed, empirical studies of practice-in-action are important elements in the assessment of effective instructional methodologies (Gage, 1985; Marzano, 2003; Nelson, 1966). What's more, as the medical field has found, sometimes it is valuable to consider the best designed empirical studies *together*, within the broader context of the available research (Farrow, 2003). This amalgamated analysis can be at once qualitative and quantitative, comparing, amongst other factors, sample sizes, study design, and subject area but also merging effect sizes for certain interventions (for instance, two different instructional techniques) on particular outcomes (for example, test scores, transfer of problem-solving skill, attendance). Quantitative results can be measured and collated for a set of studies to paint a broader empirical picture of the efficacy of a particular curricular strategy and pedagogical practices to produce a certain learned outcome for a certain type of student.

The Problem

Given the increasing willingness of educators to turn to PBL as a primary instructional intervention, and the fact that a large amount of the evidence used to support PBL comes from the field of medical education - not elementary, middle, or secondary education - it is worthwhile at this time to conduct a review and analysis of the existing

empirical research comparing PBL to direct instruction (DI), at the secondary level and below. To that end, this study examines:

1. What does the current quantitative research say about the efficacy of PBL vs. DI when considering student acquisition of content knowledge in science?
2. Can any statistically meaningful interpretation be made of the existing empirical data concerning PBL vs. DI and, if so, what does this analysis tell us?
3. What are the most significant limitations in the research pertaining to PBL vs. DI in grade school education?
4. Where and how should we focus our empirical efforts in the future?

Methodology

Selection and description of studies for targeted review. My analysis is designed to address these four questions which are pertinent to a number of studies of PBL. The hypothesis I have formulated at this point is that PBL is superior to DI for increasing student conceptual knowledge. I have selected a narrow number of outcomes amongst an even narrower subject class; specifically, I wanted to examine the studies in light of the following considerations: 1) the study must compare PBL to DI; 2) the study must be quantitative and provide statistical results capable of further analysis; 3) the students studied must be in elementary, middle, or secondary school; and 4) only studies on science classes will be considered.

In light of these considerations, this investigation was limited to seven studies (see Table 1), a description of which constitutes the subject of chapter two. And when examining each study I will use the following descriptive framework: 1) who are the students; 2)

who are the teachers; 3) what is the curriculum or subject being taught; 4) what outcome(s) were studied; and 5) what was the effect of PBL on the outcome(s)?

Characteristics of PBL. Given the variety of definitions of PBL, it was judged useful to narrow the character of this curricular and pedagogical intervention. Accordingly, the characteristics of PBL defined by Barrows' (1994) will be used:

1. Ill-structured problems are presented as unresolved so that students will generate not only multiple thoughts about the cause of the problem, but multiple thoughts on how to solve them.
2. A student-centered approach is one in which students determine what they need to learn. It is up to the learners to derive the key issues of the problems they face, define their knowledge gaps, and pursue and acquire the missing knowledge.
3. Teachers act as facilitators and tutors, asking students the kinds of meta-cognitive questions they want students to ask themselves. In subsequent sessions, guidance is reduced.
4. Authenticity forms the basis of problem selection, embodied by an alignment with professional or 'real world' practice.

Evaluation of the targeted studies and their results. Evaluating the quality of the studies is of particular importance since the reliability and validity of the studies must be high. For example, whether the studies include the randomization of their subjects or not is of special interest. It has been noted that there can be significant problems in considering non-randomized samples in systematic reviews of the literature (Linde,

Scholz, Melchart, Willich, 2002). In general, we want to draw meaningful conclusions about the quality of the existing body of research and what might be done in the future to help define more accurate distinctions between the efficacy of PBL and DI. McMillan and Wergin (2006) delineated a comprehensive set of criteria for evaluating quantitative educational research. In this thesis I employed four of their criteria;

1. Is the method of sampling clearly presented and could the way the sample was obtained influence the results?
2. Is there anything in the procedures for collecting information, or in the instruments themselves, that could bias the results or weaken the study?
3. Is the magnitude of the correlation or difference between/among groups large enough to suggest practical significance or importance?
4. Do the conclusions and interpretations follow logically from the results presented?

The critical review and evaluation of the selected studies using these criteria is the subject of chapter 3.

Limitations

As with any investigations, there are certain problems or limitations that will be inherent in it. In particular, the small number of available research studies can be an impediment to drawing compelling conclusions about the efficacy of PBL. Another factor affecting the overall reliability of any statistical analysis is the non-randomization of the subjects in the experimental and control groups. Even where the studies selected contain flawed randomization (as most did), the purpose of the study was to examine the existing

research literature. Despite such a limitation, this investigation nevertheless provided a useful critique of the extant empirical research comparing PBL to DI.

Organization of the Thesis

The report of the thesis is organized as follows. Chapter 1 is an introduction to the investigation providing the background and significance of the study; the origin and development of PBL; the specific problem addressed; the method and procedures used to describe and analyse the selected studies; and an outline of the content of each chapter. Chapter one provides a detailed description of each of the selected studies. Chapter three provides an evaluation of these studies using the criteria selected from McMillan and Wergin (2006). Finally, chapter four provides a summary of the investigation and its findings and the implications suggested by the study for additional research and pedagogical practice.

Table 1

The Empirical Studies Considered in the Thesis

Year	Author(s)	Journal	Exp. Group N	Ctl. Group N	Outcome	Age Group	Study Type	Randomized	Conclusions
2010	Finkelstein et al.	U.S. Dept. of Education	1918	1497	Content Knowledge	Secondary	Pre/Post Test	Yes (Teachers)	Positive associations for both outcomes
2009	Drake & Long	Journal of Elementary Science Education	17	16	Content Knowledge	Elementary	Pre/Post Test	No	Positive associations for both outcomes
2008	Geier et al.	Journal of Research in Science Teaching	Two Cohorts #1 = 760 #2 = 1043	Two Cohorts #1 = 8900 #2 = 8662	Content Knowledge	Middle	Standardized Post Test	No	Positive association
2006	Liu, Hsieh, Cho, Schallert	Journal of Interactive Learning Research	464	None	Content Knowledge	Middle	Pre/Post Test	No	Positive association
2005	Maxwell, Mergendoller, Bellisimo	Journal of Economic Education	252	107	Content Knowledge	Secondary	Pre/Post Test	No	Positive association but with conditions
2002	Schneider, Krajcik, Marx, Soloway	Journal of Research in Science Teaching	142	Compared against national data	Content Knowledge	Secondary	Post Test	No	Positive association
2001	Chang	Journal of Science Education and Technology	84	75	Content Knowledge	Secondary	Pre/Post Test	No	Positive association

Chapter 2

Description of the Studies and Their Results

The purpose of this chapter is to describe in some detail each of the studies that will be critically evaluated in chapter three using the criteria of research quality derived from McMillan & Wergin (2006). In organizing the description, the size of experimental and control groups and the manner of their construction – notably random or not was given most relevance. While there are several standards that might be used to evaluate the quality of any research study, a large sample size properly drawn to represent the population generally implies that the study will have more generalizable results than a study where the number of students in the study is small and their characteristics unique. Therefore, the review and description of the studies is ordered from largest to smallest student cohort, and other important features of each study, especially distinctive ones, are highlighted as the description proceeds (see Table 1). Such relevant features include how the teachers were chosen and prepared; how the PBL curriculum was chosen; what instruments were devised to measure student achievement (specifically their grasp of pertinent content knowledge); and, of course, the researchers characterization of the results achieved by the PBL intervention compared to the DI control group, especially those effects considered statistically significant.

Definition of Specialized Terms

Throughout the description some specialized technical terms are used that are essential to understanding the studies' results and the researchers' claims. The terms are defined in accordance with Field (2005) and Katzer, Cook, and Crouch (1998):

Randomization. The selection or choice of subjects or items such that there is an equal chance of any subject or item being selected in relation to all other subjects or items in the same class or category.

Sample. A portion selected from a population the study of which is intended to provide statistical estimates relating to the whole; that is, the analysis of a small part or quantity intended to show what the whole is like. In the collection and treatment of numerical data, especially in or for large quantities, we are usually concerned with inferring proportions or characteristics from a sample that accurately represents or genuinely reflects the population.

Mean. The numerical value of the quotient of the sum of several quantities and their number is the mean. For example, in the arithmetic progression 2, 6, and 8, the arithmetic mean is 5.33; the average, in ordinary language. Scrutiny of the mean scores of two different groups, for example an experimental group and a control group, on the same test is one way of comparing the achievement of the two groups when a different treatment, such as a curriculum or teaching method, is applied to one group.

Standard deviation. An estimate of the average variability (or spread) of a set of data measured in the same units of measurement as the original data. It is the square root of the variance. The standard deviation tells the degree to which the individual scores deviate from the mean or average.

Effect size (Cohen's d). The effect size is the amount of a standard deviation that is attributable to an intervention; or to put it another way, the degree to which the null hypothesis is false. For example, an effect size of 0.5 indicates that 50% of the deviation from the mean is attributable to the experimental intervention.

Statistical significance. The extent to which a result deviates from a hypothesis (for example, the so-called null hypothesis that there is no difference between the achievement of two discrete groups) such that the difference is due to more than chance or errors of sampling or measurement. Statistical significance is usually expressed as a fraction where a value at or below 0.05 is taken to be significant.

Finkelstein et al. (2010)

Split between grades eleven and twelve, 12% in the former grade and 88% in the latter, this study had a total of 2,502 students in the experimental group and 1,848 students in the control group, all attending high-schools in either Arizona or California. Almost forty percent of the students were eligible for free or reduced-price school meals. (a proxy for students' socio-economic status) and a significant minority (37%) were of Hispanic origin.

Finkelstein et al.'s (2010) investigation of the implications of PBL for student learning in mandatory high-school economics courses is unique amongst the studies considered for two reasons. First, while students were not randomized (except for any incidental randomization invoked by the school course scheduling procedures), *teachers* were randomly assigned to either experimental or control groups. Second, this is the only study where the intervention (a PBL curriculum) was written by an education advocacy group for teachers to test in their classrooms.

An experienced (ten years) economics teacher, who was also an expert at project-based economics instruction, served as the recruiter of teachers for the study, and all schools in California and Arizona with at least 1500 students were contacted. In total,

approximately 1000 schools were contacted of which 106 employed a teacher or teachers willing and qualified to participate in the study. One of the logistical criteria for inclusion in the study was that a school had to be capable of scheduling a teacher to teach economics back-to-back over two consecutive semesters (fall 2007 and spring 2008). This requirement allowed the experimental group of teachers to practice the implementation of the PBL intervention in the first semester in preparation for the second semester's course which would provide the data to be used by the investigators. Students were placed in a particular class at their school administration's discretion with no interference from the researchers.

Teachers participating in the study were assigned to a group using a random number algorithm. Initially, 128 teachers from 106 schools were recruited for the study. However, for a variety of reasons not fully explained by the researchers, 45 teachers, by chance split almost evenly between the experimental and control groups, dropped out at some point before the conclusion of data collection. Of the 83 remaining teachers and their corresponding student cohorts, data from 64 was used in the final analysis with 35 teachers and their student cohorts assigned to the experimental group and 29 teachers and their student cohorts assigned to the control group.

Prior to the study, experimental group teachers were given five days (40 hours) of training in PBL with a focus on the materials that would form the basis for the intervention curriculum. The curriculum was distributed to the teachers and they were given expert instruction in its implementation by economics teachers who had a record of successfully using PBL in their classrooms. Both the curriculum and professional development were designed by the Buck Institute for Education (BIE). Indeed, the BIE

was a key player in the study providing not only the curriculum and initial PD but also ongoing support as a facilitator of conference calls amongst the experimental group teachers at several points during the study. Control teachers did not participate in the PBL professional development but carried on with the normal professional development activities in their schools. Control teachers were not expected to deviate from their usual, text-book based, direct-instruction approach.

As stated by the authors, "Overall, the test of the curriculum was whether students, working with well-trained and supported teachers, demonstrated a level of economic performance above that of students who took traditional economics courses" (Finkelstein et al, 2010, p. ix). The Test of Economic Literacy (hereafter TEL) was administered to both experimental and control groups prior to and after the economics course. The TEL was developed and refined by the National Council On Economic Education (NCEE) and is a reflection of the content standards that the council developed in the late nineties in response to the Goals 2000 Educate America Act. It was these standards that were used by the BIE to develop the intervention curriculum used in the study.

The experimental group demonstrated a mean score 2.6 points (out of 40) higher on the post-test than the control group with a moderate effect size (Cohen's $d = 0.32$) and a 95% confidence interval; that is, the experimental group's higher mean score was statistically significant at the 0.05 level. With these results the researchers concluded that, with sufficient instructor professional development and support, a PBL based course of study in economics was superior to a DI approach for increasing the students' performances on the TEL.

Geier et al. (2009)

Another large-sample study was conducted by Geier et al. (2009) working out of the Universities of Michigan and Arizona with The Detroit Public Schools (hereafter DPS). The DPS is a big district with a population of about 160,000 students and employing around 10,000 professionals. Its students are overwhelmingly African American (91%) with 69% of the student population eligible for federal free and reduced-price lunchs. The study compared the effects of PBL on two cohorts of students, each cohort participating in the study for one school year. The study was scaled up from 760 students in the first cohort to 1043 in the second. Control groups stayed at the same size at approximately 8800 students for each study trial.

Participating teachers averaged 11 years of teaching experience, of which eight were spent teaching primarily science. Schools were selected on several criteria including technology infrastructure, and, interestingly, equity in access to innovative programs. The DPS initiated the idea of PBL in its science classrooms as part of the efforts to improve the students' engagement and retention. Once selected for the study, teachers in the experimental group were assigned their classes by school administrations in the usual way and were expected to use the PBL curriculum in place of their "habitual" methods of instruction.

Once again, professional development sessions were provided to the experimental group teachers, primarily through week-long summer institutes. The training focused on content and enactment problems. The authors do not include data on the teachers' time-on-task with respect to professional development in the way that Finkelstein et al (2010)

did; however, we can assume that, since both studies relied on a week of summer training as their PD centerpiece, the amount of time spent training the participating teachers was comparable.

The curriculum was designed by The Center for Learning Technologies in Urban Schools (LeTUS) at the University of Michigan to mesh tightly with the DPS curriculum framework and standards. It included three eight-to-ten-week units each based on a guiding question or problem. As expected, students navigated the content of the course as they formulated, and eventually communicated, a solution or answer to the problem presented in the curriculum. The authors were interested in aiding district-wide reform that would improve the students' achievement on state-wide, standardized tests. Therefore, the post-test used in the study was the Michigan Educational Assessment Program (hereafter MEAP) test. This test measures, among other things, science content knowledge, and is administered by the district at set times, one of which is in January of the year the students' were in grade eight. This scheduling was not within the control of the researchers. As a result, some students participating in the study may have had significant stretches of time between their PBL course and the writing of the MEAP test.

Effect sizes for both cohorts were greater than those in Finkelstein et al. Although it is not clear that Cohen's d is the statistical test used, a value of 0.44 was reported for Cohort I and Cohort II registering an effect size of 0.37. Breaking the effect size down by content areas, we see effects ranging from 0.28 for knowledge of the life sciences up to 0.53 for Earth sciences knowledge. All effect sizes were statistically significant at the 0.05 level.

Liu, Hsieh, Cho, and Schallert (2006)

The call for integration of new technologies into the classroom is not new. Indeed, P21 and BIE see the increasing use of both computer simulations and media technology as an essential component of 21st century education and vital in the implementation of PBL curricula. The study by Liu, Hsieh, Cho, and Schallert (2006) used an entirely computer based, PBL experimental curriculum to teach sixth graders a space science unit.

A total of 464 students from two middle schools in a mid-sized southwestern US city were used in the study. Three quarters of the students were Caucasian. Unlike the two previous studies, not all the students were in normal classrooms. Fifteen percent of the subjects were in talented and gifted programs and 10% had special needs or were students in resource programs. There was no control group. The students were taught by five teachers, four of them had experience using this particular PBL intervention. However, all five teachers participated in a training workshop where they were introduced to the curriculum, general PBL theory, and coached on implementing the curriculum in their class.

An achievement test was created by the researchers to measure content knowledge before and after the intervention. The authors are vague about the basis for the construction of the achievement test. They state only that the test "…reflected what the designers and subject matter experts consider as important for the students to acquire after using [the PBL curriculum]" (Liu, Hsieh, Cho, and Schallert, 2006, p. 232). The test contained 25 items: 15 content questions, and 10 application questions. Separate results for these two categories of questions are not indicated in the study, only the overall mark out of 25 is reported. By this measure, the mean percentage achievement on the test

increased from 47% on the pre-test to almost 71% on the post-test. The authors do not report any effect size for the PBL curriculum, but they used this increase in mean test scores as proof of the effectiveness of the PBL curriculum.

Maxwell, Meggendoller, and Bellisimo (2005)

Although not as large as Finkelstein's study of high-school economics students, Maxwell, Meggendoller, and Bellisimo (2005) conducted a study of 252 economics students at four high-schools in Northern California. Little is known about the social, economic, or cultural background of the students beyond the schools being distributed in both urban and rural settings (two in each). The administrators in the schools placed students in classes according to their usual scheduling procedures. Prior to the participating teachers acquiring their class lists, each teacher picked a class from their assigned teaching load with which to employ the PBL curriculum and a class to teach without any PBL format. It was in this manner that students were assigned to either the experimental or control group.

The authors indicate that they chose "veteran" teachers for their study although no meaningful attempt was made to discuss or quantify teacher credentials nor was the number of years-of-experience of each teacher mentioned in the study. We do know, however, that two of the teachers had undergraduate degrees in economics and, "…seemed more fluent in the economic way of thinking than the others" (Maxwell, Meggendoller, and Bellisimo, 2005, p. 319). All of the teachers had experience teaching economics using both PBL and DI. Still, they were required to attend a week-long training workshop, hosted by the researchers, to become familiar with the PBL

curriculum and its supporting documents and to learn the fundamentals of implementing PBL. The two most experienced teachers (coincidentally the teachers with economics degrees) at times served as facilitators of the sessions.

The curriculum was based upon the voluntary standards established by the National Council on Economics Education (NCEE) which were the same standards used by Finkelstein et al. (2010). As with all the studies considered, the curriculum was based on ill-conceived problems with the teacher acting as the facilitator to groups of students who worked towards a solution to the problems. The Test of Economics Literacy (TEL), in combination with a test bank from a popular economics text served as the basis for devising a pre-test/post-test instrument. The two sources of questions were used jointly to fashion a 16-item, multiple-choice test. While measuring economics content knowledge, the test used questions spanning all levels of Bloom's taxonomy from the knowledge level to the evaluation level.

Once again, using Cohen's d, effect sizes were in the medium range at 0.54. While a critique of the study results is the subject of chapter three, it is worth noting at this point the overall poor performance of the both experimental and control groups on the test. Both groups scored failing mean marks almost 1.5 points below passing in the pre-test. Despite taking the full course with experienced teachers, both groups experienced marginal gains with neither group attaining a passing mean of eight out of sixteen. Admittedly, no comparison exists between these results and a much larger, state-wide cohort of student (which arguably could constitute a control group in a quasi-experimental study design). The post-test standard deviation for the experimental group was 2.73 meaning that, with the reported effect size of 0.54, the PBL program explained

around 1.5 marks of the standard deviation. This result is consistent with the other studies considered thus far.

Schneider, Krajcik, Marx, and Soloway (2002)

Common to the research studies considered in this project, national standards and students' standing on national measures of achievement were often not far from the researchers' minds. This was certainly the case with Schneider, Krajcik, Marx, and Soloway's (2002) study of 142 students at an alternative high-school in the US mid-west. The school in question was unique in so far as its school culture emphasized independence in student thought and action. That said, it is difficult to compare this school to others in other studies because school culture was not usually mentioned by the other researchers. However, we know that those who conducted this particular study chose a small, public high-school serving a mostly "middle to upper middle class" group of students. (Schneider, Krajcik, Marx, and Soloway's 2002) State-wide assessments showed that students in the school scored comparably to the state population on tests examining proficiency in science, reading, and writing. However, very little was reported about the teachers in this study other than that there were four teachers whose experience teaching high-school science ranged from two to 25 years.

The Foundations of Science (hereafter FOS) curriculum was developed and implemented by both the researchers and the participating teachers. The FOS curriculum addresses the earth sciences, physical sciences, and life sciences and is a typical PBL curriculum in that it includes broad, ill-conceived problems that must be solved by groups of students. During the course of the study, the FOS program fully replaced all other

methods of science teaching in the school. This is unique amongst the studies considered, in that an entire school program (the teaching of science) was given over to a study. It took several years to fully implement the program in stages: FOS – I (grade 9), FOS – II (grade 10), and FOS – III (grade 11).

The National Assessment of Educational Progress (hereafter NAEP) is a tool used across the United States to assess student achievement in a variety of subject areas including science. It is administered at several set times. In this case, NAEP was used as the post-test for all students participating in the study and was administered in the students' grade 12 year. The test is a mix of fact and concept-based multiple-choice questions, constructed response questions, and performance tasks. For this thesis, we are interested in the multiple-choice questions as they assess student content knowledge.

After progressing through FOS I through III, students wrote the NAEP in grade twelve and their mean performance was compared against the national grade 12 NAEP results. Looking at the overall effect size is not useful because it reflects a mixture of conceptual, performance, and content based questions. However, the researchers report that, in the case of solely the multiple-choice responses, students engaged in the PBL curriculum scored better than the national average 42% of the time. One could assume then, that 58% of the time the national average exceeded the scores of the experimental students on the same questions. However, these results seem to not tell the whole story since, when one looks at the NAEP questions that test *conceptual* understanding (closely related to content), the experimental group outscored the national average 63% of the time.

Chang (2001)

Chang's study (2001) comparing the effectiveness of problem-based computer assisted instruction (hereafter PBCAI) is topically worth including in this thesis but is problematic none the less. Chang's students were studying Earth Sciences at a high school in Taiwan. Eighty-four students comprised the experimental group and 75 students comprised the control group. There is no mention in the study of student characteristics. All the students were taught by the same teacher who had ten years teaching experience and a university degree in the Earth Sciences but the study makes no mention of any specific professional development offered to the teacher prior to or during the study.

The curriculum consisted of a computer-aided course of study in Earth Sciences. The experimental group of students navigated the course objectives with the computer based curriculum that included components in the library and the laboratory. Within this environment the students solved ill-conceived, earth sciences problems which were often of an investigative nature; for instance, drawing conclusions and making predictions about past and future geologic activity based on current, observable geologic events. The control group experienced the same course but in a text-driven, DI environment. Both groups had the same amount of instructional time. The intervention lasted two weeks after which both experimental and control groups were tested.

Chang developed his own instrument for gauging, among other things, the students' content knowledge. A panel of high-school teachers and three university professors vetted the validity of the test and it was used in a pre/post-test manner. Information is very limited regarding the nature of the instrument, including how many

items it contained. Overall, gains made by the experimental group over the control group were marginal.

Drake & Long (2009)

The final study considered in this thesis is by Drake & Long (2009) and examined a small group of grade four science students as they studied a unit on electricity. The experimental group consisted of 17 students from a school split almost evenly between White, African American, and Hispanic students. As in some previous studies, 80% of the students received free or reduced-price school lunches. The researches assert that both the experimental and control groups were representative of the school population.

This study is unique in that it was the researchers who conducted the instruction, not the regular classroom teacher. For 45 minutes per day over a two week period instruction was given in a PBL format for the experimental group and DI was provided to the control group. The PBL curriculum was centred on one ill-conceived problem focusing on power generation and distribution during a period of severe winter weather and was designed to meet the requirements of the competency goals set forth by the North Carolina Standard Course of Study (hereafter NCSCS) for grade four science education. The pre/post-test, designed by the researchers, was aligned with the NCSCS standards and tested, among other outcomes, content knowledge in science at the grade four level.

Once again, gains in content knowledge on the post-test were statistically significant but marginal. A medium effect size of 0.72 (Cohen's d) was reported for the PBL treatment. But the difference in mean, raw scores on the post-test between the

experimental and control group was only 0.57 marks. For the experimental group the standard deviation was 1.8 which considering the reported effect size accounts for a roughly one mark increase in the content knowledge score as a result of the PBL intervention.

Summary

In summary, we see statistically significant positive effects of PBL on students' content knowledge. These results are mostly reported as differences between an experimental group and a control group in pre- and post-test mean scores. In two cases, the performance of the experimental group is compared with the total student population on state or national standardized tests. The PBL units range in length from only a few weeks to entire, five-month semesters, and grade levels range from elementary (grade four) to grade twelve students. Regardless of the statistical significance, when we look at the effects of PBL on mean, raw test scores, we see only small gains in performance between the experimental and control groups. But what about the overall quality of the studies when compared with accepted standards for educational research? In the next chapter I will try to ascertain how much faith we should put in the findings of these studies.

Chapter 3

Evaluation of the Studies and Their Results

The purpose of this chapter is to evaluate the seven studies described in the previous chapter using four criteria of quality research derived from McMillan & Wergin (2006), originally set out in *Understanding and Evaluating Educational Research*. In the past, these, or very similar criteria have been used by social scientists to evaluate empirical research and continue to be employed today. While not the only standards for research quality evaluation in existence, they are one of a set of standards respected and accepted by social scientists as being capable of presenting any research study's strengths and weaknesses (Delazio 2006). Of added benefit, these standards are not statistically dense nor overly technical and, therefore, may be more accessible to the educational practitioner and even the layperson who may have limited experience with statistical methods. As such, they form a meaningful gateway to a more sophisticated, in-depth analysis of the validity, reliability, and other hallmarks of good, empirical education research. Specifically, the following criteria were applied to the seven studies previously identified (McMillan & Wergin, 2006, p.14):

1. Is the method of sampling clearly presented and could the way the sample was obtained influence the results?

2. Is there anything in the procedures for collecting information, or in the instruments themselves, that could bias the results or weaken the study?

3. Is the magnitude of the correlation or difference between/among groups large enough to suggest practical significance or importance?

4. Do the conclusions and interpretations follow logically from the results presented?

Did the Method of Sampling Weaken the Research?

The first, and arguably most important of the criteria is the method of sampling used in each study. A goal of many quality research projects is to draw conclusions about populations based on the reaction of a sample of that population to an intervention. So, if a study is going to make a contribution to our overall understanding of a population, the method of choosing the experimental and control groups must be sound; in short, the sample must accurately represent the population. Further, once those groups have been assigned, the method of implementing the intervention is important, especially in instances where the intervention is complicated. For example, if the intervention is simple, such as in a cola taste-test (the intervention is quick and relies on only one yes/no question), there are fewer variables which could be introduced by the implementer of the intervention (the person who administers the taste test on behalf of the researchers) that could influence the subject's response. However, in the case of research in pedagogy, where many teachers might participate in a complex, lengthy intervention, individual differences between each of those teachers, for example, personality, teaching style, enthusiasm, could play a critical role in influencing the effect of the intervention and the subsequent sample data obtained. Of course, it is possible to statistically control for differences between teachers, but it would seem very difficult, if not impossible, to

control for *all* of the possible differences. Thus, the selection of the teacher sample is perhaps as critical as the experimental subjects' selection given the influence of the teacher over the execution of the intervention.

Sample selection aside, the golden rule for experimental-design studies is the random placement of students in experimental and control groups. Randomization is justified because of the number of variables exhibited by the subjects that could reasonably influence or "contaminate" the measured outcome. For example, if all the subjects were identical, placing them in either the experimental or control group would require no thought whatsoever. However, as the population becomes more diverse, increasing care must be exercised in constructing the group samples to ensure that a group's characteristics are not skewed one way or another. For example, it is increasingly common to find schools offering 'sports academies' where students opt for regular academics in the morning and training in a particular sport in the afternoon. The selection of a morning science class for a study could find a disproportionate number of these students in the sample and introduce biases to the "normal" population these students might carry.

Research in pedagogy is particularly susceptible to randomization difficulties for one main reason: researchers often have no influence over how students are assigned to a particular class of students within a school. Making matters more complicated, there is no absolute consensus on the characteristics of a student (physical, emotional, social, economic, etc.) that influence academic performance since it is often the interaction between known characteristics and other unknown factors, such as the other students in the class, that influence the achievement of the students.

Typically, school administrators assign students a schedule based on many factors such as courses selected by the student, teacher availability, and anticipated completion of pre-requisites. These factors can, at times, lead to biases in samples allowing, for instance, many students who take academic (as opposed to, say, consumer) math courses to be assigned disproportionately to a particular section of a science course. In this case the 'stacking' of a particular class with above average math students is an artifact of the scheduling process. This means that school administrative practices pose a serious problem for the generalizability of research results from a sample to a population. Therefore, whether intentional or not, it is important to note that none of the studies considered in this thesis have samples randomized by the researchers. So, when considering the sampling methods in the studies, including the means of placing subjects in the experimental or control group, we must recognize that on the first criterion of research quality – sampling/randomization – all of the studies are weak. In all the studies students were grouped by school administrators, without direct consideration of the intentions of the researchers or a specific concern for the degree to which the samples represented the population.

In my view, the study by Finkelstein et al makes the best attempt to obtain the broadest, most representative sample. For sheer size, it is the largest at around 3500 students split almost evenly between the experimental and the control groups. As mentioned, the researchers focused their attention on schools with over 1500 students in Arizona and California. The threshold population was selected because, in the researcher's estimation, this number afforded a sufficient concentration of students for schools to invest in employing an economics specialist teacher. With specialist teachers,

the researchers' posit, the quality of the instruction would be higher than in schools with non-specialists economics teachers. What is surprising is how few schools and teachers ultimately chose to participate in the study. Although in many instances after contact it became evident that schools did not meet the criteria required for their participation, many schools declined to participate even though they met all the requirements for inclusion in the study.

An interesting feature of the sample of students is its homogeneity in age. Eighty-eight percent of the students were in grade twelve so, not only does this study provide a large sample, it is a large sample concentrated at a narrow age range if we consider 88% of the subjects to be between 17 and 18 years old. However, it is unlikely that the cultural and ethnic demographics of the sample are comparable on a national level. With 77% of the students being either Hispanic or non-Hispanic White, in about the same proportion – 37% and 40%, respectively - this a far different sample with respect to ethnic origin than say, Geier et al.'s study of the Detroit Public Schools which has a student body that was 91% African American. Interestingly, these two studies show similar effect sizes for the project-based curricula. One could argue that similar results for comparable interventions with ethnically different samples do not negate the effect, if any, of ethnicity on the efficacy of PBL. However, the effect of ethnicity on learning was not the focus of this thesis and was not investigated.

Of course, in addition to the student population, there is the population of teachers to consider. In each participating school, administrators assigned teachers to teach a particular section of economics. Sections were then randomly assigned to either the experimental or the control groups. Teachers assigned to each group were remarkably

similar, scoring within 1.8 points (out of 40) of each other on the TEL, differed in years in profession by only 0.1 years, and averaged 6.9 years of economics teaching experience in the experimental group and 7.6 years for the control group. Across other measures, such as university economics courses completed and confidence in teaching, the groups were also very similar. Even before using regression analysis to account for baseline differences in teacher characteristics, the evident homogeneity amongst the instructors likely encouraged the investigators to believe that any results would be reasonably free of bias resulting from teachers.

So, we can see that research into teaching practice is inherently more complex than some other types of social science research because, when studying teaching interventions, we must consider the characteristics of both the students *and* the teachers. Again, if we consider a simple study asking subjects to pick the best tasting of two colas, the researcher (or, more likely, a research assistant acting for the researcher) has a simple script for collecting data which has little possibility of influencing the subjects' choices. No control for the researcher's ability to collect data is required. Teaching on the other hand is a highly complex endeavor where the teacher's skill, content knowledge, attitude, and many other attributes can have an enormous influence on the effect of the intervention on dependent variables. Therefore, random assignment of both teachers and students must be done properly.

At the other end of the research spectrum we find studies like Drake and Long (2009). The sample size here is extremely small, only 17 students in the experimental group and 16 students in the control group. Although the researchers suggest that these samples are representative of the school population, it is unlikely that they are in any way

representative of the district, state, or the country. As emphasized previously, size matters when it comes to samples so a study with 17 experimental subjects must be considered to have very limited generalizability to any student population.

Drake and Long's study attempts to address the problem of teacher sample selection in that it is the researchers that take over the teaching duties in this study. Two researchers were assigned to each of the classes, one team to the experimental class and one to the control class. One of the researchers was responsible for teaching and the other was responsible for data collection. By implementing the instruction themselves the researchers do, technically, maintain more control over the effect of the PBL and the DI curriculum. However, in the end this approach generates a similar list of questions about the effect of the intervention: What was the teaching skill of the researchers? Did the change of instructor, from their regular classroom teacher to the researcher, have any effect on the students' learning? Did students' attitudes to the new instructor and format change over time and influence the study results? The study provides no answers to these questions nor are they even raised by the researchers in their own discussion of "study constraints".

Finally, in selecting the teacher sample it is important to mention the professional support provided to the experimental group teachers both before and during the studies. In all but two studies, Chang (2001) and Drake and Long (2009), some professional development was used by the researchers to help the teachers understand both the theory and the practice of PBL. In the best examples, this professional development was explicitly linked to the exact PBL curricula the teachers would implement in the study. This was the case in Finkelstein et al. where the professional development materials and

presentation were based on the PBL curriculum by the BIE, the same organization that developed the curriculum. Subsequent running of professional development sessions was also managed by the BIE and focused on both the theory and practice of PBL. This study proved to have the most synergistic relationship between the aims of the study and the preparation and guidance provided to the teachers participating in the study. The professional development supplied by Geier et al. to their teachers was of comparable quality to that provided by Finkelstein et al. Other, however, studies provided varying degrees of professional development. It is worth noting that the majority of studies recognized the importance of providing assistance to teachers to ensure as fair a comparison as possible with respect to the instructors' ability and preparedness, in an effort to mitigate teacher effects on the outcomes.

Did the Procedures or Instrumentation Bias the Results?

Obviously the method of assessing the effect of the intervention is of paramount importance to the legitimacy of these empirical studies. The construction of the intervention, training of the experimental group teachers, and selection of the sample are meaningless if the method of assessing the effect of the intervention is flawed. In this thesis, studies were selected that measured students' increase in science content knowledge. So, we would hope that any test of the efficacy of the PBL intervention would be valid and reliable in measuring increases in science achievement. Further, the testing instrument should contain items that directly measure curricular content encountered or taught in both the experimental and the control groups.

A variety of tests of content knowledge were used in the studies. Two studies, Schneider et al. and Geier et al., used ether state or national standardized tests to measure the students' science achievement. Schneider et al used the National Assessment of Educational Progress (NAEP) which is a national, standardized test of science content in earth sciences, physical sciences, and life sciences. Geier et al used a similar instrument designed for use in Michigan, the Michigan Educational Assessment Program (MEAP). It should be noted that these two studies used post-tests only and compared the performance of the experimental group to the performance of the entire national and state cohorts writing the tests for that year.

Other researchers created their own instruments and in each of these cases the instrument was used in a pre/post-test capacity. The advantages of a custom built testing instrument are clear in that the instrument can be tailored to fit closely with a particular intervention ensuring that the instrument is valid for the curriculum. As with the other criteria discussed so far, Finkelstein et al demonstrate a superior attention to such details compared to the other studies. Their instrument, the Test for Economic Literacy (TEL), was developed and refined in concert with both the PBL *and* the national curriculum standards for mandatory high-school economics education in the US. The fact that the curricular outcomes, the PBL intervention designed to teach the outcomes, and the instrument intended to measure the efficacy of the intervention were developed leads me to assume that the test is valid. However, this conclusion is reached without a thorough examination of the instrument itself, a task beyond the scope of this thesis. However, it should be noted that Maxwell, Mergendoller, and Bellisimo (2005) used the same voluntary national curriculum standards in designing their experimental treatment of PBL

and the same TEL as their instrument to measure pre- and post- intervention science achievement.

Other researchers, such as Chang (2001), Drake and Long (2009), and Liu et al. (2006) also used customized instruments. However, these researchers are less explicit about how these instruments were developed and offer less evidence of the validity of the instrument. Further, in the case of Liu et al., the instrument contains both content and conceptual questions, but the researchers did not separate the students' results for each type of question. As a result, it is difficult to differentiate increases or decreases in achievement as a result of PBL to the content and conceptual domains.

Is the Magnitude of Differences Among Groups Significant?

Despite its intellectualization by the academic community, teaching is a practical, applied endeavor with but one aim: helping students learn a prescribed curriculum. Engaging in one intervention or another and, in particular, switching from one intervention to another should only be done when the new intervention proves to be significantly more effective than the previous teaching method. This is particularly the case when the switch involves a significant re-think and re-organization of teaching activities coupled with new forms of assessment. For a teacher reliant on DI, a switch to PBL likely requires significant disruption and re-orientation because many of the skills used in DI may not be transferrable to PBL. Therefore, any increases in learning achieved by PBL over DI must be significant to warrant a move away from DI.

All of the studies indicated, to one degree or another, a positive, often statistically significant difference between PBL and DI for the students' content knowledge; that is,

students in the experimental PBL group scored, on average, higher. However, in all the studies, including the best designed of the group, Finkelstein et al., this increase was relatively small. For instance, let us look at Finkelstein et al.'s results within the context of actual increases in scores on the testing instrument. On the TEL, a test scored out of 40 possible marks, the experimental group post-test mean score was 22.61 and the effect size was 0.32 for a standard deviation of 8.08. The control group scored 20.01 on the same post-test. These results mean that 32% of the standard deviation can be attributed to the PBL curriculum which works out to around 2.6 marks out of the 40 possible marks on the test. This equates to roughly a 6.5% increase in the mean score on the TEL for a student in the experimental group compared with the control group.

Geier et al., the authors of the next largest study, which is comparable in quality to the study by Finkelstein et al report similar, medium, effect sizes. When we look at the results of the students' scores on the MEAP, we see that for each cohort the number of students passing is greater for the PBL group. However, again, the raw test scores differed by around 10% between the experimental and control groups. So, although more students were passing the MEAP after being in the PBL intervention, their raw scores did not increase dramatically. Maxwell, Meggendoller, and Bellisimo reported even slimmer gains for the experimental group in their study of economics students; that is, the raw mean scores increased by only 1.5 marks on the TEL. Standard deviations in the range of 2.5 allow us to infer that the intervention is only responsible for around a one mark increase in the students' achievement between the experimental and control groups. These statistically significant yet small increases in the students' achievement scores are similar across the studies analyzed in my thesis.

Are the Conclusions Warranted?

To their credit, many of the researchers acknowledged the marginal gains observed in the experimental groups. Maxwell, Meggendoller, and Bellisimo (2005) sum up their findings by stating: "Although this study found that PBL increased learning of macroeconomics in high school, results could be viewed as either a support for PBL as an instructional strategy for teaching high school macroeconomics or support for using the traditional lecture-discussion method" (Maxwell, Meggendoller, & Bellisimo, 2005, p. 324). Finkelstein et al. offer a reserved endorsement of PBL based on their findings. While they acknowledge that there was a significant, measurable increase in the achievement of students in the PBL group, they encourage further study and, in particular, the replication of the study to further refine the understanding of the interaction between the PBL curriculum, the professional development of the teachers, and students' achievement.

It is very important for the reader to accurately differentiate between the *statistical significance* of a particular result and the actual effects on the achievement tests. Geier et al. point out that their study, "…shows strong effect sizes on [standardized] measures of achievement" (Geier et al., 2008, p. 934). While this may be the case, strong effect sizes do not necessarily equate to meaningful increases in test scores. Of course, what constitutes "meaningful" in terms of a test score is arguably highly subjective. However, practitioners and researchers need to remember that *effect size* is simply a measure of the proportion of the standard deviation that is attributable to the intervention. So, if the standard deviation is small, the difference in test scores between students in the

experimental and control groups could, in fact, be quite small. In the future, it will be up to school boards to decide how much effort and resources they are willing to expend in changing classroom practices to achieve what may be very small increases in students' science achievement.

Summary

As the evaluation shows, sample selection was not randomized and was a weakness of all the studies. Further, many of the studies had small numbers of students in the two cohort groups, making generalizations to a meaningful population difficult if not impossible. The quality of the instruments used was also highly variable among the studies. Some instruments were customized and highly coherent with the PBL curriculum, while others were less strongly connected to the curriculum they were supposed to assess. While the studies did support the conclusion that PBL is more effective than DI at increasing students' achievement, the percentage scores showed there were only small gains.

Chapter 4

Conclusion

This final chapter begins with a summary of the results of my analyses of the seven empirical studies comparing PBL to DI. Then, conclusions arising from the critical analysis of these studies using the standards set out by McMillan and Wergin (2006), are presented. This is followed by the implications of the study for educational research and practice.

Summary

As outlined in chapter one, the objective of this thesis was to critically examine the body of research comparing project-based learning (PBL) with direct instruction (DI) and the resulting effects on student content knowledge. Studies were chosen based on a specific definition of PBL, a focus on the implications of it for students' content knowledge, and a limitation of the examination to research studies that compared PBL to DI science courses at the secondary school level. Further, the studies had to be quantitative and offer some empirical, statistical evaluation of the data collected. No qualitative studies were included. In total, seven studies met the criteria for selection and were examined in the thesis. I was interested primarily in whether the studies found PBL to be better than DI for improving students' content knowledge, but I was also concerned about the general overall quality of the studies when compared to accepted standards for "good" educational research.

The studies were all quasi-experimental, with the majority employing a pre-test/post-test method. Five of the studies developed specific pre/post-test instruments to

assess the students' academic achievement and used well defined experimental and control groups selected by the researchers. Two studies did not include a pre-test but compared group performance on either a state-wide or nation-wide science achievement test. The student cohorts used in these seven studies ranged from as few as 35 in one school to thousands in two state jurisdictions. All the studies reported statistically significant results for PBL and increased scores between pre-test and post-test for the experimental PBL groups. In the studies without a pre-test, the experimental PBL student cohort out-performed the district and national cohorts.

Also, it is surprising but there simply were not many studies that fit the selection criteria I specified for this thesis. The narrowness of the selection is an artifact of the debate that exists over the use of PBL in schools. Often the debate is framed as an either/or choice with PBL on one side and DI on the other. Therefore, I made a selection to reflect the debate and did not include studies that, for example, investigated a mix of PBL and DI in the same classroom. Further, the mandate of this study was to consider only secondary school science courses because this is my area of specialty, and to keep the age and maturity of the students below university level. Finally, given the wide range of project or problem based teaching techniques, it was important to address the most common, and current, definition of PBL, which happens to be very consistently the one adopted by my school division. Together, these restrictions filter out the majority of the quantitative research comparing PBL to DI leaving me with a limited, but focused, number of studies to evaluate.

Conclusions

Some key generalizations emerged from this study in the light of the evaluation of the studies in Chapter three. They highlight certain aspects of the research enterprise in education and point toward its improvement.

High quality research uses representative samples. Once the studies had been selected it became obvious that many used samples that were very small. As McMillan and Wergin (2006) emphasize, to be considered "good", "proper" or high quality research, studies ought use samples that are representative of the population under consideration. The larger and more random the sample, to a certain extent, the better chance it has of being representative of a population. Granted, some of the smaller studies used samples that were excellent representations of their *school* populations. However, such studies are of limited use to the broader population of students and teachers who want to know, with some certainty, how well PBL works. (never mind the fact that school populations can change over time potentially rendering some data irrelevant after a number of years). If the goal is to understand whether PBL is superior to DI for increasing content knowledge amongst students on district, state, provincial, or even national levels, then the samples must be large enough to represent of these larger populations.

An issue closely linked to sample selection is the randomization of that sample between experimental and control groups. As stated earlier, true randomization is often very difficult in educational research because researchers have limited influence over the administrative task of creating classes of students. Without proper oversight by the

researchers, unexpected biases can be embedded into the selection of students for the control and experimental groups as artifacts of the scheduling process. This inability to truly randomize a sample is a "poison pill" that instantly excludes a study from being "experimental" and, unless specifically scrutinized by the researchers, poses risks to the validity and reliability of the study. All of the studies I reviewed for this thesis had serious problems with randomization.

Differences in teacher training and coaching in any curricular or pedagogical intervention complicate study comparisons. It is not only the sampling of students that is important. Teachers, who must guide students through the intervention, come with varying degrees of skill and ability which must be effectively controlled in a study. One non-statistical means of control is implementing courses of professional development with a view to leveling the playing field with respect to teacher skill and preparation. Most of the studies acknowledge that PBL is technically demanding and may require special skills that are either absent or under-used in the teachers included in the study. However, the amount and type of professional development offered to teachers in each study varied between week-long summer camps combined with a few follow up sessions throughout the year to no training at all. These differences in teacher training and coaching introduce potential variations in teacher quality and further complicate the interpretation of the results.

In the interests of a sound comparison with PBL, DI requires more precise definition. Almost lost in the shadow of issues surrounding what constitutes PBL instruction is the matter of carefully defining DI. The studies considered here all pay only

passing attention to the potential differences in DI between the control group instructors. None of the studies set out more than very casual constraints on what constitutes DI with most simply assuming that DI instruction was lecture-based and involved the use of a textbook. This failure to define DI with greater precision is a weakness in these studies. There can be wide variations in teachers' use of DI and in the way that they use lectures and text-based practices. Any future research should qualify the characteristics of DI and ensure that characteristics of this mode of teaching are accurately reflected in the control groups.

Instrumentation deserves careful development to assure appropriate measurement of the dependent variable. Arguably, the legitimacy of a study can be linked to the instruments used to test the relationship between the independent and dependent variable. Pre- and post-tests that do not measure, or measure poorly, this relationship are almost useless. In two of the studies, state and national standardized tests were used as a post-test with no pre-test given. These tests were designed to assess the overall efficacy of teaching for a large population of students, and are, by design, less appropriate than instruments purposely constructed for a particular study. For instance, in the case of Finkelstein et al. the TEL was developed specifically to test specific curricular outcomes. Both the curriculum and the test were developed *in concert* by the same organization. Further, the intervention was developed to be congruent with the outcomes and this test. This was a synergistic and wise methodological approach not seen in the other studies.

That which is statistically significant is not necessarily of practical significance. Of course, at the end of the day, what teachers and school administrators are interested in is answering the question: Is PBL better than DI for increasing students' content knowledge? The answer is "yes" PBL, is better in this regard but not much better. What was statistically significant appears only marginally important for practice. While all of the studies showed a larger increase in pre- and post-test scores for students in the PBL group, when converted into percentage scores, the increase was often very incremental. Typically, the PBL curriculum only accounted for about a 5% increase in the average test scores. Overall, effect sizes for PBL and DI were moderate with PBL interventions only marginally more effective than DI at increasing scores from pre-test to post-test.

Would meta-analysis be useful? In this thesis I have chosen to study the very specific, narrowly defined situation in which PBL can be compared quantitatively to DI with respect to the effects on student science knowledge. Emphasis has been placed on this outcome in each study, but also on the *quality* of each study according to an accepted set of criteria for evaluating educational research. These standards have allowed me to make judgments about which study's results should be given more consideration and which study's findings should be viewed with caution or even skepticism. In this regard my work here resembles the first steps in the process of meta-analysis. So, it is prudent at this stage to ask if a full and proper meta-analysis of these studies would yield further insight into the efficacy of PBL over DI?

Lipsey and Wilson (2001, p. 2) define meta-analysis as "a technique for encoding and analyzing the statistics that summarize research findings as they are typically presented in research reports." In meta-analysis, empirical data from two or more studies are combined to paint a broader, more meaningful picture of the effect of an intervention on the dependent variable. This type of statistical analysis relies heavily on effect sizes (Lipsey and Wilson, 2001) and provides a way of assigning a weight or value to a study. Often these weights are assigned by considering both sample size and the effect size.

Meta-analysis is often encountered in the field of medicine. Many studies examining the effect of a particular drug or medical intervention on the treatment of disease will be subjected to meta-analysis in an effort to determine more precisely the relationship between the independent and dependent variable. Further, the medical education field has employed meta-analysis as a way of examining the effectiveness of PBL in medical schools. However, many more studies exist comparing PBL to "traditional" teaching techniques at the medical school level than at the high-school level.

The wealth of studies comparing PBL to DI in medical schools is one of the reasons that meta-analysis has been chosen by researchers of medical education as an analytic technique. This is not to say that an abundance of research is a pre-requisite to effective meta-analysis. However, the consideration of many studies (dozens or even hundreds) and the inevitable variation in the quality of those studies does pose problems for the effective summary and amalgamation of the results, problems that can be accounted for by meta-analysis.

On the other hand, one of the pre-requisites to effective meta-analysis is the comparison of studies using the same effect size statistic. In the medical field, there is a

high degree of consistency between studies and the effect size statistics used. In this thesis, we do see the consistent use of Cohen's d as the effect size statistic used which is encouraging for the use of meta-analysis in this particular instance. The use of the same effect size statistic makes the seven studies "combinable" with respect to effect size. However, the factor that is most influential in making the studies combinable is the care taken to ensure that the studies have consistency in their independent and dependent variables (Lipsey and Wilson, 2001, p. 3). That is, in each of the studies, PBL looks very much the same.

However, determining the combinability of the studies requires more than just looking at the statistical measures used. To begin, the overall quality of the studies must first be determined from a set of pre-existing standards that can be applied to the area of research being considered. In effect, that has been the purpose of this thesis. We have applied McMillan and Wergin's (2006) standards to determine if the studies considered are, in fact, "quality studies". To that end, I have found serious flaws in many of the studies; flaws that preclude meta-analysis.

For instance, "Research synthesists evaluate the methodology of studies to determine if the manner in which the data were collected might make it inappropriate for addressing the problem at hand" (Cooper, 2010, p. 16). A rigorous application of quality standards is essential to preventing the phenomenon of "garbage in, garbage out"; poor studies, meta-analyzed, lead to erroneous results. In this thesis, we have seen that the poor randomization of the studies was a serious problem. In fact, this problem was so significant that one could argue none of the studies were truly "experimental" but are

"quasi-experimental" and therefore, by virtue of their design make meta-analysis less effective (Cooper, 2010, p. 34-35).

Another reason that the studies in this thesis make poor candidates for meta-analysis is the lack of consistency between control groups. As mentioned previously, there is no consistent definition of direct-instruction and, therefore, no certainty that each of the control groups is experiencing the same instruction. Further, inconsistency between teacher professional development introduces even more variability both between the experimental groups and control groups and *within* these groups as the amount of training teachers receive must have some effect on their performance.

Finally, meta-analysis is actually a shortcut to interpreting large amounts of data spread over many studies. As Lipsey and Wilson (2001, p. 2) point out, "If the full data sets for the studies of interest are available, it will generally be more appropriate and informative to analyze them directly using conventional procedures rather than meta-analyze summary statistics." Let us imagine that each of the studies considered in this thesis were excellent matches with McMillan and Wergin's (2006) criteria. It is not impossible to imagine that each of the researchers could be contacted and the original data sets obtained for a more precise, in-depth analysis of the findings. This method would be preferable to meta-analysis as it would provide a more detailed understanding of the conditions that lead to the observed effect(s) on the dependent variable.

To answer the original question, "would meta-analysis be useful" we see that, in this case, the answer is "no". The primary reason for this conclusion is that the body of research is of too poor quality to meet the requirements for effective meta-analysis. If the research were to meet McMillan and Wergin's (2006) criteria for quality educational

research then we may be able to justify a meta-analytic approach. However, even in this case, given the small number of studies it appears that analysis of the original, complete study data sets would be more revealing than the summary of effects offered by meta-analysis.

Implications

Further research of the effects of PBL versus DI is needed. In particular, adequate samples of known populations should be considered with PBL curricula designed in concert with specific curricular objectives, teacher professional development, and effective measurement instruments. It should be the curricular outcomes that govern the populations for the research. For instance, in Manitoba science curricula are set provincially. Therefore, province-wide studies are appropriate as it is easier to ensure that one set of curricular objectives can be set to a PBL curriculum with subsequent evaluation by one instrument designed specifically from that curriculum. This method is opposed to the use of a national instrument that may not be tied exactly to the provincial curriculum. The assessment instrument is important and arguably more difficult to construct than the PBL curriculum. Therefore, the pre- and post-test should be constructed by experts in psychometrics and its development must be a critical part of the study. Without trustworthy instruments it is difficult, if not impossible, to make causal claims or valid interpretations of the research.

Even with the obvious weaknesses, PBL has garnered wide support in Canada and the United States. My school division is firmly pushing to have PBL along with other modes of "inquiry learning" set as the division-wide standard of instruction. I believe that

all instructional policies must have their origin rigorous research and be supported by the empirical evidence. Therefore, a richer body of research needs to compare PBL and DI before instructional policies are imposed on teachers. In particular, there does not yet exist a group of "gold standard" studies using the same curriculum, PBL intervention, professional development regimes, and test instruments. At least one high-quality study that can be replicated a number of times and form the basis for the acceptance or refutation of PBL over DI is warranted. Ultimately, it is by ensuring a close link between practice, curriculum, professional development, and the study instruments that will lead to the most revealing, valid, and reliable research.

Obviously, the focus of this thesis is narrow, and admittedly it does not paint a complete picture of the potential benefits of PBL over DI. It is clear that, beyond a certain point, gains in student learning are marginal with many factors combining to achieve small gains. Although teacher training and enthusiasm, curriculum design, and effective administration of schools have effects on student levels of success there are many other factors that likely affect students' learning. The current educational research, although far from conclusive, seems to indicate that employing PBL provides only incremental improvement in learning. This is to say nothing of outcomes not examined in this research literature such as the engagement or overall enjoyment and satisfaction both students and teachers experience in a PBL classroom. Finally, these small positive effects must be viewed in light of the complexities that come with implementing a PBL curriculum in an individual classroom, division, state or province. Perhaps the cost of implementing such a program with so little support from the research literature is too high for the potential gains that could possibly be achieved.

References

Akinoglu, O., Tandogan, R. (2007). The effects of problem-based active learning in science education on students' academic achievement, attitude and concept learning. *Eurasia Journal of Mathematics, Science & Technology Education,* 3(1), 71-81.

Bakula, N. (2010). The benefits of formative assessments for teaching and learning. *Science Scope,* 34(1), 37-43.

Barrow, R. (2006). Empirical research into teaching. *Interchange,* 37(4), 287-307.

Barrows, H. (1994). *Practice-based learning: Problem-based learning applied to medical education.* Southern Illinois University School of Medicine. Springfield, Illinois, USA.

Bowman, N. (2012). Effect sizes and statistical methods for meta-analysis in higher education. *Research in Higher Education,* 53, 375-382.

Bruce, C., & Ross, J. (2007). Teacher self-assessment: A mechanism for facilitating professional growth. *Teaching and Teacher Education,* 23, 146-159.

Buck Institute for Education. (2009) Accessed August 12, 2012 http://www.bie.org/research/study/does_pbl_work

Canadians for 21st Century Learning and Innovation. (2012). *Shifting Minds: A 21st Century Vision for Public Education in Canada.* Accessed September 20, 2012. http://www.c21canada.org/wp-content/uploads/2012/10/Summit-design-English-version-Sept.-26.pdf

Chaddock, G. R. (1998, August 25). Perils of the Pendulum: Resisting Education's Fads. *The Christian Science Monitor*.

Chang, C-Y. (2001). Comparing the impacts of a problem-based computer-assisted instruction and the direct-interactive teaching method on student science achievement. *Journal of Science Education and Technology,* 10 (2), 147-153.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach.* London, Sage.

Cottrell, R., & McKenzie, J. (2010). *Health Promotion & Education Research Methods: Using the Five Chapter Thesis/Dissertation Model: Using the Five Chapter Thesis/Dissertation Model*. Jones & Bartlett Learning.

Crosthwaite, C., Cameron, I., Lant, P., & Litster, J. (2006). Balancing curriculum processes and content in a project centered curriculum: In pursuit of graduate attributes. *Chemical Engineering Research and Design,* 84, 619-628.

DeWitt, P. (2012). Partnership for 21st Century Skills: An Interview with Tim Magner. Retrievedfromhttp://blogs.edweek.org/edweek/finding_common_ground/2012/07/partnership_for_21st_century_skills_an_interview_with_tim_magner.html?qs=tim+magner+interview

Drake, K. & Long, D. (2009). Rebecca's in the dark: A comparative study of problem-based learning and direct instruction/experiential learning in two 4th-grade classrooms. *Journal of Elementary Science Education,* 21, (1), 1-16.

Delazio, J. (2006). Theory into practice: A bridge too far?. *Association for the Advancement of Computing In Education Journal*, 14 (3), 221-233.

Farrow, R. (2003). The effectiveness of PBL: the debate continues. Is meta-analysis helpful? *Medical Education*, 37, 1131-1132.

Field, A. (2005). *Discovering statistics using SPSS*. London: Sage Publications Ltd.

Finkelstein, N., Hanson, T., Huang, C.W., Huang, M., & Hirschman, B. (2010). *Effects of problem based economics on high school economics instruction.* Report prepared for the Institute of Education Sciences and the U.S. Department of Education.

Gage, N. (1985). *Hard gains in the soft sciences: The case for pedagogy.* Bloomington, IL: Phi Delta Kappa.

Geier, R., Blumenfeld, P., Marx, R., Krajcik, J., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008) Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. *Journal of Research in Science Teaching*, 45 (8), 922-939

Health Canada. (2001). *How drugs are reviewed in Canada*. Accessed October 23, 2012. http://www.hc-sc.gc.ca/dhp-mps/prodpharma/activit/fs-fi/reviewfs_examenfd-eng.php

Katzer, J., Cook, K., & Crouch, W. (1998). *Evaluating information: A guide for users of social science research*. Boston, MA: Mcgraw-Hill.

Linde, K., Scholz, M., Melchart, D., & Willich, N. (2002). Should systematic reviews include non-randomized and uncontrolled studies? The Case of Acupuncture for Chronic Headache. *Journal of Clinical Epidemiology*, 55, 77-85.

Liu, M., Hsieh, P., Cho, Y., & Schallert, D. (2006). Middle school students' self-efficacy, attitudes, and achievement in a computer-enhanced problem-based learning environment. *Journal of Interactive Learning Research*, 17(3), 225-242.

Lunenberg, F. (2011). Theorizing about curriculum: Conceptions and definitions. *International Journal of Scholarly Academic Intellectual Diversity,* 13(1), 1-6.

Marzano, R. (2003). *What works in schools: Translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Maxwell, N., Mergendoller, J., & Bellisimo, Y. (2005). Problem-based learning and high school macroeconomics: A comparative study of instructional methods. *Research in Economic Education,* 36, 315-331.

McMillian, J.H., & Wergin, J. F. (2006). *Understanding and evaluating educational research*. Columbus, OH: Pearson Prentice Hall.

Nelson, L. (1966). Teacher leadership: An empirical approach to analyzing teacher behavior in the classroom. *Journal of Teacher Education,* 17(4), 417-425.

Partnership for 21st Century Skills. http://www.p21.org/, Accessed October 15, 2012.

P21 Common Core Toolkit: A Guide to Aligning the Common Core State Standards with the Framework for 21st Century Skills, Accessed October 1, 2012 http://www.p21.org/index.php?option=com_content&view=article&id=1005&Itemid=236 The Partnership for 21st Century Skills. 2011

Partnership for 21st Century Skills. Our Mission. Retrieved from www.p21.org 21st Century Learning Initiative, February 12, 2000.

Schneider, R., Krajcik, J., Marx, R., & Soloway, E. (2002). Performance of students in project-based science classrooms on a national measure of science achievement. *Journal of Research in Science Teaching,* 39 (5), 410-422.

Slavin, R. E. (1989). PET and the pendulum: Faddism in education and how to stop it. *Phi Delta Kappan*, *70*(10), 752-758.

Stanovich, P. J., & Stanovich, K. E. (2003). Using research and reason in education: How teachers can use scientifically based research to make curricular & instructional decisions. Portsmouth, NH: RMC Research Corporation,

Stronge, J. (1997). *Evaluating teaching: A guide to current thinking and best practice*. Thousand Oaks, CA: Corwin.

Zwaagstra, M.C., Clifton, R.A., & Long, J.C. (2010), *What's wrong with our schools: And how we can fix them.* Lanham, MD. Rowman & Littlefield.