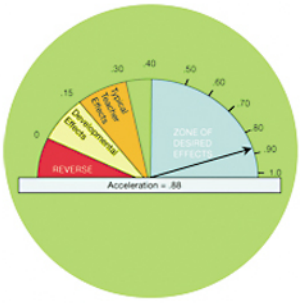
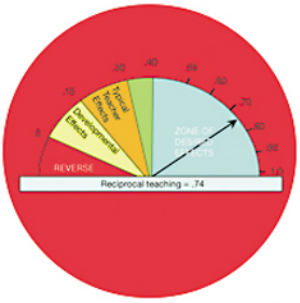
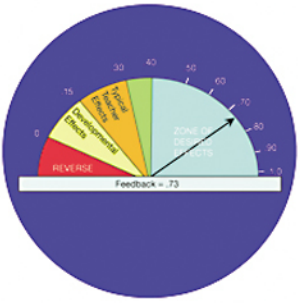
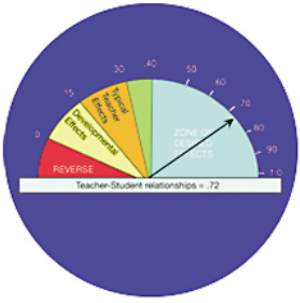
VISIBLE LEARNING

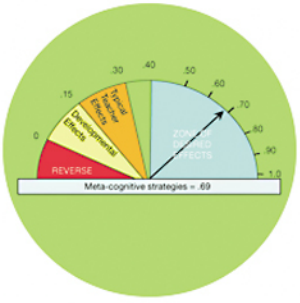
A SYNTHESIS OF OVER 800 META-ANALYSES RELATING TO ACHIEVEMENT

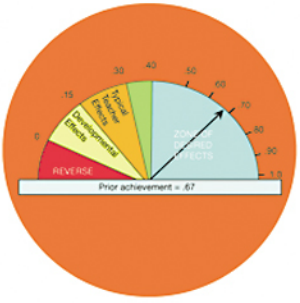


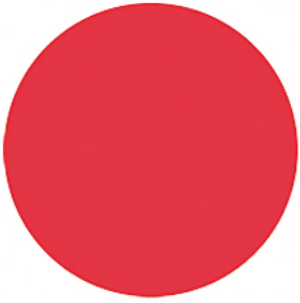












Visible Learning

This unique and ground-breaking book is the result of 15 years' research and synthesises over 800 meta-analyses relating to the influences on achievement in school-aged students. It builds a story about the power of teachers and of feedback, and constructs a model of learning and understanding.

Visible Learning presents research involving many millions of students and represents the largest ever collection of evidence-based research into what actually works in schools to improve learning. Areas covered include the influences of the student, home, school, curricula, teacher, and teaching strategies. A model of teaching and learning is developed based on the notion of visible teaching and visible learning.

A major message within the book is that what works best for students is similar to what works best for teachers. This includes an attention to setting challenging learning intentions, being clear about what success means, and an attention to learning strategies for developing conceptual understanding about what teachers and students know and understand.

Although the current evidence-based fad has turned into a debate about test scores, this book is about using evidence to build and defend a model of teaching and learning. A major contribution to the field, it is a fascinating benchmark for comparing many innovations in teaching and schools.

John Hattie is Professor of Education and Director of the Visible Learning Labs, University of Auckland, New Zealand.

Visible Learning

A synthesis of over 800 meta-analyses relating to achievement

John A. C. Hattie

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2009
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
270 Madison Avenue, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group, an informa business

This edition published in the Taylor & Francis e-Library, 2008.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk."

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
Hattie, John.
Visible learning: a synthesis of meta-analyses relating to achievement/John A. C. Hattie.
p. cm.
Includes bibliographical references.
1. Learning—Longitudinal studies. 2. Teaching—Longitudinal studies. 3. Effective teaching—Longitudinal studies. 4. Teacher effectiveness—Longitudinal studies. I. Title.
LB1060.H388 2008
370.15'23—dc22 2008021702

ISBN 0-203-88733-6 Master e-book ISBN

ISBN10: 0-415-47617-8 (hbk)
ISBN10: 0-415-47618-6 (pbk)
ISBN10: 0-203-88733-6 (ebk)

ISBN13: 978-0-415-47617-1 (hbk)
ISBN13: 978-0-415-47618-8 (pbk)
ISBN13: 978-0-203-88733-2 (ebk)

Contents

Tables

Figures

Preface

Elliott is my hero. On his fifth birthday he was diagnosed with leukemia, and this past year has been his *annus horribilis*. On the day of the diagnosis, it was impressive to see the medical team immediately begin interventions. While they aimed to make Elliott stable, the diagnosis regime burst into action. They knew which tests were needed to make the correct diagnosis and when they were satisfied with the initial diagnosis they immediately moved to interventions. Thus began a year of constant monitoring and feedback to the medical team about Elliott's progress. All throughout they collected evidence of progress, they knew what success looked like, and kept all informed about this evidence. Elliott went through many ups and downs, lost his hair (as did I when he gave me a No. 1 cut as his Christmas present, although I drew a line when he asked to shave my eyebrows off as well), and had daily injections in the front of his legs, but he never balked, and throughout the treatment maintained his sparkly personality. The family was never in the dark about what was happening, books were provided, sessions offered, and support for treatment was excellent. The messages in this book owe a lot to Elliott.

This book started in Gil Sax's office in 1990 searching and coding meta-analyses. Motivation to continue the search was inspired by Herb Walberg, and continued in Perth in Australia, North Carolina in the US, and finished here in Auckland in New Zealand. It is a journey that has taken 15 years. The messages have been questioned, labelled provocative, liked, and dismissed, among other more positive reactions. The typical comments are: "the results do not mirror my experience", "why have you not highlighted my pet method", "you are talking about averages and I'm not average", and "you are missing the nuances of what happens in classrooms". There are many criticisms and misunderstandings about what I am *and* am not saying.

So let me start with what this book is not.

1 It is *not* a book about classroom life, and does not speak to the nuances and details of what happens within classrooms. Instead it synthesizes research based on what happens in classrooms; as it is more concerned with main effects than interactions. Although I have spent many hundreds of hours in classrooms in many countries, have observed, interviewed, and aimed to dig quite deeply into the nuances of classrooms, this book will not show these details of class living.
2 It is *not* a book about what cannot be influenced in schools—thus critical discussions about class, poverty, resources in families, health in families, and nutrition are not included—but this is NOT because they are unimportant, indeed they may be more

important than many of the influences discussed in this book. It is just that I have not included these topics in my orbit.

3 It is *not* a book that includes qualitative studies. It only includes studies that have used basic statistics (means, variances, sample sizes). Again, this should not mean qualitative studies are not important or powerful but just that I have had to draw some lines around what can be accomplished over a 15-year writing span.

4 It is *not* a book about criticism of research, and I have deliberately not included much about moderators of research findings based on research attributes (quality of study, nature of design) again not because these are unimportant (my expertise is measurement and research design), but because they have been dealt with elsewhere by others (e.g., Lipsey & Wilson, 1993; Sipe & Curlette, 1996a, 1996b).

Rather this is a book about synthesizing many meta-analyses. It is based on over 50,000 studies, and many millions of students—and this is a cut down version of what I could have included as I also collected studies on affective and physical outcomes and on many other outcomes of schooling. I occasionally receive emails expressing disbelief that I have had the time to read so many studies. No, I have not read all primary studies, but as will be seen I have read all meta-analyses, and in some cases many of the primary studies. I am an avid reader, thoroughly enjoy learning the arts of synthesizing and detecting main ideas, and want to create explanations from the myriad of ideas in our discipline. The aim of this book is not to overwhelm with data—indeed my first attempt was discarded after 500 pages of trenchant details; who would care about such details? Instead this book aims to have a message, a story, and a set of supporting accounts of this story.

The message about schools is a positive one. So often when talking about the findings in this book, teachers think I am attacking them as below average, non-thinking, boring drones. In New Zealand, for example, it is clear to me why we rank in the top half-dozen nations in reading, mathematics, and science—we have a nation of excellent teachers. They exist and there are many of them. This book is a story of many real teachers I have met, seen, and some who have taught my own boys. Many teachers already think in the ways I argue in this book; many are seeking to always improve and constantly monitor their performances to make a difference to what they do; and many inspire the love of learning that is one of the major outcomes of any school. This is not a book claiming that teachers are below par, that the profession is terrible, and that we all need to "put in more effort and do better". Nearly all studies in the book are based on real students in front of real teachers in real schools—and that so many of the effects are powerful is a testament that excellence is happening. *The* major message is that we need a barometer of what works best, and such a barometer can also establish guidelines as to what is excellent—too often we shy from using this word thinking that excellence is unattainable in schools. Excellence is attainable: there are many instances of excellence, some of it fleeting, some of it aplenty. We need better evaluation to acknowledge and esteem it when it occurs—as it does.

Acknowledgments

There are so many who have contributed to the data, the book, and the message, and who have provided feedback over these past 15 years: Nola Purdie, Krystoff Krawowski, Richard Fletcher, Thakur Karkee, Earl Irving, Trisha Lundberg, Lorrae Ward, Michael Scriven, Richard Jaeger, Geoff Petty, and Russell Bishop. I am especially indebted to Janet Rivers for her attention to the details and to Debbie Waayer for her remarkable skills in finding articles, referencing, and data skills, and in ensuring that I completed this book. Others have been critics and this is among the more welcome contributions for any author: Lexie Grudnoff, Gavin Brown, Adrienne Alton-Lee, Christine Rubie-Davis, Misty Sato, David Moseley, Heidi Leeson, Brian Marsh, Sandra Frid, Sam Stace, and John Locke. I particular thank Gene Glass for his development of meta-analysis that allowed me and many others to stand on his shoulders to peer into what makes a difference to teaching and learning.

But most of all I thank my family—they have endured this book, shaped the many versions of the message, and provided the feedback that only a loving family can give. Unlike most children who are asked about their day at school each night at the dinner table, my boys have endured the same interrogation every night of their school years: What feedback did you receive about your learning today? Thanks to my boys—Joel, Kyle, Kieran, Billy, Bobby, and Jamie—you are my inspirations for living. And most of all to Janet—the one who has given unconditional positive regard through the ups and downs of moving a family across many countries, putting up with "yet another study", and being the love of my life. The size of your effect on my life exceeds any reported in this book.

Chapter 1

The challenge

In the field of education, one of the most enduring messages is that "everything seems to work". It is hard to find teachers who say they are "below average" teachers, and everyone (parent, politician, school leader) has a reason why their particular view about teaching or school innovation is likely to be successful. Indeed, rhetoric and game-play about teaching and learning seems to justify "everything goes". We acknowledge that teachers teach differently from each other; we respect this difference and even enshrine it in terms like "teaching style" and "professional independence". This often translates as "I'll leave you alone, if you leave me alone to teach my way." While teachers talk to their colleagues about curriculum, assessment, children, and lack of time and resources, they rarely talk about their teaching, preferring to believe that they may teach differently (which is acceptable provided they do not question one another's right to teach in their particular ways). We pass laws that are more about structural concerns than about teaching concerns: such as class size, school choice, and social promotion, as if these are clear winners among the top-ranking influences on student learning. We make school-based decisions about ability grouping, detracking or streaming, and social promotion, again appealing to claims about influences on achievement. For most teachers, however, teaching is a private matter; it occurs behind a closed classroom door, and it is rarely questioned or challenged. We seem to believe that every teacher's stories about success are sufficient justification for leaving them alone. We will see throughout this book that there is a good reason for acknowledging that most teachers can demonstrate such success. Short of unethical behaviors, and gross incompetence, there is much support for the "everything goes" approach. However herein lies a major problem.

It is the case that we reinvent schooling every year. Despite any successes we may have had with this year's cohort of students, teachers have to start again next year with a brand new cohort. The greatest change that most students experience is the level of competence of the teacher, as the school and their peers typically are "similar" to what they would have experienced the previous year. It is surely easy to see how it is tempting for teachers to re-do the successes of the previous year, to judge students in terms of last year's cohort, and to insist on an orderly progression through that which has worked before. It is required of teachers, however, that they re-invent their passion in their teaching; they must identify and accommodate the differences brought with each new cohort of students, react to the learning as it occurs (every moment of learning is different), and treat the current cohort of students as if it is the first time that the teacher has taught a class—as it is for the students with this teacher and this curricula.

As will be argued throughout this book, the act of teaching reaches its epitome of

success after the lesson has been structured, after the content has been delivered, and after the classroom has been organized. The art of teaching, and its major successes, relate to "what happens next"—the manner in which the teacher reacts to how the student interprets, accommodates, rejects, and/or reinvents the content and skills, how the student relates and applies the content to other tasks, and how the student reacts in light of success and failure apropos the content and methods that the teacher has taught. Learning is spontaneous, individualistic, and often earned through effort. It is a timeworn, slow and gradual, fits-and-starts kind of process, which can have a flow of its own, but requires passion, patience, and attention to detail (from the teacher and student).

So much evidence

The research literature is rich in recommendations as to what teachers and schools should do. Carpenter (2000), for example, counted 361 "good ideas" published in the previous ten years of *Phi Delta Kappan* (e.g., Hunter method, assertive discipline, Goals 2000, TQM, portfolio assessment, essential schools, block scheduling, detracking, character education). He concluded that these good ideas have produced very limited gains, if any. Similarly, Kozol (2005, p. 193) noted that there have been "galaxies of faded names and optimistic claims," such as "Focus Schools", "Accelerated Schools", "Blue Ribbon Schools", "Exemplary Schools", "Pilot Schools", "Model Schools", "Quality Schools", "Magnet Schools", and "Cluster Schools"—all claiming they are better and different, with little evidence of either. The research evidence relating to "what works" is burgeoning, even groaning, under a weight of such "try me" ideas. Most are justified by great stories about lighthouse schools, inspiring principals and inspiring change agents, and tales of wonderful work produced by happy children with contented parents and doting teachers. According to noted change-theory expert, Michael Fullan, one of the most critical problems our schools face is "not resistance to innovation, but the fragmentation, overload, and incoherence resulting from the uncritical and uncoordinated acceptance of too many different innovations (Fullan & Stiegelbauer, 1991, p. 197). Richard Elmore (1996) has long argued that education suffers not so much from an inadequate *supply* of good programs as from a lack of *demand* for good programs—and instead we so often *supply* yet another program rather than nurture *demand* for good programs.

There is so much known about what makes a difference in the classroom. A glance at the journals on the shelves of most libraries, and on web pages, would indicate that the state of knowledge in the discipline of education is healthy. The worldwide picture certainly is one of plenty; we could have a library solely consisting of handbooks about teaching, most of which cannot be held in the hand. Most countries have been through many waves of reform, including new curricula, new methods of accountability, reviews of teacher education, professional development programs, charter schools, vouchers, and management models. We have blamed the parents, the teachers, the classrooms, the resources, the textbooks, the principals, and even the students. Listing all the problems and all the suggested remedies could fill this book many times over.

There are thousands of studies promulgating claims that this method works or that innovation works. We have a rich educational research base, but rarely is it used by teachers, and rarely does it lead to policy changes that affect the nature of teaching. It may be that the research is written in a non-engaging style for teachers, or maybe when research is presented to teachers it is done in a manner that fails to acknowledge

that teachers come to research with strong theories of their own about what works (for them). Further, teachers are often very "context specific", as the art for many of them is to modify programs to fit their particular students and teaching methods—and this translation is rarely acknowledged.

How can there be so many published articles, so many reports providing directions, so many professional development sessions advocating this or that method, so many parents and politicians inventing new and better answers, while classrooms are hardly different from 200 years ago (Tyack & Cuban, 1995)? Why does this bounty of research have such little impact? One possible reason is the past difficulties associated with summarizing and comparing all the diverse types of evidence about what works in classrooms. In the 1970s there was a major change in the manner that we reviewed the research literature. This approach offered a way to tame the massive amount of research evidence so that it could offer useful information for teachers. The predominant method had always been to write a synthesis of many published studies in the form of an integrated literature review. However in 1976 Gene Glass introduced the notion of meta-analysis—whereby the effects in each study, where appropriate, are converted to a common measure (an effect size), such that the overall effects could be quantified, interpreted, and compared, and the various moderators of this overall effect could be uncovered and followed up in more detail. Chapter 2 will outline this method in more detail. This method soon became popular and by the mid 1980s more than 100 meta-analyses in education were available. This book is based on a synthesis (a method referred to by some as meta-meta-analysis) of more than 800 meta-analyses about influences on learning that have now been completed, including many recent ones. It will develop a method such that the various innovations in these meta-analyses can be ranked from very positive to very negative effects on student achievement. It demonstrates that the reason teachers can so readily convince each other that they are having success with their particular approach is because the reference point in their arguments is misplaced. Most importantly, it aims to derive some underlying principles about why some innovations are more successful than others in influencing student achievement.

An explanatory story, not a "what works" recipe

The aim is to provide more than a litany of "what works", as too often such lists provide yet another set of recommendations devoid of underlying theory and messages, they tend to not take into account any moderators or the "busy bustling business" of classrooms, and often they appeal to claims about "common sense". If common sense is the litmus test then everything could be claimed to work, and maybe therein lies the problems with teaching. As Glass (1987) so eloquently argued when the first *What Works: Politics and research* was released, such appeals to common sense can mean that there is no need for more research dollars. Such claims can ignore the realities of classroom life, and they too often mistake correlates for causes. Michael Scriven (1971; 1975; 2002) has long written about mistaking correlates of learning with causes. His claim is that various correlates of school outcomes, say the use of advance organizers, the maintenance of eye contact, or high time on task, should not be confused with good teaching. While these may indeed be correlates of learning, it is still the case that good teaching may include none of these attributes. It may be that increasing these behaviors in some teachers also leads to a decline in other attributes (e.g., caring and respect for students). Correlates, therefore, are not to be confused with the causes.

For example, one of the major results presented in this book relates to increasing the amount of feedback because it is an important correlate of student achievement. However, one should not immediately start providing more feedback and then await the magical increases in achievement. As will be seen below, increasing the amount of feedback in order to have a positive effect on student achievement requires a change in the conception of what it means to be a teacher; it is the feedback to the teacher about what students can and cannot do that is more powerful than feedback to the student, and it necessitates a different way of interacting and respecting students (but more on this later). It would be an incorrect interpretation of the power of feedback if a teacher were to encourage students to provide more feedback. As Nuthall (2007) has shown, 80% of feedback a student receives about his or her work in elementary (primary) school is from other students. But 80% of this student-provided feedback is incorrect! It is important to be concerned about the climate of the classroom before increasing the amount of feedback (to the student or teacher) because it is critical to ensure that "errors" are welcomed, as they are key levers for enhancing learning. It is critical to have appropriately challenging goals as then the amount and directedness of feedback is maximized. Simply applying a recipe (e.g., "providing more feedback") will not work in our busy, multifaceted, culturally invested, and changing classrooms.

The wars as to what counts as evidence for causation are raging as never before. Some have argued that the only legitimate support for causal claims can come from randomized control trials (RCTs, i.e., trials in which subjects are allocated to an experimental or a control group according to a strictly random procedure). There are few such studies among the many outlined in this book, although it could be claimed that there are many "evidence-informed" arguments in this book. While the use of randomized control trials is a powerful method, Scriven (2005) has argued that a higher gold standard relates to studies that are capable of establishing conclusions "beyond reasonable doubt". Throughout this book, many correlates will be presented, as most meta-analyses seek such correlates of enhanced student achievement. A major aim is to weave a story from these data that has some convincing power and some coherence, although there is no claim to make these "beyond reasonable doubt". Providing explanations is sometimes more difficult than identifying causal effects.

Most of these claims about design and RCTs are part of the move towards evidence-based decision making, and the current debate about influences on student learning is dominated by discussion of the need for "evidence". Evidence-based this and that are the buzz words, but while we collect evidence, teachers go on teaching. The history of teaching over the past 200 years has attested the enduring focus of teachers on notions of "what works" despite the number of solutions urging teachers to move in a different direction. Such "what works" notions rarely have high levels of explanatory power. The model I will present in Chapter 3 may well be speculative, but it aims to provide high levels of explanation for the many influences on student achievement as well as offer a platform to compare these influences in a meaningful way. And while I must emphasize that these ideas are clearly speculative, there is both solace and promise in the following quotation from Popper:

> Bold ideas, unjustified anticipations, and speculative thought, are our only means for interpreting nature: our only organon, our only instrument, for grasping her. And we must hazard them to win our prize. Those among us who are unwilling to expose their ideas to the hazard of refutation do not take part in the scientific game.
>
> (Popper, K. R., 1968, p. 280)

While we collect evidence, teachers go on teaching

As already noted, the practice of teaching has changed little over the past century. The "grammar" of schooling, in Tyack and Cuban's (1995) terms, has remained constant: the age-grading of students, division of knowledge into separate subjects, and the self-contained classroom with one teacher. Many innovations have been variously "welcomed, improved, deflected, co-opted, modified, and sabotaged" (p. 7), and schools have developed rules and cultures to control the way people behave when in them. Most of us have been "in school" and thus know what a "real school" is and should be. The grammar of schooling has persisted partly because it enables teachers to discharge their duties in a predictable fashion, cope with the everyday tasks that others expect of them, and provide much predictability to all who encounter schools.

One of the "grammars of schooling" is that students are to be made responsible for their learning. This can easily turn into a conception that some students are deficient in their desire for, and achievements from teaching. As Russell Bishop and his colleagues have demonstrated, such deficit thinking is particularly a problem when teachers are involved with minority students (e.g., Bishop, Berryman, & Richardson, 2002). From their interviews, they illustrated that the influences on Māori students' educational achievement differed for each of parents, students, principals, and teachers (Figure 1.1). Students, parents, and principals see the relationships between teachers and students as having the greatest influence on Māori students' educational achievement. In contrast, teachers identify the main influences on Māori students' educational achievement as being Māori students themselves, their homes and/or the structure of the schools. Teachers engage in the discourse of the child and their home by pathologising Māori students' lived experiences and by explaining their lack of educational

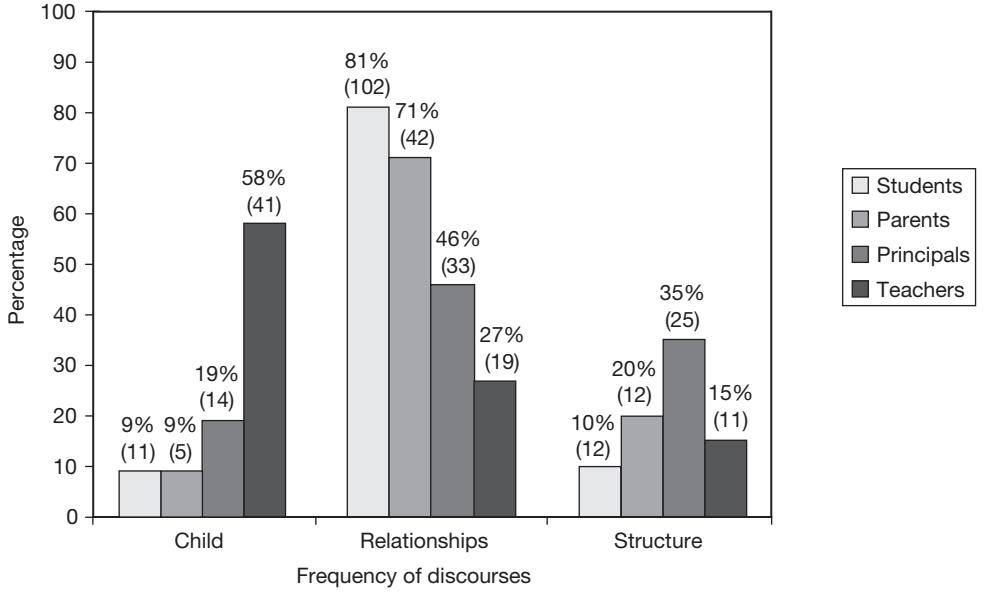


Figure 1.1 Percentage of responses as to the claimed influences on student learning by students, parents, principals, and teachers

achievement in deficit terms. My colleague Alison Jones calls this type of thinking a "discourse of disadvantage" (Jones & Jacka, 1995). They do not see themselves as the agents of influence, see very few solutions, and see very little that they can do to solve the problems.

From their extensive classroom observations, analyses of achievement results, and working with teachers of minority students, Bishop *et al.* have devised a model of teaching Māori students based on caring for all students, and the primacy of the act of teaching. The major features of Bishop's model include the creation of a visible, appropriate context for learning such that the student's culture is involved in a process of co-learning, which involves the negotiation of learning contexts and content. The teacher provides supportive feedback and helps students to learn by acknowledging and using the students' prior knowledge and experiences, and monitoring to check if students know what is being taught, what is to be learnt, or what is to be produced. It involves the teacher teaching the students something, instructing them how to produce something, and giving them instructions as to the processes of learning. This is a high level of teaching activity, indeed.

Concluding comments

This introduction has highlighted the amazing facility of those in the education business to invent solutions and see evidence for their pet theories and for their current actions. Everything seems to work in the improvement of student achievement. There are so many solutions and most have some form of evidence for their continuation. Teachers can thus find some support to justify almost all their actions—even though the variability about what works is enormous. Indeed, we have created a profession based on the principle of "just leave me alone as I have evidence that what I do enhances learning and achievement".

One aim of this book is to develop an explanatory story about the key influences on student learning—it is certainly not to build another "what works" recipe. The major part of this story relates to the power of directed teaching, enhancing what happens next (through feedback and monitoring) to inform the teacher about the success or failure of their teaching, and to provide a method to evaluate the relative efficacy of different influences that teachers use.

It is important from the start to note at least two critical codicils. Of course, there are many outcomes of schooling, such as attitudes, physical outcomes, belongingness, respect, citizenship, and the love of learning. This book focuses on student achievement, and that is a limitation of this review. Second, most of the successful effects come from innovations, and these effects from innovations may *not* be the same as the effects of teachers in regular classrooms—the mere involvement in asking questions about the effectiveness of any innovation may lead to an inflation of the effects. This matter will be discussed in more detail in the concluding chapter, where an attempt is made to identify the effects of "typical" teachers compared to "innovations" in teaching. Indeed, the role of "teaching as intervention" is developed throughout the chapters in this book.

Chapter 2

The nature of the evidence

A synthesis of meta-analyses

> It is the mark of an educated man … that in every subject he looks for only so much precision as its nature permits.
>
> (Aristotle, 350BC)

This chapter outlines the methodology relating to the evidence used in the remainder of this book. The fundamental unit of analysis is 800+ meta-analyses and how the major results from these studies can be placed along a single continuum. The chapter then outlines some of the problems of meta-analyses, discusses some of the previous attempts to synthesize meta-analyses, and then introduces some of the major overall findings from the synthesis of the 800+ meta-analyses.

Would it not be wonderful if we could create a single continuum of achievement effects, and locate all possible influences of achievement on this continuum? Figure 2.1 shows one possible depiction of this continuum.

Influences on the left of this continuum are those that decrease achievement, and those on the right increase achievement. Those near the zero point have no influence on achievement outcomes.

The next task was to adopt an appropriate scale so that as many outcomes as possible from thousands of studies are converted to this single scale. This was accomplished using effect sizes, and this scale has been among the marvelous advances in the analysis of research studies over the past century. An effect size provides a common expression of the magnitude of study outcomes for many types of outcome variables, such as school achievement. An effect size of $d = 1.0$ indicates an increase of one standard deviation on the outcome—in this case the outcome is improving school achievement. A one standard deviation increase is typically associated with advancing children's achievement by two to three years, improving the rate of learning by 50%, or a correlation between some variable (e.g., amount of homework) and achievement of approximately $r = 0.50$. When

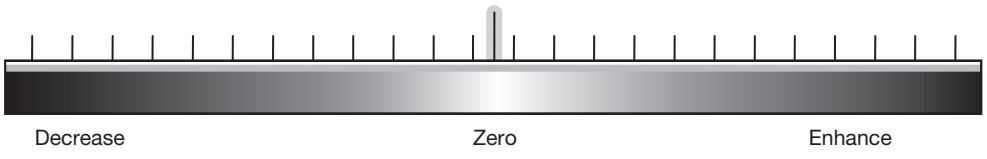


Figure 2.1 An achievement continuum

implementing a new program, an effect size of 1.0 would mean that, on average, students receiving that treatment would exceed 84% of students not receiving that treatment.

Cohen (1988) argued that an effect size of $d = 1.0$ should be regarded as a large, blatantly obvious, and grossly perceptible difference, and as an example he referred to the difference between the average IQ of PhD graduates and high school students. Another example is the difference between a person at 5'3" (160 cm) and 6'0" (183 cm)—which would be a difference visible to the naked eye. The use of effect sizes highlights the importance of the magnitude of differences, which is contrary to the usual emphasis in much of our research literature on statistical significance. Cohen (1990) has commented that "under the sway of the Fisherian scheme [or dependence on statistical significance], there has been little consciousness of how big things are … science is inevitably about magnitudes … and meta-analysis makes a welcome force toward the accumulation of knowledge" (pp. 1309–1310).

Thus, we have a continuum and a scale (effect size) to ascertain which of the many possible influences affect achievement. Many textbooks detail how effect sizes can be calculated from various summary statistics such as *t*-tests, ANOVAs, repeated-measures (e.g., Glass, 1977; Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985). Statistically, an effect size can be calculated in two major ways:

$$\text{Effect size} = [\text{Mean}_{\text{treatment}} - \text{Mean}_{\text{control}}]/\text{SD}$$

or

$$\text{Effect size} = [\text{Mean}_{\text{end of treatment}} - \text{Mean}_{\text{beginning of treatment}}]/\text{SD}$$

where SD is the pooled sample standard deviation. There are many minor modifications to these formulas, and for more detail the interested reader is referred to Glass, McGaw, & Smith (1981); Rosenthal (1991); Hedges & Olkin (1985); Hunter & Schmidt (1990); and Lipsey & Wilson (2001).

As an example of synthesizing meta-analyses, take an examination of five meta-analyses on homework: Cooper (1989; 1994); Cooper, Robinson, & Patall (2006); DeBaz (1994); Paschal, Weinstein, & Walberg (1984). Over these five meta-analyses there were 161 studies involving more than 100,000 students, which investigated the effects of homework on students' achievement. The average of all these effect sizes was $d = 0.29$, which can be used as the best typical effect size of the influence of homework on achievement. Thus, compared to classes without homework, the use of homework was associated with advancing children's achievement by approximately one year, improving the rate of learning by 15%, about 65% of the effects were positive (that is, improved achievement), 35% of the effects were zero or negative, and the average achievement level of students in classes that prescribed homework exceeded 62% of the achievement levels of the students not prescribed homework. However, an effect size of $d = 0.29$ would not, according to Cohen (1988), be perceptible to the naked eye, and would be approximately equivalent to the difference between the height of a 5'11" (180 cm) and a 6'0" (182 cm) person.

Thus it is possible to devise a unidimensional continuum such as shown in Figure 2.1 that can allow the various effects on achievement to be positioned as they relate to each other. The scale is expressed in effect sizes (or standard deviation units) such that 1.0 is an unlikely—although a very obvious—change in achievement, and 0.0 is no change at all.

This continuum provides the measurement basis to address the question of the relative effects of many factors on achievement.

An alternative way of considering the meaning of an effect size was suggested by McGraw and Wong (1992). They introduced a measure called the common language effect size indicator, which is the probability that a score sampled from one distribution will be greater than a score sampled from some other distribution. Consider as an example the difference in height of the average woman (5'4"/162.5 cm) and the average male (5'10"/177.5 cm), which is a d of 2.0. This d translates into a common language effect (CLE) of 92 percent. Thus we can estimate that in any random pairing the probability of the male being taller than the female is $d = 0.92$; or that in 92 out of 100 blind dates the male will be taller than the female. Now, using the example above, consider the $d = 0.29$ from introducing homework (throughout this book effect sizes are abbreviated, following tradition, to d). The CLE is 21 percent, so that in 21 times out of 100, introducing homework into schools will make a positive difference, or 21 percent of students will gain in achievement compared to those not having homework. Or, if you take two classes, the one using homework will be more effective 21 out of a 100 times. In all examples in this book, the CLE is provided to assist in interpreting the effect size.

We do need to be careful about ascribing adjectives such as small, medium, and large to these effect sizes. Cohen (1988), for example, suggested that $d = 0.2$ was small, $d = 0.5$ medium, and $d = 0.8$ large, whereas the results in this book could suggest $d = 0.2$ for small, $d = 0.4$ for medium, and $d = 0.6$ for large when judging educational outcomes. In many cases this would probably be reasonable, but there are situations where this would be just too simple. Consider, for example, the effects of an influence such as behavioral objectives, which has an overall small effect of $d = 0.20$ (see Chapter 9), and reciprocal teaching, which has an overall large effect of $d = 0.74$. It may be that the cost of implementing behavioral objectives is so small that it is worth using them to gain an influence on achievement, albeit small, whereas it might be too expensive to implement reciprocal teaching to gain the larger effect. Instead of considering only the size of an effect, we should be looking for patterns in the various effect sizes and the causal implications across effect sizes, and making policy decisions on an overall investigation of the differences in effect sizes.

Further, there are many examples that show small effects may be important. A vivid example comes from medicine. Rosenthal and DiMatteo (2001) demonstrated that the effect size of taking low dose aspirin in preventing a heart attack was $d = 0.07$, indicating that less than one-eighth of one percent of the variance in heart attacks was accounted for by using aspirin. Although the effect size is small, this translates into the conclusion that 34 out of every 1,000 people would be saved from a heart attack if they used low dose aspirin on a regular basis. This sounds worth it to me.

Meyer *et al.* (2001) list other seemingly small effect sizes with important consequences: the impact of chemotherapy on breast cancer survival ($d = 0.12$), the association between a major league baseball player's batting average and success in obtaining a hit in a particular instance at bat ($r = 0.06$), the value of antihistamines for reducing sneezes and a runny nose ($d = 0.22$), and the link between prominent movie critics' reviews and box office success ($d = 0.34$).

Even more interestingly, it can be possible to identify various moderators that may enhance or detract from the overall average effect. For example, to use the homework case discussed above, it may be that males have greater improvements (i.e., have a higher effect

size) than females, younger students' achievement gains may be different from older ones', the effects may be greater in mathematics than reading. And indeed, the effects do decrease with age: primary students gain least from homework ($d = 0.15$) and secondary students have greater gains ($d = 0.64$, see Chapter 10).

Also, the nature of the achievement outcome may turn out to be critical. That is, when one is seeking influences on a very specific, narrow outcome (e.g., improvement in addition, understanding of phonics), then it may be likely that the effect size will be greater than when one is seeking influences on a more generalizable, wider concept (e.g., numeracy or reading achievement). While the synthesis of research on the effects of narrow or wide influences (Hattie, 1992) did not find such differences, it is still important to be aware of the potential of this moderator.

Problems with meta-analysis

Glass (2000) celebrated the 25th anniversary of the invention of the term "meta-analysis" (see also Hunt, 1997) by noting the growth of interest in meta-analysis shifting from an original "preoccupation of a very small group of statisticians" to a current "minor academic industry" (Glass, 2000, p. 1). About 25 percent of all articles in *Psychological Bulletin* have the term "meta-analysis" in the title, and he particularly noted the adoption of the method in medicine. Not surprisingly, given this growth, there remain many criticisms of meta-analysis. A common criticism is that it combines "apples and oranges" and such combining of many seemingly disparate studies is fraught with difficulties. It is the case, however, that in the study of fruit nothing else is sensible. The converse argument is absurd: no two things can be compared unless they are the same! Glass argued that "The question of 'sameness' is not an *a priori* question at all; apart from being a logical impossibility, it is an empirical question" (2000, p. 2). No two studies are the same and the only question of interest is how they vary across the factors we conceive as important.

Another criticism, which Cronbach (1982) referred to as the "flat earth society", is that meta-analysis seeks the *big facts* and often does not explain the complexity nor appropriately seek the moderators. However, meta-analysis indeed can seek moderators, and, as will be seen throughout this book, classrooms are places where complexities abound and all participants constantly try to interpret, engage or disengage, and make meaning out of this variegated landscape. While there are many common themes, sometimes "averages do not do it justice" (Glass, 2000, p. 9). However, the issue (which will be discussed throughout this book) is that the generalizability of the overall effect is an empirical issue, and, as will be seen, there are far fewer moderators than are commonly thought.

A further criticism is that the findings from meta-analysis are based on historical claims—that is, they are based on "past" studies, and the future is not so bound by what worked yesterday. It is critical to always appreciate that the meta-analyses in this book are indeed historical—that is what a research review is: a synthesis of published studies. The degree to which these past studies influence today's or tomorrow's schools is an interpretative issue for the reader.

Eysenck (1984) has been particularly critical of the use of low quality studies in any synthesis, promoting the cliché "garbage in—garbage out". In meta-analysis, it is possible to address this question by ascertaining if the effects are affected by quality, and in general they are not. For example, Lipsey and Wilson (1993) summarized 302 meta-analyses in psychology and education, and used a number of outcomes (besides achievement) in

their analyses (the overall effect was $d = 0.50$, $SD = 0.29$). They found no differences between studies that only included random versus non-random design studies ($d = 0.46$ vs. $d = 0.41$), or between high ($d = 0.40$) and low ($d = 0.37$) quality studies. There was a bias upwards from the published ($d = 0.53$) compared to non-published studies ($d = 0.39$), although sample size was unrelated to effect size ($d = -0.03$). Sipe and Curlette (1996) found no relationship between the overall effect size of 97 meta-analyses ($d = 0.34$) and sample size, number of variables coded, type of research design, and a slight increase for published ($d = 0.46$) versus unpublished ($d = 0.36$) meta-analyses. There is one exception that can be predicted from the principles of statistical power: if the effect sizes are close to zero, then the probability of having high confidence in this effect is probably related to the sample size (see Cohen, 1988; 1990).

There is every reason to check the effects of quality, but no reason to throw out studies automatically because of lower quality. An excellent example is the recent synthesis by Torgerson *et al.* (2004), who identified 29 studies from a total of 4,555 potentially relevant papers reporting evaluations of interventions in adult literacy and/or numeracy that were published between 1980 and 2002. Their criterion of acceptance was that only "quality" studies—that is, those studies that used randomized controlled trials—were selected. To decide that it is worthwhile to include only certain types of designs or only studies meeting some criteria of quality presupposes that the studies using only the specified designs or levels of quality are the best representatives of the population estimates. This is speculation, and by using meta-analysis these concerns are subject to verification.

When the studies from Torgerson *et al.* (2004) are examined, it is clear that many of their randomized control studies were of low quality. The median sample size was only 52, and given there were at least two groups (experimental and control) the "typical" study had only 26 people in each group. The average attrition rate was 66 percent, so two-thirds of each sample did not complete the study. It would have been more defensible to include all possible studies, code them for the nature of the experimental design *and* for the quality of the study, and then use meta-analysis techniques to address whether the effects differed as a consequence of design and quality. The aim should be to summarize all possible studies regardless of their design—and then ascertain if quality is a moderator to the final conclusions (see Benseman, Sutton, & Lander, 2005 for a full analysis).

As noted in Chapter 1, Scriven (2005) has argued that a more critical criterion for all scientific conclusions is "beyond reasonable doubt (BRD)", and in some cases randomized studies do not come close to being beyond reasonable doubt. "It seems more appropriate to think of 'gold standard' designs in causal research as those that meet the BRD standard, rather than those that have certain design features … The existence of more threats to internal or external validity in quasi-experimental designs does not entail a reduction of validity for well-done studies below BRD levels" (pp. 45–46). Scriven noted that one of the advocates of random controlled designs, Cook (2004), claimed that "Interpreting [randomized control trial] results depends on many other things—an unbiased assignment process, adequate statistical power, a consent process that does not distort the populations to which results can be generalized, and the absence of treatment-correlated attrition, resentful demoralization, treatment seepage and other unintended products of comparing treatments. Dealing with these matters requires observation, analysis and argumentation." As this last sentence notes, there may be many other research designs that can address

critical education questions. Design method and quality of studies are mediators, not prior conditions for choosing studies in a synthesis of studies.

A more statistical concern is that there can be quite a difference depending on whether the author of the meta-analysis used a random or a fixed model to calculate the effect sizes. The fixed effects model can be viewed as a special case of the random model where the variance of the universe effect size is zero; the random model allows generalization to the entire research domain whereas the fixed model allows an estimate of *one* universe effect size underlying all studies available (Kisamore & Brannick, 2008; Schulze, 2004). Typically, but not necessarily, the mean effect size from estimates based on the random model can be appreciably higher than when a fixed model is used. Hence, combining or comparing effects generated from the two models may differ solely because different models are used and not as a function of the topic of interest. Given that the majority of meta-analyses so far published have used the fixed effect model, then this fixed model has been used in this book. Where effects have been based on the random model and this seems to make a difference to the means, this is noted.

Previous attempts at synthesizing meta-analyses

There have been previous attempts to synthesize across meta-analyses. For example, I have published a study based on 134 meta-analyses of studies of educational innovations (Hattie, 1987; 1992). This research concluded that educational innovations can be expected to change average achievement outcomes by 0.4 standard deviations and affective outcomes by 0.2 standard deviations. Some overall findings were drawn about the factors above and below this average benchmark. Innovation, for example, was a theme underlying most of these positive effects. That is, a constant and deliberate attempt to improve the quality of learning on behalf of the system, the principal, and the teacher, typically related to improved achievement. The implementation of innovations probably captures the enthusiasm of the teacher implementing the innovation and the excitement of the students attempting something innovative. Often this has been explained as an experimental artifact in terms of a Hawthorne effect. However, another reason is that when teachers introduce innovation there can be a heightened attention to what is making a difference and what is not, and it is this attention to what is not working that can make the difference—feedback to the teacher about the effects of their actions!

I realized that the most powerful single influence enhancing achievement is feedback. This led me on a long journey to better understand this notion of feedback. After researching and reviewing feedback from a student's perspective (e.g., help-seeking behaviors) and from a teacher to student perspective (e.g., better comments on tests, increasing the amount of feedback in a class), it dawned on me that the most important feature was the creation of situations in classrooms for the teachers to receive more feedback about their teaching—and then the ripple effect back to the student was high (Hattie & Timperley, 2007). Indeed, my team and I have devised a computer-based classroom assessment tool primarily focused on enhancing such feedback (see www.asTTle.org.nz)—but that is another story.

When investigating the continuum of achievement, there is remarkable generality—remarkable because of the preponderance of educational researchers and teachers who argue for treating students individually, and for dealing with curriculum areas as if there were unique teaching methods associated with English, mathematics, and so on. The findings

from this synthesis apply, reasonably systematically, to all age groups, all curriculum areas, and to most teachers. It did not seem to matter whether the achievement outcomes were broad or narrow. The average effects of broad constructs and narrow outcomes were slightly lower ($d = 0.23$) compared with those categorized into broad constructs and broad outcomes ($d = 0.43$), narrow constructs and narrow outcomes ($d = 0.37$), and narrow constructs and broad outcomes ($d = 0.35$). Generality is the norm, but, as with many things, there are exceptions.

The majority of the findings of the meta-analyses were derived from studies conducted in English-speaking, highly developed countries (particularly, but not exclusively, the United States). We should not generalize the findings of these meta-analyses to non-English speaking, or non-highly developed countries. Note, for example, the results of a study by Heyneman and Loxley (1983), drawing on 52,252 elementary school-age pupils, 12,085 teachers, and 2,710 classes from 29 developing countries. They concluded that, relative to high-income countries, academic achievement in low-income countries was affected more by pupils' social status and less by teacher quality.

Kulik and Kulik (1989) reviewed more than 100 meta-analyses, including those relating to instructional methods and design. They concluded that "most of the well-known systems devised for improving instruction have acceptable records in evaluation studies" (p. 289). This was an appropriately cautious claim at that early stage of synthesizing across meta-analyses. They concluded that there were promising effects from curricular innovations (especially in science), and suggested that it was important to be cautious about effects from teacher education programs (which were lower than anticipated). They claimed that large effects were not the norm, although there were few negative effects. The major message for teachers was that there were many advantages with providing clear definitions of learning tasks for students, having a requirement of mastery on class activities and quizzes, and providing increased feedback, but policies relating to reorganizing classrooms did not get much support. These messages of learning intentions, success criteria, direct teaching, and the power of feedback—rather than being concerned with structural adaptations—are still powerful two decades later.

Walberg used my earlier synthesis (Hattie, 1987) to defend his nine-factor "Education Productivity" model, which he argued incorporated the three major psychological causes of learning (Reynolds & Walberg, 1998). The first was student aptitude (prior achievement, $d = 0.92$; age or maturation, $d = 0.51$; motivation, self-concept, willingness to persevere on learning tasks $d = 0.18$). The second was instruction (time in learning, $d = 0.47$; quality of teaching, $d = 0.18$). The third was psychological environments (morale or student perceptions of classroom social group, $d = 0.47$; home environment, $d = 0.36$; peer group outside school, $d = 0.20$; minimal leisure-time mass media exposure, particularly television, $d = 0.20$). More recently he argued that "each of the first five factors—prior achievement, development, motivation, and the quantity and quality of instruction—seems necessary for learning in school. Without at least a small amount of each, the student may learn little … (each) appears necessary but insufficient by itself for effective learning" (Walberg, 2006, pp. 103–106). Quality is an important enhancement of study time, and the four psychological environments expand and enhance learning time.

Marzano (1998) was critical of these attempts by me, Walberg, and others, claiming that basing a synthesis on "brand names" could be misleading. For example, he argued that the categories we used in our syntheses were too broad and included too many varied treatments, and that instead the categories used should be specific and functional

enough to provide guidance for classroom practice. He used four basic building blocks in his synthesis: knowledge ($d = 0.60$), the cognitive system ($d = 0.75$), the meta-cognitive system ($d = 0.55$), and the self-system ($d = 0.74$). Marzano used 4,057 effect sizes and found an overall effect size of $d = 0.65$. (This overall effect is somewhat larger than reported later in this book, as Marzano did not include many of the school and structural influences.) He reported on eight moderators:

1 whether the technique was designed for use by the teacher ($d = 0.61$) or the student ($d = 0.73$);
2 the degree of specificity of the influence (he argued that the more specific the influence, the higher the effect, although the means were $d = 0.67$, $d = 0.64$, $d = 0.64$ for least to most specific);
3 grade level of students (no differences);
4 student ability (low $d = 0.64$; middle $d = 0.70$; and high $d = 0.91$);
5 duration of treatment (shortened programs < 3 weeks $d = 0.69$ vs. > 4 weeks $d = 0.52$);
6 the specificity of dependent measures in the treatment (very specific $d = 0.97$, appropriate $d = 0.91$, and very general $d = 0.55$);
7 methodological quality (no difference);
8 publication type (published $d = 0.72$ vs. unpublished $d = 0.64$).

From his very systematic review he concluded that the "best way to teach organizing ideas—concepts, generalizations, and principles—appears to be to present those constructs in a rather direct fashion" (Marzano, 1998, p. 106) and then have students apply these concepts to new situations. He regarded the meta-cognitive system as the "engine" or primary vehicle for enhancement of the mental processes within the cognitive system and recommended providing students with clear targets of knowledge and skills, and strategies for the processes involved with what they are learning. Marzano, Gaddy, and Dean (2000) outlined an excellent and extremely fascinating set of implications for teachers and the learning processes deriving from these analyses. In a further re-analysis of these effect sizes, Marzano (2000) argued that 80 percent of the variance in achievement could be accounted for by student effects, 7 percent by school effects, and 13 percent by teacher effects. He then used these estimates to evaluate the effects on student achievement of an ineffective, an average, and an exceptional teacher in an ineffective, an average, and an exceptional school respectively. Average schools and average teachers, although he said they did little harm, also did little to influence students' relative position on the distribution of achievement for all students; ineffective teachers, no matter how effective the school, had a negative impact on the standings of all students, whereas students of exceptional teachers, even in ineffective schools, either maintained or increased achievement, many quite substantially. "Exceptional performance on the part of teachers not only compensates for average performance at the school level, but even ineffective performance at the school level" (Marzano, 2000, p. 81).

Synthesizing the meta-analyses

This book is not another meta-analysis. There are hundreds of those. Instead, this book aims to synthesize over 800 meta-analyses about the influences on achievement to present a more global perspective on what are and what are not key influences on achievement.

The project started by collecting 134 meta-analyses and proposing a set of common themes as to why some influences were more or less influential than others (Hattie, 1992). Since 1992, this collection of meta-analyses has been supplemented with a large number of other meta-analyses; over the past few years these have all been coded—at the study level—and the current database has a line for each meta-analysis that summarizes and categorizes the study and notes the effect sizes and standard errors which are needed for the calculations reported in this book.

It was possible to locate a total of about 800 meta-analyses, which encompassed 52,637 studies, and provided 146,142 effect sizes about the influence of some program, policy, or innovation on academic achievement in school (early childhood, elementary, high, and tertiary). Topics *not* included are those concerning English as a second language, affective or physical outcomes, and meta-analyses where the number of studies was fewer than four. When the same meta-analysis has been published multiple times, (e.g., when dissertations are rewritten as articles), only the most recent or most accessible is included.

As can be imagined, these effects cover most school subjects (although the majority are reading, mathematics, science, and social studies), all ages, and a myriad of comparisons. These effects are based on many millions of students across the main areas of influence—from the student, the home, the effects of schools, teachers, curricula, and teaching methods and strategies. The total number of students identified in the meta-analyses is large. Only 286 of the meta-analyses included total sample size but together these alone totaled 83 million students. Using the average sample size per study, this would multiply out to about 236 million students in total. However, it is likely that many students would have participated in more than one study, and thus this is a gross estimate of sample size. Even so, it would be safe to conclude that these studies are based on many millions of students.

Appendix A lists all the meta-analyses included in this book, provides the number of studies, people, and effect sizes, along with the average effect size, standard error (if provided), and common language effect. The meta-analyses are listed by the chapters they are referred to in this book. Appendix B lists these influences in their rank order.

The distribution of effect sizes

To start, let us see an overall distribution of all the effect sizes (Figure 2.2) from each of the 800+ meta-analyses. The bars that indicate points on the y-axis represent the number of effects in each category, while the x-axis gives the categories of effect sizes.

There are six immediate implications from Figure 2.2 that are critical to the arguments in this book:

1 The effects follow a normal distribution. To those immersed in large-scale statistics, this would not be surprising: normality is often, but not necessarily, present when there are large sample sizes. The normal distribution, however, is a consequence of the data and not imposed on it. Given this normal distribution, there are as many influences above the mean effect size as there are below it, and, most importantly, the mean is a reasonably good indicator of all the influences on achievement.
2 Almost everything works. Ninety percent of all effect sizes in education are positive. Of the ten percent that are negative, about half are "expected" (e.g., effects of disruptive students); thus about 95 percent of all things we do have a positive influence on

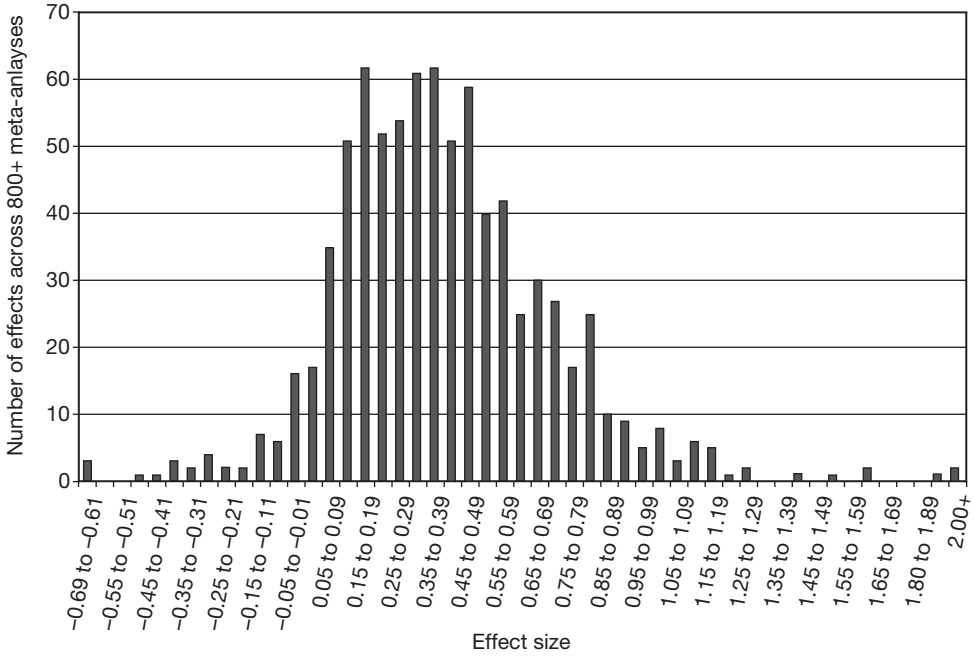


Figure 2.2 Distribution of effect sizes across all meta-analyses

achievement. When teachers claim that they are having a positive effect on achievement or when a policy improves achievement this is almost a trivial claim: virtually everything works. One only needs a pulse and we can improve achievement.

3 Setting the bar at zero is absurd. If we set the bar at zero and then ask that teachers and schools "improve achievement", we have set a very very low bar indeed. No wonder every teacher can claim that they are making a difference; no wonder we can find many answers as to how to enhance achievement; no wonder every child improves. As noted at the outset of this book, it is easy to find programs that make a difference. Raising achievement that is enhancing learning beyond an effect size of $d = 0.0$ is so low a bar as to be dangerous and is most certainly misleading.

4 Set the bar at $d = 0.40$. The average effect size is $d = 0.40$. This average summarizes the typical effect of all possible influences in education and should be used as the benchmark to judge effects in education. Effects lower than $d = 0.40$ can be regarded as in need of more consideration, although (as discussed earlier) it is not as simple as saying that all effects below $d = 0.40$ are not worth having (it depends on costs, interaction effects, and so on). Certainly effects above $d = 0.40$ are worth having and a major focus of this book is trying to understand the common denominators of what makes a difference (i.e., the effect sizes above compared to those below $d = 0.40$). Throughout this book this $d = 0.40$ effect size is referred to as the hinge-point or h-point, as this is the point on the continuum that provides the hinge or fulcrum around which all other effects are interpreted.

5 Innovations are more than teaching: Teachers average an effect of $d = 0.20$ to $d = 0.40$ per year on student achievement. This h-point of $d = 0.40$ does *not* mean that this

is the typical effect of teaching or teachers. It does not mean that merely placing a teacher in front of a class would lead to an improvement of 0.40 standard deviations. In most studies summarized in this book, there is a deliberate attempt to change, improve, plan, modify, and innovate. The best available estimate as to the effects of schooling is based on longitudinal studies. For example, the National Assessment of Educational Progress (NAEP, Johnson and Zwick, 1990) surveyed what students in American schools know and can do in the subject areas of reading, writing, civics, United States history, mathematics, and science. The students were sampled at ages 9, 13, and 17, and the testing has been repeated every two years. The average effect size across the six subject areas was $d = 0.24$ per year. In our own New Zealand studies, we have estimated that the yearly effect in reading, mathematics, and writing from Years 4 to 13 ($N = 83{,}751$) is $d = 0.35$—although this is not linear: in some years and for some subjects there is more or less growth. The inference for the argument in this book is that teachers typically can attain between $d = 0.20$ to $d = 0.40$ growth per year—and that this is to be considered *average*. They should be seeking greater than $d = 0.40$ for their achievement gains to be considered above average, and greater than $d = 0.60$ to be considered excellent.

6 The variance is important. This typical effect size of $d = 0.40$ may not be uniform across all students or all implementations of any influence. There may be many moderators. For example, the typical effect size of homework is $d = 0.29$, but the effects are greater for high school students and closer to zero for elementary school students. The major point of this "achievement barometer" or "achievement continuum" is to provide a basis to interpret the effects of change, both the overall effects and effects broken down by important moderators.

The typical effect: the hinge-point

The effect size of 0.40 sets a level where the effects of innovation enhance achievement in such a way that we can notice real-world differences, and this should be a benchmark of such real-world change. It is not a magic number that should become like a $p < 0.05$ cut-off point, but a guideline to begin discussions about what we can aim for if we want to see students change. It provides a "standard" from which to judge effects: it is a comparison based on typical, real-world effects rather than based on the strongest cause possible, or with the weakest cause imaginable. It is not unreasonable to claim that at least half of all implementations, at least half of all students, and at least half of all teachers *can and do* attain this h-point of $d = 0.40$ change as a consequence of their actions.

An aim of this book is to position the various influences along this continuum, relative to the typical $d = 0.40$ effect. The fundamental claim is that influences in education are relative: we should judge the success of an innovation relative to $d = 0.40$ (and certainly not $d = 0.0$). To return to the homework example used earlier, the typical influence after introducing homework was just below the typical effect across all possible influences. Thus, when the influence of homework is compared to the more usual zero point, those who argue that homework is effective would say "yes", but when the effects from classes without homework are compared to the typical effect across all other influences, then homework is well below an average effect—there are many more innovations that have greater effects. Maybe it is not so surprising that teachers have found that the effect of prescribing homework is not as dramatic as many advocates

and researchers promised. The advocates and researchers compare the outcome to zero, but we should be comparing the effect to alternative innovations. The null hypothesis ($d = 0.0$) is not the question of interest, so it is no wonder that the answer is misleading; introducing nearly any innovation is better than its absence. The null hypothesis is virtually certain to be false before analysis commences and thus it is uninformative (see Novick & Jackson, 1974).

The main contributors

Table 2.1 presents the average effect for each of the major categories of contributors to learning. The averages of all effects are quite similar with the exception that school differences are far less critical to enhancing achievement: take two students of the same ability and it matters less to which school they go than the influences of the teacher, curricula program, or teaching they experience.

Figure 2.3 presents the number of meta-analyses from these 800+ meta-analyses relative to average $d = 0.40$ h-point. There are just as many home, student, curricula, and teaching effects above as below the average, more teaching effects above 0.40, and many more school effects below $d = 0.40$. But averages can hide too much. The remainder of this book works through each of these major categories of influences, chapter by chapter, and aims to more deeply evaluate the underlying causes of what specific innovations and influences are above and below average. Each chapter will work through a number of innovations and influences; sufficient detail is given for each of these innovations to give a sense of the claims, but the primary aim is to draw inferences for the overall model outlined in Chapter 3.

The barometer of influences

We seem to have no barometers of success or failure to show what works and what does not work in education. Yes, we do have tests, lots of them, which we use to evaluate whether students have gained sufficiently. But this is not enough. An influence may "work", but by how much, and how differently from other influences? Some innovations or actions are more influential than others. Instead of asking "What works?" we should be asking "What works best?" as the answers to these two questions are quite different. As has been indicated already, the answer to the first questions is "Almost everything" whereas the answer to the second is more circumscribed—and some things work *better* and some work *worse* relative to the many possible alternatives.

Table 2.1 Average effect for each of the major contributors to learning

Contribution	*No.*	*Studies*	*People*	*Effects*	*d*	*SE*	*CLE*
Student	139	11,101	7,513,406	38,282	0.40	0.044	29%
Home	36	2,211	11,672,658	5,182	0.31	0.058	22%
School	101	4,150	4,416,898	13,348	0.23	0.072	16%
Teacher	31	2,225	402,325	5,559	0.49	0.049	35%
Curricula	144	7,102	6,899,428	29,220	0.45	0.076	32%
Teaching	365	25,860	52,128,719	55,143	0.42	0.071	30%
Average	816	52,649	83,033,433	146,626	0.40	0.062	28%

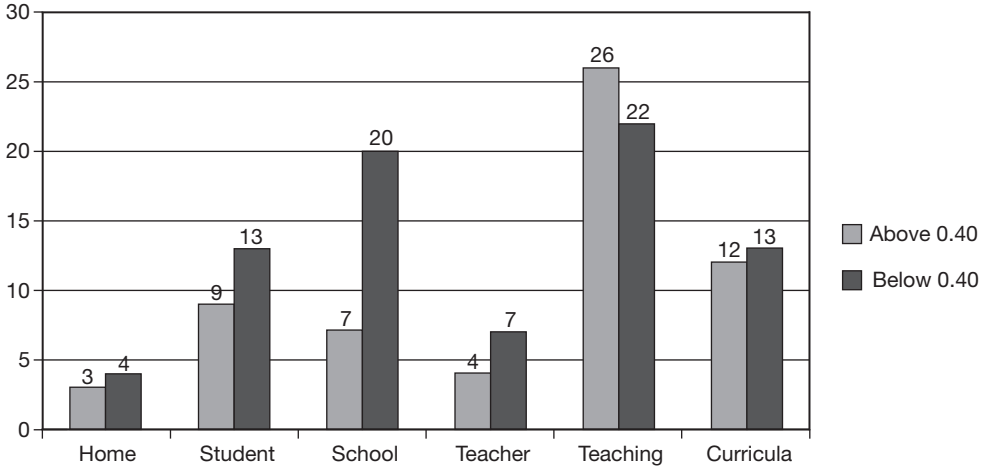
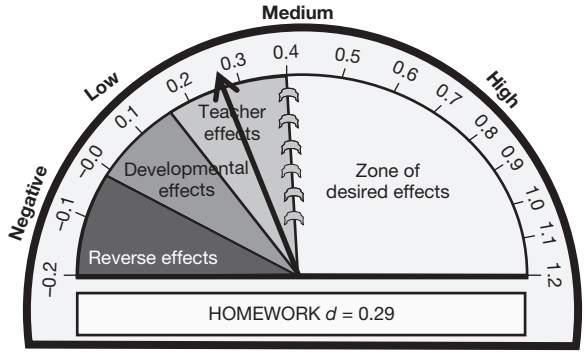


Figure 2.3 Number of meta-analyses above and below the h-point

We need a barometer that addresses whether the various teaching methods, school reforms, and so on are worthwhile relative to possible alternatives. We need clear goalposts of excellence for all in our schools to aspire towards, and most importantly, for them to know when they get there. We need a barometer of success that helps teachers to understand which attributes of schooling assist students in attaining these goalposts.

Figure 2.4 outlines one such barometer that has been developed for use throughout this book. The development of this barometer began not by asking whether this or that innovation was working, but whether this teaching worked better than possible alternatives; not by asking whether this innovation was having positive effects compared to not having the innovation, but whether the effects from this innovation were better for students than what they would achieve if they had received alternative innovations.

For each of the many attributes investigated in the chapters in this book, the average of each influence is indexed by an arrow through one of the zones on the barometer. All influences above the h-point ($d = 0.40$) are labeled in the "Zone of desired effects" as these are the influences that have the greatest impact on student achievement outcomes.



KEY	
Standard error	0.027 (Low)
Rank	88th
Number of meta-analyses	5
Number of studies	161
Number of effects	295
Number of people (4)	105,282

Figure 2.4 A typical barometer of influence

The typical effects from teachers are between $d = 0.15$ and $d = 0.40$, as identified from the longitudinal studies discussed above. Any influences in this zone are similar to what teachers can accomplish in a typical year of schooling. The zone between $d = 0.0$ and $d = 0.15$ is what students could probably achieve if there was no schooling (and is estimated from the findings in countries with no or limited schooling). Maturation alone can account for much of the enhancement of learning (see Cahan & Davis, 1987). Thus, any effects below $d = 0.15$ can be considered potentially harmful and probably should not be implemented. The final category includes the reverse effects—those that decrease achievement—and these are certainly not wanted.

The arrow points to the average effect of the various meta-analyses on the particular topic (in the above it is $d = 0.29$ for the five homework meta-analyses. The variability (or standard error) of the average effect sizes from each meta-analysis is not always easy to determine. Often the information is not provided, and it is well known that the variance is very much related (inversely) to the sample size of studies—the more studies there are, the greater the variance. Across all 800+ meta-analyses the typical standard error of the mean is about $d = 0.07$—and to provide a broad sense of variance, any influence where the average "spread of effects" is less than $d = 0.04$ is deemed low, between $d = 0.041$ and $d = 0.079$ is deemed medium and greater than $d = 0.08$ is deemed large. While these are crude estimates, it is more important to read the discussion about each influence to ascertain whether important sources of variance could be identified to explain differential effects within that influence. In many cases there is insufficient information to estimate the standard error and thus it is not provided in the summary information. The information under the barometer allows an interpretation of how confident we can be about this summary information: the number of meta-analyses on each category (five in the above case), based on 161 studies, and 295 effect sizes. There were 105,282 students in the four meta-analyses that provided information about sample size. The average effect is $d = 0.29$, with a standard error of 0.027 (considered "low" relative to all meta-analyses). The effects of homework, in this example, rank 88th of all 138 meta-analyses (see Appendix B).

Relation between effect size and sample size

A funnel plot is often used to examine whether a meta-analysis is based on a biased sample of studies (Light & Pillemer, 1984). The funnel plot is a scatterplot of effect sizes versus the number of studies (in this case), with each data point representing one study. Because meta-analyses with a larger number of studies are more likely to better estimate the effect size, they tend to lie in a narrow band at the top of the scatterplot, while the smaller meta-analyses (with expected greater variation in results) fan out over a larger area at the bottom—thus creating the visual impression of an inverted funnel. As can be seen in Figure 2.5, the results from this synthesis show a reasonably symmetric funnel, indicating a lack of publication bias.

Concluding comments

This chapter sets the scene for the interpretation of the 800+ meta-analyses, which form the fundamental dataset used throughout this book. An achievement continuum has been developed along which the many effects can be located, and the importance of the h-point of $d = 0.40$ has been emphasized. The barometer of achievement can be used to assist in seeking the explanation of what leads to successful learning that exceeds the $d = 0.40$ hinge-point.

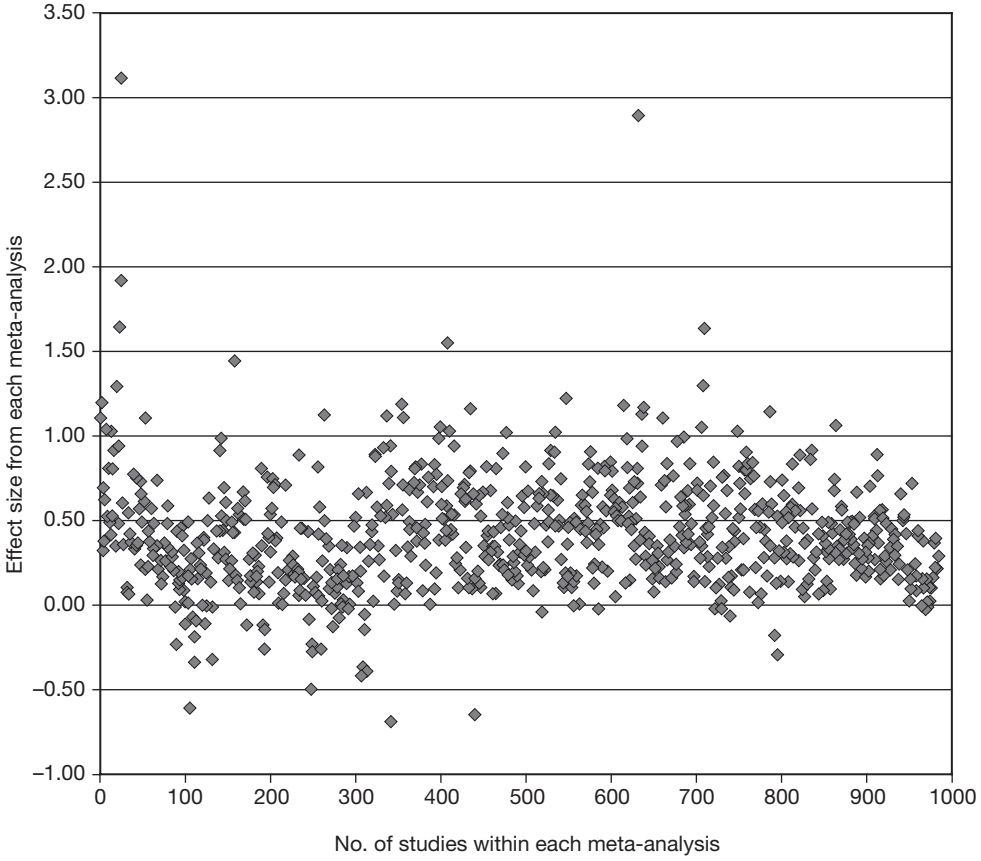


Figure 2.5 Funnel plot of the effect size and sample size from each meta-analysis

Chapter 3

The argument

Visible teaching and visible learning

> We think in generalities, but we live in details.
>
> (Whitehead, 1943, p. 26)

This chapter introduces the major findings that will be elaborated in the following chapters. The aim of this book is not to overwhelm with detail from the more than 50,000 studies and 800+ meta-analyses that form the basis of the discussion. Instead, the aim is to build an explanatory story about the influences on student learning and then to convince the reader of the nature and value of the story by working through the evidence to defend it. It is as much theory generation as it is theory appraisal. The art in any synthesis is the overall message, and the simple adage underlying most of the syntheses in this book is "visible teaching and learning". Visible teaching and learning occurs when learning is the explicit goal, when it is appropriately challenging, when the teacher and the student both (in their various ways) seek to ascertain whether and to what degree the challenging goal is attained, when there is deliberate practice aimed at attaining mastery of the goal, when there is feedback given and sought, and when there are active, passionate, and engaging people (teacher, student, peers, and so on) participating in the act of learning. It is teachers seeing learning through the eyes of students, and students seeing teaching as the key to their ongoing learning. The remarkable feature of the evidence is that the biggest effects on student learning occur when teachers become learners of their own teaching, and when students become their own teachers. When students become their own teachers they exhibit the self-regulatory attributes that seem most desirable for learners (self-monitoring, self-evaluation, self-assessment, self-teaching). Thus, it is visible teaching and learning by teachers and students that makes the difference. The following chapters provide the evidence to defend this overall message.

What teachers *do* matters

The major message is simple—what teachers *do* matters. However, this has become a cliché that masks the fact that the greatest source of variance in our system relates to teachers—they can vary in major ways. The codicil is that what "some" teachers do matters—especially those who teach in a most deliberate and visible manner. When these professionals see learning occurring or not occurring, they intervene in calculated and meaningful ways to alter the direction of learning to attain various shared, specific, and challenging goals. In particular, they provide students with multiple opportunities and alternatives

for developing learning strategies based on the surface *and* deep levels of learning some content or domain matter, leading to students building conceptual understanding of this learning which the students and teachers then use in future learning. Learners can be so different, making it difficult for a teacher to achieve such teaching acts—students can be in different learning places at various times using a multiplicity of unique learning strategies, meeting different and appropriately challenging goals. Learning is a very personal journey for the teacher and the student, although there are remarkable commonalities in this journey for both. It requires much skill for teachers to demonstrate to *all* their students that they can see the students' "perspective, communicate it back to them so that they have valuable feedback to self-assess, feel safe, and learn to understand others and the content with the same interest and concern" (Cornelius-White, 2007, p. 23).

The act of teaching requires deliberate interventions to ensure that there is cognitive change in the student: thus the key ingredients are awareness of the learning intentions, knowing when a student is successful in attaining those intentions, having sufficient understanding of the student's understanding as he or she comes to the task, and knowing enough about the content to provide meaningful and challenging experiences in some sort of progressive development. It involves an experienced teacher who knows a range of learning strategies to provide the student when they seem *not* to understand, to provide direction and re-direction in terms of the content being understood and thus maximize the power of feedback, and having the skill to "get out the way" when learning is progressing towards the success criteria.

Of course, it helps if these learning intentions and success criteria are shared with, committed to, and understood by the learner—because in the right caring and idea-rich environment, the learner can then experiment (be right and wrong) with the content and the thinking about the content, and make connections across ideas. A safe environment for the learner (and for the teacher) is an environment where error is welcomed and fostered—because we learn so much from errors and from the feedback that then accrues from going in the wrong direction or not going sufficiently fluently in the right direction. In the same way, teachers themselves need to be in a safe environment to learn about the success or otherwise of their teaching from others.

To facilitate such an environment, to command a range of learning strategies, and to be cognitively aware of the pedagogical means to enable the student to learn requires dedicated, passionate people. Such teachers need to be aware of which of their teaching strategies are working or not, be prepared to understand and adapt to the learner(s) and their situations, contexts, and prior learning, and need to share the experience of learning in this manner in an open, forthright, and enjoyable way with their students and their colleagues.

We rarely talk about passion in education, as if doing so makes the work of teachers seem less serious, more emotional than cognitive, somewhat biased or of lesser import. When we do consider passion, we typically constrain such expressions of joy and involvement to matters unrelated to our teaching (Neumann, 2006). The key components of passion for the teacher and for the learner appear to be the sheer thrill of being a learner or teacher, the absorption that accompanies the process of teaching and learning, the sensations in being involved in the activity of teaching and learning, and the willingness to be involved in deliberate practice to attain understanding. Passion reflects the thrills as well as the frustrations of learning—it can be infectious, it can be taught, it can be modeled, and it can be learnt. It is among the most prized outcomes of schooling and, while rarely studied

in any of the studies reviewed in this book, it infuses many of the influences that make the difference to the outcomes. It requires more than content knowledge, acts of skilled teaching, or engaged students to make the difference (although these help). It requires a love of the content, an ethical caring stance to wish to imbue others with a liking or even love of the discipline being taught, and a demonstration that the teacher is not only teaching but learning—typically about the students' processes and outcomes of learning.

Learning is not always pleasurable and easy; it requires over-learning at certain points, spiraling up and down the knowledge continuum, and building a working relationship with others in grappling with challenging tasks. This is the power of deliberative practice. It also requires a commitment to seeking further challenges—and herein lies a major link between challenge and feedback, two of the essential ingredients of learning. The greater the challenge, the higher the probability that one seeks and needs feedback, but the more important it is that there is a teacher to provide feedback and to ensure that the learner is on the right path to successfully meet the challenges.

The key to many of the influences above the $d = 0.40$ h-point is that they are deliberate interventions aimed at enhancing teaching and learning. But the message is to not merely *innovate*—but for us to learn from what makes the difference when teachers innovate. When we innovate we are more aware of what is working and what is not, we are looking for contrary evidence, we are keen to discover any intended and unintended consequences, and we have a heightened awareness of the effects of the innovations on outcomes. In these situations teachers become the learners about their own effects! In any innovation there is deliberate attention to implementation and its effects, there is a degree of challenge, and a valuing of feedback. It is critical that teachers learn about the success or otherwise of their interventions: those teachers who are students of their own effects are the teachers who are the most influential in raising students' achievement. Seeking positive effects on student learning (say, $d > 0.40$) should be a constant theme and challenge for teachers. As this does not accrue by serendipity or accident, then the excellent teacher must be vigilant to what is working and what is *not* working in the classroom.

A concept of excellent teaching

A story serves to illustrate my claims about excellent teaching. Some time ago one of my Master's students completed a meta-analysis on the effects of various programs on self-concept of children and adults (Clinton, 1987). The most successful programs were the Outward Bound or Adventure programs. There are four major features of these programs that led to their positive influence. First, Outward Bound programs have an emphasis on the immediate quality of the experience, as well as aiming to have these immediate experiences have an effect on later experiences. That is, there is a planned and intentional transfer of experiences, knowledge, and decisions during the earlier learning experiences to later experiences (see Hattie, Marsh, Neill, & Richards, 1997 for more details). Second, Outward Bound programs set difficult and specific goals that are known to the learner, and then tasks are structured so that participants can attain these goals. The program provides challenging and specific goals (e.g., successfully negotiating a 60-foot cliff by abseiling or rappelling) and then structures situations (e.g., adequate preparation, social support, etc.) so that participants share a commitment to reaching these goals. Third, the program increases the amount and quality of feedback, which is vital to the learning process. The often dangerous and risky situations demand feedback, and the learning intentions and

success criteria are crystal clear. A major function of challenging and specific goals is that they direct attention and effort, and thus the learner is more aware and keen for feedback related to attaining these goals. Fourth, in the Outward Bound program, the instructor is keenly aware of the need to understand and, if necessary, reassess and redirect an individual's coping strategies. Such coping strategies can be cognitive (learning strategies), personal (building of self-efficacy, perseverance in the face of challenge), and social (help seeking, cooperative learning). These four major features are the keys of successful teaching and learning.

As another example, take my involvement in a Bush Search and Rescue Squad which involved teaching the skills and fun of cliff rescues. Consider the following: I am going to take you to the top of a three-storey building and teach you to rappel down the outside of this building. Typically, I would then demonstrate to you how to put on a safety harness, tie the rope in a bowline, and then show you how to lean backwards to commence the descent. In line with the principles of good teaching, I would then ask you, the student, to implement this learning. Typically, such a learning situation leads to much care by the students, an enhanced level of interest in what peers are doing, and high levels of help-seeking behaviors to ensure the knowledge of rope-work is correct and harnesses are correctly positioned. The goals are challenging, specific, and visible, and the learners are committed to them! The learning is actively visible and there are high levels of feedback and monitoring. The learner typically "seeks" the feedback. When a novice first gets to the edge, there is a remarkably high level of peer teaching and learning: it is not natural to fall backwards when descending as it is more typical for the feet to precede the head. When finally the student reaches the bottom there is a surge of excitement appreciating that the challenging goal has been reached (it is abundantly clear what the success criteria are!), the experience was exhilarating, and the learning absorbed in the experience itself. Most want to repeat the experience and continue to enjoy meeting the challenging goals. Moreover, all these acts and most of the "thinking" about the task are visible to the teacher and to the learner. This is the heart of the model of successful teaching and learning advocated in this book.

Visible teaching

It is critical that the teaching and the learning are visible. There is no deep secret called "teaching and learning": teaching and learning are visible in the classrooms of the successful teachers and students, teaching and learning are visible in the passion displayed by the teacher and learner when successful learning and teaching occurs, and teaching and learning requires much skill and knowledge by both teacher and student. The teacher must know when learning is correct or incorrect; learn when to experiment and learn from the experience; learn to monitor, seek and give feedback; and know to try alternative learning strategies when others do not work. What is most important is that teaching is visible to the student, and that the learning is visible to the teacher. The more the student becomes the teacher and the more the teacher becomes the learner, then the more successful are the outcomes.

This explanation of visible teaching relates to teachers as activators, as deliberate change agents, and as directors of learning. This does not mean that they are didactic, spend 80 percent or more of the day talking, and aim to get through the curriculum or lesson come what may. Effective teaching is not the drilling and trilling to the less than willing.

When I reviewed the videotapes of many of the best teachers in the United States (via the National Board for Professional Teaching Standards video assessment task) it was stunning how active and involved the best teachers were in the classrooms—it was clear who was in control in those classrooms. The activity was visible and "in the air"; passive was not a word in the vocabulary of these accomplished teachers—learning was not always loud and heated but it was rarely silent and deadening, and it was often intense, buzzing, and risky.

The model of visible teaching and learning combines, rather than contrasts, teacher-centered teaching and student-centered learning and knowing. Too often these methods are expressed as direct teaching versus constructivist teaching (and then direct teaching is portrayed as bad, while constructivist teaching is considered to be good). Constructivism too often is seen in terms of student-centered inquiry learning, problem-based learning, and task-based learning, and common jargon words include "authentic", "discovery", and "intrinsically motivated learning". The role of the constructivist teacher is claimed to be more of facilitation to provide opportunities for individual students to acquire knowledge and construct meaning through their own activities, and through discussion, reflection and the sharing of ideas with other learners with minimal corrective intervention (Cambourne, 2003; Daniels, 2001; Selley, 1999; von Glasersfeld, 1995). These kinds of statements are almost directly opposite to the successful recipe for teaching and learning as will be developed in the following chapters.

A model of learning

The major point here is that constructivism is *not* a theory of teaching, but a theory of knowing and knowledge, and it is important to understand the role of building constructions of understanding. Bereiter (2002) used Popper's three worlds to make sense of much of what we strive for in school: the physical world, the subjective or mental world, and the world of ideas. These three worlds have major parallels with the three worlds of achievement: surface knowledge of the physical world, the thinking strategies and deeper understanding of the subjective world, and the ways in which students construct knowledge and reality for themselves as a consequence of this surface and deep knowing and understanding. This third world, often forgotten in the passion for teaching facts and thinking skills, is entirely created by humans, is fallible but capable of being improved, and can take on a life of its own. Students often come to lessons with already constructed realities (third worlds), which, if we as teachers do not understand them before we start to teach, can become the stumbling blocks for future learning. If we are successful, then the students' constructed realities (based on their surface and deep knowing) and keenness to explore these worlds are the major legacy of teaching. The contents of this third world are not concrete like books, statues, and teapots (see Bereiter, 2002, pp. 62–63): they are more conceptual. It is certainly the case, as Bereiter documents, that "much of what is meant by the shift from an industrial to a knowledge society is that increasing amounts of work are being done on conceptual objects rather than on the physical objects to which they are related" (Bereiter, 2002, p. 65).

The distinctions are not clean cut, as at all three levels we often learn in a haphazard manner. So much teaching is aimed at the first world—the world of ideas and knowledge, and there is also much discussion about the importance of deep knowledge and thinking skills (the second world). But the task of teaching and learning best comes together when we attend to all three levels: ideas, thinking, and constructing.

> In many situations, it will be impossible to make a clear distinction between knowing or thinking about the conceptual artifact and knowing or thinking about the material world the conceptual artifact applies to […] What matters is that we recognize conceptual artifacts as real things, recognize creating and improving them as real work, and recognize understanding them as real understanding.
>
> (Bereiter, 2002, p. 67).

The real work of schooling is to create or add value to conceptual artifacts in the same way that builders add value to building artifacts. It is a world of conjectures, explanations, proofs, arguments, and evaluations.

Similarly, there can be many cultural artifacts particular to a culture and an important aspect of education is to teach these artifacts. For example, in New Zealand Māori culture, there is much importance attached to whānau (family), history, and cultural norms. A major focus of schooling is to therefore enable learners to adopt these cultural artifacts as a key part of their own conceptual artifacts—a way to see their world in a similar manner to how the culture has learnt to see its world, and communicate its worldviews and values. The importance is that this three-level view of achievement allows relations between theory and observation, personal and cultural belief and observation, and between personal belief and theory.

> … there are the relations between different theories, different phenomena, and different people's readings of the same phenomena. None of these relations are easy. They are all inferential and highly problematic. But they are what people work on when they are building scientific knowledge.
>
> (Bereiter, 2002, p. 91)

Bereiter claimed that there are a number of commitments to knowledge improvement: the third world is not limited to accepted, verified, or important knowledge objects. It can include discredited theories, crank notions, unsolved problems, and new ideas that may or may not gather a following. In this respect, "the third world is more inclusive than the canons of liberal education. This inclusiveness goes a long way toward eliminating the split between established knowledge and students' constructive efforts because it places the ideas created by students in the same world as the ideas handed from authoritative sources" (p. 237). "Knowing one's way around in the world of conceptual artifacts affords a wealth of possibilities not open to people who know that world only from a distance, if at all." (p. 238). Knowledge building is an activity directed toward the third world. It is doing something to a conceptual artifact. Bereiter claimed that knowledge building includes thinking of alternatives, thinking of criticisms, proposing experimental tests, deriving one object from another, proposing a problem, proposing a solution, and criticizing the solution. It is more than knowing, mistakenly believing, or doubting some knowledge object.

Educating is more than teaching people to think—it is also teaching people things that are worth learning. Good teaching involves constructing explanations, criticizing, drawing out inferences, finding applications, and there "should never be a need for the teacher to think of ways to inject more thinking into the curriculum. That would be like trying to inject more aerobic exercise into the lives of Sherpa porters" (Bereiter, 2002, p. 380). If the students are not doing enough thinking, something is seriously wrong with the instruction. "If the only justification for an activity is that it is supposed to encourage or improve

thinking, drop it and replace it with an activity that advances students' understanding that increases their mastery of a useful tool" (Bereiter, 2002, p. 381).

Surface, deep, and constructed understanding

There needs to be a major shift, therefore, from an over reliance on surface information (the first world) and a misplaced assumption that the goal of education is deep understanding or development of thinking skills (the second world), towards a balance of surface and deep learning leading to students more successfully constructing defensible theories of knowing and reality (the third world).

For many students, success at school relates to adopting a surface approach to understanding both how and what they should learn, whereas many teachers claim that the goal of their teaching is enhancing deep learning (Biggs & Collis, 1982). Brown (2002), for example, investigated the beliefs about learning of more than 700 15-year-old New Zealand students and 71 of their teachers of English, mathematics, and science. Students argued that learning for them primarily meant exhibiting surface knowledge involving the reproduction of taught material in order to maximize achievement in assessments. In contrast, teachers of these same students claimed that they were teaching towards deep learning outcomes. The students were more governed by the tasks and examinations set by teachers and schools, so, despite claims by teachers, the students were very strategic in concentrating on acquiring sufficient surface and whatever deeper understanding was needed to complete assignments and examinations. (The same phenomenon is especially evident when comparing conceptions of learning by academics and their students; Purdie, 2001.)

Students can be strategic in their approach because most questions and examinations (verbal and written) relate to surface knowledge (Marzano, 1991). For example, Gall (1970) claimed that 60 percent of teachers' questions required factual recall, 20 percent were procedural, and only 20 percent required thought by the students. Other studies have found the proportion of surface thinking questions can be in the order of 80 percent or more (Airasian, 1991; Barnette, Walsh, Orletsky, & Sattes, 1995; Gall, 1984; Kloss, 1988). Teachers' questioning may not elicit deep thinking from students because students understand that questioning is how teachers lead and control classroom activity; in other words, students know that the teacher already knows the answer to the questions (Gipps, 1994; Torrance & Pryor, 1998; Wade & Moje, 2000). So much of daily classroom life is "knowledge telling", and thus surface knowledge is sufficient. Students soon learn that studying or learning with surface strategies or methods (i.e., revision, re-reading, and reviewing of the year's work) leads to success. In contrast, teachers claim to prefer a deep view of learning, usually focused on academic and cognitive development, while at the same time they emphasize surface methods of teaching, usually with the defense that this is what is required in order to prepare students for high-stakes qualification examinations or assessments. This emphasis on surface approaches means that students tend to experience very few opportunities or demands for deep thinking in contemporary classrooms.

To be more specific, *surface* learning involves a knowing or understanding of ideas or facts. In contrast, the two *deep* processes—relational and elaborative—constitute a change in the quality of thinking that is cognitively more challenging than surface questions. Relational responses require integration of at least two separate pieces of given knowledge, information, facts, or ideas. In other words, relational questions require learners to impose