1

Mapping the Territory

Computer science (CS) education research is an emergent area and is still giving birth to a literature.

While scholarly and scientific publishing go back to *Philosophical Transactions* (first published by the Royal Society in England in 1665), one of our oldest dedicated journals *Computer Science Education* was established less than two decades ago, in 1988.

Growth which has led to the emergence of CS education research as an identifiable area has come from various places. Some from sub-specialist areas: the *Empirical Study of Programming* (ESP) and *Psychology of Programming Interest Group* (PPIG) series of workshops; some from the major practitioner conference which have included research papers—the Innovation and Technology in CS Education (ITiCSE) conference and the SIGCSE Symposium (now in its 36th year) run by the Special Interest Group in Computer Science Education of the ACM.

Following on from these starting points, there has been a burgeoning growth of publications appearing, perhaps opportunistically, in diverse locations, such as the IEEE *Frontiers in Education* (FiE) and the *American Society for Engineering Education* (ASEE) conferences, the ACM OOPSLA etc. Simultaneously within this same time there have emerged a number of CS education research groups within academic institutions.

Despite this growth—and because of it—we are struggling to find the shape and culture of our literature. The task is difficult not only because the literature is distributed (there is no CS education *research* conference or publication) but also because our researchers and writers come from many established fields of scholarship and research—at least from education, psychology, computer science, technology and engineering. We have different intellectual traditions, and different conceptual frames, not to mention methodological differences and different reporting and citation styles.

What gets published from these different disciplinary areas, in diverse venues, are papers of very different types. However, in a simple-minded way, whatever their tradition, they can be thought of as having two components: a dimension of rationale, argumentation or "theory", and a dimension of empirical evidence. If we think of these dimensions as plotting a space, then four quadrants can be defined. On the top left, we have papers that have lots of argument, but little empirical evidence (although they may draw on other sorts of evidential material, similar to other disciplinary areas such as history). The bottom left quadrant *should* be empty; this is the home of papers with no evidence and no argument. The bottom right quadrant represents papers which are constructed around evidence—most often empirical—but are not strong on argumentation: here is where descriptive, practice-based, "experience" papers are found, probably the most common type of paper in the area today. Finally, the top right quadrant where papers contain both evidence and argument is where, we contend, most CS education research papers should be found.

Partly, the concentration of papers in the lower right hand quadrant is representative of the state and status of CS education research today.

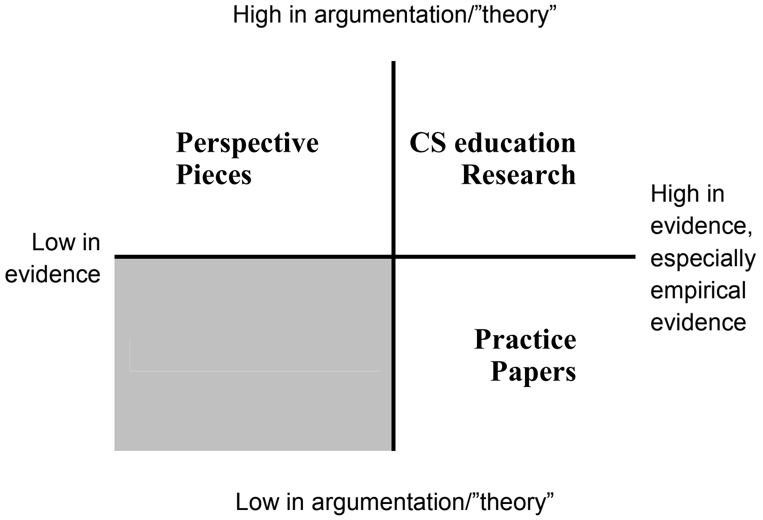

High in argumentation/"theory"
Perspective Pieces
CS education Research
Low in evidence
High in evidence, especially empirical evidence
Practice Papers
Low in argumentation/"theory"

Figure 1: Diarammatic representation of the types of paper found in CS education research (after *Pasteur's Quadrant* (Stokes, 1997))

If we espouse a scientific approach to investigation in this area (and we do) then this includes the general trend and growth of scientific knowledge over decades. The process is, in general, an alternation of inductive reasoning (the discovery of general propositions) with deductive reasoning (the application of general propositions—once discovered—to particular cases considered to be included within their scope). The small-scale, empirical studies each designed to purposefully observe a phenomenon build up to allow generalisations—"theories"—to be constructed. These can, in their turn, be applied in new situations. "These waves of alternate induction and deduction are superimposed, as it were, on a rising tide of which the general direction is inductive: a heaping up of a vast swell of generalisations which constitutes scientific knowledge as a whole" (Holmstrom, 1947). CS education research is not at a stage of development where many generalisations of this nature are possible: it is currently theory-scarce.

Motivating Areas for Computer Science Education Research

This section is principally a work of cartography. Its first aim is to map a territory of topics encountered in CS education research. Similar to all such attempts, there are some areas better known and better charted than others. Equally, this map will of necessity be historical, in that it must represent what people have found interesting enough to pursue to create a field. These topics and interests may not persist into the future, but taken together they should give a researcher new to the field a good idea of the current areas of investigation and interest.

We identify ten broad areas that motivate researchers in CS education. We use “motivate” intentionally here because we believe that these areas reflect key topics and highlight *why* researchers become interested in the first place. Each area has its own, sometimes small, tradition. Some we shall explore here, others are developed more fully in later chapters. Sometimes these areas, and their literatures, are encompassed within a single disciplinary genre; at other times they may have been examined from many approaches, and their only commonality is the interest inherent in the questions they pose or the nature of the phenomena they represent.

The ten areas are: *student understanding, animation/visualisation/simulation systems, teaching methods, assessment, educational technology, the transfer of professional practice into the classroom, the incorporation of new development and new technologies into the classroom, transferring to remote teaching (“e-learning”), recruitment and retention of students*, and, finally, *the construction of the discipline* itself.

Student Understanding

The area of student understanding is characterised by investigation of students’ mental and conceptual models, their perceptions and misconceptions. The kinds of question that researchers find motivating in this area are concerned with *why* students have trouble with some of the things they have trouble with, what *distinguishes* good students from bad students, and what the differences are between how students understand things and how experts understand things. This area of interest encompasses investigations at a wide variety of scale from very broad topics, such as “What design behaviors do students exhibit?” and “How do students learn in particular programming paradigms?” to very specific questions such as “How do students learn recursion?” Because of the nature of the evidence required in this area, given that it focuses on internal phenomena, some work has been conceived as CS education research, but there is also a lot of work of this kind that that occurs within psychology, or sociology. A good overview of some of this literature can be found in later chapters in Part Two, particularly *Research on Learning to Design Software* and [**Clancy** chapter]

Animation, Visualization and Simulation

There is a strand of research that draws on one of the peculiar facets of CS education research, which is that, as computer scientists, we can build things. We can devise and develop systems. We can do this to serve our research questions, to allow us to investigate and follow up conjectures. This has given rise to an area of research which uses software tools and environments to affect student learning. These tools broadly fall into the categories of animation (often algorithmic), visualization (often of processes within the machine) and simulation (of differing situations and conditions, traffic of differing network configurations, or simulating the results of using different search techniques). The kinds of question that researchers find motivating in this area concern the changes in teaching and learning when students can explore, enhance and even construct their own understandings. One of the unusual aspects of this is that these tools, these systems, are often the sorts of tool that educational researchers in *other* fields are interested in. It may be that in this area CS education research has a unique export. A good overview of literature in this area can be found in [**Stasko/Hundhausen** chapter]

Teaching Methods

It is clear that this is a very broad topic that could be broken down further. However, there are two kinds of question that researchers find motivating in this area.

Firstly, the concerns are with how do teachers can "build bridges" for students, how they can scaffold their students' learning, helping them to make sense, of the subject. Sphorer and Soloway's work on programming plans (Sphorer, Soloway, & Pope, 1985) is an early example of work in this area, which has proved to be influential in the long term. Linn & Clancy, too, have had great impact with their work on case studies (Linn & Clancy, 1992).

Secondly, there is a body of work regarding how teachers control the dynamics of the teaching interaction to make it profitable. This often highlights activities and presentation methods constructed to advance student learning, e.g., (Astrachan, 1998; Astrachan, Wilkes, & Smith, 1997).

There is also some work that is theoretically motivated, building on findings from other disciplines (largely in psychology) with regard to issues like "active learning", "cognitive styles" and "learning styles". We discuss the basis for some of these ideas later in the *link research to relevant theory* section.

Other aspects of teaching methods research are those studies that take a single approach and trace it through many instantiations, sometimes many institutions. One such example is the Effective Projectwork in Computer Science project (EPCoS) (Fincher, Petre, & Clark, 2001). This examined the place of projectwork in the computing curriculum, identifying different kinds of projectwork; how and where it occurs in the curriculum; and how teachers discover ideas about the topic, how they utilise it, and how they transfer it amongst themselves.

Assessment

Assessment is another broad area, which we break down in terms of categories of research question: types of assessment, validity of assessment, automated grading.

Some questions address distinctions among *types of assessment,* trying to understand which types are most suitable for particular assessment aims or contexts, and what makes them effective. Assessment may formative (conducted at stages during the course of a project in order to contribute information that will influence, i.e., help to 'form', subsequent stages) or summative (conducted at the end of a project in order to assess the project as a whole, i.e., to 'sum it up'). It may address different kinds of learning, such as acquisition of factual knowledge, change in conceptual understanding, acquisition of skills. A good example is Lister and Leaney's work applying Bloom's taxonomy to CS education (Lister & Leaney, 2003)

Some research is aimed at understanding *whether the assessment is valid*, whether it represents the kinds of knowledge one wants it to assess. For example, results of conventional testing might be compared to conclusions of in-depth investigations of students understanding of the material tested, as was done at the Open University (Petre, Price, & Carswell, 1996).

Issues associated with *automating grading* form another strand. Among these issues are the assignment of partial credit, handling responses to open questions, verification that the examination has actually been taken by the named student, and plagiarism (e.g., (Lancaster & Culwin, 2004) and Dick (Sheard, Dick, Markham, Macdonald, & Walsh, 2002)). Work in this area started very early; a more recent example is the Ceildh project (Foxley, 2003), a significant attempt to develop and research automated grading.

There are also studies that consider assessment in a broader context, examining assessment from a curricular or cross-institutional perspective. For example, one ITiCSE working group (McCracken et al., 2001) conducted a multi-institution study which examined what proportion of 'first competency' programmers could actually the solve the sort of programming problems their teachers thought them capable of.

Educational Technology

Many researchers find this area motivating, and not always for disciplinary reasons. Some of the work that occurs here is generic. In that way, it is similar to the work in visualization (above), drawing upon our disciplinary skills and interests, to investigate questions in the world. Just as every inventor can see opportunities to build a better mousetrap, every computer scientist can see opportunities to build a better application. Some of this work in this area leverages the advantages that new devices and technologies offer. There is work on presentation systems, using tablet PCs (Anderson, Anderson, VanDeGrift, Wolfman, & Yashuhara, 2003), and Personal Digital Assistants (PDA) (B. A. Myers, 2001). There is also an area of interest in whole environments, instrumented learning spaces, and "smart" classrooms (Abowd, 1999).

Of course, there are also examples of educational technology being applied to CS subject matter. Common examples are pedagogic environments for initial language learning, for example BlueJ (www.bluej.org) and DrScheme (www.drscheme.org), but this area, too, can involve hardware, for example the "smart" whiteboard *Ideogramic* for creating UML diagrams (www.ideogramic.com). A more detailed overview of the environments literature can be found in [Guzdial chapter].

Transferring Professional Practice into the Classroom

CS is a vocational discipline. Unlike some other academic disciplines, this means that there is a cadre of professionals who are developing and expanding (at least) the practices of the discipline, in parallel with academia.

There is a research strand that takes as its focus the transfer of professional practice into the classroom. The motivations are clear. Academic researchers observe expert practice and say, "Ooh, our students need to know about that because they're going to have to make the transition into that environment" or, sometimes, "Gee, maybe if we taught it that way, our students' understanding would improve". Academics seek this in the work of professionals because some professional methodologies are specifically about good practice, about scaffolding people to "walk the walk" of software development in an efficient and effective way. A good example here is XP (which typically manifests in the classroom as pair programming) (Williams & Kessler, 2001).

Incorporating new developments and new technologies

Running alongside *transferring professional practice*, but separate from it, is an area which concerns itself with incorporating new developments and new technologies into the classroom. Some of the most ephemeral research in CS education research is in this domain, because industry moves fast. Educators see new developments and work to incorporate them. Sadly, only some of those developments last. For those that do, the research often ends up simply reporting on how to make a transition from old to new.

Transferring from campus-based teaching to distance education

Like all other disciplines, CS is increasingly taught in "non-traditional" settings. In the twenty-first century, "non-traditional" almost always means "at-a-distance-and -computer-mediated". Of course, there are many generic issues here about Web-based learning and what is an appropriate transfer of educational interaction from a radically co-located environment to a remote one. An early pioneer in this field has been the RUNESTONE project. Within RUNESTONE students work on a CS project (involving real-time systems) in teams, under academic supervision. The especially interesting feature of RUNESTONE is that half the students in each team are from Sweden and half in each team are from the US (Hause, Almstrum, Last, & Woodroffe, 2001). They live and work in different time zones, and never meet face to face. Yet they work on the same project, for which they are assessed as in any other similar piece of academic work.

Recruitment and Retention

Issues of recruitment and retention are a motivating area, and this includes a considerable interest, and body of work, in diversity and gender issues. There are real questions about what makes students come into CS and what makes students stay in CS. And the other side of the coin: what makes them not bother, and what makes them leave. Usually, there's a diversity perspective on this, but there doesn't have to be. It can be just one of the questions that concern the discipline, for example, what innate abilities contribute to performance in CS? What are the kinds of skills you can engender?

Construction of the discipline

The final category is of a different kind, concerning questions about the construction of the discipline. In some other domains, for example mathematics, there is a *didactics*, a sense of what it is we're supposed to teach, an acknowledgement of what we should cover as fundamental principles, and an associated understanding of which curricular areas are advanced and which are optional. There can also be a sense of how subjects should be taught, how they should be delivered. We clearly don't have agreement on that in CS, although the ACM computing curricula (http://www.acm.org/education/curricula.html) and, in particular CC2001 (http://www.computer.org/education/cc2001/final/index.htm), has generated a real discourse around these areas.

As well as curricula constructions, this area encompasses questions concerning the nature of the discipline: is it an engineering discipline? Is it mathematics? Is it design? Is it business? Is it something else altogether? And this leads to discussions of interpretation (Fincher, 1999) and scope; of how many things this discipline actually embraces.

What Computer Science Education Research isn't

"Computer Science" is itself a new discipline, created out of many others. The debt to mathematics is clear, and often seen as the "core". Formal methods, theoretical CS, algorithms etc., all use mathematical methodology. Hardware interfacing is clearly akin to electronic engineering. Software engineering methodologies have come into academia from industrial practice; operational research and business processes from business schools.

CS education research is also informed by other disciplines—much theory is from education and the learning sciences: experimental and analysis techniques are drawn widely, from statistics to empirical studies to social science methodologies such as ethnography. In this way its development has been similar to many of the other research areas within CS which have developed from other subjects.

The key to viewing CS education as a distinct area must surely be in the questions that we ask. If we ask questions that are generalist ("Do students learn better from face-to-face or Web-based interaction") or facile ("Do students learn better if A or B is their first programming language"), then perhaps practice descriptions may be written, perhaps more pedagogic environments and visualization tools may be built, but perhaps what we are doing (although valuable) is not CS education research.

But if we ask questions that can only be addressed from within CS, perhaps "Does a knowledge of computer architecture make better—more expert—programmers?" or "Does student understanding of programming concepts differ with language of first instruction?" then these questions *cannot* be addressed by someone outside of CS. What meaning, after all, would such questions have to someone who could not program? Who did not understand what the quality of the relationship between, say, functional and object-oriented programming and its import for first instruction?

There is a lot of material published in the area of "computer science education", and much of it is very good. However, the majority of it is not computer science education *research*, and the ways in which "computer science education" papers are good and bad are different from the ways in which "computer science education research" papers are good and bad.

CS education research is new. It co-exists in places with other sorts of publication (like SIGCSE), and where it starts and stops, where the edges of the endeavor are, are not yet entirely clear.

A Preface to Pragmatics

> Science is not a cut-and-dried body of knowledge which someone has collected once and for all: it is an attitude of mind, a way of finding out. Unless these facts are appreciated science degenerates into mere scholarship and its study has a narrowing instead of a broadening effect on the mind. (Holmstrom, 1947)

Science is a discourse. The "stuff of science" resides not solely in data and variables, in hypothesis and observation—but in ideas and reasoning, in reflection and critique, and in conversational interactions among scientists who aspire to explain the world.

Method of science vs. scientific method

> At a workshop I attended not long ago, my colleague Matthew Chalmers made the observation that computer science is based entirely on philosophy of the pre-1930s. Computer-science in practice involves reducing high-level behaviors to low-level, mechanical explanations, formalizing them through pure scientific rationality; in this, computer science reveals its history as part of a positivist, reductionist tradition. (Dourish, 2001)

We're familiar with thinking about science in terms of "scientific method", an approach to theory validation based on Karl Popper's (Popper, 1959) hypothetico-deductivism. Popperian science proceeds by refuting hypotheses. Hypotheses are specific operational predictions that can be tested *empirically*. He distinguished science from "pseudo-science" by the notion of falsifiability, the doctrine that hypotheses are tested in order to demonstrate that they are wrong.

Experimenters can (and did) say "as impossible as a black swan"—for, in their world, all swans were white. Because one has only seen white swans, does that mean all swans are white—and that they are *all* white, forever? Well, maybe. But if you see just one black swan, then the theory is disproved. It doesn't matter how many white swans you see before that or after that one black swan, it doesn't make the theory any stronger, or truer. In other words, in a Popperian world, we can't have true theories, only theories that haven't yet been disproved.

Common thinking has adopted a “shorthand science”[2], shifting the focus from the *meaning* of rigor in science to an *adherence to procedure*: as if following a set of rules will produce the desired result. The shorthand is that “science” is seeking knowledge through scientific method. By viewing science as a discourse, we need to put scientific method into perspective. Precise, controlled experiments are *one* approach to rigor, but not the only one, certainly with regard to the human sciences. There is no such thing as a “blank state human”, a person without experiences, abilities, and knowledge. Hence we cannot control all human variables, and a strict experimental approach can exclude large portions of human experience. Human behavior—governed by fallible human perception and variable human cognition, and occurring in complex social contexts—presents problems for the pursuit of a straightforward one-to-one relationship between things and events in the outside world and people’s knowledge of them. The complexity of human social reality makes it impossible to establish “facts” about behavior unequivocally. Falsification allows us to eliminate some theories, but we can’t be sure that what survives is a “true” theory. We work with “best approximations”, in which we have more or less confidence, depending on what evidence supports them.

In research involving human beings, it can be useful to think more broadly, in terms of a “method of science”, characterized by principles such as articulation, validation, exposure to falsification, and generalization. “Method of science” demands rigor and seeks to contribute to empirically-founded theory, but it does not view “scientific method” as the only way of gaining knowledge. Rather, it seeks information gathered using a variety of methods. This “method of science” admits a broader approach to rigor, and sees the construction and discussion of theory as closer to the ideas of Thomas Kuhn (Kuhn, 1970), who described science as a social endeavor which passed through a series of “paradigms”, each paradigm persisting until a more comprehensive (and popular) explanation supplants it: the canonical example is the way that (in Physics) Einstein’s relativistic paradigm supplanted the Newtonian one.

Method of Science

Induction	Deduction
↓	↓
“questions”	“answers”
Identification of regularities	method of testing hypotheses
generation of theory	scientific method
↓	↓
utility	replicability/repeatability

Table 1: Method of Science

“Method of science” values description as well as hypothesis generation. It sees the relationship between theory and evidence as two-way, allowing both theory-driven and data-driven investigations. As shown in the *Method of Science* figure, above, “method of science” embraces both inductive and deductive reasoning. Deduction tests the predictions or hypotheses derived from a theory in ways that allow them to be disproved; induction draws inferences from observations in order to make generalizations. “Shorthand science” emphasizes deductive approaches at the expense of inductive approaches, but they *both* have value and *both* contribute to the discourse.

A combined approach can strengthen an investigation and make it less likely that critical factors will be overlooked, particularly in investigations of complex systems. For example, using induction to observe patterns in programmers' use of representations (e.g. graphical *vs* textual; data-flow based *vs* control-flow based), Thomas Green generated a conjecture about the influence of representation on task performance. Using deduction, he predicted that people using representations would perform best when the information made accessible by the representation matched that required by the task. He and others tested the conjecture empirically by measuring programmers' performance.

At the heart of the distinction between "scientific method" and "method of science" is the role of theory. Scientific method uses theory as a driver and a goal: theory generates hypotheses that can be falsified, and scientific enquiry produces predictive theory. With "method of science" what we are seeking is not formalism *per se*, but articulation. "Method of science" recognizes that the process of articulation is a crucial part of science: making explicit in sufficient explanatory detail. Articulation is needed for assumptions, meanings, constructs and their relationships in order to help clarify the contribution to the discourse.

Epistemology

We need to be aware of what we understand to be the substance of science, of what we believe constitutes "knowledge", what we mean by "truth", and what we allow as "theory". In other words, we need to be aware of epistemology, of the philosophical stances underlying our research. Many different words and phrases are used in this area. Some are: *theory, explanatory theory, empirical laws, models* (sometimes "*explanatory models*"), and *conceptual frameworks* (sometimes "*conceptual lenses*"). Exploring what they might mean for CS education research and how we can use them to strengthen our work helps frame our endeavor.

"Theory" can be a problematic term for researchers whose background is in analytic disciplines. For the natural sciences, "theory" means "understanding that will generalize across situations or events", and it necessarily has a predictive and causal quality. "If I do *this*, then *that* will happen". Often it is bound up with notions of hypothesis, experimental setup and scientific method, and with representations that are not only precise, but also capable of formal expression—often mathematical. "Theory" in this sense works extremely well, and is responsible for many of the great advances of the last few centuries. However, it relies on certain fundamental circumstances:

- it relies on an unchanging natural world (the speed of light is always the same in Brisbane as it is in Berlin);
- it relies on accepted experimental procedures (external variability can be controlled so that what the experimenter observes actually is caused by what they think is the cause and not some extraneous, unaccounted factor[3]), and

- it relies on accepted methods of analysis, so that the “truth” of a knowledge claim can be assessed (you can/can’t tell that from Bayesian analysis).

CS education researchers who start from this perspective, often flounder and “give up before they begin” in part because of the difference in research conditions. Learning environments are continually changing, are hard to control, and in CS education research we do not have an established “common ground” of methods of investigation and analysis. Naïve researchers either try to tie down the world to a ridiculous extent in the hope that their experience will generalize (pre- and post-tests; teaching two sections with the same teacher, same aims but different materials/approaches) or give insufficient (or the wrong sort of) evidence to convince: they describe inessentials of the course setting in detail and then say “the students seem to like it”. Considering a separate notion of “theory” can alleviate some of these problems.

> To explain the phenomena in the world of our experience, to answer the question ‘why?’ rather than only the question ‘what?’ is one of the foremost objectives of empirical science (Hempel & Oppenheim, 1965)

Explanatory Theories (also called “mid-range” theories)

Explanatory theories are characteristic of knowledge in the social and human sciences. The purpose of an “explanatory theory” is to explain observed human behavior (i.e., it is not predictive)—although generalized theoretical understanding is still a goal. The human sciences seek to encompass the complexity of human systems, actions, and experiences, understanding them as influenced not just by what can be observed and measured but also by intentions, motives, attitudes, and beliefs. Explanatory theories may not seek to account for cause and effect, but rather to tease out and explicate factors and conventions that mediate human action and understanding in particular situations.

Those with an analytic background often deny that there can be any truth or utility in such an approach, but partly this is an under-representation of the word “explanatory”. Work in the human sciences is often based on small numbers of very specific situations that are explored in a “deep” fashion. That is, many aspects are explored in many layers (explaining the Second World War, for example, requires many different layers of explanation, which take into account many different influences and affects; detailing the cultural norms of “gift-giving” require understanding of anthropological theory, specific cultural understanding and the vision to relate the two; the question “Do prisons work?” necessitates a mix of political, economic and psychological approaches, together with their respective theory bases.). However “theory” in this context is still *outside* of the circumstances in which observation/evaluation/experimentation take place. It is external, added-value intellectual effort that the researcher brings to their work. All too often, however, in CS education research, this is quite absent: the work is simply descriptive and in no way explanatory.

So, a “theory” in the natural sciences is concerned with cause-effect and predictability; a “theory” in the social sciences is concerned with the reasons for effects, the causes of which may not be determined, or determinable.

Empirical Laws

Given these intellectual contradictions, many CS education researchers strive to generalize to "empirical laws". These are well-understood, simple quantitative predictions of human performance. Empirical Laws gain their validity from statistical significance derived from sampling large populations, they may or may not apply to any given individual.

Largely, for our purposes, these are derived from cognitive science. Limits on short-term memory, for example, clearly have implications for complex cognitive tasks (such as programming) and on the design of systems to support such activities. Cognitive science has demonstrated that most human beings have a short-term memory capacity of seven, plus or minus two, things (Miller, 1956). Other disciplines also strive to discover (and utilize) empirical laws. There is no doubt that much medical research is devoted to discovering how the "average" human will respond to a specific intervention (normally drugs). Like "plus or minus two" this data comes with error boundaries—hence the list of "possible side effects" that come with every bottle of aspirin. Epidemiology, too, also relies on most people being "the same", so if there is a peak in, say, infant deaths in Canterbury in 1597 then it can be taken as an indicator of something out of the ordinary. Educational psychology also accrues data of this kind. SATs, IQ tests and other forms of standardized testing rely on a) the quantity of people who take the tests and b) the fact that what is tested is an ability ("intelligence") that every person possesses in some measure.[4]

Models

A common feature of theory and theoretical work, of whatever nature, is the formation of models. Models are generalized, hypothetical descriptions of something that is not directly observable. Sometimes, the generation of a model is a sufficient end of a theoretically-inspired investigation; mostly it occurs as part of the process of research.

> Models are inevitably simplified versions of reality. They can only aim to select the more important variables in a complex problem, so as to provide a manageable insight into the issues. (Kember, 1995)

Very often models are based on an analogy: an atom is model (the most common analogy is to the solar system, with the electrons "orbiting" the nucleus); culture is a model. In every analogy some things are the same, and some things are different. It is worth thinking about what is important to have as similar, and what can safely be different. Some researchers consider models to be "second best". If we can explain things by cause-and-effect, that's best: otherwise we'll make do with comparisons and analogies.

Conceptual Frameworks

A conceptual framework defines a particular point of view within a discipline from which the researcher focuses his or her study. This "theoretical perspective" identifies underlying assumptions from which particular kinds of questions are generated.

For example, when researching ecological impact and sustainability there are a variety of conceptual frameworks that a researcher can work within, including “net primary productivity” (NPP) and the “ecological footprint” framework (OECD, 1995).

NPP is the amount of energy left after subtracting the respiration of primary producers (mostly plants) from the total amount of energy (mostly solar) that is fixed biologically. The researcher who takes an NPP framework also takes the underlying assumptions that human co-option of terrestrial resources contributes to the extinction of species and will shut out a number of options for humanity. Consequently the measures (or indicators) that they use are constrained by the framework and its assumptions.

Ecological footprints are calculated by a population’s demand for domestic food, forest products and fossil energy consumption, converted into the required area of eco-productive (agricultural and forested) land. An ecological footprint provides an area-based indicator of the physical limits to material growth. The researcher who adopts an “ecological footprint” framework takes on the key assumptions that industrial economies currently survive through importing the “surplus” carrying capacity of developing countries. This pattern of consumption activity implies (1) that developing countries are restricted in their own development (insufficient carrying capacity available) and (2) that developing countries’ desire to emulate Western living standards cannot be fulfilled since there is insufficient global carrying capacity: the footprint already covers the earth.

So, in this example, we see that both researchers are interested in human/environment interaction, but, depending on their conceptual frameworks, they ask different questions and use different methods and indicators to provide evidence to answer those questions.

Relationship between phenomena and evidence: two examples

It is clear that much of science is concerned with the observation of circumstance, the recording of phenomena. However, what is the nature of observation? What does it mean to look at (or for) something particular? For something that will support (or disprove) what we believe to be the case? What is it that turns the *observation of a phenomenon* into *evidence that supports a theory* or theoretical approach? To illustrate the problems of phenomena and evidence we take two examples, one from physics and one from medical history

Physics

Since 1910, physicists have (with increasing sophistication) been able to record the “trails” made by sub-atomic particles in cloud chambers. Sub-atomic particles had been *theoretically* postulated previously, of course, and there were *models* of what atoms were and they were made up of. But in 1910 photographs were available for the first time. But what was the epistemological status of these pictures? Are they simply recording a phenomenon that happens in the world anyway, whether we can directly see it or not? Or do those trails provide evidence of the very existence of those particles, which were only previously theorized about? (“This is not a photograph of a single, possibly *singular*, event, but evidence of great regularity, of how the world works")

Medical History
Edward Jenner (1749-1823) was a doctor in a rural English setting. He noticed that, within the community he served, people who worked closely with cattle and had caught a mild infection from them (called cowpox) did not get smallpox—at that time a disfiguring, often deadly, and much-feared disease. He found this curious, and compiled extensive case studies and observations to find out if this was, in fact, the case (Jenner, 1798). He found that this was so, and went on to invent the process of "vaccination" protecting people from smallpox.

But what is the epistemological status of his case studies? Are they simply recording a *phenomenon* that happens in the world anyway, whether we can see it or not? ("Oh yes, Sarah had cowpox and didn't get smallpox") Or do those compilations provide evidence of the very existence of something only previously theorized about? ("There is a relationship, not between specific people and circumstances, but between these diseases").

CS Education Research
Much of what is published as CS education (called "research" or not) has been concerned with *noticing phenomena*: "This is what happens in my classroom", "This is what happens when you teach *x* in this way", "If I teach *x* differently, something else happens". What moves recognition of phenomena to *evidence* is purposeful investigation and a relationship of that to theory.

Truth Claims
Underlying these issues—of epistemology, of theory, of phenomenon and evidence—is how we can know something is "true" and how we share that knowledge. Different parts of life have different systems of creating and understanding knowledge, which are not transferable to other systems. For example, religious truth-claims find their validity in the epistemology of religions and judicial truth-claims find their validity in the epistemology of the judicial system, but we can't apply the way we believe in a God to the way we judge a criminal. The evidence is different, the "burden of proof" is different and the way we share the results with others is different.

Scientific truth-claims find their validity in the epistemology of "science" which we in our consideration of "method of science" take to be meaning-filled public discourse regarding replicable experience. Validity (or "truth") in the discourse science involves a combination of factors, including the accuracy of our observation, the quality of our reasoning and the completeness of our explanation

The kinds of research questions that can be asked are (partly) dictated by the researcher's epistemology. This underpins (and constrains) what *kinds* of question can be asked, what are *legitimate* questions, what are *appropriate* questions, and even what questions are *allowable*. Epistemology provides ways of conceiving and seeking knowledge which focus endeavor. Theories and models provide reasoning frameworks, which highlight important relationships, focus enquiry and hence simplify.

Pragmatism: striking a balance

Our approach to CS education research is pragmatic: demanding rigor, aspiring to generating fact and theory, but accepting "best fit" *en route*, and using whatever research methods contribute to the process and discourse. As pragmatists, we needn't argue about whether we can derive general, predictive models for CS education or we can only explicate specific evidence – we can use explications as a step toward explanatory theory preliminary to predictive theory, and not worry about whether the predictive theory will actually be achieved. We strive constantly for rigor, but we seek rigor *in terms of* whichever approach we're taking.

Each discipline has its own rigor and standards; each study must be rigorous in its own terms. It is essential to understand the underlying premises of any approach employed. Importing methods alone is insufficient—the researcher must understand the knowledge concerns, and the associated assumptions and limitations, not just the application of techniques. There can be multiple ways of studying a phenomenon, but that doesn't mean, "anything goes". The conceptual framing of each method must be understood and respected. This is especially true if different approaches are to be combined or compared. This is at the heart of the "method of science". As Liam Bannon said "Science is not just a set of methods, but a way of reasoning."

The CS education research endeavor

Our approach is pragmatic. We do not advocate research driven by any one theory, discipline, or methodology. Rather, what we intend to do in the next sections is outline the necessary process for conducting effective empirical research in CS education. We concentrate on what to do, not how to do it.

We shall take as a framework for the next sections, "six guiding principles" for research that have been formulated by the *Committee on Scientific Principles for Education Research* and which were published in their book *Scientific Research in Education* (Shavelson & Towne, 2002) . There are several reasons why we do this: Firstly, the principles they have identified are good ones. Secondly, they are formulated at a sufficient level of abstraction so that they are not bound to any specific research stance, or tradition. Thirdly, they embody "method of science".

In each section we link these principles to CS education research, its nature and status, and illustrate some of the problems that are bound with research in CS education.

The six principles

- Pose significant questions that can be answered empirically
- Link research to relevant theory
- Use methods that permit direct investigation of the question
- Provide a coherent and explicit chain of reasoning
- Replicate and generalize across studies
- Disclose research to encourage professional scrutiny and critique

Pose significant questions that can be answered empirically

What we mean to offer is a pragmatic approach to empirical research in CS education, because CS education research will necessarily involve compromises. The pragmatic approach centers on understanding the value of evidence and its fitness for purpose—its utility.

Designing empirical studies: 1 – 2 – 3

Our advice about how to design empirical studies for CS education research always follows the same formula:

1. Figure out what the question is

Figuring out what the question is, is probably the most intensive step. It requires identifying what is important for you to know, out of what might be known. It requires an assessment of whether what you want to know is something that can feasibly and reasonably be investigated. It typically involves an analysis of whatever motivated you to ask the question in the first place, and it often involves resolving an initial question into a smaller, more tractable question.

2. Decide what sort of evidence will satisfy you

The next step is deciding what sort of evidence will satisfy—actually, will satisfy a reasonable, skeptical colleague—*in addressing or answering the question.* What would a sufficient answer look like? Determining the evidence requirements involves figuring out how the phenomenon of interest might be manifest in the world and hence how to 'operationalise' your question, to phrase it in terms of things you can observe directly. It also involves learning the value of different types of evidence and assessing how strong the evidence must be to serve your purpose.

only then…

3. Choose a technique that will produce the required evidence

In this pragmatic approach, methods or techniques follow from the question and the evidence requirement. This is equally true of theory-driven and inductive studies.

Researchers who start by asking *how can I design an experiment in order to prove X* are starting from the wrong place, because they've chosen a technique before they've really considered what the question is. They're seeking proof, when experiments can only disprove. Even well-designed experiments are of little utility if they address the wrong question—and it's difficult to design an experiment well if the question is not clearly in focus. Good experiment design requires an understanding of the key variables, as well as a precise question.

Sorting out what the question is

> "A question well-stated is a question half-answered." (Isaac & Michael, 1989)

Formulating the question well sets up the rest of the study design. Some questions arise from theory, some from observation or from conjecture based on observation. Question formulation is itself a crucial reasoning skill. There is a refinement path from an initial, intuitive question to a well-specified question worth asking, a question that relates to what is already known and understood, is significant, and can be investigated.

Why ask?

A good first step in formulating a research question from an intuitive curiosity is to consider what kind of evidence made you think the question was worth asking in the first place. The question is not just why ask, but why bother asking further? Considering both why that evidence (be it introspection, anecdote, a classroom observation, a line of discussion in the literature) was enough to make you ask – and what is missing from that evidence in providing an answer – can provide insight into specifying the question.

What would not suffice?

A good second step is to consider what sort of answer would be inadequate – what a 'non-answer' would look like. This can help to clarify the evidence requirements. It can also help in distinguishing the question of interest from other, related questions.

Counter-examples?

A good third step is to consider what a counter-example or a contradiction might look like. How might the conjecture be falsified? People tend to seek confirmatory evidence, to try to demonstrate what they believe to be true. But often insight lies in the 'surprises', the unexpected, the contradictions. Considering the nature of counter-evidence can be a way of reflecting on the basis of the question: the observation, conjecture, or hypothesis underlying it. It can be a way of exposing alternative accounts, which in turn can be a way of exposing inadequacies in the question formulation.

Work back from the analysis
Another useful approach is to think 'backwards', to consider what relevant data might look like and how it might be analyzed. Plan the analysis with the study.

So what?
One of the tests of a worthwhile question is 'so what', or what will I do with the answer if I get it? In formulating a research question, it's important to consider the question not just in its own terms, but also in the context of the discourse to which the study might contribute. Hence, it will be important to establish the:

- *importance* of the question
- *significance* of the findings
- *implications* for theory
- *limitations* to generalization

These are appropriate considerations when framing the research question, because they distinguish questions worth asking.

Questions to avoid:
It's important to identify questions that are worth answering. Which means that some questions are better avoided:

- the unanswerable: There are questions that, although compelling, are too abstract, too elusive to operationalise, or simply far too costly to address effectively.
- overworked topics: Some questions have received considerable attention from others, limiting the likely impact of any one study. (e.g., "Can mathematics grades be used to predict success in CS1?") Unless a researcher can bring a surprising novelty and insight to an overworked topic, or provide a definitive study, the topic is best avoided. Sometimes, a parallel inquiry in a different context or addressing a different but structurally similar area can unlock an overworked topic.
- trivial topics: These are questions whose scope is too small to make them generaliseable, relevant or interesting to others. (e.g., "Do colored mice improve student productivity?")
- "one-shots": These are questions that don't aggregate, that don't contribute to an accumulation of evidence.

Evidence

The key to empirical studies is in knowing the value of *evidence*. Evidence is at the heart of the scientific discourse. Theories are supported or contradicted on the basis of evidence. Rather than thinking of studies as 'good' or 'bad', it's more productive to think of them as producing 'strong evidence' or 'weak evidence'. Whether or not a study is 'good' depends on whether it produces sufficiently sound and convincing evidence *for its purpose*.

Some purposes need only weak evidence or contrary examples. For example, a researcher might be content to make a minor interface design decision based on the performance of a handful of subjects on a simple task. When one is trying to dispute a universal claim, one only needs a single counter-example, one 'black swan'. Other purposes require strong evidence. For example, deciding whether to re-vamp an entire curriculum around project work suggests a substantial investment and probably demands compelling evidence of efficacy.

In CS education we can't hold enough of the world stationary to achieve precise control. 'Proof', therefore, is impossibility. Rather, the discourse examines the relative strengths of competing claims in terms of the evidence that supports them—and the evidence that contradicts them. So, we must decide what matters and how it can be investigated, whether by counting or through a qualitative method; i.e., we must understand the value of evidence.

A second key is to remember that research studies do not stand-alone; they must be assessed in the context of other studies that provide relevant evidence. Results accumulate, and studies may be repeated with different subject groups or with slight variations in order to explore the reliability (consistency across repetitions) and robustness of the findings (consistency across slightly different settings). Different studies may combine methods, or 'triangulate', in order to overcome the shortcomings of any one method.

Research is about *learning* (i.e., adding to knowledge), not proving. Discourse and scrutiny are as important as outcomes in developing theory. The purpose of empirical research is not only to observe behavior, but to *think about* behavior. Empirical science in young domains such as CS education is not so much a process of getting answers as one of finding ever better questions. We are unlikely to achieve total accuracy; total generalizeability; realistic integration, comprehensiveness, or completeness. But *so what*? Will we be able to ask better questions?

Data is not necessarily evidence

> There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment in fact. (Mark Twain)

Data is not necessarily evidence. Data *becomes* evidence when its relevance to the conjectures being considered is established. [give ref; evidentiary reasoning man] The path from data to evidence is *inference*—reasoning about what is observed and what it means. Inference is reasoning from what we know and observe to conclusions, explications and, possibly, predictions. The value of evidence relies both on the quality of the *data* (deriving from the appropriateness of its selection, the quality of its capture, and its representativeness) and on the quality of the *inference* that connects the data to the phenomenon of interest.

Knowing the value of evidence:
quality of data + quality of inference

Evidence is not proof

Evidence is not proof. In general, it is whatever empirical data is sufficient to cause us to conclude that one account is more probably true than not, or is probably more true than another.

In order to substantiate evidence, we must establish its:

- plausibility (is it likely, given existing knowledge)
- validity (that it is a true reflection of the phenomenon under investigation)
- relevance (the data relates to the research question)
- credibility (whether the researcher's interpretation is likely to be accurate)
- inferential force (the legitimacy of the chain of reasoning).

We need to understand the value of different forms of evidence and how they fit together. We need to understand how reliable the evidence is likely to be (how consistent the outcomes will be given repetition by different researchers, at different times, with a different sample of the same population) how robust (how consistent the outcomes will be across different related tasks, across different environments, across different related contexts), what margin of error it entails. In the same way that we report the standard deviation associated with a mean, we must report the uncertainty and error associated with evidence—hence enabling assessment of its fitness for purpose.

Richness/Rigor

We need to strike a balance between the richness of realism and the precision of controlled studies, exploiting the advantages and compensating for the disadvantages of each. As Bill Curtis has said "We need experience in a real environment to figure out what the critical variables are…practical experience leads to a better development of relevant theory".

Table 2: Some characteristics of evidence: richness and rigor (Informed by ideas from Marilyn Mantei, D.A. Schkade, and Bill Curtis)

<table>
<tr><th>Richness/
Realism</th><th>Rigor/
Control</th></tr>
<tr><td>Reflecting reality</td><td>Replicable</td></tr>
<tr><td>(Natural phenomena)</td><td>Repeatable</td></tr>
<tr><td>'Practical'</td><td>Reliable</td></tr>
<tr><td>Valid</td><td></td></tr>
<tr><td colspan="2">BUT</td></tr>
<tr><td>"Real data is real dirty."</td><td>Sand through the fingers</td></tr>
</table>

‘Richness’ indicates the number and generality of questions we can attempt to address using the available evidence. The very process of squeezing reality into the requisite structure of a controlled experiment often strips it of some of its worldly richness—and hence constrains the scope of the question it can answer. Equally, the very richness of realistic investigation makes it hard to draw conclusions that are not context dependent, and possibly specific to the particular instance.

So, on the one hand designing experiments in a CS education context can be like ‘sand through the fingers’: we try to grasp an issue, but it slips away as we sift through factors in a search for control. On the other, as Bill Curtis so eloquently phrased it: “Real data is real dirty.” In effect, realism trades off with control, richness with precision.

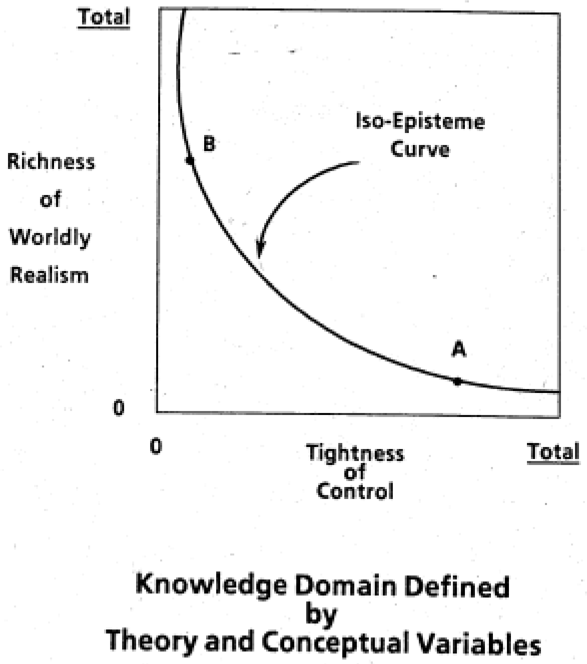

Total
Richness of Worldly Realism
0
B
Iso-Episteme Curve
A
0
Tightness of Control
Total
Knowledge Domain Defined by Theory and Conceptual Variables

Figure 2: Richard Mason (Mason, 1989) has characterized the richness/rigor tradeoff as a space populated by iso-episteme curves, each curve representing points of equal knowledge generated by various forms of investigation. In the figure, points A and B represent equal knowledge, but they are qualitatively different. A might represent a tightly controlled experiment, whereas B represents a study that incorporates considerable reality. Mason notes that the amount of knowledge generated depends on the skill and care of study design and execution; a well-conducted study would be placed on a higher curve than a sloppy one.

Richard Mason has characterized the richness/rigor tradeoff as a space populated by iso-episteme curves “which represent the fundamental tradeoffs that must be made between these dimensions in conducting research” (p. 10), as shown in Figure XX. His analysis, too, is pragmatic. He observes that all empirical studies can be improved, but that improvements are made at a cost.

Granularity and focus are important concepts here. The granularity of the evidence must relate to the focus of the research question. Very precise studies may leave too much of the question unaddressed; very realistic studies may illuminate too many factors to provide insight about the particular question—in either case, the mismatch between study and question mean that the resultant evidence does not serve its purpose adequately.

There are many forms of evidence we tend to think of as weak. A good example is 'anecdote', which is often used to motivate studies, but is rarely accepted as sufficient evidence. But a *collection* of anecdotes, gathered from a variety of independent sources, can provide much stronger, more convincing evidence (as in Eisenstadt's study of 'tales of debugging' (Eisenstadt, 1993), sufficient to refocus research questions and provide a basis for comparison to other forms of evidence. The point is that purpose (what are the findings to be used for?) gives insight into pragmatics (what sort of evidence is sufficient).

Our perspective is that we need *both* realism and control in our studies of human behavior and reasoning, in our case the behavior and reasoning of students learning computer science, but not always in the same study. Concept formation and definition are a key part of science, and of the scientific discourse: they are preliminary to the operational definition needed for experimentation. Our focus is not on techniques per se, but on the accumulation of evidence fit for our purposes.

Triangulation and Multi-method Research

A way to strike a balance between realism and control is through the use of multiple methods over a series of studies that accumulate evidence. *Triangulation* is a name for combining techniques (for example in a series of studies of differing design) in order to shed light on an issue from different perspectives and thus overcome the limitations of any single technique. It can also be a method of challenging or corroborating findings, by gathering and comparing multiple data sources against each other.

Multi-method research is a way of both increasing the richness and increasing the rigor, sometimes within a study, but more often across a series of studies, e.g.:

- multiple factor studies (e.g., system design and task)
- multiple 'dependent variables' (e.g., not just performance, but change of performance with experience) within experimental and constrained task designs
- multiple measurement methods
- multiple research methods

Of course multiple-method research doesn't come 'for free'. Each of the methods must be understood in its own terms: the assumptions that attach to it, the level of resolution of the data it generates, its constraints and limitations, and so on. Its appropriateness in generating data relevant to the research question must be considered; the method must be fit for its purpose. The consequences and challenges of combining methods must be understood, lest inappropriate comparisons or faulty inferences be made. The application of each method must be valid, and the analysis and interpretation that links them must be sound.

Multi-method research carries a burden of responsibility: we must take care that when we ‘connect the dots’ we do so legitimately.

- respecting the epistemology and traditions that influence (and constrain) the methods we borrow
- articulating the reasoning by which we establish associations and make comparisons among disparate data, and taking care when drawing inferences beyond the granularity and focus of a given study and its evidence
- identifying constructs and developing operational definitions that are valid
- explicitly considering constraints, threats to validity, and possible alternative accounts
- confining conclusions to what is supported directly by the data, noting the limitations of our results, and taking care in recognizing when generalization is conjecture only
- indicating the level of uncertainty and error associated with the evidence

Multi-method research can increase rigor through triangulation, both by providing opportunities to challenge findings and to expose alternative accounts and by the accumulation of evidence, using different techniques to provide different approaches to a phenomenon. Pragmatically, rigor comes from using the best procedures we can devise to make the *best effort* to avoid error (such as observing principles of repetition, exposure to falsification) in order to get the best information/knowledge that we currently can.

Utility

At the heart of this perspective on evidence is a concept of ‘utility’, that the ‘goodness’ of evidence relates not to some absolute standard, but to its ‘fitness for purpose’, to its relevance and strength in terms of the use to which it is to be put, to its usefulness for understanding some phenomenon of interest. The utility of evidence is what establishes its value.

Utility applies throughout the empirical chain, from determining the focus and granularity of the question, through the collection of data, through the analysis. Each stage must match the evidence requirement, be it strong or weak. Sometimes there is a need for definitive, generaliseable answers. At other times, it doesn't matter if a phenomenon is universal, only that there's evidence it exists.

The Priority of Question Formulation

Of course, the 1 – 2 – 3 scheme is a simplification. It is rarely possible to specify the question without also thinking about evidence, which entails thinking about what data might be gathered, and how it might be analyzed. In practice, design of empirical studies is a highly iterative process running round the question, evidence, and data collection/analysis loop many times. Even so, there are times when a constrained opportunity demands action before adequate planning is possible: ‘have data set, will investigate’. Sometimes one seizes the opportunity before the question comes into focus. This is a dangerous, albeit sometimes necessary, route.

From the outset, it is important to *priorities* the framing of the question, and to use consideration of evidence and techniques as a way to *refine* the question, not as questions in their own right. The corollary is to avoid commitment to techniques until the question comes into satisfactory operational focus, and especially to beware the 'tail wagging the dog': premature commitment to a technique in the absence of a well-formed question.

leaping before looking

Failure to reflect.

Failure to recognize available evidence.

Failure to consider conflicts, confounds, representativeness, limitations, etc.

premature experimentation

For a precise study, you need a precise question.

If your starting point is too complex, broad, or poorly articulated, your question will disappear ‘as sand through the fingers’ as you try to refine an experiment design.

scarcity of theory

Failure to explicate conceptual underpinnings.

Failure to consider alternatives.

Failure to contribute evidence that can accumulate or be compared.

lack of situation

Ignorance and isolation are the enemies of discourse.

This is not just ‘bad form’, it can lead to ‘re-inventing the wheel’.

A day in the library can save six months of redundant research.

borrowing methods out of context

Can lead to major oversights, and to mis-matches between method and needed evidence.

Need to understand the underlying stance and assumptions associated with a given method. Are you applying it as it was intended? Is it able to uncover the sort of evidence you need?

putting the cart before the horse

Choosing techniques before understanding the question.

great expectations – taking too big a bite

‘A life’s work takes a lifetime, but it is achieved one step at a time’

‘How can one eat an elephant?’ If the question is intractable; ask a smaller question.

confusing anecdote with fact

What ‘everyone knows’ is not always accurate or valid.

confusing statistics with rigour

Einstein: “Not everything that counts can be counted, and not everything that can be counted counts.”

The point is to know what can and cannot be shown with different sorts of evidence.

The false seduction of the definitive experiment – experimentation is inappropriate when the questions are not yet precise enough.

death by surprise

Lack of respect for failure leads to false claims, distorted reports, and loss of crucial information.

A good study is one that is *informative*, even if it doesn’t go as expected.

Some of the most valuable results are surprises and side-effects.

Consider in advance what ‘failure’ will indicate, what will happen if the study goes wrong.

Link Research to Relevant Theory

Interdisciplinarity and the "Trading Zone"

CS education research is inevitably interdisciplinary. The nature of the field, and of the knowledge we aim to transmit in the course of education, is rooted in mathematically-derived, computational, natural science. However, the circumstances of the classroom, the nature of education, and models of teaching and learning, are areas that are amenable to investigation only through the human sciences. This means that our specific area is theory-scarce and we have to look to other disciplines for a theory-base. The tensions of different perspectives make coherent integration of the components of research—question, theory, method—tricky. At worst, this can mean inappropriate use of "borrowed" ideas and techniques. At its best, however, CS education can resolve these tensions into a new and distinctive way of working. Peter Galison (Galison, 1997) has a construction of how new areas of working can arise:

> I intend the term "trading zone" to be taken seriously, as a social, material, and intellectual mortar binding together the disunified traditions of experimenting, theorizing, and instrument building [in subcultures of physics]. Anthropologists are familiar with different cultures encountering one another through trade, even when the significance of the objects traded- and of the trade itself-may be utterly different for the two sides

He persuasively makes the case that such trading zones have built (and diminished) within physics[5], and identifies whole disciplinary areas that have emerged and flourished over time.

An example of such a trade, which illustrates the separation of the parties' points of view, is the selling of Manhattan, New York for a few dollars. The evidence for this trade resides in a letter in the archives in the Rijksarchief in The Hague. Peter Schaghen wrote this letter in 1626 to his employers, the West India Company. In it he reports (amongst other things):

> … our people are in good spirit and live in peace. The women also have borne some children there. They have purchased the Island Manhattes from the Indians for the value of 60 guilders.

So, it would seem from the perspective of the Dutch settlers, Peter Schaghen, the West India Company, and, perhaps especially from our historical perspective, that this was a very good trade for the Dutch. However, the idea of land ownership did not exist among Native Americans. The Lenape (the Native American tribe in the area) were "trading" the right to use the land – which everyone had *as a right* – for money and goods. The trade goods were valuable; they used uncommon raw materials and were troublesome and time-consuming to produce. From their point of view, the Lenape too made a very good trade. Here is a clear example of the significance of the things traded, and the trade itself, being totally different from the two sides.[6]

It is not just in the interdisciplinary areas of physics that the exchange of tools, methods and ideas occurs. Ann Brown, a "classic" psychologist who made great contributions to educational research, identifies similar problems. She describes some of the tensions within herself and her own research.

> As a design scientist in my field, I attempt to engineer innovative educational environments and simultaneously conduct experimental studies of those innovations. This involves orchestrating all aspects of a period of daily life in classrooms, a research activity for which I was not trained. My training was that of a classic learning theorist prepared to work with "subjects" (rats, children, sophomores), in strictly controlled laboratory settings. The methods I have employed in my previous life are not readily transported to the research activities I oversee currently (Brown, 1992)

She also describes how problems of the perception of the "place" in which she chose to work was perceived by others:

> Indeed, the first grant proposal I ever had rejected was about 10 years ago, when anonymous reviewers accused me of abandoning my experimental training and conducting "Pseudo-experimental research in quasi-naturalistic settings" This was not a flattering description of what I took to be microgenetic/observational studies of learning in the classroom (Brown, 1992)

The Trading Zone & CS education

> For a practicing scientist, even though he cannot be familiar with more than a small territory of science, must be prepared at any time to make excursions into collateral regions to find what he wants; must know, therefore, their main landmarks and enough of their languages to ask his way in them (Holmstrom, 1947)

It may be that every interdisciplinary field is a "trading zone" (or grows from one). For CS education, we must learn to speak with our trading partners: our use of theory is largely from the social and learning sciences—if we call upon "constructivism" would an educationalist recognize the concept? Or, if we wanted to "make an excursion" (as Holmstrom would have it) into education, would we know the terrain, would we recognize the important landmarks? Would we be able to tell what were important results from mediocre ones?

Our methods for empirical study are widely drawn—If we use quantitative methods, would a statistician recognize the "truth" of our conclusions? If we use ethnographic investigation, would an anthropologist recognize the validity of our methodology?

A key aspect of this is the necessity to play back into our own discipline. We must engage with the validity of methods in terms of the originating discipline, but also (when we engage with our CS colleagues) the validity of the investigation (and its methods and conclusions) within CS: can we tell a "CS story" about research that we're doing? So one sense of trading zones, of trying to build up a legitimate, well-founded, thoughtful use of interdisciplinarity, means that we have to establish standards. Part of what we need to do in order to ensure that CS education research is effective and is useful is to ensure that we understand what rigour means for us. If we understand *why* the methods that we're using are valid methods, if we understand *how* we're constructing knowledge in this way, then we can establish standards and our own distinct disciplinary norms and practices. David Schkade outlines this necessity well, in regard to his own "interdiscipline" of Information Studies (IS):

> … an important issue is the concept of "reference disciplines". Researchers in other disciplines cannot be relied upon to develop theories that are directly relevant to IS out of the goodness of their hearts. Their objectives are different. The IS field must develop at least some researchers who are competent theory builders in their own right, so that existing theories in reference disciplines can be rigorously adapted, or new theories developed, as needed. There is a lot of bad psychology, economics, etc. out there as well as good and useful work. (Schkade, 1989)

When human cultures engage in trade they often do not speak a common language, but a derivative of the language of the more dominant partner, which has a much reduced vocabulary and simplified grammar. Such trading languages are called "pidgins". A pidgin language, when it develops native speakers—when children are born who use it as their first language—grows in sophistication and complexity, developing new vocabulary, structures and idioms, and is then called a Creole. CS education research is, today, a pidgin. The outstanding question is whether we can develop distinctive areas of working – moving towards a Creole: whether we can "grow up". We conclude with David Schkade: "Research imported from other disciplines should be viewed with a healthy scepticism until … researchers, developing the reasoning on their own, see the rationale and applicability themselves."

Theories of learning: a reference discipline & a trading partner

If we are to take seriously the guiding principle "link research to relevant theory", and if CS education is theory-scarce, then we must be familiar with the construction and investigation of theory in other, relevant, areas. One of the most obvious territories to explore is that of learning theories. These have themselves been formulated in several separate disciplines: from cognitive and educational psychology, certainly, but also from sociology and social policy and from developmental and childhood studies. Also, historically, there have been many influential individuals who have founded schools—and schools of thought—and accrued followers. We need to be aware of their histories and traditions before we claim them as trade goods.

There are different ways in which researchers can think about education and learning, and different ways in which they might impact on research agendas. Some of these ways come directly from theories/theorists. Others come from the ways in which theories have been instantiated in educational environments. This separation is crude but indicative of one of the ways learning theories have been used. We explore a few possibilities further.

Learning

Empirical Laws

There is considerable work within "classic" cognitive and developmental psychology that examines and defines broad categories of human cognitive capacity. These investigations and results have a significant effect on human learning, even if there is little we can do address them with any specific educational (instructional) context. The finding that human short-term (or "working") memory is more-or-less bound to the limit of seven (plus or minus two) things (Miller, 1956) is obviously of interest in an educational setting. Equally the acquisition of implicit learning requires large numbers of instances with rapid feedback about which category the instance fits into (Seger, 1994); wittingly or unwittingly, we can produce an environment, especially in software, that affects how students learn. In an analogous way, the work of Mihaly Csikszentmihalyi on the notion of "flow", the necessary conditions for human beings to be fully engaged in an activity (Csikszentmihalyi, 1991), informs the milieu of electronic game design, without actually providing a one-to-one relationship with specific elements.

Theories

There are many educational theories, from a variety of sources. For example, the work of Jerome Bruner (which builds upon the structured stages of cognitive development outlined by Jean Piaget) and emphasises the relationship of cognitive structure to the structure of disciplinary content "What are the implications of emphasizing he structure of a subject, be it mathematics or history—emphasizing it in a way that seeks to give a student as quickly as possible a sense of the fundamental ideas of a discipline?" (Bruner, 1960). Bruner is often assoicated with Lev Vygotsky.

Vygotsky's ideas centered on the notion that knowledge and learning are culturally and societally constructed. In particular, his idea of the "zone of proximal development" (ZPD) has been enormously influential. ZPD states that students have limitations in the amount of progress they can make from their starting point, but, with the help of a teacher giving appropriate interventions and scaffolding, their understanding can expand further than it would if left to alone. "…the distance between the actual level of development as determined by independent problem solving [without guided instruction] and the level of potential development as determined by problem solving under adult guidance or in collaboration with more capable peers". (Vygotsky, 1962)

The general ideas represented by these (and other) theorists taken together lead to an understanding of what happens in teaching and learning that has been labelled "constructivist": encapsulating the idea that learners actively construct knowledge, rather than stand as passive recipients, as vessels to be filled. This can lead to direct classroom implications, entailing specific practices such as "reciprocal teaching" (Brown & Palincsar, 1989; Palincsar & Brown, 1984) and "jigsaw instruction" whereby students learn through constructing their knowledge in order that they can teach others; or to more general approaches, such as outlined by Moti Ben-Ari (Ben-Ari, 2001).

Another collection of approaches, known broadly as "behaviourist", grow from the work of B.F.Skinner (Skinner, 1938) and focus only on objectively observable behaviours, therefore discounting internal mental activities. In a behaviourist environment, learning is considered to be the acquisition of new behaviour, and "conditioning" for teacher-approval, high marks, or other reward, is the process by which learning occurs. Although these find some use in the classroom, these ideas are more often seen incorporated into various on-line pedagogic environments (Skinner, 1968).

For CS education, the work of Seymour Papert, co-founder of the MIT Media Laboratory and collaborator of Jean Piaget, is influential (Papert, 2003). His theoretical approach holds that things that are readily available in the everyday environment provide relevance and concrete experience on which learning is constructed. In order to develop mathematical reasoning, therefore, it is important to provide relevant stimuli in the environment, together with language for discussion of the resultant concepts. The tradition of instructional approach which has developed from this is focussed around the use of LEGO in the classroom, although there are other "constructionist" approaches which do not rely on proprietary manipulables.

More recently, the work of Jean Lave and Etienne Wenger (Lave & Wenger, 1991; Wenger, 1998) has extended the notion of the social nature of learning with ideas that learning is always situated within authentic situations, and takes place within communities of practice. With this is associated larger notions of the how communities are structured and how learning occurs in them, how learning and membership of community are closely identified with each other and how knowledge cannot be separated from practice. As a discipline that has a clear set of vocational communities and constituencies this is an intriguing theory for CS, which has begun to be thoughtfully explored within the CS classroom (Kolikant, in press).

Models and taxonomies

As well as broadly-conceived theories which influence at the most general level, there are also more narrowly-drawn concepts, more precisely targeted at areas or types of education:

Bloom (et al)'s Taxonomy first codified in *Taxonomy of educational objectives, handbook 1: The cognitive domain* described a range of cognitive behaviours found in educational assessment. The taxonomy comprises six levels, arranged along a continuum of complexity, vis: knowledge, comprehension, application, analysis, synthesis, evaluation. It was devised in order to describe educational objectives that went beyond mere recall of fact and was aimed at "teachers, administrators, professional specialists and research workers" and was "especially intended to help them discuss these problems [of educational objectives] with greater precision" (Bloom, 1956). It has been abidingly influential, and has a recent renaissance in CS education research (Lister & Leaney, 2003).

The use of Kolb's Learning Cycle (Kolb, 1984) on the other hand, is to structure instruction so that experience is seen as the source of learning and development

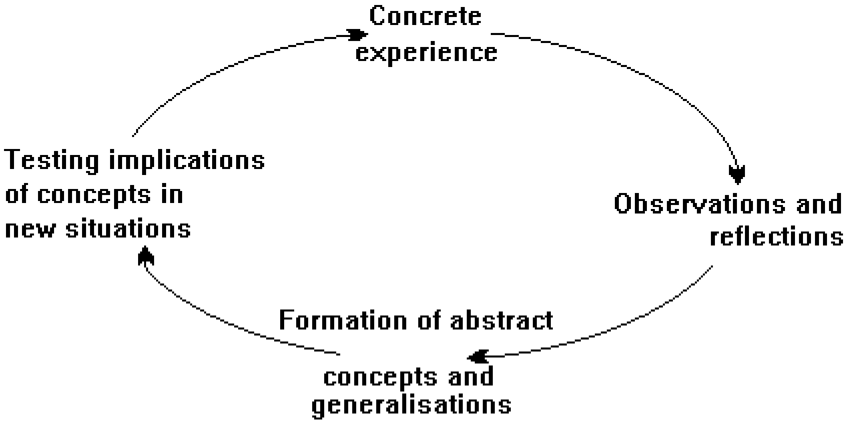

Concrete experience
Observations and reflections
Formation of abstract concepts and generalisations
Testing implications of concepts in new situations

Figure 3: The four stages of Kolb's Learning Cycle

In contrast to these cognitive models, William Perry relates learning to a model of intellectual development maturity. (Perry, 1981; Perry & Harvard University. Bureau of Study Counsel, 1970) Perry claims that college students (and others, too) "journey" through nine "positions" with respect to intellectual and moral development. These stages can be characterized in of terms the student's attitude towards knowledge. There are nine levels, grouped into four stages: dualism, multiplicity, relativism, commitment[7].

Table 4: Perry's Model of Intellectual and Ethical Development

Stage	Characterisation	Students' View
A. Dualism/Received Knowledge:	There are right/wrong answers, engraved on Golden Tablets in the sky, known to Authorities	The students' task is to learn the Right Solutions learn the Right Solutions and ignore the others

B. Multiplicity/Subjective Knowledge:	There are conflicting answers; therefore, must trust one's "inner voice", not external Authority	The students' task to learn how to find the Right Solutions
C. Relativism/Procedural Knowledge:	There are disciplinary reasoning methods	The students' task is to learn to evaluate solutions
D. Commitment Constructed Knowledge:	Integration of knowledge learned from others with personal experience and reflection	The student explores issues of responsibility The student realizes commitment is an ongoing, unfolding, evolving activity

The journey is sometimes repeated; and one can be at different stages at the same time with respect to different subjects.

The aspect of learning that they have chosen, what they have chosen to emphasize in the creation of their model, what to simplify, and what to discard all contribute to the differences between these constructs, but Bloom, Kolb and Perry have all devised *models*.

Models have also been developed with regard to situated classroom instruction. One of the most enduring in recent times has been the model of *Problem Based Learning*, first instantiated at McMaster University Medical School as a method of helping students learn skills of medical diagnosis (which were poorly served by lectures and other, formal, classes) problem-based learning has spread.

.

Instruments

Finally, at the lowest level of granularity, there are specific instruments devised to expose particular aspects of the instructional situation. Mostly these are manifest as questionnaires or scales which describe students. For example there are inventories on learning styles (Kolb, 1984), on approaches to studying (Entwistle & Tait, 1995) and on personality types (I. B. Myers, 1998, 2000). Their use within the classroom is predicated by the instructor being interested in a very specific idea, for example, in adjusting materials to groups of student with differing learning styles, or for assuring well-formed groups for project work. The step that follows—of evaluating whether there is a difference between the groups distinguished by the instruments—is the step towards their use in CS education research.

Summary

We believe that the way in which CS education *links research to relevant theory* is via trading zones and reference disciplines. It would be impossible to be comprehensive about sources of theory; indeed, it is difficult to be comprehensive about material within a single source. Our limited survey above indicates of some of the ways in which CS education as a research area overlaps with, draws upon and "trades" with education.

The theories, methods, instruments and measures located within our reference disciplines have different traditions and uses. How they are represented and utilised within our disciplinary domain, a necessarily interdisciplinary context, is a source of both problem and opportunity for CS education research.

Provide a coherent and explicit chain of reasoning

It is not enough that we feel confident in our work; we have to be able to explain it to others—in the classroom probably, in the laboratory possibly, in print certainly. "Science" is about sharing, and if the people you want to share your material with cannot follow the progression of your argument, cannot understand the reason that you chose the question in the first place … cannot understand the reason you think your choice of method is going to provide compelling evidence (even *sufficient* evidence) … cannot understand the relationship you claim for your intervention and the cited theoretical tradition … then this not just bad form, it is bad science.

So we need to "provide a coherent and explicit chain of reasoning". *Coherent* and *explicit* are straightforward and understandable terms.

Our accounts must be coherent because incoherent accounts are difficult to follow and inherently suspect.

Our accounts must be explicit for two reasons. Firstly, because we seek to be rigorous and precise. If we do not describe—precisely—what we did, and provide—precisely—the background information given, and detail the assumptions we built on, then our work cannot be judged fairly. It may not be bad work, but it will be impossible to tell. Our accounts must be explicit, too, because we aim to make them comprehensive. What is un-stated will be unclear, and so will be susceptible to many different interpretations.

However, *chain of reasoning* has different implications when used in different contexts, and if we ask what a chain of reasoning *is*, the reality of working in a trading zone may cause problems. This is graphically illustrated by Marian Masterman's comment on Thomas Kuhn's *The Structure of Scientific Revolutions*:

> Insofar, therefore, as his material is recognizable and familiar to actual scientists, they find his thinking about it easy to understand. Insofar as this same material is strange and unfamiliar to philosophers of science, they find any thinking that is based on it opaque. (Masterman, 1970)

Whatever view you take on Kuhn's book, the chain of reasoning is the same in both cases; yet it demonstrably causes problems for an unfamiliar audience.

However, if we turn our heads slightly and ask what a chain of reasoning *does*, the meaning becomes clearer. A *chain of reasoning* is the strong intellectual filament upon which we can string the pearls of our work. In that sense, it becomes possible to determine some fairly standard constituent components.

Relationship to theory.

Theory can, in broad terms, be used in two ways. We can situate our work in exploration of a theory-base (or theoretical position), or we can use a theoretical perspective to inform our investigations. There is a difference between seeing the world/classroom from a certain perspective and the world/classroom being designed according to particular perspective. For example, if our work was an exploration of theory, we could imagine asking questions (and seeking associated evidence) along the lines of "Are CS educators behaviorists?" (or constructivists, or whatever), or perhaps "Are their practices behaviorist—with or without their intention?" If, however, we were conducting work that was informed by these same theories then we would be using them as a conceptual lens and, in looking from a behaviorist perspective, we might expect to illuminate certain qualities in any teachers' practice.

The first step in our chain must be to articulate the relationship that our work has to any theory, or theoretical assumptions, and how our work uses or interprets those.

The "chain" in chain of reasoning

In talking about research studies, there are several components, at different levels of granularity that have to be considered. (See figure four)

The chain starts, as explored above, with theory. However, theories don't stand alone and are related to disciplines (as well as each other). Theory and discipline taken together can inform the research questions we ask. Often, situation within a disciplinary context brings with it an associated methodology, which prescribes the selection, combination and sequencing of the methods and techniques we employ. So, asking questions within an economic context will force the use of a different methodology than asking them within a psychological or sociological context.

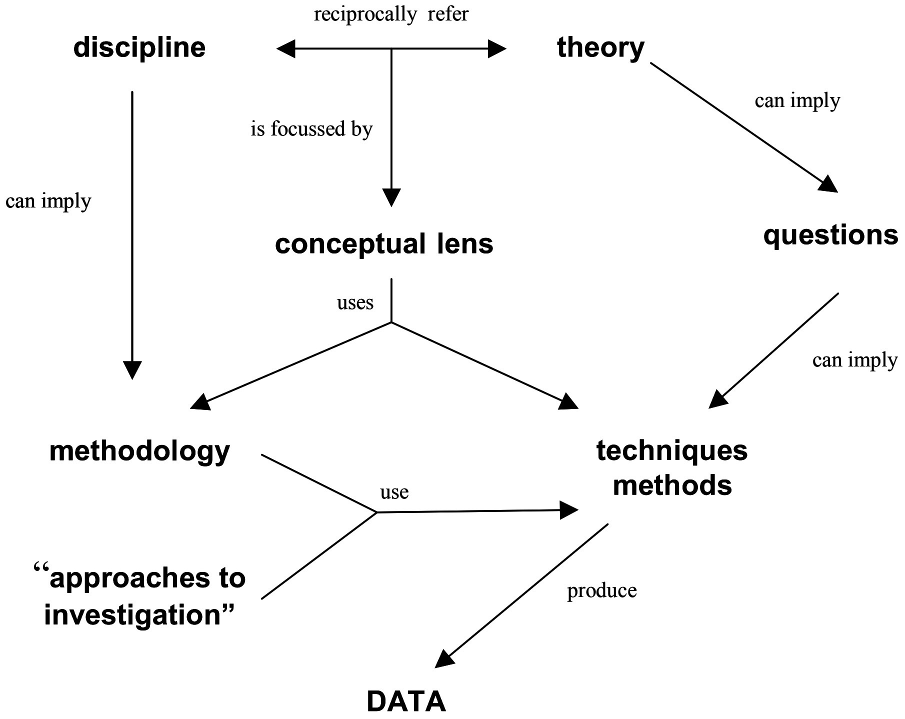


Figure 4: Some components (the labels) and influences (the labelled arrows) to consider in a chain of reasoning.

A methodology is not the same as any specific method or technique. For example, "diary studies" or "questionnaires", both useful methods, may be included in several distinctly different methodologies.

There are also some specific aggregations of methods that might be termed "approaches to investigation". Phenomenography (Research, 1995), activity theory (Engestrom, Miettienen, & Punamaki, 1999; Nardi, 1996) and grounded theory (Glaser & Strauss, 1967) are all examples of such approaches. These are not tied to a particular discipline. They are characterized by a stance which focuses the nature of inquiry (e.g., phenomenographic investigation seeks patterns across individuals via in-depth qualitative techniques), a selection of methods oriented to the stance, and often a descriptive language or framework. They are "approaches to investigation" because they are a package; they cannot be used in a "pick and mix" way, taking just one part of them. Use of such an approach constrains both data gathering and analysis.

So, what the "chain" in "chain of reasoning" *does* is to detail and describe each of the possible points of situation, implication and dependency where choices, or assumption, have been made. The end result of all these implications, choices and dependencies is data. A strong chain of reasoning conveys many benefits onto our resultant data: assurances of validity, replicability and representation for example, as well as higher confidence that the study has not been biased.

Even the most rigorous empirical methodology is no substitute for careful development of the reasoning that underlies the hypothesis (Schkade, 1989)

Use methods that permit direct investigation of the question

The first key principle is *pose significant questions that can be answered empirically*. In our pragmatic approach, the question and evidence requirement taken together inform the choice of research method or technique (the terms are, for pragmatic purposes, interchangeable). The technique must deliver data of a sort that can answer the evidence requirement—that is fit for the purpose. This means that the technique must generate data (and hence evidence) that is *sufficiently rich*, that it must provide enough information to address the research question *at the right resolution*, and that it must be *feasible* within available resources.

Richness of Data

One of the factors that characterizes an empirical technique is the richness of the data it is likely to produce. *Richness* here is used to mean the number and generality of questions that the researcher can attempt to answer using that data. For example, examination scores offer single-point data that can only be used to answer limited questions, whereas the examination questions and student answers taken together would offer much richer data, and the questions and answers taken in conjunction with interviews with the students about why they answered as they did, would be richer still.

Richness of data potentially translates to richness of evidence, depending on the method of analysis and on the interpretation based on those analyses—on inferences made from the analyzed data. Alternatively, the data collection technique and method of analysis may reduce the richness of the data, by aggregation or selection. For example, an interview, a method which yields inherently rich data, can be combined with a coding scheme which records selected data, such as categories of activity against time. Rich data may also be constrained by aggregation, for example averaging across subjects or sub-populations.

Level of Resolution

The nature and focus of the research question implies an equivalent level of detail and specificity in the data. *Fitness for purpose* requires a match between granularity of the research question and the level of resolution of the data collection, and hence requires a method that can generate data of appropriate resolution. For example, studies of individual behavior give insufficient insight into social interactions, and surveys of group process give little insight into individual cognition.

Mantei (1989) distinguishes five levels of resolution of question and data, as summarized in the 'Levels of Resolution' table .

Table 5: Levels of resolution

Level	Research question focus	Resolution of data
micro-micro	questions about internal structures and processes of the human mind (e.g., memory, cognitive load)	performance measures reflecting internal cognitive mechanisms, (e.g., response times in microseconds)
micro	questions that focus on the individual's interaction with the external environment (e.g., how individuals use tools to solve problems)	specific data about individual behavior, such as sequences of decision making and problem solving (e.g., , how individuals perform given tasks)
standard	questions about regularities associating individual characteristics with individual behavior (e.g., effects of personality differences on productivity)	micro-level data, as well as data about attitude, style, and preference, aggregated for a given individual at the time of data collection, individual data aggregated over a group of individuals (e.g., averages of individuals' performance)
macro	questions about group properties, behaviors, and processes (e.g., group creativity, leadership, cohesiveness)	group-generated data on group behavior
macro-macro	questions about systems, networks, and organizational behavior (e.g., impact of computer-supported meetings on organizational communication patterns)	data aggregated over a group of people who do not interact with each other during the data collection; aggregation of responses from multiple subgroups and individuals

Analysis methods must match the data resolution in order to provide meaningful evidence at a level appropriate to the research question. A mismatch between analysis and question can produce findings of dubious utility; the findings may have some meaning, but they are unlikely to address the question effectively. Mantei [ref] gives an example: "Cognitive style measures in IS have been criticized because they serve as the wrong level of data collection for answering questions about the design of an information system's interface, which affects behavior at the micro level."

Costs

A number of factors characterize a technique. Among the most important are:

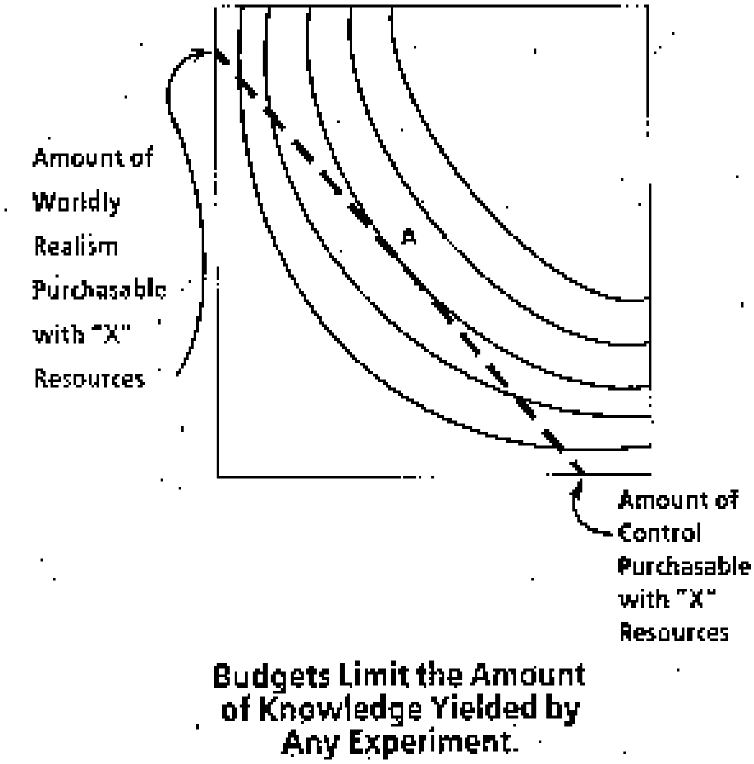
- location (in situ, in lab)
- what data (and richness of data)
- how much data
- level of resolution of data collected
- number of subjects
- representativeness of subjects (which has implications for generalizeability of results)
- whether the research strategy is theory-driven or data-driven
- basis of analysis: descriptive, statistical, etc.
- precision of question (maturity of investigation)
- constraint of conditions / number of variables

Unfortunately, data doesn't come "for free". Any data implies data collection and analysis costs. The more and richer the data, the higher the cost. Costs arise at all stages of an empirical study: from design and planning through implementation and execution to analysis. Control, validation of instruments, and pilot studies all contribute to planning cost, and all influence the quality of the data collected. The preparation of study materials, gaining access to subjects and settings, and conducting the study, all contribute to implementation and execution costs. All resources incur costs: time, subjects, settings, researchers, instruments, equipment, expertise, etc. Cost implies constraint: few researchers have unlimited resources, and hence budgets constrain research designs.

Tradeoffs

Richness and resolution trade off with cost. Mason summarizes: "Ultimately, the total resources available to conduct an experiment delimit the amount of knowledge that can be obtained from it." (Mason, 1989). Hence, within a given level of resource individual factors in the study design trade off against each other. For example, the cost of data collection and analysis in a case study may trade off against the number of cases that can be considered. Increasing the precision of the research question can afford an increase in sample size.

A balance of tradeoffs in study design is assessed in terms of fitness for purpose. If the evidence is required to be a statistically significant correlation, then the sample size must be sufficiently high, and the number of variables under consideration may be constrained. If the evidence is required to be a contextualized account of a problem-solving process, then data collection and analysis costs may be unavoidable but fewer subjects may be sufficient. The aim must be to amass evidence of


Amount of Worldly Realism Purchasable with "X" Resources
A
Amount of Control Purchasable with "X" Resources
Budgets Limit the Amount of Knowledge Yielded by Any Experiment.

sufficient utility within the constraints of cost—and the aspiration must be to amass evidence of the highest possible utility within the budget

Figure 5: Mason's (Mason, 1989) depiction of how studies are constrained by costs. Mason uses iso-episteme curves to illustrate how realism (y axis) trades off with control (x axis) and suggests that budgets (the dotted line) limit the amount of realism or control that can be achieved. He suggests that studies that lie at point "A" on the curve will be most cost-efficient. However, "fitness for purpose" may demand a different strategy.

Methods / Techniques

The purpose of this section is simply to indicate the range of research methods or techniques available, not to provide a complete catalogue nor to endorse any particular approach. A discipline-based methodology might specify techniques. In CS education, in the absence of such a methodology, the choice of technique is influenced by the research question. Whether the research is theory-driven (testing hypotheses derived from theory which predicts the outcome) or inductive (seeking emergent patterns in the data) can shape the techniques, but the techniques themselves do not pre-suppose a given approach or methodology.

Case studies

Case studies are in-depth, descriptive examinations, usually of a small number of cases or examples. They provide an intensive, holistic description of a single phenomenon, investigated in situ. Case studies usually encompass a variety of data collection techniques, potentially ranging from ethnographic and participant observer methods, through artifact analysis, through interviews, to constrained tasks. Case studies are appropriate especially when the boundaries between the phenomenon and the context are not clear, when the objective is to tease out as many factors contributing to some phenomenon as possible. Because of the numbers and sensitivity to context, there are limitations to generalization of findings. Analysis tends to be inductive reasoning based on multiple data sources. Case studies are demanding and intensive.

Table 6: Case study tradeoffs

Good for:	Case studies are appropriate especially when the boundaries between the phenomenon and the context are not clear, when the objective is to tease out as many factors contributing to some phenomenon as possible.
Bad for:	Generalization, given small numbers and sensitivity to context.
Kind of evidence:	Very rich and contextualized. Group or individual. Possibly historical. Small numbers.
Cost of planning:	Can be low, depending on how focused the study is.
Cost of Data collection:	Tends to be high, because it they involve in-depth interviews and observation, although some types of data (such as existing records and artifacts) can be low-cost.
Cost of analysis:	High. A good case study requires considered analysis, and integration of evidence from multiple data sources. There is often no pre-specified protocol.

Diary studies

Diary studies rely wholly on self-report; individuals are asked to keep recorded accounts of their behavior over time. Diaries may be structured, to focus reports on key issues and to facilitate comparisons. Diary studies afford a glimpse into subjects' introspection over time. Introspective accounts can provide considerable insight, but they suffer a number of limitations. For example: individuals vary enormously in their value as diarists; reflecting on behavior can influence subsequent behavior, changing the very phenomena under observation; diaries are selective, usually retrospective and rationalized. Hence, diary studies are usually combined with other data sources.

Table 7: Diary study tradeoffs

Good for:	Insights into individual experience, perceptions, and beliefs. Afford good potential for longitudinal views.
Bad for:	Dependent on individual skill, and often selective, retrospective and rationalised accounts. Hence, they present a high chance of distortion. Can be intrusive on natural behaviour.
Kind of evidence:	Rich, individual, often longitudinal accounts. Numbers usually small.
Cost of planning:	Low to medium. Planning cost depends on whether the diary is structured.
Cost of Data collection:	Low. Collection is via self-report, so the cost derives mainly from whatever incentives or spurs are provided to keep the diarists active.

Cost of analysis:	Very high, as with any analysis of rich qualitative data.

Constrained tasks, quasi-experiments, and field experiments

Constrained tasks are intended to bridge between observation and laboratory experiment, providing some constraints on subjects' activities (and hence some basis for comparison), while maintaining the richness of context. The level of constraint varies: typically, specified tasks are carried out in situ (hence the term 'field experiments'). The constraint is usually on the task, which is chosen to represent some aspect of natural activity, in order to investigate some phenomenon of interest – usually one identified in previous observation. Sometimes the constraint is on the environment, in order to tease out factors contributing to behaviors or processes of interest. The constraint may be stronger, as well, so that the study draws on experimental techniques, and seeks quantitative data, but without full experimental control (hence the term 'quasi-experiments'). For example: a classroom-based study styled on experimental comparison but undertaken in situ. One of the limitations of quasi-experiments is that numbers may be insufficient to exclude environmental factors through statistical analysis. Another variant is to try to reproduce a typical environment in a laboratory, for example providing a representative designer's library and typical range of tools and media, and specifying a design task. Constrained tasks offer some increased control over observation and hence provide some basis for comparisons and for validation of observations – while maintaining some realism. However, they do not have the power or precision of laboratory experiments, and generalization is limited.

Table 8: Constrained task tradeoffs

Good for:	Bridging between observation and experiment, for example in order to investigate an observed phenomenon in more depth and with more control, without stripping away context. Helpful in focussing in on key factors and their inter-relationships. Can provide a basis for comparison among subjects.
Bad for:	Generalisation is limited, due to limited control and preservation of context. Often numbers are small.
Kind of evidence:	Regularities associated with the particular task, whose identification may be supported with statistical analysis.
Cost of planning:	Medium to high, approaching the planning cost associated with controlled experimentation. Care is required for the selection of the task and subjects, the constraint of the environment, and the specification of the protocol.
Cost of Data collection:	Variable, depending on what data is collected. Constrained tasks may use quantitative measures (which have a relatively low collection cost) or rely on qualitative data (which can entail higher collection cost).
Cost of analysis:	Variable. Can be low for quantitative measures, high for inductive analysis of rich qualitative data, or medium for a focused, mixed analysis.

Document studies

Existing records, logs of electronic communication, individual notes and diaries, sketches and diagrams—various written or recorded artifacts provide a naturally-occurring source of information, often closely allied with a phenomenon of interest, and potentially offering insights into processes, interactions, organizational character and culture, and individual experience that may be hard to capture otherwise. They are steeped in the context (environment and language) in which they are produced, and they can provide a longitudinal view. They can provide an unobtrusive form of data collection. The utility of such 'documents' depends on their completeness, authenticity, accuracy, and representativeness. Documents may well be colored by the purpose for which they were originally produced; this can be an advantage, or a limitation, depending on the focus of the study. The analysis of a 'corpus' or collection of documents is time-consuming and demanding, and it may be difficult to assemble an appropriate, representative corpus.

Table 9: Document study tradeoffs

Good for:	Unobtrusive, longitudinal, context-sensitive views of phenomena of interest.
Bad for:	Limited by what documents are available, with implications for completeness, accuracy, and representativeness. There is often no access to interpretation by the originators of the documents, and hence gaps may be hard to fill.
Kind of evidence:	Very rich, contextually steeped, qualitative material often affording a longitudinal view and multiple perspectives.
Cost of planning:	The planning cost is associated with planning the analysis and is dependent on the nature and variety of documents under consideration.
Cost of Data collection:	Can be very low, because corpora may be available for 'harvest'. However, the challenge lies in acquiring an appropriate, representative corpus for the purpose.
Cost of analysis:	Very high, although it can be ameliorated by pertinent automated tools.

Automated logging

Any computer-mediated activities can be recorded automatically, for example as activity logs, streams of keystrokes, or sequences of electronic communication. Such logs can be comprehensive (for the data they collect), precise, and accurate. They can include precise timing information. Electronic logs in appropriate formats are readily amenable to automated analysis of many varieties, from performance measures to linguistic profiles. From automated logs, it can be possible to reconstruct with detail and accuracy the conduct of a task by many subjects, for example to analyze sequences of actions, associations between actions and errors or actions and outcomes, and time spent on different task components. Automated logs facilitate performance comparisons between subjects or activities in terms of speed, accuracy, and outcomes. Their disadvantage is that, although they record precisely what people do while they interact with the system, they offer no direct information about what they intended, or where they looked, or what they did when they weren't interacting with the system. Although they can record the electronic context well, they record only the electronic context, potentially omitting factors important in the phenomena of interest. However, automated logging combines well with interview techniques which give insight into intentions and personal experience.

Table 10: Automated logging study tradeoffs

Good for:	Unobtrusive, accurate capture of electronic communication and interactive behaviours, including precise timing of actions.
Bad for:	Relating behaviour to intention.
Kind of evidence:	Can be qualitative or quantitative accounts of behaviour, communication and interaction, usually supported with statistical analysis.
Cost of planning:	High. Logging easily produces a flood of precisely detailed data; planning the collection and analysis of that stream requires careful reasoning about how to interpret research questions and how to filter and manipulate that data relevantly.
Cost of Data collection:	The collection cost is associated with the creation of logging tools, and subsequently with the cost of data storage. Once the tools are in place, the collection cost is minimal.
Cost of analysis:	Can be low, depending on the question and the available tools. Can be high, depending on the level of human intervention and interpretation, and on the need for new or customized tools.

Observation

Observation is an extremely broad category of investigation, ranging from intensive ethnographic methods through targeted, short-term approaches. The common theme is the watching—and recording—of behavior in context, usually in a natural situation and environment. Observation can produce data that is descriptive (e.g., a record of behavior, possibly within a descriptive scheme), inferential (considering the intentions behind observed behavior), or evaluative (assessing or measuring characteristics of observed behavior). Hence, two key aspects that distinguish different observation approaches are the level of participation of the observer, and the nature of the records kept (and whether the records preserve the richness of the setting or focus on selected phenomena).

Records can take many forms, from ethnographic field notes, through verbatim contemporaneous notes, to audio and video recording. The data can be descriptive or quantitative. The obligation is for records to be complete and accurate. Although some records (such as field notes) might be made immediately after-the-fact, they must ultimately stand on their own, and hence must provide a sufficient record without reliance on additions from memory.

Observation can produce very rich, highly situated data reflecting behavior in context. It can provide opportunities to identify important factors which were previously un-remarked. It can capture complex interactions in a rich social, physical, and activity environment. However, it is demanding both in terms of data collection and in terms of analysis. Selection by the observer (or by observation protocol or the recording scheme) or expectations the observer brings to the setting may color the data, and the mere presence of the observer may influence the behaviors observed.

What follows is an indication of the variety of observation strategies that might be adopted.

Table 11: Some observation strategies

Participant observation:

The observer participates in the respondent's natural activities (e.g., becoming a member of a design team) for first-hand experience, in order to become integrated in the social interaction and immersed in the culture. In effect, the observer becomes a collaborator or an apprentice of the informant. Insight may arise from shared or common activities. The impact of the participation may 'cut both ways': on one hand, it may distort the activity and interaction; on the other, it may reduce bias by making the interaction with the informant more naturally a part of the task.

Ethnographic observation:

The aim is to understand the activities within the informant's frame of reference. Questions concern the correct identification of behavior. Typically, the observer comes prepared with a theoretic framework for describing what happens (for example, concepts of kinship and ritual).

Unobtrusive observation:

The aim is to observe (and possibly question) with as little impact as possible on the informant's activity. The observations gathered tend to be descriptive, unless observation is paired subsequently with interview.

Structured observation:

The 'unobtrusive' observer codes the informant's behaviour in terms of pre-defined categories or scales.

Systematic observation:

In this quantitative approach, observations are captured in terms of existing schema, for example, behaviour might be coded behaviour in terms of a set of categories, or rated on a scale.

Observation of constrained tasks:

The aim is usually to control what the informant does, and possibly constrain the environment in which it is done, and thereby set a task which will expose some interesting or obscure part of the informant's behaviour, or provide a basis for comparison between informants.

Observation of tasks with concurrent verbalization:

The informant, having been instructed in 'thinking-aloud' or articulating normally silent processes, is asked to verbalize while performing some task.

Observation of working in pairs:

One technique for drawing verbalizations (including explanations or articulations of reasoning or other internal processes) from the informant is to ask informants to work in pairs, so that communication about the performance of the task is inherent in the task. One assumption is that the two informants will share the same frame of reference.

Fundamentally, the quality of observation depends on the quality of the observer. Observation requires skill: in attending, in filtering, and in recording. Sometimes it also requires domain knowledge, in order to comprehend what is being observed.

Whether observation is open and descriptive, or guided by a theoretical framework or by an observation protocol (a script which identifies which information is to be gathered and what criteria are to be applied), the observer needs training and experience in order to gather data consistently and accurately.

Table 12: Observation tradeoffs

Good for:	In-depth views of real phenomena as it occurs naturally. It can provide opportunities to identify important factors which were previously unremarked. It can capture complex interactions in a rich social, physical, and activity environment.
Bad for:	Selection by the observer (or by observation protocol or the recording scheme) or expectations the observer brings to the setting may color the data, and the mere presence of the observer may influence the behaviors observed. Limited generalisation.
Kind of evidence:	Rich data reflecting behavior in context.
Cost of planning:	Low. Planning cost is associated mainly with the preparation of any observation protocol, if one is used.
Cost of Data collection:	Very high. Data is collected only while the researcher is present; there are no shortcuts.
Cost of analysis:	Very high, although some approaches employing coding schemes or ratings reduce analysis costs significantly, at the expense of richness.

Interview

Interviews are guided dialogues, valuable in eliciting subjects' experiences, perceptions, opinions, attitudes, intentions, and beliefs. They allow subjects to respond in their own words, to explain behaviors in terms of their own values, goals, and expectations, to assign their own meanings, and to provide clarification. Interviews can elicit affective responses as well as cognitive processes. They can range from open-ended, in-depth probing of key topics to highly structured 'oral questionnaires' which emphasize uniformity of the interview 'script' for all subjects. Interviews are normally conducted face-to-face, but telephone and even electronic interviews can provide useful data.

The strength and weakness of interviews resides in the interaction between the interviewer and the respondent. The interview can be influenced by the quality of the rapport between the two, by the compatibility of their frames of reference, and by the skill and knowledge of the interviewer. The potential for bias or distortion is high. The questions themselves may influence responses, depending on phrasing, on individual interpretation, and on associations they may trigger. Subjects may try to please the interviewer, or to anticipate the 'correct' or desired response. The quality of an interview is influenced by the subject's ability as a self-reporter: on recall, selection, and accuracy.

Yet interviews also have high potential to provide insight into people's thinking and feeling. They can be combined with other techniques in order to compare what a respondent reports in interview to what the respondent does in practice, and hence to corroborate the accuracy of reports and provide insight into behavior, motivation, and perception.

What follows is an indication of the variety of interview strategies that might be adopted.

Table 13: Some interview strategies

Ethno-methodological interviews:

Interviews which take an ethnographic stance: the interviewer is "as a Martian", arriving (notionally) without preconceptions and seeking to elicit the respondent's meanings. There is no assumption of a shared frame of reference; the point is to elicit the respondent's frame of reference. The interviewer is non-directive; the interview is largely directed by the respondent, who maps out the topic. Probes are used to verify the interviewer's understanding.

Ethnographic interviews:

Ethnographic interviews are concerned with eliciting the respondent's frame of reference. The interviews are open; the respondent maps out the topic, and probes verify interviewer's understanding. However, the interviewer brings to bear theories of society and interaction, and may therefore structure understanding in terms of the framework provided by the theory.

Semi-structured interviews:

The overall structure of semi-structured interviews is planned by the interviewer in advance, with a script of main questions. The order of questions may be altered to adapt to the subject's responses; the respondent is given considerable freedom of expression, but the interviewer controls the interview to ensure coverage. Prompts (open questions encouraging breadth) and probes (focussed questions which seek to clarify or specify, to explore depth) fill in the structure.

Structured interviews:

Structured interviews are organised according to a fixed script of carefully-phrased questions. The order of questions is fixed, and follow-up questions are minimised. The script ensures coverage and comparability across multiple interviews with different respondents.

Oral questionnaires:

Oral questionnaires are formal, highly-structured interviews, largely comprised of closed or focussed questions presented in a fixed order. There is no additional or follow-up questioning, no deviation from the question script.

Group interviews or group elicitations:

Group interviews add social context to the interview, allowing group dynamics to play a role in eliciting data through interactions within the group. Such interviews are usually semi-structured, with open discussion questions or group tasks. Group dynamics cut both ways: they can draw out differing perspectives and challenge individual thinking, but they can also exert peer pressure that inhibits or distorts individual response.

Focus groups:

Focus groups target specific sub-groups, examining their responses to products, processes, arguments, etc. They are typically used in market research, where the 'focus' is on different kinds of customers. There is typically also a 'focus' on particular topics or objectives. Focus groups involve semi-structured group interviews, and they use group interaction explicitly to generate data. Participants make individual responses, but they hear and can react to others' responses as well. The interviewer acts as a moderator who keeps the discussion focussed and ensures that all voices are heard.

Table 14: Interview tradeoffs

Good for:	Eliciting subjects' experiences, perceptions, opinions, attitudes, intentions, and beliefs. Permit in-depth probing and elicitation of detail. Powerful when it is important to understand the interaction between attitudes and behaviours.
Bad for:	The interview can be influenced by the skill and knowledge of the interviewer, as well as by the recall, perception, and self-reporting ability of the respondent. The potential for bias or distortion is high.
Kind of evidence:	Rich data reflecting what people think and feel.
Cost of planning:	Low to medium. Planning cost is associated mainly with the interview script and the analysis.
Cost of Data collection:	Collection costs increase with the number of interviews. Skilled interviewers are required.
Cost of analysis:	Very high for open interviews, given the volume of qualitative information. Can be low for highly structured interviews, which limit the richness of the data.

Survey research and questionnaires

"Survey research involves gathering information for scientific purposes from a sample of a population using standardized instruments or protocols. Ultimately, the purpose of survey research is to generalize from the sample to the population about some substantive issue." [Kraemer, 1991]

Kraemer identifies three characteristics of survey research:

- It is a quantitative method requiring standardized information designed to produce quantitative descriptions of some aspects of a study population.
- The principal means of collecting data is by asking structured, pre-defined questions.
- Data is collected from or about a sample of the study population, but is collected in such a way as to support generalization to the whole population.

Questionnaires (or surveys), then, are a method of data collection within "survey research", as are structured interviews. They have the potential to generate large volumes of data at relatively low collection cost. Surveys can be descriptive (fact-finding, enumerating, characterizing a population) or analytic (seeking associations or causal relationships among variables).

Questionnaires typically rely on self-report: subjects' own responses to questions about their own behavior, attitudes, perceptions, etc. Questions may be 'open' (offering a wide scope in answering) or 'closed' (requiring constrained answers within a specific formulation, e.g., placement on a scale, selection from a list, yes/no). Fact-finding surveys, such as background questionnaires, may use open questions and qualitative analysis. Survey research relies on closed questions and statistical analysis.

The success and utility of survey research hinges on the validity of the questionnaires and other survey instruments: that they do measure or capture what they intend to, and that what they measure or capture represents the construct under consideration. High-quality survey research makes a substantial planning investment, working carefully on research designs (strategy, constructs and operationalization), validating questionnaires and other survey instruments through pilot studies, and designing the sample. Reliability is increased through the use of sets of questions, which minimize the impact of wording.

The utility of survey research is enhanced by combination with other methods, such as observation and interview, which provide depth and additional perspectives.

Table 15: Survey research tradeoffs

Good for:	Obtaining consistent profiles of the characteristics of a population in terms of the constructs under scrutiny. Allows systematic, generalizeable investigation of associations among variables in a social context, when controlled laboratory experiments are not feasible. Can encompass affective, social, and cognitive factors.
Bad for:	
Kind of evidence:	Quantitative measures and statistical analysis.
Cost of planning:	High. The success and utility of surveys relies on the design, questionnaire preparation, and pilot testing.
Cost of Data collection:	Medium, depending on the extent of the survey.
Cost of analysis:	High, given the potential quantity of data..

Controlled experiments

Experiments are the systematic manipulation of variables under controlled conditions, in order to test hypotheses generated from theories. Hence experimentation is theory-driven, and is characterized by:

- a setting controlled by the researcher
- systematic selection of a representative sample of subjects, and assignment to treatment conditions
- the manipulation of one or more independent variables, in order to observe their effect on the dependent variables

For effective experimentation, the researcher requires control of control of variables: of the independent variables, and of all intervening variables that might affect the dependent variables. The internal validity of an experiment depends on the chain of inference between the hypothesis and the conclusion. An advantage of experimentation is that the high level of control should help reduce threats to validity and hence lend strength to the inferential chain. Another is that it facilitates accumulation of evidence. The external validity reflects how representative the setting and sample are of the target population, and hence the extent to which findings from the experiment can be generalized to other settings, and populations. A disadvantage of experimentation is that the control it exerts reduces the relevance of its findings by stripping away the correspondence between natural events and those in the laboratory. Hence utility of evidence may be limited. Further, there are some factors that cannot be manipulated. Crucial to the utility of experimental findings is the operationalization, the way a construct is 'made usable', in the form of phenomena that can be observed (and measured) in the world.

There are two classic models of human experimentation which address issues of human variability:

Between-subjects design: Different groups of subjects are assigned to the different treatments. Hence the comparison is between groups or between subjects.

The advantage is that subjects come fresh to the treatment; there is no learning or order effect. The disadvantage is the impact of individual differences, which may skew variability in the study.

Within-subjects or repeated measures design: The same subjects are used for all experimental treatments. hence the comparison is within the same group of subjects. Each subject is measured repeatedly, for each treatment, hence 'repeated measures'. The advantage is that individual differences are equalised across the conditions. The disadvantage is the potential for 'order effects' or 'learning effects' (variations in performance due to the order in which treatments are experienced)

Table 16: Experiment tradeoffs

Good for:	Control, statistical analysis.
Bad for:	Questions that aren't precise enough yet. Experiments involving human beings are problematic, because people are not fully controllable – it is impossible to eliminate all individual variability. Highly controlled experiments may not have sufficient richness for compelling generalisation to real-world settings.
Kind of evidence:	Quantitative data reflecting performance.
Cost of planning:	High.
Cost of Data collection:	Can be low, and is related to the number of subjects.
Cost of analysis:	Low to medium, depending on the breadth of the statistical analysis.

Any of the above techniques may be used in a variety of settings and in a variety of configurations:

Settings

Techniques can be applied *in situ* (in the normal or 'natural' environment) for example, studying professional programmers in their workplace, or studying students in the classroom; *under constraints* (in a natural environment on which selected limitations have been imposed) for example, studying students in their familiar classroom setting using a particular programming environment; or *in a laboratory* (a highly-controlled environment).

Configurations

Methods are not only used in a "one-off" manner, but can be employed (and re-employed) in various configurations:

A single technique may be applied in different ways to the same research question, refined through successive iterations. The refinement may be, for example, by way of sample size, or choice of task, or method of data recording.

A single technique may be applied at many levels: the scope of research can vary from investigation of a tool, to a course, or a whole curriculum.

A single technique may be employed in longitudinal fashion; for example, concept acquisition might be studied at intervals within an entry-level programming class. The cohort might also be re-visited a year later, perhaps after a comparative languages course.

Analysis

Like data collection techniques, analysis techniques have purposes which they suit, costs, and conditions for their application. Which analysis is chosen is shaped by the question to be addressed, and by the evidence sought. But it is also constrained by the data collected: how it is selected and how it is recorded. On the one hand, the nature of the data demands or excludes particular analysis. On the other, the nature of the analysis puts minimum requirements on the data.

Two analysis examples are discussed here, to highlight this inter-relation between question, evidence, data, and analysis. They are outlined in order to indicate both how the question may shape the analysis, and how the way the data is selected and recorded constrains the analyses which may be applied.

Protocol (transcript) analysis

Protocol analysis is a general term for the systematic analysis of transcribed speech from empirical studies (e.g., interviews, “think-aloud” monologues, discussions during activities by pairs or groups). Observation, case studies, interviews, open questions on questionnaires—all amass transcripts or written material which must be analyzed. Analysis can be approached in a variety of ways:

- quantitative (based on what can be quantified through counting or measurement),
- qualitative (based on identification of non-numeric patterns and on interpretation of meaning and usage, possibly pre-defined),
- theory-driven (drawing categories from theory, testing hypotheses derived from theory which predicts the outcome)
- data-driven (the data is examined for emergent patterns; such analyses do not —or can not—anticipate outcomes, but rely on finding what can be found in the data that is collected)
- comprehensive (seeking to characterize all of the collected data)
- vectored (having a particular focus, and seeking only specific phenomena within the data)

Hence, one thing that distinguishes approaches is the focus: what is of interest, and at what level of granularity.

The approaches are not mutually exclusive: they may be combined (subjecting one data set to different analyses) or may be used in sequence (with output of one analysis feeding into the next). For example, a data-driven analysis of interview transcripts may identify emergent categories. Those categories may be used as the basis for a coding scheme, and the data may be analysed afresh by applying that scheme. Alternatively, the data may be divided into sub-sets, with patterns emerging from an inductive analysis of one set tested through their application to another subset or to the whole data set. The findings of an analysis of one data set may be tested by applying that analysis to a different data set, e.g., data collected later or from a different subject sample.

Regardless of approach, the best analyses keep an "audit trail" between the primary data (the actual utterances) and the coded or analysed forms, so that contexts and sequences can be re-established or re-examined, as needed. It is wise to let the informants 'speak for themselves' and hence to maintain the links between excerpts and conclusions.

Because the analysis of qualitative data is a matter of judgment, the researcher must decide how an utterance or action is to be described. A number of techniques are employed to reduce the subjectivity of the process. For example, all of the data can be encoded independently by more than one researcher, resolving discrepancies through discussion and refinement of the coding scheme until an acceptable level of 'cross-coder consistency' is achieved. Coding can be done by researchers external to the project, so that they come 'fresh' to the analysis scheme. Alternatively, independent coders can 'calibrate' to each other through practice and negotiation, and then work on divisions of the data, subject to spot checks.

Particularly in data-driven analyses, it is advisable to review the entire corpus seeking counter-examples, gaps in the patterns, and other evidence that would suggest an alternative interpretation of the data. An important concept in such a 'counter-evidence review' is 'significant absence': the absence of a pattern or a phenomenon which one might reasonably expect to see.

What follows is an indication of the variety of analysis strategies that might be adopted.

Table 17: Some Analysis Strategies

Trawling:

The richest possible data is collected (and usually transcribed). Analysis (which can be qualitative or quantitative) is data-driven, seeking emergent patterns or organizing concepts. The aim is usually to determine what's important in some situation—possibly to find out what the important questions are, for subsequent investigation. One initial trawling technique is to "skim the cream": to mark important or compelling passages.

2-pass analysis:

Requires a reasonably large corpus of data. Data is subdivided, and one subset is analysed in order to identify emergent patterns, from which a formal analysis scheme is derived. The analysis scheme is then applied to the remaining data (and possibly to all of the data as well).

Pre-determined categories:

Tasks and a data coding scheme are determined based on theory or on previous studies. New transcripts are analysed in accordance with the scheme. This can transform transcript data into a variety of forms, such as quantitative data, process descriptions, instance collections, etc.

Bottom-up analysis:

Break data into 'units'; then systematically code and collate the lower-level categories. Group progressively into higher-level, more encompassing aggregates.

Top-down analysis:

Abstract emergent or organizing concepts from the data. Work down, to create outlines of the data, sorting phenomena within the concept divisions.

Analysis need not rely wholly on human interpretation. Once data is in electronic form, it is amenable to *automated analysis*, again in various forms. The simplest form is mechanistic counts, for example of occurrences of words or phrases. But computational linguistics affords a wealth of techniques for characterizing texts. And, again, techniques may be combined. For example, an initial manual coding of features can be augmented by application of automated analysis to the coded data.

These approaches have been described in as 'generic' a form as possible, in order to reveal some basic analysis strategies.

Statistical analysis

Statistical techniques, similarly, have purposes, costs, and conditions for their application. For example, statistical tests for association or co-relation are familiar in the context of experimental techniques. *Non-parametric* tests are suitable for experiment designs which test only one independent variable. *Parametric tests* can handle experiment designs which vary more than one independent variable and hence which require more complex statistical treatment. Requirements for statistical significance and power determine minimum numbers of subjects, and different statistical tests have different pre-conditions. For example, parametric tests require interval measurement, normal distribution, and homogeneity of variance. Such conditions have implications for the ability of a given technique to address the complexity of human behavior—the requirements for a particular test may be too constraining to fit the purpose of the research question (for example the assumption of homogeneous variance)—and for the ability of a given technique to be applied within the pragmatic constraints on data collection (for example, limitations on the numbers of available subjects may exclude some tests). Hence, the experiment design shapes the analysis through its focus, the nature of the data constrains which analysis technique may be applied, and the minimum requirements of the statistical technique limit which sorts of situations it may address.

Statistical methods can also be applied to quasi- and non-experimental data, sometimes as a test of association, but more often as a descriptive tool. Again, the question and its evidence requirements shape the analysis desired. Data and technique make demands of each other, the nature of the data constraining which techniques may be applied, and the desired techniques setting requirements for the data to be collected.

Summary

Effective research requires methods which generate data relevant to the research question. Our pragmatic approach hinges on formulating the research question in a way that encompasses not just *what* is asked, but *for what purpose*—and hence establishes what sort of evidence is fit and sufficient to address the question. These point the way to the choice of method. The "what" suggests the sort of data required and hence the sort of method needed, and the “for what purpose” influences the choices about how the method will be applied, in order to maximize utility within the constraints of cost. Study design is a matter of tradeoffs, between richness, resolution and costs; among the costs of different stages of design, implementation, execution and analysis; and among resources (such as numbers of subjects, amount and richness of data, time, and equipment) constrained within a budget.

Replicate and generalize across studies

In order to contribute usefully to the discourse, our research findings must be valid, relevant, and important. These qualities are the drivers for the attention to replication and generalization. We need to establish that our findings and conclusions are 'true', that they are neither chance findings nor distortions. One mechanism for doing so is to expose the work to validation—to *replication* or *repetition* and investigation—by others. We need to establish that our questions are significant, and that our findings address those questions usefully. Hence, we hope that the findings *generalize*, that they apply beyond our particular study to reveal some underlying 'truth' applicable to a larger population, set of tasks, or context. We also need to clarify how our findings are bounded—and also what the limits are of the theory that explicates them.

Replication and repetition

Replication and repetition are means for testing validity, in terms of the reliability and robustness of the findings. Replication and repetition are closely related. 'Replication' is the 'verbatim' reproduction of a study by another researcher, that is, using the same protocol under the same conditions. Replication tests how 'reliable' the findings are, that is, how consistent the outcomes of a given study will be given repetition by different researchers, at different times, with a different sample of the same population. Reliability contributes to the strength of evidence. We seek replication in controlled experimentation.

Replication is not necessarily feasible in educational research, which is set in a complex and dynamic social environment that may defy reproduction of conditions. Hence we seek 'repetition' in studies other than experiments, reproduction by another researcher using the same protocol under similar conditions in a similar setting, e.g., moving it from one classroom to a similar classroom with a similar learning context. Repetition tests reliability and also, because of the small differences in context and setting, the 'robustness' of the findings. That is, repetition can show how consistent the outcomes of a given study are across different related tasks, across different environments, across different related contexts. Repetition also exposes study design and conduct to the scrutiny of more minds, and hence puts the inference chain to the test and may help to draw out alternative interpretations.

Generalization and representativeness

Repetition, in offering an indication of consistency of outcomes across slightly different conditions, may help us understand how well findings generalize—or help to indicate a margin of error and to establish what limits might apply to the findings. It is part of the nature of empirical study that we investigate a particular example in the hope that it represents a more general phenomenon, and in the hope that any understanding we derive from the particular may extend to the general phenomenon. In seeking to generalize, we must also seek the boundaries of the generalization and understand that it encompasses some margin of error.

Empirical study is characterized by selection: the selection of subjects, of tasks, of time, of setting, of data collection. Every time a selection is made, its ability to represent what it is selected from (whether population, repertoire of activities, environment, etc.) must be questioned. Representativeness is the key to generalization: if the study is representative, then its outcomes can be generalized to the greater population, to other settings, and so on. The particularity of a research outcome, that is, its lack of representativeness, constrains its utility in the research discourse.

Selection of samples

If we want research to be representative, then we must attend to how we make our selections, to how we 'sample' from the world in order to focus an investigation. We usually use the term 'sample' as shorthand for 'sample of subjects'; it usually refers to a selection of people intended to represent a defined population. But sample may equally refer to a defined population (or set) of artifacts, events, or tasks. For each, the representativeness of the selection must be considered.

The first step in sample selection is the characterization of the population which the sample is to represent—the population to which the results of the research are meant to generalize. The characteristics of the population relevant to the phenomenon of interest must be identified, in order to consider *in what ways* the sample must be representative. (The catch is that this may be difficult to do in advance—the phenomenon might be influenced by population factors you may not consider to be relevant.)

Another step is to decide how large the sample must be. A good 'rule of thumb' is to use the biggest sample one can afford and obtain. This is true of tasks and artifacts, as well as of subjects. Early consideration of the analysis strategy will influence this decision: statistical power depends on sample size, and some statistical treatments require minimum sample sizes. Qualitative analysis is expensive and time-consuming and may suggest depth of treatment rather than breadth in sample selection. Hence the evidence requirement will influence sample size.

There are a number of indicators for a large sample size, such as:

- requirement for a high level of statistical significance (the probability that a result is not due to chance), statistical power (the probability that, if an effect exists, it will be found), or both
- likelihood of high attrition rate
- need to sub-divide groups
- many uncontrolled variables
- likelihood that effect sizes will be small
- population is highly heterogeneous with respect to the variables being studied

In short, a large sample is called for when the likelihood of drawing wrong conclusions from a small sample is high. On the other hand, sample selection can also be matter of diminishing returns. It is worth considering what utility will be gained by, say, doubling a sample. There are times when a small sample will do as well as a large one—it is a matter of fitness for purpose, of the evidence requirements of the research.

Strategies for sample selection may be random or non-random, and may be based on the individual or on groups (that is, sub-groups of the population). In this context, it is important to distinguish between *random* samples (in which all individuals in a population have an equal and independent chance of being selected) and *arbitrary* samples (in which the selection is made on some basis notionally irrelevant to the study; 'any one will do'). The two are not equivalent and have different implications for representativeness. Further, there are *non-random* samples, selected on a basis intended to maximize representativeness of the sample for the purpose of the study. Some general methods are indicated here:

Sampling methods

Simple random sampling:

All individuals in a population have *equal* and *independent* chance of being selected. Entails a measurable degree of uncertainty.

Systematic sampling:

Devise a procedure for selecting every *nth* member of a given list of members of the population.

Stratified sampling:

Assure that subgroups in the population will be represented in proportion to the numbers in the population; select randomly from within the subgroups.

Cluster sampling:

The unit is not an individual but a naturally-occurring group; all members of randomly selected groups are included.

volunteer sampling:

Subjects select themselves. *Note:* Volunteers have been shown to differ from non-volunteers in important ways; therefore, use of volunteers constrains generalization.

Empirical study design, particularly in the context of CS education research, is not a pure exercise, but a pragmatic one, in which factors such as access, cost, and ethics can put sharp constraints on design decisions. The opportunistic nature of much research (e.g., times of transition, briefly available resources) means that practical decisions often intrude. However, opportunism can jeopardize representativeness, and many of the common errors in selecting participants for a study arise from practical compromise, for example:

- selecting people because they're available and appropriate sampling is not convenient;
- selecting participants who are not in an appropriate population;
- using volunteers but failing to ascertain how they may differ from non-volunteers on crucial characteristics or abilities;
- selecting a sample that does not provide for attrition and may be too small by the end of the study.

It is essential, therefore, to consider the limitations attendant on such decisions, and to consider their implications for the value of the evidence gathered. There is no utility in selecting a sample by a method which fails to meet the needs of the research design.

Validity

Validity is the extent to which an account accurately represents the phenomenon to which it refers. More generally, validity is the ability of the research to provide accurate and credible conclusions, building on evidence that is sufficient to warrant the interpretation made. The validity of research is established (or threatened) at many levels, and it affects the value of the results, their representativeness, and the legitimacy of generalization from them.

Table 18: Types of validity

internal validity: addresses how consistently similar results can be obtained for these subjects, for this setting, using these techniques; addresses the quality of inference and conclusions within the study	**e.g.** →	**construct validity:** whether the constructs related to a phenomenon are valid, whether the operationalisation (the mapping of the construct onto manifestations in the world) is valid, the comparability of that operationalisation with other studies of the same construct
	→	**validity of measures:** whether measures measure what they claim to, and whether they do so reliably
external validity: addresses whether the study provides a true reflection of the phenomenon as it occurs the world, hence the generaliseability of the conclusions to other times, settings, and populations	→	**ecological validity:** whether the setting is representative of settings of the same type and of such settings in the world, and hence whether the findings within one setting are generaliseable to other similar settings
	→	**population validity:** whether the sample is representative of the greater population to which the results are generalized

Bias

Bias threatens the validity of research.

Consider a laboratory experiment that compared two solvents. The setting is controlled: it is a ‘fair test’, with identical environment, materials, and protocol for the two conditions. Temperature, the nature and amount of material to be dissolved, the application of the solvents – all are identical. And all are arguably representative of natural environmental conditions for the task. One solvent is demonstrably more effective than the other. What conclusions might one draw, how convincing would they be, and how safe would they be to generalize?

But what if the laboratory were also a television studio, and the two solvents were dishwashing detergents. Advertising product comparisons are presented as controlled experiments, but do you consider them to be ‘fair tests’? Advertising standards require that the control of variables in product tests be genuine. But they allow the control—the choice of temperature, nature and amount of material to be dissolved, and means of application of the detergents—to be optimized for one of the detergents.

In product comparisons, the comparison is controlled, and the results are reliable, but bias is built into that control. Now, what if the bias were not intentional? Might the appearance of control mask the limitations of the result, and their implications for restrictions on the conclusion?

‘Bias’ is when things creep in unnoticed to corrupt the evidence. It is the distortion of results due to factors that have not been taken into consideration, e.g.,

- extraneous or latent influences
- unrecognized conflated variables
- selectivity in a sample which renders it unrepresentative

The very act of experimenting introduces the potential for bias. This is referred to as the Heisenberg principle: you can’t observe without influencing what you’re observing. The fact of observing phenomena changes them.

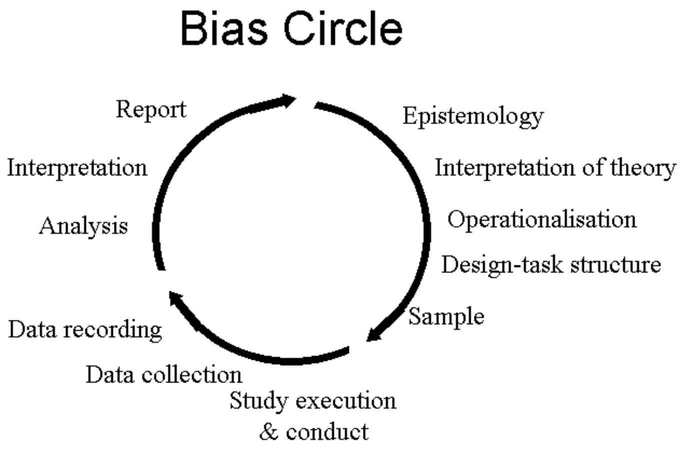


Figure 6 : Bias Circle. After James Powell (Powell, 1998)

Bias can creep in at any point in research, from the earliest planning through each reasoning and implementation step, through execution, data collection, analysis, and even reporting. In CS education research, fallible, variable humans are both the subjects and the instruments of research, providing multiple opportunities for error and distortion. Rigor demands vigilance against bias, with implications for the design of empirical studies, and for the design and execution of data collection. There are 'dangers around every corner'.

Dangers in operationalisation

A crucial link in the chain of inference is 'operationalisation', linking the concept or construct of interest to an observable indicator—to something that can be investigated empirically. The construct is mapped onto one or more manifestations in the world, things that can be observed, recorded, and ultimately measured in some way. The validity of the study rests on that operationalisation, on that mapping from construct to observable phenomenon to measure. If the reasoning that associates the measure with the construct is faulty, then the data may be irrelevant or misleading.

Operationalisation is important, too, in the accumulation of evidence. Not only is the construct mapped onto some manifestation in the world, but also the mappings applied in different studies must be compared. Is the construct interpreted in the same way? Are the manifestations comparable? Are the measures applied to the manifestations actually measuring the same thing? How well do the measures reflect the manifestation, and how well does the manifestation represent the construct?

The difficulties of finding relevant measures are many:

- Hard to find a measure.
- Hard to be sure it measures what's wanted.
- Hard to be sure it reflects enough of the story.

We've already discussed the difficulty of achieving precision and retaining relevance, characterized as 'sand through the fingers'.

Time and error measures are often used. But what is the meaning of time? Typically, the measure is of performance time, but 'time off task' or 'fiddling time' apparently spent in distraction tasks such as tidying or playing may have a bearing on performance, and they are difficult to measure. What is the meaning of error? Experts tend to make more errors than journeymen (experienced non-experts), but their overall performance is better, because they are better able to recognize and correct their errors, and journeymen expend more time and effort fending off error during initial generation. Time and error are accessible, but they may not be able to account for human perceptions and behavior of interest. For example, motivation can lead people to spend disproportionate time on a task, and yet to perceive it as quick.

Measures are shorthand, a compact expression or reflection of a phenomenon. But they're often *not* the phenomenon—the measure is typically a simplification. Some things are hard to measure, to quantify. For example, making continuous phenomena discrete can distort them. Experimental techniques have us focused on surface features, and the quest for measures can distract us from what is relevant with what is readily measured—sometimes we need other techniques to investigate deeper issues, before we can seek relevant measures.

Dangers in interpretation

The difference between 'data' and 'evidence' is interpretation. Evidence is data, plus the meaning we ascribe to it. Therefore, our reasoning about data is crucial, and it must take into account a variety of dangers in interpretation. There are many dangers, but some are more common than others: selectivity, flaws in reasoning, failing to make alternative accounts, falsely comparing heterogeneous evidence and failing to distinguish "frame of reference".

Selectivity

The danger of selectivity in interpretation is that any description we make of our observations, phenomena, processes, etc., whether to ourselves or in print, is selective. Through the process of research we gather data, but the phenomenon of interest is always more general than the data we choose to collect. Any measure we use to characterize and compare our results is shorthand for some feature of the world.

Flaws in the reasoning chain

The relationships between the phenomena of interest, the research design which aims to capture information about them, the constructs by which we describe them, the ways we operationalise those constructs, and the measures that capture them, are linked by a chain of reasoning, expressed through a chain of argument. Weakness in that chain potentially impairs the relevance and value of the data.

Alternative accounts

Data often admits more than one interpretation, more than one account of its meaning. Alternative accounts should be sought and given due consideration. If possible, alternative accounts should be investigated empirically to establish if they are valid. Alternatives should be ruled out systematically.

Comparing heterogeneous evidence

Some of the most informative studies combine techniques. But one burden of multi-method research is the difficulty of aggregating or comparing heterogeneous evidence. False similarity can lead to false conclusions. Dangers of interpretation are exacerbated when we are reasoning across a number of studies, or across a variety of data.

An exaggerated illustration of the danger is the way evidence was used in a popular TV programme that addressed controversial issues. The program presented heterogeneous evidence in a progression that tended to lead the viewer to draw false conclusions:

Progression	**Fictional illustration**
0. identify a controversial issue	There is concern that fluoride in the water makes us crazy.
1. report the results of a survey (the usual stratified sample format)	Do you think it reasonable to believe that fluoride in the water may have unexpected side-effects?
2. extract the interesting statistic	65% of the population thinks that fluoride may have unexpected side-effects.
3. find a couple of extreme cases and interview them	Jonny and Jimmy think that fluoride made them crazy; let's talk to Johnny and Jimmy and see how strangely they behave
4. add a dose of authority by interviewing scientists	Yes, some studies have been conducted in Europe into the side-effects of fluoride.
5. draw an unsupported causal inference	With such wide-spread concern about fluoride making people crazy...
(6. create panic)	People stop letting their children drink tap water.

Frame of reference differences

Borrowing methods without understanding the disciplinary, methodological, and conceptual framework is dangerous. For example, software engineering is task-oriented, whereas psychology of programming is human-oriented. Both use tasks, and indeed may investigate comparable tasks, but their interpretation may differ because of the disciplinary orientation. Frame of reference differences can distort data collection and lead to specious conclusions.

For example, "A question from [a] language development test instructs the child to choose the 'animal that can fly' from a bird, an elephant, and a dog. The correct answer (obviously) is the bird. Many first grade children, though, chose the elephant along with the bird as a response to the question." (Mehan, 1973)

Any child familiar with the 'Dumbo' film featuring a flying elephant might answer in this way. Test materials do not always have the same meaning for the tester and the subject, i.e., test scoring is interpretive.

There is danger in taking things out of context, and hence losing the original frame of reference. One classic example is 'seven plus or minus two', a limit on working memory established by George Miller in some classic psychology experiments (Miller, 1956). HCI designers have taken the finding up and applied it to interface design, using it as a limit on the number of items in a menu. However, selecting from a menu requires recognition, not recall; the finding is irrelevant to the application.

Danger in naïve appeals to scientific method

One artifact of the dominance of the 'scientific method shorthand'—of the appeal to method without a sufficient perspective on evidence—is a confusion of form with rigor. In fact, naïve approaches to scientific method can produce misleading or false insights. There is no hope of achieving the precision required for controlled experimentation before one understands what the question is, and what evidence is required to address it, and hence what constraints, simplifications, and trade-offs are acceptable for the purpose.

Among the dangers of 'premature experimentation' are:

- ill-formed hypothesis (hence lack of precision, confirmatory bias, danger of uninformative results)
- lack of control (don't know which variables are likely to be important, and hence which to control for)
- uncertain operationalisation (the relationship between the constructs being examined and the particular variables under observation is not established; are the manifestations in the world true reflections of the phenomenon of interest?)
- inadequate measures (the measures are insufficient to capture what they're meant to capture)
- inappropriate expectation (inappropriately seeking proof or conclusive evidence)

The consequences are spurious data, flawed analysis, and false conclusions.

Well-designed experiments are a powerful research tool. 'Scientific method' achieved dominance for good reason. But the 'if I find the right experiment I can do a statistical proof' model of empirical study design is often a case of 'trying to run before one can walk'. In a theory-scarce domain, one needs enough disciplined observation to provide a reasonable basis for identifying important factors and relationships, in order to distil well-founded conjectures (pre-theories?), from which one can generate the sort of precise hypotheses which are worth the cost of experimentation. Premature experimentation can narrow the focus too soon, and miss important phenomena entirely.

The definitive experiment clearly has its place in theory validation. The accumulation and valuation of evidence through a variety of methods is the preparation for theory generation, the obvious prerequisite for theory validation. The definitive experiment is a fine aspiration, but it is perhaps the wrong mechanism for CS education research, when what we need is better questions.

Danger in naïve appeals to metrics and statistics

> The measurement of the 100-yard dash is trivial… Measurement of intellectual artifacts is in its infancy (Curtis, 2000)

Beside the 'scientific method shorthand' walks an uncritical veneration for metrics "Numbers are good. Numbers are objective.", for numerical data, and for statistical analysis. The 'method of science' focuses our attention on questions that can be addressed empirically. The shorthand confuses that with 'what can be measured' or 'what can be addressed experimentally', hence potentially overlooking crucial factors and phenomena. The dangers here are captured in the McNamara Fallacy:

> The first step is to measure whatever can be easily measured. This is OK as far as it goes. The second step is to disregard that which can't be easily measured or to give it an arbitrary quantitative value. This is artificial and misleading. The third step is to presume that what can't be measured easily really isn't important. This is blindness. The fourth step is to say that what can't be easily measured really doesn't exist. This is suicide. (Handy, 1995)

Well-founded, valid metrics are powerful instruments. The key is to find a measure for what is important, rather than to make important what is measurable. How good is the evidence provided by a given metric? From it flows a series of related questions: What does the metric measure? How reliably does it measure it? How does what it measures relate to what we want to know? What *doesn't* it measure that might be important? The power of numbers (or words) in capturing phenomena lies in the validity of the measures (or constructs), in the chain that connects question to operationalisation to data to interpretation. Measures are context-dependent.

All too many studies simply measure the wrong thing. An example comes from software visualization. A researcher had devoted considerable energy to developing a debugging tool, applying a cocktail of metrics in order to select the most complex segments of code. Unfortunately, the work overlooked a pertinent characteristic of programmer behavior: programmers focus their analytic skills, tools, and time on the complex segments during development. Hence, the killer bug is more often in the simple code, the bits that programmers take for granted while they are focusing on the complex code they're worried about.

Researchers often confuse form with rigor not just in their data collection, but also in their analysis—in their appeals to statistics. Statistics 'feel' precise, but that doesn't mean that they are. From statistics, people hope to gain:

- rigor,
- a 'conclusive' demonstration of an effect,
- objectivity.

But, as Huff phrased it: "A difference is a difference only if it makes a difference." (Huff, 1954) Application of statistics without sufficient statistical insight can be meaningless or misleading.

Hence, we offer some cautions against common errors in statistical argument (drawn from Huff):

- Statistics work best in simple cases.
- In a statistical analysis, notions of 'trends' and 'influences' are meaningless if they are not supported by statistically significant results.
- In assessing the strength of statistical evidence, we must consider not just result of the test, but also the levels of significance (the probability that a result is not due to chance) and power (the probability that, if an effect exists, it will be found).
- The failure of data to pass a statistical test doesn't necessarily mean that the effect doesn't exist, only that it wasn't detected in this sample.
- An association between two factors is not proof that one has caused the other. Co-variation often reflects influence from a third factor.
- It is dangerous to infer beyond the data.
- A correlation may be real and based on real cause-and-effect — and still have little utility in addressing the research question.
- "The trend-to-now" may be a fact, but the future trend represents no more than an educated guess.

Tools are as good as the use we make of them. At their best, statistics are an incisive research tool (or collection of tools) that can be used in a variety of ways, e.g.:

- to describe,
- to compare,
- to detect patterns or relationships.

Nevertheless, their status is subject to interpretation: "Statisticians believe that the validity of the statistics can be proven mathematically; whereas mathematicians believe that the validity of statistics can be proven empirically."

Pilot studies

Pilot studies are the first defense against oversight (or stupidity) and the bias it may invite. They help to establish credibility, feasibility, and comprehensibility in advance of the data collection. A good pilot study provides a chance to debug the protocol, to expose frame of reference problems, to test the analysis on genuine data. It can expose design flaws, hidden assumptions, and unexpected problems.

So what makes a good pilot study? It must be a genuine 'dress rehearsal', using the full protocol with subjects representative of target population. Every aspect of the study must be tested out beforehand. The protocol, instruments, and procedures must be tried out, debugged, and tried out again until it is clear that they will work as intended, and that they will generate data which will be pertinent and amenable to analysis. Pilot studies are expensive of time and resources, but the consequences of inadequate testing are likely to be even more expensive. Short-cuts can be catastrophic.

It is crucial that the sample used for the pilot studies be representative of the target population. For example, British academics cannot be taken as representative of European academics (a short-cut that cut one of us short); they may have significantly different interpretations of taken-for-granted terminology. It is also important to pilot the analysis; working back from the analysis can reveal fundamental inadequacies in the study design. The data needs of the statistical tests may expose shortcomings in the data collection. Working back from the analysis may expose gaps in the chain of inference. Better to spot them early than to collect inadequate or irrelevant data.

Accumulation of evidence

Replication is one way of testing the strength of evidence—and potentially of contributing to its strength. Repetition is another, with the additional potential to extend the evidence by accumulating related findings from comparable but differing studies. A condition for replication or repetition is that the study be made accessible, that its definitions, protocols, links to theory, reasoning, and reporting be thorough, accurate, and public. This is also a condition for accumulation of evidence across a number of studies; full access is necessary for the assessment of the relatedness and comparability of constructs and findings.

One of the advantages of standard procedures (of adherence to form) is that it facilitates accumulation of *comparable* data and evidence. Those operating within a given set of standards share epistemology, terminology, conceptual frames, ways of reasoning, ways of reporting, and even assumptions, and this allows them to think about and compare each other's work readily. They get to compare 'apples to apples'.

Hence, one of the burdens of a triangulation approach is to accommodate heterogeneous data, somehow rendering it into comparable forms, or finding means of recognizing regularities. In other words, the challenge is to connect variables among studies so that inferences can be made with increased realism and increased control. For example, field experiments can counter-balance laboratory experiments, if they address the same constructs interpreted in comparable ways—or if the two share enough essential features to be similar.

Any comparison that is made among heterogeneous data must take into account the way that data is colored by how it was collected and interpreted: by the epistemology and disciplinary traditions influencing the casting of the question and the study design, by the assumptions and limitations that attach to the conceptual frames employed in the data's interpretation, by the selections made in the operationalisation and instrumentation, and by the selections and simplifications employed in the description and report. In comparing 'apples with oranges', we need to find a means of reasoning about fruit, while maintaining awareness of the particularities and differences.

There are many issues to consider in making sense across studies:

- *Terminology*: are words used to mean the same things, with the same granularity?
- *Conventions/standards*: what is implied by the conventions and standards observed in the different studies; is some data or reasoning excluded by one and not the other? are there differences in the standards of reporting that may have consequences for the completeness of the accounts? what is considered to be acceptable practice?

- *Assumptions*: are different assumptions implicit in the techniques applied or the theories brought to bear?
- *Conceptual frames/ ways of reasoning*: what assumptions are implicit in the conceptual frames? Are the levels of granularity and abstraction comparable and is the interpretation of concepts or constructs comparable? do differences in reasoning about data lead to differences in legitimacy?
- *Time*: might the effects of time (history, changes over time, fluctuations, patterns or variations in phases) influence the quality of the evidence?

Time is a key issue, often overlooked. Given human memory and psychology, the impact of time—or rather of limitations or considerations associated with time, our perception of it, and the way our perception of time influences our interaction with the world—can be profound in CS education research. For example:

- Initial use does not necessarily generalize to evolved use.
- Single use does not necessarily generalize to repeated use.
- Time is reflected in sequences, processes, antecedents, and context.
- History may have an impact on current phenomena.
- Phenomena may change over time.
- Phenomena (patterns, variations) may occur in phases or have periodic fluctuations.

.

The focus provided by theory makes it natural to pursue cumulative research. But, in the absence of well-founded theory, attention to the accumulation of evidence contributes to a pragmatic approach to theory-building and theory use. In either case, theory (in the role of the driver, or the goal) provides a focus, making accumulation easier. Accumulation of evidence over a number of studies provides a means of addressing the difficulties of achieving a critical mass of work on a given topic. Attention to accumulation mitigates against isolated and esoteric studies.

The need for honesty

With so many vectors of bias and threats to validity, vigilance is a constant necessity. But so is honesty. The impact of evidence in the discourse depends on people’s ability to assess its strength. Good evidence presentation requires clear description of data collection and analysis, an explicit account of the chain of reasoning from study design through data interpretation to conclusions, and an assessment of the reliability and margin of error. Through honest reporting, evidence is exposed to scrutiny, to test and possible falsification.

Disclose research to encourage professional scrutiny and critique

CS education research is …

As we view—and practice—it, CS education research is not just “scientific method”, nor is it confined to the natural world. It borrows from other areas and traditions, in terms of theory, method and approach. We adopt what we term “method of science”, a principled and rigorous articulation of observation and explanation.

Which is almost, but not quite, enough. Because science is a discourse, and articulation is chiefly about reading and writing.

Reading is important, because it helps direct research purposefully, providing others’ work to build on, indicating which avenues to avoid, showing where contributions are needed. We are fuelled by our scrutiny, critique, and use of others’ work. If we don’t know what others have done, we stand a good chance of wasting our time by “re-inventing the wheel”, unknowingly re-creating work that makes no contribution to knowledge. Whether we use others’ work to provide situation and context for our own, or, more closely, as a study to replicate or generalize from, we owe the researchers whose work we use a duty of care, and should practice basic academic skills of reference and report. Naturally, this means giving proper credit. It also entails ensuring that we use others’ ideas and work as the originators intended, and not for what we would like them to be or for what we would like them to say.

Writing is important, because otherwise our work is invisible and unscrutinised. We can pose significant questions, choose appropriate methods, operationalise them scrupulously avoiding all possible bias, to uncover evidence which has an impeccable chain of inference. But if we don’t then write about it, we might as well not have bothered.

The discourse puts obligations on what we write and how we write it. Research papers are not just telling a story or making a report: they must provide an audit trail of the work and thought that lead to our claims and conclusions. In this way our work can be scrutinized (examined for accuracy) and critiqued (probed for weakness) by our colleagues and peers. If our work is good, then we can expect to be read by others, and perhaps used to situate their own work, or perhaps be replicated by them. We owe them a duty of care to be honest in the framing, situation, conduct, and reporting of our work.

Reading and writing together are about “joining the discourse”

References

Abowd, G. D. (1999). Classroom 2000: An Experiment with the Instrumentation of a Living Educational Environment. *IBM Systems Journal Special Issue on Pervasive Computing, 38*(4), 508-530.

Anderson, R. J., Anderson, R., VanDeGrift, T., Wolfman, S., & Yashuhara, K. (2003). *Promoting Interaction in Large Classes with Computer -Mediated Feedback.* Paper presented at the Computer Supported Collaborative Learning, Bergen, Norway.

Astrachan, O. (1998). *Concrete teaching hooks and props as instructional technology.* Paper presented at the ITiCSE, Dublin.

Astrachan, O., Wilkes, J., & Smith, R. (1997). *Apprenticeship Learning in CS2.* Paper presented at the 28th SIGCSE Technical Symposium on Computer Science Education, San Jose, CA.

Ben-Ari, M. (2001). Constructivism in Computer Science Education. *Journal of Computers in Mathematics and Science Teaching, 20*(1), 45-73.

Bloom, B. S. (1956). *Taxonomy of educational objectives; the classification of educational goals* (1st ed.). New York,: David McKay.

Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *The Journal of the Learning Sciences, 2*(2), 141-178.

Brown, A. L., & Palincsar, A. S. (1989). Guided, co-operative learning and individual knowledge acquisition. In L. B. Resnick (Ed.), *Knowing, learning and instruction: Essays in honor of Robert Glaser*. Hillsdale NJ: Lawrence Erlbaum Associates.

Bruner, J. (1960). *The Process of Education*. Cambridge MS: Harvard University Press.

Csikszentmihalyi, M. (1991). *Flow - the Psychology of Happiness*: Rider.

Curtis, B. (2000, 5 June). *Borrow or steal? Using Multidisciplinary Approaches in Empirical Software Engineering Research (Keynote talk).* Paper presented at the International Conference on Software Engineering, Limerick, Ireland.

Dourish, P. (2001). *Where the Action Is: The Foundations of Embodied Interaction*. Cambridge MA: MIT Press.

Eisenstadt, M. (1993). *Tales of Debugging from the Front Lines.* Paper presented at the Empirical Studies of Programmers.

Engestrom, Y., Miettienen, R., & Punamaki, R.-L. (Eds.). (1999). *Perspectives on Activity Theory*. Cambridge, UK: Cambridge University Press.

Entwistle, N. J., & Tait, H. (1995). *The Revised Approaches to Studying Inventory*. Edinburgh: Centre for Research on Learning and Instruction, University of Edinburgh.

Fincher, S. (1999). *What are we doing when we teach programming?* Paper presented at the Frontiers in Education, San Juan, Puerto Rico.

Fincher, S., Petre, M., & Clark, M. (Eds.). (2001). *Computer Science Project Work: Principles and Pragmatics*. London: Springer-Verlag.

Foxley, E. (2003). *Ceilidh*, from http://www.cs.nott.ac.uk/~ceilidh/

Galison, P. L. (1997). *Image and logic : a material culture of microphysics*. Chicago: University of Chicago Press.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago.

Handy, C. (1995). *The Age of Paradox*. Cambridge, MA: Harvard Business School Press.

Hause, M. L., Almstrum, V. L., Last, M. Z., & Woodroffe, M. R. (2001). *Interaction Factors in Software Development Performace In Distributed Student Groups In Computer Sceince.* Paper presented at the 6th Conference on Innovation and Technology in Computer Science Education, Canterbury, England.

Hempel, C. G., & Oppenheim, P. (1965). Studies in the Logic of Explanation. *Philosophy of Science, 15*, 135-175.

Holmstrom, J. E. (1947). *Records and Research in Engineering and Industrial Science: A guide to the sources, processing and storekeeping of technical knowledge with a chapter on translating* (Second ed.). London: Chapman and Hall Ltd.

Huff, D. (1954). *How to Lie with Statistics*. London: Penguin Books.

Isaac, S., & Michael, W. B. (1989). *Handbook in Research and Evaluation for Education and the Behavioural Sciences*. San Diego, CA: EdiTS Publishers.

Jenner, E. (1798). *An inquiry into the causes and effects of the Variolae Vaccinae, a disease discovered in some of the western counties of England, particularly Gloucestershire, and known by the name of the cow-pox*.

Kember, D. (1995). *Open Learnig Courses for Adults: A Model of Student Progress*. Eaglewood Cliffs, NJ: Educational Technology Publications.

Kolb, D. A. (1984). *Experiential learning: experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice-Hall.

Kolikant, Y. B.-D. (in press). Learning Concurrency as an Entry Point to the Community of CS Practitioners. *Journal of Computers in Mathematics and Science Teaching*.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Lancaster, T., & Culwin, F. (2004). A Comparison of Source Code Plagiarism
Detection Engines. *Computer Science Education, 14*(2).

Lave, J., & Wenger, E. (1991). *Situated learning : legitimate peripheral participation*. Cambridge England ; New York: Cambridge University Press.

Linn, M. C., & Clancy, M. J. (1992). The Case for Case Studies of Programming Plans. *Communications of the ACM, 36*(3), 121-132.

Lister, R., & Leaney, J. (2003). *Introductory Programming, criterion-referencing, and Bloom.* Paper presented at the 34th SIGCSE Technical Symposium on Computer Science Education, Reno, NV, USA.

Mason, R. O. (1989). MIS experiments: a pragmatic perspective. In I. Benbasat (Ed.), *The Information Systems Research Challenge: Experimental Research Methods, Vol. 2* (pp. 3-20): Harvard Business School.

Masterman, M. (1970). The Nature of Paradigm. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

McCracken, M., Almstrum, V., Guzdial, M., Hagan, D., Kolikant, Y. B.-D., Laxer, C., et al. (2001). A multi-national, multi-institutional study of assessment of programming skills of first year CS students. *SIGCSE Bulletin, 33*(4), 125-180.

Mehan, H. (1973). Assessing Children's Language Using Abilities. In J. M. Armer & A. D. Grimshaw (Eds.), *Methodological isses in compartative sociological research*. New York, USA: John Wiley and Sons.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review, 63*, 81-97.

Myers, B. A. (2001). Using Hand-held Devices and PCs Together. *Communications of the ACM, 44*(11), 34-41.

Myers, I. B. (1998). *MBTI manual : a guide to the development and use of the Myers-Briggs Type indicator* (3rd ed.). Palo Alto, Calif.: Consulting Psychologists Press.

Myers, I. B. (2000). *Introduction to Type: A Description of the Theory and Application of the Myers-Briggs Type Indicator*. Oxford: Oxford Psychologists Press.

Nardi, B. (Ed.). (1996). *Context and Consciousness: Activity Theory and Human-Computer Interaction*. Cambridge, MA: MIT Press.

OECD. (1995). *Background Paper to the OECD Workshop: Sustainable Consumption and Production: Clarifying the Concepts*, from http://www.sustainableliving.org/appen-e.htm

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension -fostering and monitoring activities. *Cognition and Instruction, 1*(2), 117-175.

Papert, S. (2003). *Seymour Papert*, from http://www.papert.org/

Perry, W. G. (1981). Cognitive and Ethical Growth: The Making of Meaning. In A. W. Chickering (Ed.), *The Modern American College* (pp. 76-116). San Francisco: Jossey-Bass.

Perry, W. G., & Harvard University. Bureau of Study Counsel. (1970). *Forms of intellectual and ethical development in the college years; a scheme*. New York,: Holt Rinehart and Winston.

Petre, M., Price, B. A., & Carswell, L. (1996, April). *Moving programming teaching onto the Internet: experiences and observations.* Paper presented at the 8th Workshop of the Psychology of Programming Interest Group, Ghent.

Popper, K. (1959). *The Logic of Scientific Discovery*. New York: Basic Books.

Powell, J. (1998). In M. Petre (Ed.).

Research, C. f. A. E. a. S. E. (1995). *Phenomenographic Research: An Annotated Bibliography* (Occasional Paper 95.2). Brisbane, Australia: QUT Publications and Printing.

Schkade, D. A. (1989). Prospects for Experiments Focusing on Individuals in IS Research. In I. Benbasat (Ed.), *The Information Systems Research Challenge: Experimental Research Methods* (Vol. 2, pp. 49-52). Boston, MS: Harvard Business School.

Seger, C. A. (1994). Implicit Learning. *Psychological Bulletin, 115*, 163-196.

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific Research in Education*. Washington DC: National Academy Press.

Sheard, J., Dick, M., Markham, S., Macdonald, I., & Walsh, M. (2002). *Cheating and plagiarism: perceptions and practices of first year IT students*. Paper presented at the Proceedings of the 7th annual conference on Innovation and technology in computer science education, Aarhus, Denmark.

Skinner, B. F. (1938). *The behavior of organisms; an experimental analysis*. New York,: Appleton-Century-Crofts.

Skinner, B. F. (1968). *The technology of teaching*. Englewood Cliffs, N. J.: Prentice-Hall.

Sphorer, J. C., Soloway, E., & Pope, E. (1985). A goal/plan analysis of buggy Pascal programs. *Human-Computer Interaction, 1*, 163-207.

Stokes, D. E. (1997). *Pasteur's Quadrant: Basic Science and Technological Innovation*: The Brookings Institution.

Vygotsky, L. S. (1962). *Thought and Language*. Cambridge MS: MIT Press.

Wenger, E. (1998). *Communities of practice : learning, meaning, and identity*. Cambridge: Cambridge University Press.

Williams, L., & Kessler, R. (2001). Experimenting with Industry's "Pair-Programming" Model in the Computer Science Classroom. *Computer Science Education, 11*(1).

[2] Richard Feynman uses a similar term, “cargo-cult” science in analogy to the behaviour of certain remote peoples, who built runways in order to tempt airplanes to land.

[3] Richard Feynman describes an iconic example of the control of variables “ …”

[4] There are other well-known and well-explored factors here, of course. Age, culture and preparedness will affect performance. Perhaps more interesting is the idea that the value that we put upon these indicators is extrinsic to what they measure. Because our society values “high IQ” then performance on this scale is more valued than being on the end of other scales of empirical law: being tall, perhaps, or being able to store 12 things in short term memory

[5] Unusually, we also have a strong tradition of instrument-building, with simulations, visualisations, algorithm animation and construction of whole environments to have an effect on the teaching and learning of computing concepts. In a CS education research context these must be grounded in theory and refined with experiment, of course—but it might be that they are a unique contribution from CS education to other disciplinary areas, trade goods of value.

[6] Of course, subsequently, with historical perspective, the value of the trade to both parties can be seen to be different again. We would now feel that the Lenape made a very bad trading deal.

[7] This description taken from: William J. Rapaport *William Perry's Scheme of Intellectual and Ethical Development*, `http://www.cs.buffalo.edu/~rapaport/perry.positions.html`