Multi-Institutional, Multi-National Studies in CSEd Research: Some design considerations and trade-offs

Sally Fincher
Computing Laboratory
University of Kent
Canterbury, CT2 7NF, England
+44 1227 824061
S.A.Fincher@kent.ac.uk

Raymond Lister
Faculty of Information Technology
University of Technology, Sydney
Broadway, NSW 2007, Australia
+61 2 9514 1850
raymond@it.uts.edu.au

Tony Clear
School of Comp. & Inf. Sciences
Auckland University of Technology
Private Bag 92006, Auckland 1020,
New Zealand
+64 9 917-9999
tony.clear@aut.ac.nz

Anthony Robins Computer Science Department, University of Otago,
PO Box 56
Dunedin 9015, New Zealand
+64 3 479 8314
anthony@cs.otago.ac.nz

Josh Tenenberg University of Washington, Tacoma Computing and Software Systems
1900 Commerce St. Tacoma,
WA 98402-3100 USA
+1 253 692 5800
jtenenbg@u.washington.edu

Marian Petre
Faculty of Mathematics & Computing
The Open University
Milton Keynes, MK7 6AA, England
+44 1908 65 33 73
M.Petre@open.ac.uk

ABSTRACT

One indication of the maturation of Computer Science Education as a research-based discipline is the recent emergence of several large-scale studies spanning multiple institutions. This paper examines a "family" of these multi-institutional, multi-national studies, detailing core elements and points of difference in both study design and the organization of the research team, and highlighting the costs and benefits associated with the different approaches.

Categories and Subject Descriptors

K.3.2 [Computers and Education]: Computer and Information Science Education – computer science education, curriculum.

General Terms

Management, Measurement, Documentation, Experimentation.

Keywords: Multi-Institutional, Empirical, Education, Research

1. INTRODUCTION

The scope of educational research studies spans a broad spectrum. At one extreme there are small studies involving a few participants, usually at a single institution. These are often associated with "qualitative" methods such as case studies or interviews, or the use of specific tools. At the other extreme there are very large-scale studies comparing sizeable populations, usually over many institutions and possibly international in scope. These are often associated with "quantitative" data and methods such as the gathering of demographic information and the use of surveys. Such studies are typically carried out (or funded by) government organisations. They often have an explicit focus on benchmarking and are used to inform policy, for example surveys of student achievement have figured prominently in debates about educational standards in the United States since the 1980s [1].

Presented at ICER'05, October 1–2, 2005, Seattle, Washington, USA

Research in the comparatively young field of Computer Science Education (CSEd) consists almost exclusively of small-scale local studies. Many individual studies are of high quality and present significant and useful results. Overall, however, it is probably fair to say that the field of CSEd lacks a foundation of established theory and methods, is characterised by isolated findings that are difficult to assemble into a coherent whole, and thus has little impact on practice.

The purpose of this paper is not to criticise small-scale studies (which will probably always constitute the bulk of productive CSEd research): our goal is rather to draw attention to and explore an emerging trend in CSEd towards studies which are larger in scope, to characterise the different models for carrying out these larger-scale studies, and to provide a set of axes of comparison among these different models. Each of the examples discussed in this paper involves a larger than usual pool of subjects (of the order of 100 to 300) with all but one drawn from multiple institutions in at least two countries, hence we characterise them as "Multi Institutional Multi National" (MIMN) studies. Published MIMN studies in CSEd of which we are aware are [2-7]. The authors of this paper have all been involved—as organisers—in one or more of these studies.

Our characterisation of MIMN studies is deliberately general. The examples reviewed here employ different modes of organisation and realisation, explore a diverse range of issues in CSEd, and use a range of tools (both qualitative and quantitative). One factor that they have in common, however, is that MIMN studies are more complex, expensive and difficult to administer than typical small-

scale studies. What are the advantages? Why undertake a MIMN study? There are several possible motivations:

Statistical power. If it is properly designed and executed then the larger subject pool of a MIMN study will (in many cases) increase the power of tests used to establish the significance of effects and interactions. The effort of collecting data for many subjects is distributed over a number of participating researchers.

Richness. A MIMN study can have a richer structure (involve more conditions) than a typical small-scale study. Hence it is possible to address a broader range of issues. For example, MIMN studies can explicitly compare different institutions, and hence the effects of different educational environments. This makes it possible to identify and explore effects which are shaped by educational experiences (such as different teaching practices) vs. effects which are independent of them or shaped by demographic or developmental factors. Variation across institutions and cultures constitutes a "natural laboratory" [8] within which the effects of different hypotheses about teaching and learning can be observed.

Hypothesis generation. Conceding that in some cases sources of possible bias in large-scale studies may make it difficult to establish causal connections, [1], p15 note that "the value of [large-scale comparative] international studies may lie more in their potential for generating hypotheses about causal explanation than in their use as platforms for testing hypotheses".

Improved methodology. Experience with MIMN studies will almost certainly improve the methodology of such studies, and may contribute to improving practice within CSEd research generally, [1], p15 note for example that "Four decades of experience with large-scale cross-national surveys have led to substantial improvements in methodology, including better tests, better samples, better documentation, and better statistical analyses".

Accounting for background factors. Background factors such as culture or socioeconomic status may be of direct interest for their impact on teaching and learning, and such factors clearly lend themselves to MIMN investigation. Even when they are not the focus of direct interest, cognition is inextricably set in the context of such factors.

In short, MIMN studies can make many contributions, both specific and general, to CSEd research. Such studies will have an important role to play as the field matures. There is every reason to expect that they will significantly contribute to establishing both a common conceptual and methodological framework and a growing body of practical results and observations that can be used to improve teaching practice and learning in CS.

Yet this very relationship with practice means that CSEd research as a young field, suffers from many of the problems of practice based research [9, 10]. Enthusiastic CSEd practitioners develop an innovation in their teaching practice and typically evaluate its impact through some form of reflective practice. This model is not unknown. As Taylor [11] has noted in respect to flexible learning initiatives, many innovations in education have arisen from the work of "lone rangers"—individual academics who are energetic, early adopters of innovation, and who are motivated by a desire to improve the accessibility and quality of their teaching. Yet such work is frequently characterized by a practice focus, limited evaluation of its effectiveness, lack of shared knowledge building, and "failure to institutionalize the outcomes" [11]. In CSEd likewise, we see examples of local initiatives driven by the enthusiasm of specific educators, often written up as reflective practice pieces and shared at CSEd conferences, but without significant impact upon the practice of fellow CS educators.

In much the same way that action researchers suffer from criticisms of their work as "mere consultancy", the practitioner and researcher roles need to be consciously separated [10] so both practice questions and research questions are addressed with appropriate methods to produce solid and credible conclusions. Thus much work in CSEd research to date has been isolated, has had limited impact on practice, has not contributed to a research tradition and has not necessarily generated generalisable and reusable findings.

Multi-institutional CSEd studies offer one mechanism to demonstrate the effectiveness of certain forms of practice which are generalisable beyond the single institution or the "lone ranger" study.

2. TWO ORIGINAL MODELS

Most of the recent MIMN studies were modelled on the two projects discussed briefly in this section. These two models are introduced here, and discussed again later in the paper.

2.1 The McCracken Working Group

In one sense, MIMN studies are familiar within CSEd. From its inception (1996), the ITiCSE conference has had working groups associated with it. These have taken the following form:

a) topics are proposed, and peer-reviewed

b) one or more topics are selected for presentation

c) the topic is posted with an invitation for others to join in the work specified

d) the resulting group(s) work electronically before the conference, then work at the conference (and often for a day or more in advance)

e) the group(s) write a paper detailing their results. This is peer-reviewed and, if accepted, published in the SIGCSE Bulletin

f) the group disbands

Most working groups have not undertaken empirical research. Instead, most working group topics produce a report that either:

- distils collected resources and experiences on an issue of direct relevance to practicing teachers, for example Resources, Tools, and Techniques for Problem Based Learning in Computing [12] or A Road Map for Teaching Introductory Programming Using LEGO Mindstorms Robots [13], or
- addresses common problems that benefit from the application of collective intellectual and analytical effort, for example: How shall we assess this? [14] or Evaluation: turning technology from toy to tool [15].

The first ITiCSE working group to change this pattern was the "McCracken Group" [2]. The ten group members were from eight different institutions across five countries. What brought them together was not the collection of resources on a theme, or a common interest in thinking about an aspect of the computing

curriculum but the gathering of empirical data in response to a question. The question was (in our vulgar construction) Are your students as bad at programming as mine? Students at participating institutions (four collected data) were given a programming problem. While the problem was not the same at all institutions, all the problems involved evaluating an arithmetic expression input as a line of text. "The opinion of the working group's participating schools was that a student at the end of the first year of study should be able to solve the most difficult exercise of the three in about an hour and a half." [2] p.4. Most students did much more poorly than their instructors expected.

The impact of this study rests upon its multi-institutional nature, its focus on a question, and the purposeful gathering of empirical data. Whereas a similar report from a single institution might be dismissed as a consequence of poor teaching at that institution, it is difficult to dismiss the remarkably consistent results from multiple institutions. Thus the "McCracken Group" contributed the first model for MIMN studies to the CSEd canon.

2.2 "Bootstrapping"

In 2002 the US National Science Foundation funded the project Bootstrapping Research in Computer Science Education [16]. The aims of this project were: "to improve the state of Computer Science education research—and thereby ultimately to improve the state of CS education—by developing skills (in the design, conduct and management of research) of Computer Science educators and by exposing them to relevant theory and methods, and to facilitate the establishment of research relationships that extend beyond the duration of the workshops, contributing to a research community able to sustain a constructive discourse as well as ongoing collaboration." The project took the following form:

a) the PIs design and pilot a MIMN study

b) participation in the project is solicited (in the original form, participation was funded); participants are selected

c) the group meets for a four-day workshop, where (amongst other activity) the MIMN study is presented

d) the group works over the intervening year, each in their own universities, to gather data for the study to a common protocol

e) the group meets for a second four-day workshop to analyse the results in aggregate and write a paper

The research study relied on all participants gathering the same data each in their own classrooms. In this way, they experienced common practice and contributed to a common artefact. The aims of the Bootstrapping model were therefore quite different from the McCracken group. There was expectation of extended relationships between the participants, and the project model was designed to maximise these community aspects.

2.3 Adaptations

Since these initial instantiations, there have been six further studies that have, in one way or another, been influenced by these two models (often by individual researchers moving from study to study). Some have closely followed the originals: Scaffolding Research in Computer Science Education [4] (also funded by the NSF), Building Research in Australasian Computing Education (BRACE)[5] and the "South Carolina group" [17]. Others have adapted to evolve new forms: the 2004 ITiCSE working group [6], BRACElet [18] and ExploreCSEd. The relationship between the different studies is depicted below in figure one, and each of these models is described further in section 4, below.

3. IMPORTANT CONSIDERATIONS IN MIMN WORK

Although the two models can be seen as quite different, and the adapted models even more so, MIMN studies have aspects in common. They all have to deal with coordination of institutions, participants, researchers and data: issues which are at best implicit and often invisible in single-institutional studies. MIMN studies display varying tightness of control over these areas, with various trade-offs

3.1 Scale

3.1.1 Multiple researchers (at multiple institutions)

MIMN studies involve large, distributed, teams of researchers, and while this underlies their strengths, it also contributes to their cost and complexity. Distributed teams need to have effective mechanisms for coordinating their activities. These can include clearly defined roles and responsibilities, shared resources and protocols, a clear timetable, and effective means of communication.

A major resource for guiding and coordinating the activities of the research team will be the documents that define the questions of interest, the tools and protocols used to collect and analyse data, and so on. In Bootstrapping-model studies this material was collected into a single document called the "Experiment Kit" [19].

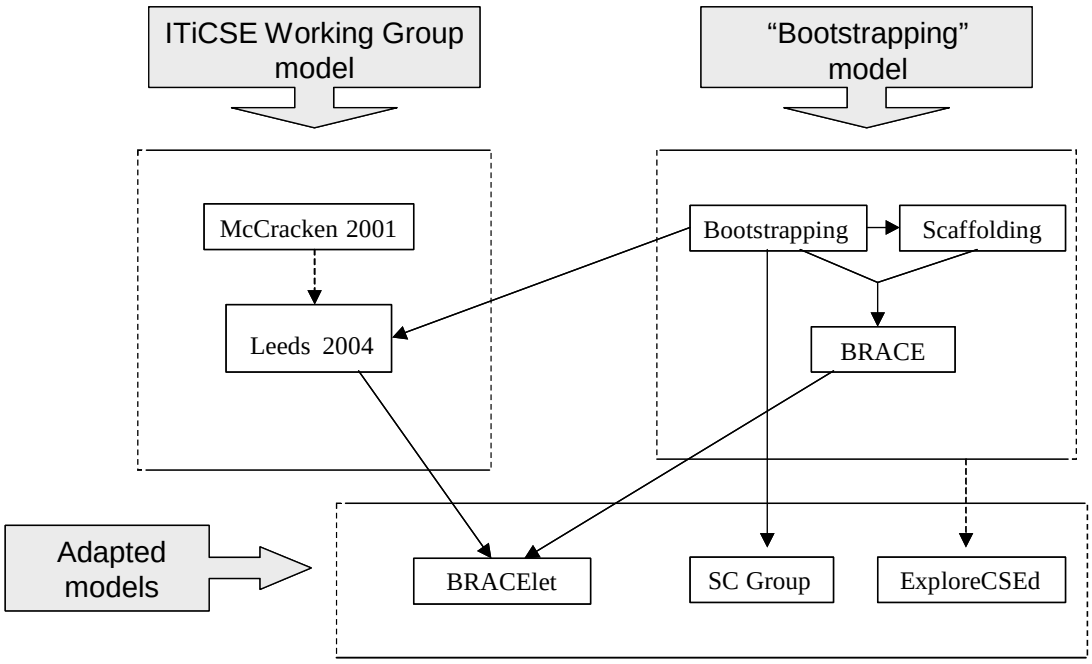

Figure one: MIMN studies and their influence. A solid arrow indicates the studies' leadership were researchers from a previous study, a dashed arrow indicates leadership was influenced by the previous studies

Experiment Kits: An Experiment Kit (or equivalent) is constructed during the planning stages of the study. Ideally the development of the kit is an iterative process involving extensive pilot testing. In our collective experience this pilot testing is indispensable, as it is never possible to anticipate every possible problem or the rich variety of participating subject's responses.

1. Question formulation
2. Protocol
a. Data collection specification
b. Human Subjects materials
c. Background questionnaire
d. Discriminator question
e. Specification of set-up
f. Experimenters script (including guidance on notes/diagramming)
g. Participant design brief
h. Design criteria elicitation Stimuli set
i. Design criteria elicitation Recording Sheet
3. Analysis protocol
4. Background
5. Literature

Figure two: A typical table of contents from a Bootstrapping-model Experiment Kit. Copies of core literature papers would also be included.

The Experiment Kits used in studies of the Bootstrapping-model included material which describe the study's focal questions and the reasoning behind them. It describes the pilot studies undertaken (usually by the PIs) and situates the work in the context of relevant literature and underlying theoretical and methodological approaches and assumptions. An Experiment Kit should contain everything an individual researcher needs to understand and undertake their portion of the study, including all material to be given to the study participants, copies of papers which are core reference material for the investigation (as well as pointers to further reading), specification of the format in which data is to be collected, a specification of the information required about each participating institution (see the discussion of institutional characterisations in Section 3.1.2) and some indication of the types of analysis that will be undertaken.

An Experiment Kit of this type has pragmatic benefits in providing a communication tool between a distributed team of researchers, and in being detailed enough to allow relatively inexperienced researchers to participate. It has conceptual benefits in providing a common frame of reference for individual researchers, and the practical benefit of forming a useful foundation for any further writing arising from the study.

Other communications: Methods of communication for the research team will almost certainly include email and the distribution of resource material electronically (e.g. via a dedicated web site). They may also include meetings such as the workshops which characterised the Bootstrapping model studies, and the analysis phase of the Working Group model. The issue of long lead times, which is important with respect to planning and piloting MIMN studies, is also significant when it comes to communications. The timetables and processes of individual institutions vary widely, particularly over different countries and, especially, hemispheres. Consequently, in MIMN work, individual researchers may at any given calendar time be at very different stages of the study process – some perhaps finished collecting data and enjoying Summer Vacation before others have even begun. These factors emphasise the requirements for both well planned communications and a clearly specified timetable for the study.

Other practical considerations which may be significant for larger teams include the possible constraints imposed by participating institutions. These can range from the minor, such as requirements for multiple acknowledgments, to the major, such as ethical procedures requirements (including constraints on the use of data). Issues relating to the authorship of any papers resulting from the

study, and to the "ownership" and possible further use of the study data must also be discussed and resolved.

In short, the involvement of multiple researchers underlies many of the strengths, and many of the weaknesses of MIMN studies. More researchers extend the range of institutions covered and add to the number of participating subjects, but they also add to the complexity and communication overheads of the study. In this respect it may be interesting to compare MIMN studies to other large distributed team processes, such as perhaps the open source software development model, or distributed research projects such as the human genome project.

Collective Sensemaking: When knowledge about the study is distributed across researchers, how do individuals and the collective of researchers, come to make sense of their data? Answering this question involves the pragmatics of data collection and analysis, and organization of the research team. While many aspects of data collection can be taken for granted by the lone researcher, virtually all aspects of the data are subject to differences in interpretation among the collaborating researchers. Points where different interpretations are possible include:

- Which subjects qualify to participate?
- How will material be presented to subjects, and what follow-ups will be made. Will some data in the corpus be translated from other (natural) languages?
- What sorts of events, utterances, behaviours, and observations are to be recorded out of the "blooming and buzzing confusion" that characterizes human activity? And if part of this activity is an interaction with other people or technologies, how much of this interaction will be captured? If a researcher has to make a choice about what is important in an interaction, do all agree on what the important points are?
- If subjects require clarification, how much information will be provided, of what sort? Will all of the "marks on paper" that subjects make be saved and disseminated, and if so will the translation to electronic form result in data loss? How should "field notes" be taken, and should they be transcribed and disseminated?

Carrying out MIMN studies requires researchers to make these data characteristics explicit, (although data characterisation explicitness of this nature has significant advantages for the lone researcher as well). When reported along with the study data it provides crucial knowledge to others about both the study's limits and its generalisability. What one researcher might take for granted and hence not explicitly describe (e.g. that everyone knows what "CS1" means, or that all questions that participants asked have been dutifully recorded) become problematic almost immediately in MIMN studies. Discussion of these issues can highlight the ambiguity inherent in many single-researcher studies, i.e. are the results specific to a particular institution, to students having taken a particular course with a particular instructor, or do they generalize across most or all individuals learning to program?

In looking at the design of the MIMN studies to date, the main tradeoffs with regard to collective sensemaking concern agreed protocols, roles within the team and longevity of the collaboration.

Agreed protocols: The protocols trade-off concerns the effort that goes into specifying the data and analysis protocol prior to data collection versus the time taken afterward to make sense of it (where data might have been collected under different assumptions). Associated with this is the cost of identifying, and discarding, incomparable data. It is likely that variation from one MIMN study to the next is a result of learning on the part of the principal investigators more than differences in research philosophy. For example, McCracken et al. ([2] p.136) caution that "Another important challenge is making the exercises sufficiently general so that they are neutral with respect to both culture and the university." Heeding this advice, the Bootstrapping PIs took considerable pains to provide explicit instructions for data characterisation for the Bootstrapping researchers, included in the Experiment Kit. Despite efforts such as this, in most of the MIMN studies to date, some of the data has not been usable as a result of ambiguity and misunderstanding among the collaborating researchers.

Roles within the team: Sensemaking depends on who is available to do this work. In the MIMN studies conducted thus far, certain key roles have emerged, though not all of the studies have involved all roles.

- The Principal Investigators The principal investigators (PIs) take primary responsibility for determining the study's focal question, for designing and pilot testing the data collection and analysis protocols, for coordinating the research team's activities, and for overseeing dissemination of research results. Other responsibilities might include recruitment of the research team and obtaining research funding.
- Data Coordinator The data coordinator collects the data, checks the data for integrity (Do all required fields have valid data?), and maintains a documented data archive. The data coordinator might also write or use specialized software tools for converting from one format (e.g. SQL queries) to another (e.g. comma-separated text files). Finally, the data coordinator will often provide up-to-date descriptive statistics on demographic data of the entire population and specific subpopulations.
- Individual Investigators Individual researchers conduct the research at each location. Typically they will administer the tools or treatments of the study, and manage interventions or record observations with participants. They will probably, though not necessarily, be engaged in subsequent analysis and interpretation of the resulting data. In larger institutions a site co-ordinator may be necessary to oversee the work of individual investigators.

Longevity of the collaboration: Investment in organizational infrastructure is strongly affected by the duration that the research team is anticipated to work together. In the Bootstrapping model, the research commitments were for a minimum of two years, and so there was considerable investment in development of the research team, and some amount of fluidity between different tasks and the individuals who oversaw them to completion. Some of the responsibilities for writing, data coordination, recruitment, and fund raising could be distributed among a number of the participants, especially as participants developed increasing expertise. In the Working Group model, the short, focused nature

of the interaction means that leadership is more centralized, and a single individual carries more of the roles identified above.

There are further tradeoffs in terms of sensemaking dependent on the relationship between these roles. No project to date has had all researchers participate in the design phase: PIs have always been a limited subset. Sometimes PIs have also participated as investigators, sometimes not. In some projects (ExploreCSEd) individual investigators just fill the role of data-gathering research assistants and play no part in the analysis and write-up of the work (and are also not available to disambiguate problems). In other projects (Bootstrapping model) individual investigators may begin in a supplementary role to the PIs, but by the time the data has been collected, joint analysis is undertaken, and results collectively written up, the role of the individual investigators has grown in to that of a genuine collaborator. This allows the common identification of appropriate models and theory, the collective agreement on aspects of importance to be emphasised when writing, and an evolutionary development of the sense of the data.

3.1.2 Multiple participants (at multiple institutions):

The complexities of managing large numbers of participants at multiple institutions shape many of the requirements of MIMN studies. In this section we will focus on matters relating to planning and preparation. Matters relating to collecting and analysing the data from multiple participants are discussed in Section 3.2 below.

Ethical approval: The process of obtaining ethical ("IRB") approval for the involvement of human subjects is, in our experience, made much more complicated when multiple institutions are involved. The process at each institution is naturally geared towards studies based solely at that institution, and naturally focuses on its own particular requirements. There is typically a substantial emphasis on how the data will be collected and subsequently protected. The specific requirements of institutions can vary wildly however. For example, in some institutions, once audio recordings have been transcribed, it is a condition that they be destroyed, while at other institutions it is a condition that the recordings be retained for several years. The requirements of all institutions must be met or in some way resolved if the study is to proceed as planned. As a matter of course it will be prudent for the principal investigators to collect all approval letters from each of the individual investigators.

Initiating the ethical approval process at their institution should be the first task undertaken by each individual investigator as they join the study team. This approval is a critical path item and can cause delays of a semester or more in commencing the study at each location. The requirement for informed consent can also dictate whether students perform the study as an integral component of the course delivery and assessment or as a voluntary and somewhat peripheral extra activity. This can significantly impact on the quality of the results from the study, especially in institutions where students are motivated primarily by summative assessment.

To safeguard the privacy of participants and institutions the principal investigator will need to design and promulgate a coding system to guarantee anonymity of respondents from the outset. This may as simple as assigning a one-letter code to each participating institution. While institutions that are part of the collaboration may know the codes of the other institutions, this information is not divulged outside the group. Even within the collaboration, the identity of the individual participants will not be divulged outside their own institution.

To avoid pressure on researchers from outside sources, it is also advisable that it be made an explicit condition of IRB approval that the data not be used for inter-institutional comparison and external promotion via "league tables".

Institutional characterisations: In order to help interpret the data from participants at each institution we need to know background details such as the kind of qualifications offered, the nature of the student population, the numbers of students enrolled in the relevant courses, the grading system used, and perhaps relevant details of specific courses (such as the language taught in a CS1 course, "objects early" vs. "objects late" and so on). Such institutional characterisations are useful for both the investigators, to help deal with issues of replicability and generalisabilty of the results; and for the audience for the study, to allow practitioners to assess whether the results are relevant to their own context.

Selection of participants: Most studies will involve volunteer participants. Depending on the size of the institution and the motivation of the students some investigators may have more volunteers than they need (or are able to cope with), but in our experience it is more likely that some investigators will have too few volunteers. In practical terms these investigators may need to offer a form of inducement or payment to participants (which should be recorded in the institutional characterisation), or draw on participants from other nearby institutions.

Most CSEd studies are aimed at students at a certain stage in their development. However, in a MIMN study it is not usually possible to simply equate a given stage of development with a specified stage of a degree program (e.g. "finished their first CS course") due to national and institutional variations in the organisation of the curriculum. For this reason, many MIMN studies to date have specified "stage of development" operationally, for example that students should have achieved some level of competence such as being capable of writing a program of some well-defined level of complexity. Some studies have attempted to reduce the significance of institutional context by using data collected in one part of the study to characterize students into groups (such as low, medium and high performance), then used that characterization as a basis for analysing performance on other tasks contained in the study: this method has become known as the "two task" approach.

Grades: Grades are often used as a measure of performance (typically the variable that we are trying to predict or influence), and can also be used as a basis for dividing students into groups. Once again, institutional variations make achieving consistency difficult. A specific grade like "B+" or "4.0" may have different interpretations in different national and institutional contexts. Even with pragmatic definitions such as quartiles there is no guarantee that top quartile students at one institution are equivalent to top quartile students at another. Here again a rich institutional characterisation may help to determine appropriate interpretations.

Clearly in general the more subjects at each institution the better, but there are trade-offs with the work required of the investigators and the complexity of the data processing and analysis. While one advantage of a MIMN study is that the costs of collecting data on many subjects are distributed, if the study design compares populations across institutions then this still requires as large a sample size as possible at each location.

3.2 Nature of MIMN data

3.2.1 Collection

Reliability: MIMN studies naturally highlight issues of inter-rater reliability in the collection of data. Some tools, such as questionnaires, are comparatively easy to administer reliably, while others, such as recording observations of behaviour, are notoriously difficult. The reliability of the observations collected by the investigators can be improved by training, by iteratively developing the study tools, by the used of a detailed "script" describing the data collection process, and by the use of explicit checks for inter-rater reliability wherever possible in the data collection and / or analysis process.

The training process may lead to further iterative development of the tools or the script for applying them, but this can only occur as part of the preparation / pilot study phase. Once data collection begins in earnest it becomes very difficult to alter the tools or processes without invalidating the data already collected.

Data cleanliness: Data collected may be of many different types, such as background demographic information, recorded times, completed materials or artefacts produced by the participants, or the observations and notes of the investigator. The techniques involved in managing, organising, coding and preparing large amounts of diverse data for analysis require considerable discipline on the part of the investigators. All notes and artefacts should be labelled with the participant's code. Specified formats for recording data should be followed exactly. Any ambiguities noted should be resolved right away in consultation with the PIs, and relevant decisions disseminated to the whole team if necessary. Communication within the research team is vital, to share experiences, ask questions, and come to a common understanding of the data collection process.

Not only is primary data a source of ambiguity, in MIMN studies management of secondary data is equally problematic. When faced with writing up results several months (or even years) after data collection, the lone researcher can rely upon idiosyncratic mnemonics for locating the important scraps of papers and files that provide the audit trail for specific inferences. Searches through particular piles on the desk, or in appropriately named directories, often locate the files holding secondary data analysis, perhaps with cryptic notes that trigger memories about assumptions hastily made. But a MIMN study cannot rely upon the idiosyncrasies of individual investigators. The trade-off is between keeping an explicit audit trail for all secondary analysis (and archiving it along with primary data to the central data repository), versus having to redo analysis work later should the need arise, perhaps based on different assumptions than the original analysis.

In short, managing MIMN data involves constant trade-offs between the effort and discipline of the individual investigators and the reliability and integrity of the final data set. The looser the processes the less confident one can be about the accuracy of the data. This trade-off is by no means unique to a MIMN study, but given the size and complexity of the data set, and particularly the involvement of many different investigators, the effects of the trade-off are greatly magnified.

3.2.2 Analysis

Character of the data MIMN data often has characteristics that affect analysis choice. Without adequate characterization of the institutional and instructional context, analysis through identification of sub-populations by external variables (such as age, gender, academic performance, institution etc.) becomes problematic, and researchers must then rely on internal variables (e.g. performance on the study task) or within-corpus characteristics (such as the "two task" approach, as above).

MIMN studies frequently use multiple instruments to gather data. However, equally frequently, not all study participants complete all parts of the study, whether by attrition (being present at the start of the study, but not at the end), inclination (declining to participate in certain tasks) or researcher focus (gathering survey data from a larger population than interview data). For whatever reason, care must be taken in analysis that the appropriate section of the cohort is being studied, and especially if comparative judgements are drawn.

Choice of analysis techniques In MIMN studies, there is a danger that both the size of the project and the lack of shared paradigms may drive the study toward statistical inference, upon counting and comparing decontextualized study data. This stems not from an epistemological commitment of statistical analysis over other kinds of analysis, but more from the pragmatics associated with the issues of scale detailed above. Alternative methods (e.g. ethnographic, observational, unstructured interviews) might yield deep insights into student thinking, such as the metaphors of computation that novices bring to programming. However these alternative methods also require negotiation and interaction between the researchers, as well as iterated interaction with the data, often through several theoretical lenses. However, this communication may exceed the bandwidth available that must be maintained for negotiated understandings among a large group of researchers. Statistical analysis and inference, on the other hand, require far less communicative overhead among the researchers—no messy data coding, no inter-rater reliability—and far less interaction with the data. Additionally, statistical analysis has well-defined analysis procedures: my t test will give the same answer as your t test. A multi-institutional study can easily and implicitly become a matter of collecting and statistically analysing large amounts of data but with little understanding of what these results mean, simply because this might be the only way that a set of researchers with few shared beliefs about method can see to proceed, especially if faced with a large amount of data to analyse in a short period of time.

Several of the authors of this paper experienced these difficulties with the Bootstrapping study. The initial statistical analyses of the study data (reported in [20]) were insufficient to answer the study questions. It was only the subsequent qualitative analysis performed by the South Carolina group, using data collected via the same types of instruments as the original Bootstrap study — but fewer researchers interacting over an extended period of time, both face-to-face and via email—that provided the richer kinds of data needed to answer the original study's questions.

Data ownership: At the end of a project's life, what happens to the data? To some extent this is constrained by ethical approval considerations. There are several possibilities: (1) data remains within the research team, never to be released to others, (2) the investigators can recruit outside collaborators for "break off" studies, or (3) suitably anonymised data is released publicly. Over time, the data may move from tighter to weaker restrictions. For example, one year after the creation of the joint dataset, the

restriction could move from (1) to (2). Whatever the option, the collaborators need to have an agreed policy on ownership, either defined within the Experiment Kit by the PIs, or discussed and agreed to by all investigators.

4. CASE STUDIES

In this section, we take each model identified in figure one. We examine the particular study that was undertaken and discuss the costs and benefits of each approach. Where a model has evolved (or been adapted from) a previous version, we exemplify the points of distinction.

4.1 Working Group studies (McCracken group and Leeds group)

Whilst the McCracken group examined programming performance (the model is described in section 2.1, above), the Leeds group examined programming comprehension. The test instrument for the Leeds group study consisted of twelve multiple choice questions that had been used in past exam papers at one of the institutions. Three types of data were collected: (1) performance data, the answers chosen by each student; (2) doodle data, the calculations made by each student as they answered the questions; (3) transcript data, from students who thought out loud as they solved the multiple choice questions. The analysis of the data collected was a mixture of quantitative and qualitative.

2004 adaptations: Like the Bootstrapping project, but unlike the McCracken working group, all members of the Leeds working group were required to collect data. This entry criterion ensured that all members shared a strong commitment to the project, and they all had a "feel" for the data that comes from having collected data. And although the project team met for the first (and only) time over several days at the ITiCSE conference in Leeds in June 2004, of the twelve participants in this study, eight had already participated in the Bootstrapping, Scaffolding, or BRACE projects. These eight collaborators brought to the working group a substantial set of shared beliefs about CSEd research and the execution of a MIMN study.

Costs: The Working Group model had a relatively high cost on PIs, as they have to do design, data cleanliness and most of the writing.

Benefits: There is a relatively low-cost for individual participants, but all share the rewards of MIMN data. With appropriate agreements, there can be follow-on studies from any member of the group.

References: [2, 6, 21]

4.2 Bootstrapping-model studies (Bootstrapping, Scaffolding and BRACE)

The form of this model is described in section 2.2, above. The content of the Bootstrapping project was a single instrument (an open card-sort of 26 programming concepts) to elicit "first competency" programmers' construction of programming concepts. Each individual researcher decided when students at their own institution were capable of successfully completing the programming task set in the McCracken Working Group: this determined the point of "first competency". (Researchers did not administer the task, they just had to identify the point they believed that the students were capable of completing it).

Scaffolding adaptations: the Bootstrapping-model was instantiated a second time, one year later, in the Scaffolding Computer Science Education Research project. The second instantiation was, to all intents and purposes, a replication: the organisers, workshop leaders, location and funding body were the same. However, a different MIMN project was devised. In Scaffolding, multiple instruments were used to elicit students understanding of software design. Students were recruited at the point of "first competency" and at the point of graduation. Instruments included: a design diagram (created to a brief), a design criteria prioritisation task and semi-structured interviews. These multiple instruments provided diverse types of data that afforded richer opportunities for analysis. Additionally, for Scaffolding, the role of dedicated data-coordinator was identified.

BRACE adaptations: Building Research in Australasian Computing Education was the third instantiation of the Bootstrapping-model, and shared the same roots. (One of the workshop organisers was a graduate of Bootstrapping, the second of Scaffolding). A surface difference was that participants in BRACE were self-funded. As had happened in Scaffolding, the instruments for the MIMN study were expanded. For BRACE, they included attitudinal, behavioural and cognitive tasks: the attitudinal and cognitive tasks were previously validated instruments affording an additional point of comparability and reliability to the results.

Costs: The Bootstrapping-model has high costs for PIs, who have to design and pilot the MIMN study, write the Experiment Kit which contains it, and structure and populate the workshops in which it is presented—amongst other activities. For participants, the focus on other workshop activities can detract from the results and reporting of the MIMN study. For the participants there is also a considerable chronological and intellectual commitment.

Benefits: The extended period of interaction (one-to-two years) and the reliance of individual researchers on each other to gather data and conduct analysis, allows extended research relationships to develop. These frequently extend beyond the life of the project and form seedcorn for a researcher's community.

References: [3-5, 16, 19-27]

4.3 BRACElet

As its name suggests, the BRACElet project was inspired by BRACE. However, in some respects BRACElet is closer to the Working Group model, thus providing a hybrid form. The first formal meeting of BRACElet occurred in December 2004. A call for participation in the two-day meeting was distributed among New Zealand tertiary education institutions, and 11 institutions sent representatives. There was also one invited attendee from Australia. Two of the attendees were veterans of both BRACE and the Leeds Working Group.

The (currently ongoing) project has the goal of understanding the process of program comprehension by novice programmers, in order to provide a sound base for the subsequent investigation of program writing skills.

The study adopted the Leeds Working Group study design. But the group together decided to adopt the revised Bloom's taxonomy [28] as both a means for analyzing the instrument used by the Leeds group, and as mechanism for generating a new instrument. It is intended that use of the revised Bloom's taxonomy may help to "unpack" the constructs inherent in program comprehension.

	McCracken WG	Leeds WG	Boot/Scaff/BRACE	SC	BRACElet	ExploreCSEd
Researcher Recruitment	Selected colleagues and normal WG solicitation	Normal WG solicitation & colleagues from previous study	Funded participation,. Selection (except BRACE)	Colleagues from previous study	Sectorally based – (NACCQ) - NZ Tertiary computing educators & colleagues from Leeds WG	Open Call
Introduction of study to research team	Remote	Remote	Guided (capstone to 4-day workshop)	Already familiar	Presented 2004, refined 2005 during pilot studies	On request
Data	Single source (programming task)	Multiple (answers to MCQs, "doodles" made by students, and "think out loud" transcripts).	Initially single source. Over time, move to mixed instruments	Two related sources (constrained and open card sorts)	Multiple (categorised MCQ's, short answer questions and "doodles")	Two instruments
Analysis	Statistical followed by qualitative (to gain insight into the reasons why particular statistical regularities were observed)	Statistical and qualitative	Statistical and qualitative	Statistical and qualitative	Statistical and qualitative	Only statistical, only undertaken by research leaders. Participants may have access to own data.
Follow-up work	No	Some additional analysis with a sub-set of participants, and BRACElet	Yes, lots. Benefit of "community building" approach	Yes, and is itself a follow up (from Bootstrapping)	Is itself an adaptation/extension of Leeds WG	
Model strengths & weaknesses	Relatively low-cost on individual participants, but all share the rewards of MIMN data. Relatively high cost on PI, as they have to do design, data cleanliness and most of writing.		Interdependence builds collegiality & common view. Relatively high cost on individual participants (two-year commitment).	Builds on shared method and collegiality from Bootstrapping. High cost to carry out the qualitative research and to write at-a-distance, especially with no pre-defined PI.	Builds upon prior work and shared expertise with key common participants. Enables mixture of novice and experienced researchers to work together. Costs for PI's in coordinating, hope to share data analysis load and writing.	Very low cost on participation, but very few benefits of participation. Very high cost on PI, have to do all design, data cleanliness, all analysis and writing.

Figure three: Some dimensions of comparison across some CSEd research MIMN studies

The study also adapted the Leeds instrument to cover multiple programming languages and their distinct idioms, and added some short answer questions to gain added insight.

Costs: The Working Group model has had a relatively high cost on PIs in coordinating the study and its design, but it is hoped that the data analysis and writing load will be shared between the project members.

Benefits: As with the Bootstrapping model, participants benefit from extended interaction, over multiple meetings. Unique to BRACElet is the individual researchers co-designing the instruments, and thus aims, of the project.

References: [18]

4.4 South Carolina Group (or "BootTwo")

This group self-organized at the end of the Bootstrapping workshops with the aim of comparing the Bootstrapping data on first-competency students with the performance of graduating seniors on the same task: an open card sort on a set of 26 terms representing programming concepts. The study question was identical to that of Bootstrapping. The researchers also added an additional task of constrained card sorts on the same 26 terms using researcher-provided category names relating to the point at which the student first encountered and mastered the programming concept and the perceived level of difficulty. Different researchers from among the participants played leading roles at various times throughout the collaboration, though none served as PI throughout the project in the sense described above.

Costs: Costs for individual researchers has been relatively high. Although the group was able to leverage the shared framework and methodology from the Bootstrapping project, they undertook a challenging qualitative data analysis on a large data corpus. This required a new round of data collection on a different study population, a dedicated week face-to-face (at a rented house in South Carolina, hence the group name), as well as a large number of iterations of data interpretation and writing in a distributed, electronically mediated fashion over several months. An increased overhead stemmed from the group's not having explicit, pre-defined PIs since this was a collaboration "among equals".

Benefits: The benefits have been relatively high for the individual researchers, both in strengthening ties among the collaborators, and in the consistent research results that this group continues to produce.

References: [7, 17]

4.5 ExploreCSEd

ExploreCSEd, funded by a small grant from the UK Higher Education Academy, is the most recent model which we examine. The study is currently ongoing, and has been structured to gather considerable quantities of data. The study has been designed by four PIs and uses two instruments (1) a skills survey to "discover the skills that students and educators believe are the most important in learning to program" and (2) a difficulties questionnaire "to identify difficulties faced by students in their programming module". The difficulties questionnaire is adapted to each course, with regard to programming language taught, order in which concepts are presented etc. Additionally, background data is collected on each participating institution (including information on the structure and content of the programming module) and on the background of each student participant. All information is gathered via online questionnaires.

The project has a central, coordinating document, detailing background, protocols, ethics approval etc. (here called a "Toolkit"). An interesting feature of ExploreCSEd is that the Toolkit is disseminated and distributed at disciplinary conferences and workshops, inviting a wide participation. However, participating institutions may only have access to their own data (if ethical approval documents are completed) either raw, or as analysed by the PIs.

Costs: The ExploreCSEd model has a very high cost on the PIs who are solely responsible for design, data cleanliness, all analysis and writing. There is little incentive for individual institutions to participate—only access to their own raw data, and the results of the PIs analysis of their contribution

Benefits: For the PIs, there are benefits of collating a MI(MN) dataset for their own use. There is no especial time pressure on analysis, and, as only the PIs are involved, only slight overhead for collective sensemaking,

References:
http://mathstore.ac.uk/news/april05/ExploreCSEdToolkit.pdf,
http://www.rebeccamancy.net/ExploreCSEd/Index_E.php

5. A WIDER CONTEXT

We have, in this paper, taken a deep look at a particular "family" of MIMN studies, as they have emerged in the CSEd research context. Yet there is a wider context in which these reside.

There are parallel multi-institutional (although not, as yet, multi-national) developments in Engineering Education research, which have different forms and emphasis. For example, the Conducting Rigorous Research in Engineering Education workshops have a single co-located meeting where participants develop individual studies. They continue working on these after the workshop on a one-to-one basis with a research mentor and a grant of $2,000. The CAEE Institute for Scholarship on Engineering Education has a mix of faculty and graduate students who each undertake individual studies set in the context of the investigation of a learning issue derived from their own teaching.

Beyond disciplinary-specific education research, there are multi-national studies that necessitate large distributed team processes, such as the Open Source software development model, and distributed research projects that require similar co-ordination of researchers and data, such as the human genome project.

As CSEd MIMN studies mature, it is likely the forms we have documented here will continue to hybridize and adapt with influences from these, and other, disciplines and other endeavours.

6. ACKNOWLEDGEMENTS

The authors thank their collaborators in the MIMN studies that led to the writing of this paper. Part of this material is based upon work supported by the National Science Foundation (NSF) under grants numbered DUE-0243242 and DUE-0122560. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Our thanks to Dennis Bouvier who (in Scaffolding) made it obvious that the data coordinator role was pivotal for success, and who so successfully made it obvious.

BRACE was supported, in part, by a small project grant from ACM SIGCSE. For the BRACElet project the contributions from Jacqui Whalley co-principal are gratefully acknowledged, as is the financial support from AUT's School of Computer & Information Sciences in enabling the inaugural workshop to be held.

7. REFERENCES

[1] Porter, A.C. and A. Gamoran, eds. Methodological Advances in Cross-National Surveys of Educational Achievement. 2002, National Academy Press: Washington DC.

[2] McCracken, W.M., et al., A multi-national, multi-institutional study of assessment of programming skills of first-year CS students. SIGCSE Bulletin, 2001. 33(4): p. 125-180.

[3] Petre, M., et al., "My criterion is: Is it a Boolean?": A card-sort elicitation of students' knowledge of programming constructs. 2003, University of Kent: Canterbury. p. 37.

[4] Fincher, S., et al., Cause for alarm? A multi-national, multi-institutional study of student-generated software designs. 2004, Computing Laboratory, University of Kent: Canterbury.

[5] Fincher, S., et al., Programmed to succeed? A multi-national, multi-institutional study of introductory programming courses. 2005, Computing Laboratory, University of Kent: Canterbury.

[6] Lister, R., et al., A multi-national study of Reading and Tracing Skills in Novice Programmers. SIGCSE Bulletin, 2004. 36(4): p. 119-150.

[7] McCauley, R., et al., What do successful computer science students know? An integrative analysis using card sort measures and content analysis to evaluate graduating students' knowledge of programming concepts. Expert Systems, 2005. 22(3): p. 147-159.

[8] Gilford, D.M., ed. A Collaborative Agenda for Improving International Comparative Studies in Education. 1993, National Academy Press: Washington DC.

[9] Mathiassen, L., Collaborative practice research. Information Technology and People, 2002. 15(4): p. 321-245.

[10] McKay, J. and P. Marshall, The dual imperatives of action research. Information Technology and People, 2001. 14(1): p. 46-59.

[11] Taylor, P., Institutional Change in Uncertain Times: Lone ranging is not enough. Studies in Higher Education, 1998. 23(3): p. 269-279.

[12] Eliis, A., et al. Resources, Tools, and Techniques for Problem Based Learning in Computing. in 3rd annual SIGCSE/SIGCUE ITiCSE conference on Integrating technology into computer science education. 1998. Dublin, Ireland: ACM Press.

[13] Lawhead, P., et al., A Road Map for Teaching Introductory Programming Using LEGO Mindstorms Robots. SIGCSE Bulletin, 2003. 35(2): p. 191-201.

[14] Carter, J., et al., How shall we assess this? SIGCSE Bulletin, 2004. 35(4): p. 107-123.

[15] Almstrum, V., et al., Evaluation: turning technology from toy to tool. SIGCSE Bulletin, 1996. 28(S1): p. 201-217.

[16] Bootstrapping, Bootstrapping Research in Computer Science Education. 2002. http://depts.washington.edu/bootstrp/

[17] Murphy, L., et al. A multi-institutional investigation of computer science seniors' knowledge of programming concepts. in SIGCSE Symposium. 2005. St Louis, MO.

[18] Whalley, J., Report on the BRACElet Workshop. 2004, Auckland University of Technology: Auckland.

[19] Experiment-Kits, Experiment Kit Page. 2005. http://www.cs.kent.ac.uk/people/staff/saf/experiment-kits/index.html

[20] Sanders, K., et al., A multi-institutional, multinational study of programming concepts using card sort data. Expert Systems, 2005. 22(3): p. 121-128.

[21] McCartney, R., et al. Questions, Annotations, and Institutions: observations from a study of novice programmers. in Kolin Kolistelut - Koli Calling. 2004. Koli, Finland.

[22] BRACE, Building Research in Australasian Computing Education. 2004. http://www.cs.otago.ac.nz/brace/

[23] de Raadt, M., et al. Approaches to Learning in Computer Programming Students, and its Effect on Success. in Annual International Conference of the Higher Education Research and Development Society of Australis (HERDSA 2005). 2005. Sydney, Australia.

[24] Deibel, K., R. Anderson, and R. Anderson, Using edit distance to analyze card sorts. Expert Systems, 2005. 22(3).

[25] Fincher, S. and J. Tenenberg, Making sense of card-sorting data. Expert Systems, 2005. 22(3).

[26] Fossum, T. and S. Haller, Measuring card sort orthogonality. Expert Systems, 2005. 22(3).

[27] Tenenberg, J., et al., Students designing software: a multi-national, multi-institutional study. Informatics in Education, 2005. 4(1): p. 143-162.

[28] Anderson, L.W., et al., eds. A taxonomy for learning, teaching, and assessing : a revision of Bloom's Taxonomy of educational objectives. Complete ed. 2001, Longman: New York. xxix, 352.