Characterizing Research in Computing Education: A preliminary analysis of the literature

Lauri Malmi
Aalto University, Finland
Lauri.Malmi@tkk.fi

Judy Sheard
Monash University, Australia
judy.sheard@infotech.monash.edu.au

Simon
University of Newcastle, Australia
simon@newcastle.edu.au

Roman Bednarik
University of Eastern Finland
roman.bednarik@cs.joensuu.fi

Juha Helminen
Aalto University, Finland
juha.helminen@cs.hut.fi

Ari Korhonen
Aalto University, Finland
archie@cs.hut.fi

Niko Myller
University of Eastern Finland
nmyller@cs.joensuu.fi

Juha Sorva
Aalto University, Finland
Juha.Sorva@iki.fi

Ahmad Taherkhani
Aalto University, Finland
ahmad@cc.hut.fi

ABSTRACT
This paper presents a preliminary analysis of research papers in computing education. While previous analysis has explored what research is being done in computing education, this project explores how that research is being done. We present our classification system, then the results of applying it to the papers from all five years of ICER. We find that this subset of computing education research has more in common with research in information systems than with that in computer science or software engineering; and that the papers published at ICER generally appear to conform to the specified ICER requirements.

Categories and Subject Descriptors
K.3.2 [Computers and education]: Computer and Information Science Education – computer science education; A.0 [General]: conference proceedings.

General Terms
Measurement.

Keywords
Classifying publications, computing education, research methods.

1. INTRODUCTION
Computing education is a relatively new area of research with a growing body of literature. It is multifaceted and draws on a range of disciplines with more established traditions of research (e.g. cognitive psychology, education, and behavioral science). As computing education research matures as a discipline it is important to understand what characterizes this area of research and where it is positioned in the research milieu. This will help establish computing education as an identifiable and acknowledged field of research, increasing the credibility and acceptance of its research findings. For this to happen we must be aware of the type of research that is conducted in computing education and the theoretical bases upon which this research is founded. One way to achieve this awareness is through analysis of the computing education research literature.

In recent years a number of studies have examined the literature of computing education research (CER). Some of them have concentrated on identifying the subfields of CER [13, 23]. Later work investigated the characteristics of CER, forming an overview of what kind of work is being carried out in the field and commenting on the quality of work [16, 27-29, 33-36]. Some meta-analysis has focused more narrowly on the research literature on the teaching and learning of programming [24, 32, 38]. All of these studies have provided valuable overviews of research in computing education and have contributed to our understanding of the current state of the art of CER.

An important motivation for this kind of work is to improve the quality of research by recognizing good practices and exploring what might qualify as rigorous research in different contexts. Analysis of this sort also assists in the making of a clearer distinction between practice reports and research papers. Practice reports focus on disseminating new ideas, tools and approaches among computing educators. Some research papers aim at rigorously testing such novel ideas to evaluate their effectiveness and generalizability, while others set out to build a deeper understanding of students’ and teachers’ conceptions and working processes.

In this paper we continue this classification work by looking more deeply at the research processes documented in CER papers. We present a purpose-built classification scheme and demonstrate its use with analysis of papers from five years of ICER workshops. Specifically, we ask:

- What kind of data has been collected for the research and how has it been analyzed?
- What theories, models, frameworks, instruments, technologies or tools have been used, built upon, or extended by the research?
- What other disciplines does the research link to?
- What was the general purpose of the research: describing something new, formulating new tools or methods, or evaluating them?
- What research frameworks have been used?
- What kind of data has been collected for the research and how has it been analyzed?

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 2010 ACM 978-1-4503-0257-9/10/08...$10.00.
Through this analysis we hope to get a better understanding of how computing education research builds on previous work, what features it uses from other disciplines, what purposes the researchers have, and how they address their research questions.

2. Research Method and Analysis Approach

Our investigation of the CER literature entailed the development of a purpose built classification scheme. We based our scheme on an existing one designed by Vessey, Ramesh and Glass [40]; even so, developing and implementing it proved to be a non-trivial task.

The two principal authors developed a draft classification scheme of seven dimensions and presented it to the other authors at a workshop. Particular papers were discussed in depth at the workshop, culminating in consensus classifications of those papers in all seven dimensions. Participants then classified a trial set of papers which were discussed at two further workshops, resulting in a number of clarifications and several modifications to the scheme. A second trial set of papers was then classified and discussed at a fourth workshop.

All classifications after the initial round were done in pairs, with each pair classifying a set of papers individually, meeting to compare and discuss, and then submitting an agreed list of classifications. Previous work by Simon et al. [35] found classifying in pairs to be more reliable than classifying singly, as it helped to eliminate simple errors and focused discussion on papers that were genuinely difficult to classify.

Following the initial classifications we conducted an inter-rater reliability test, using the Fleiss-Davies kappa [11] for the case where all of the raters are rating all of the items. To be amenable to such a test, a dimension must consist of a fixed set of categories and each item being classified should be placed into just one category. Not all of our dimensions meet these requirements, but we were able to adjust four of the seven dimensions for the purposes of the test.

On this and other kappa measurements, an agreement of less than 40% is generally considered to be poor; between 40% and 75% is considered fair to good; and more than 75% is rated excellent [3]. With four pairs of raters rating the same set of 28 papers, our agreements on four dimensions were 68%, 60%, 55%, and 65%. These values are generally closer to the top than to the bottom of the fair to good range and suggest that in these dimensions we are classifying the papers fairly reliably. In the absence of a formal measure for the other three dimensions, we can only assume that our classification in those dimensions is as reliable as in the ones that we were able to test.

Simon et al. [35] reported inter-rater reliability of 55% to 80% on different dimensions when classifying in pairs. Vessey et al., classifying individually, reported reliability “roughly at the 60% level” in 2002 [15], improving to agreements in the 75%-85% range in 2005 [40]. Our results are comparable with those of Simon et al. and with the early results of Vessey et al., suggesting that our system is comparable in reliability with theirs. It seems reasonable, too, to expect that our inter-rater reliability will improve as we become more familiar with our scheme and its application.

We now considered that the classification scheme was stable and we had sufficient agreement and understanding of the scheme to begin our study. From this point onward, each pair classified a different set of papers.

2.1 Selection of Papers for Analysis

While CER papers are found in many journals and conference proceedings, for this study we decided to focus on long papers, as opposed to the 5-page papers found in SIGCSE and ITICSE. The extra length of an 8-to-12-page paper encourages discussion of theoretical background and methodological issues as well as presentation of experimental results. Longer papers can also allow reporting of qualitative studies that typically use more space than quantitative papers to explain their background and approach. This decision limited our initial pool of data sources to journals, ICER, the Australasian Computing Education Conference, and Koli Calling.

We decided to apply our classification scheme to an analysis of papers from the International Computing Education Research workshop (ICER). In the five years it has been running, ICER has clearly established itself as a leading forum for CER. A previous analysis of ICER papers [35] found that ICER had the highest percentage of research papers among computing education conferences. The call for papers for ICER states that papers for the ICER workshop should, as appropriate, display “A clear theoretical basis, drawing on existing literature in computing education or related disciplines” and “a strong empirical basis, drawing on relevant research methods”. Such expectations are not made explicit for other computing education conferences such as SIGCSE and ITICSE. Our study will thus offer some insight into how these instructions have been followed in the papers accepted to ICER. Obviously we cannot say anything about the papers not accepted to the conference, though analyzing such data would be interesting from the point of view of investigating the review process and what kind of impact the above instructions have on the acceptance decisions.

3. CLASSIFICATION SCHEME

The classification scheme we have devised has seven dimensions, enabling the classification of computing education papers based on Theory/Model/Framework/Instrument, Technology/Tool, Reference Discipline, Research Purpose, Research Framework, Data Source and Analysis Method.

In developing our classification scheme we drew upon the work of Vessey, Ramesh and Glass [40], who devised a scheme for the classification of research literature across the different disciplines in the computing field [15, 26, 39]. Our intended focus had a great deal in common with this scheme, so we adopted it, but modified it to give specific details of linkages to other disciplines and more detail of the research methods.

From the Vessey et al. [40] scheme of five dimensions we used the dimensions of reference discipline, research approach and research method. We modified slightly their categories for research approach and renamed this dimension “Research Purpose” as we felt this better describes the categories in this dimension. We considered their research method dimension too broad and so divided it into Research Framework, Analysis Method and Data Source.

To identify specific ways in which the reported work connects to other disciplines we added the dimensions of Technology/Tool and Theory/Model/Framework/Instrument.

Our scheme was designed to be used in conjunction with that of Simon et al. [35], whose dimensions of Theme and Scope were very similar to the Topic and Level of Analysis of Vessey et al. Therefore we chose not to use the latter two dimensions. However,
Table 1: Descriptive, Evaluative, and Formulative categories of the Research Purpose dimension

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive-information/human system (DI)</td>
<td>Describing an existing information system or human system, e.g. a classroom environment, teaching approach or course offering.</td>
</tr>
<tr>
<td>Descriptive-technical system (DT)</td>
<td>Describing an existing technical system or implementation. This may include rationale for development, explanation of functionality and technical implementation.</td>
</tr>
<tr>
<td>Descriptive-other (DO)</td>
<td>Describing something other than an existing technical/information/human system, e.g. an opinion piece, a description of a proposed system or a paper whose purpose is to review or summarize the literature.</td>
</tr>
<tr>
<td>Evaluative-positivist (EP)</td>
<td>Research that takes the positivist perspective of an objective reality that can be measured and determined. Typically employing quantitative methods and deductive in nature; however, data mining, which is often inductive in nature, will also be included in this category.</td>
</tr>
<tr>
<td>Evaluative-interpretive (EI)</td>
<td>Research that takes the interpretive perspective that reality is subjective. Typically employing qualitative methods, interpretive research attempts to understand phenomena from the subjective meanings that participants assign to them. May be inductive or deductive in nature.</td>
</tr>
<tr>
<td>Evaluative-critical (EC)</td>
<td>Research that takes the critical perspective. The purpose of critical research is to understand and expose false beliefs or ideas for the purpose of explaining and perhaps transforming social systems. An example of a paper with this research purpose is Schwartzman [31].</td>
</tr>
<tr>
<td>Evaluative-other (EO)</td>
<td>Research with an evaluative component that does not belong in any of the other Evaluative subcategories above, e.g. an opinion piece survey that does not use a construct-based questionnaire [40].</td>
</tr>
<tr>
<td>Formulative-concept (FC)</td>
<td>The paper formulates a novel concept.</td>
</tr>
<tr>
<td>Formulative-model (FM)</td>
<td>The paper develops a model, which is a simplified view of a complex reality. The model may be usable as a framework (a basic conceptual structure used to solve or address complex issues) or as a classification or taxonomy.</td>
</tr>
<tr>
<td>Formulative-process, method, algorithm (FP)</td>
<td>The paper formulates a process, method or algorithm for accomplishing something (other than quality evaluation, which is covered by FS below).</td>
</tr>
<tr>
<td>Formulative-standards (FS)</td>
<td>The paper formulates guidelines, standards or criteria for evaluating the quality of something. It may also formulate a process for using the guidelines, standards or criteria.</td>
</tr>
</tbody>
</table>

As with the TMFI dimension, we are looking for an explicit level of use which is more than just a reference to the technology or tool, and we do not include TT that are developed in the paper itself. We only include technologies whose use is significant for the purpose of the paper. For example, if a paper investigates the effects of teaching Java in CS1 using BlueJ as the IDE, BlueJ will be listed as a TT if its features are explicitly relevant to the research, but not if the use of BlueJ appears to be incidental.

As with most of the ICER papers have already been classified according to Simon’s scheme [35], these two dimensions and two more of Simon’s have been omitted from the analysis presented here.

The remainder of this section explains the dimensions that we used and the specific categories within each dimension, and concludes with illustrative classifications of a handful of papers.

3.1 Theory / Model / Framework / Instrument

This dimension is used to show linkages to prior work. Here we list the TMFI – theories, models, frameworks and instruments (established questionnaires) – used in the research reported in the paper. Note that we include only those TMFI that are used explicitly rather than just referred to in motivating or positioning the work. For example, we include TMFI that are applied to a research design or used in an analysis, and we include TMFI that are modified or extended by the reported work. We do not include TMFI that are developed in the paper itself.

TMFI may range from those that are well known and identifiable by name to those that are unnamed and the result of a single study, such as a theory generated from a phenomenographical study. Note that we do not include technical frameworks (e.g. a Model-View-Controller GUI framework) or research frameworks (e.g. phenomenography) in this dimension. The former will often be represented in the Technology/Tool dimension (section 3.2) and the latter in the Research Framework dimension (section 3.5).

3.2 Technology / Tool

Here we list any technologies or tools (TT) used in the work reported in the paper; for example, a visualization tool, a class library, software or hardware.

As with the TMFI dimension, we are looking for an explicit level of use which is more than just a reference to the technology or tool, and we do not include TT that are developed in the paper itself. We only include technologies whose use is significant for the purpose of the paper. For example, if a paper investigates the effects of teaching Java in CS1 using BlueJ as the IDE, BlueJ will be listed as a TT if its features are explicitly relevant to the research, but not if the use of BlueJ appears to be incidental.

3.3 Reference Discipline

In this dimension we list the disciplines that the reported work is linked to through use of theories, models, frameworks, instruments, technologies or tools. These are commonly termed reference disciplines [40]. All of the papers we analyze are positioned in computing education research, so we only note reference disciplines outside of computing education.

3.4 Research Purpose

The research purpose is concerned with the goals of the research. This dimension is based on the ‘research approach’ dimension of Vessey et al [39], which was developed from the work of Morrison and George [20]. The three categories are:

- **Descriptive** – description of a tool, technology or system. This may involve detailed explanation of features, functionality and rationale for development.
- **Evaluative** – assessment of a tool, method or situation, typically through a systematic process involving data gathering, analysis and reporting. This may involve hypothesis testing and may be exploratory or investigative in nature.
- **Formulative** – development and/or refinement of a theory, model, standard, or process, or proposition of a new concept.
papers report studies with several parts, we also list further research. The categories of this dimension are explained in terms depends on the perspective of the researcher: a taxonomy may be considered a model and may also be used as a framework for further research. The categories of this dimension are explained in Table 1.

For every paper we identify at least one research purpose. As many papers report studies with several parts, we also list further research purposes where appropriate.

### 3.5 Research Framework

A research framework (see Table 2) is an overall orientation or approach that guides or describes the research, as opposed to a specific method or technique. A research framework may have associated theoretical, epistemological, and/or ontological assumptions (e.g. phenomenology), may prescribe or suggest the use of particular methods (e.g. grounded theory), or may simply be a descriptive term for a kind of research activity that has certain characteristics (e.g. action research, case study). Not all papers will have a research framework.

### 3.6 Data Source

The data source dimension describes the nature of the data and how it was gathered in the reported research. Most research papers will have at least one data source (see Table 3).

### 3.7 Analysis Method

An analysis method describes how empirical data was analyzed or what other means were used to draw conclusions. Practically all papers will have at least one kind of analysis method. Table 4 lists the categories in this dimension. If a paper has a research framework, that framework might well direct the analysis method that is used; but the same analysis method can of course be found in a paper that is not applying a specified research framework.

### 3.8 Sample classifications

A brief illustration of the application of our scheme will examine five papers from ICER 2008, the first set of papers that we classified.

Bennedsen et al [4] is a paper with three distinct TMFI: the theory of cognitive development, the pendulum test, and SOLO. It has a reference discipline of psychology, and a single research purpose of evaluative positivist. Its research framework is survey, its data sources are research specific and naturally occurring data, and its analysis method is statistical analysis.

Chinn et al [9] has no TMFI, does not report on a technology/tool, and has no reference discipline. It has research purposes of evaluative positivist and evaluative interpretive. Its framework is experimental research, its data source is research specific data, and it uses three analysis methods: interpretive qualitative analysis, statistical analysis, and interpretive classification.

Denny et al [12] has no TFMI and no reference discipline, but reports on the technology/tool of Peerwise. Its single research purpose is evaluative positivist, its research framework is survey, its data source is naturally occurring data, and its analysis method is statistical analysis.

Moström et al [21] has the TMFI of threshold concepts, with education as reference discipline. Its research purpose is evaluative interpretive. It has no research framework, its data source is research specific data, and its analysis method is interpretive classification.

Typically this will involve synthesis and integration of information and argumentation. The subcategories we use within each category are slightly modified from those specified in Vessey et al [39]. We have expanded the subcategories of the Descriptive category to distinguish between technical systems and information/human systems. We have also consolidated three subcategories in the Formulative category as we found it difficult to distinguish between the formulative categories of model, framework and taxonomy/classification. The use of these terms depends on the perspective of the researcher: a taxonomy may be considered a model and may also be used as a framework for further research. The categories of this dimension are explained in Table 1.

<table>
<thead>
<tr>
<th>Sample classifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bennedsen et al [4]</td>
</tr>
<tr>
<td>Chinn et al [9]</td>
</tr>
<tr>
<td>Denny et al [12]</td>
</tr>
<tr>
<td>Moström et al [21]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Research Framework dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action Research (AR)</td>
</tr>
<tr>
<td>A self-reflective systematic inquiry undertaken by participants to improve practice. Typically conducted as an iterative cycle of planning, action, change, reflection [13, 41].</td>
</tr>
<tr>
<td>Case Study (CS)</td>
</tr>
<tr>
<td>In-depth, descriptive examination conducted in situ, usually of a small number of cases/examples [13].</td>
</tr>
<tr>
<td>Constructive Research (CR)</td>
</tr>
<tr>
<td>Research that aims to demonstrate and/or evaluate the feasibility of a proposed idea (concept implementation, proof-of-concept research). Revolves around the development of, e.g., software, technology, a teaching approach, or an evaluation instrument.</td>
</tr>
<tr>
<td>Delphi</td>
</tr>
<tr>
<td>Seeking consensus by showing a group of raters a summary of their ratings, with justifications, then iteratively inviting them to reconsider their ratings in the light of what the others have said [35].</td>
</tr>
<tr>
<td>Ethnography (Eth)</td>
</tr>
<tr>
<td>A branch of anthropology that deals with the scientific description of individual cultures [41].</td>
</tr>
<tr>
<td>Experimental Research (Exp)</td>
</tr>
<tr>
<td>Qualitative research based on manipulating some variables while varying and measuring others. This requires formation of control and experimental groups of participants with random assignment of participants or use of naturally formed groups.</td>
</tr>
<tr>
<td>Grounded Theory (GT)</td>
</tr>
<tr>
<td>Qualitative, data-driven research in the tradition of Glaser and/or Strauss [37] which aims to formulate theories or hypotheses based on data.</td>
</tr>
<tr>
<td>Phenomenography (PhG)</td>
</tr>
<tr>
<td>Investigation of the significant variations in the ways that people experience a phenomenon (2nd order perspective).</td>
</tr>
<tr>
<td>Phenomenology (PhL)</td>
</tr>
<tr>
<td>Investigation of the richness and essence of a phenomenon by studying one’s own or others’ experiences of it (1st order perspective).</td>
</tr>
<tr>
<td>Survey Research (Survey)</td>
</tr>
<tr>
<td>Quantitative research based on exploring the incidence, distribution and/or relationships of variables in non-experimental settings.</td>
</tr>
</tbody>
</table>
Naturally occurring data (Nat) | Data pertaining to human subjects that would be available regardless of whether the research was carried out. This includes examination results, course grades, task submissions, student/teacher emails, course documentation, and published literature. It also includes data from previous research that is re-analyzed.

Research specific data (Res) | Data collection pertaining to human subjects tailored specifically for the needs of the research. This includes interviews, questionnaires, observation data and any other data collected from assignments or tasks designed specifically for the target research.

Reflection (Ref) | The researchers’ own reflections and experiences of a phenomenon serve as data. These reflections are what the paper deals with (they are analyzed and/or described); e.g. phenomenological research.

Software (Sw) | Data collected about a software system. The data primarily describes the software rather than the humans using it or their activities; e.g. performance benchmarking, software metrics.

Simon et al [35] has Simon’s system as TMFI, with no technology/tool or reference discipline. Its single research purpose is descriptive other, it uses the Delphi research framework, its data source is naturally occurring (in the literature), and its analysis methods are descriptive statistics and interpretive classification.

4. RESULTS
In total we analyzed all 72 ICER papers published to date, with the following numbers per year: 2005: 16 papers; 2006: 13; 2007: 14; 2008: 16; 2009: 13 papers. In this section we present the results for each dimension.

4.1 Theories, Models, Frameworks and Instruments
We deliberately limited the TMFI to those explicitly mentioned and applied in the paper we were classifying. A secondary reference, a TMFI mentioned in the papers referenced by the paper we were considering, might have been useful to note, but might also misrepresent the research we were classifying.

We did not distinguish between theories, models and frameworks, as the definitions tend to overlap, and the terms themselves can be used inconsistently in different papers. We also decided not to question how faithfully a TMFI had been used in the work being reported, as such criticism would be beyond the scope of this work. This also holds for methodological issues, discussed later.

In 41 of the 72 papers we identified that the work had been built on some previous work by others in the sense that some explicit theories, models, frameworks and instruments (TMFI) were used in the work. In two further cases the work was built on such TMFI developed by one or more of the authors themselves [32, 35].

The remaining 29 papers, 40% of the data, did not apply any TMFI either from the authors or from others. These included wholly standalone works presenting novel research and a fair number of papers building on previous technical work, such as evaluating the effect on learning outcomes of using some tool. Many papers also extended or built on such results from their authors’ previous work or from other authors’ results. Identifying such work was rather challenging, as few papers emphasize the TMFI they are using. While some papers might mention them explicitly in the methodology or discussion sections, others might discuss them just briefly in a section on related work, then proceed to use them with no further mention.

A general observation was that most TMFI identified in the data were not classical theories from education, such as constructivism, behaviorism or cognitive learning. They were much more specific, and in many cases referred to the theory by reference to an author rather than by its recognized name. There was great diversity. We listed in total 78 instances of TMFI, of which 68 were distinct. These numbers are clearly subject to interpretation, as the decision to list a TMFI was in many cases not clear-cut. However, they clearly indicate that a large and diverse pool of external work is being used in CER research.

<table>
<thead>
<tr>
<th>Analysis Method dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Table 4</strong></td>
</tr>
<tr>
<td>Argumentation (Arg)</td>
</tr>
<tr>
<td>Conceptual Analysis (CA)</td>
</tr>
<tr>
<td>Descriptive Statistics (DS)</td>
</tr>
<tr>
<td>Exploratory Statistical Analysis (ESA)</td>
</tr>
<tr>
<td>Interpretive Classification (IC)</td>
</tr>
<tr>
<td>Interpretive Qualitative Analysis (IQA)</td>
</tr>
<tr>
<td>Mathematical Proof (MP)</td>
</tr>
<tr>
<td>Statistical Analysis (SA)</td>
</tr>
</tbody>
</table>
Some of the well-known theories, models and frameworks that we identified are:

- Bloom’s taxonomy [7]
- cognitive apprenticeship [10]
- cognitive load theory [1, 22]
- general systems theory [5]
- schema theory [18]
- self-efficacy theory [2]
- situated learning [17]
- SOLO taxonomy [6]
- threshold concepts [19]

Very few papers used instruments, established questionnaires that measure specific behaviors or characteristics. Examples of those found include the pendulum test [8] to measure abstraction ability, the test of emotional intelligence [30], and the motivated strategies for learning questionnaire (MSQL) [25].

4.2 Technologies and Tools

Tools research is an important part of CER. We found 12 tools or technologies that played an important role in the research of 12 papers. These were Agar, ALVIS Live!, Aropä, BlueJ, Classroom presenter, Jeliot, Marksense, Peerwise, PlanAni, SALSA, tablet PCs, and WebCAT.

4.3 Reference Disciplines

The reference disciplines were determined by noting the TMFI used that were not developed in computing education research. The following reference disciplines emerged from the data (the number in parentheses is the number of papers linking to each reference discipline).

- Biology (1)
- Computing (11)
- Education (11)
- Human-computer interaction (1)
- Mathematics (1)
- Philosophy of science (1)
- Psychology (12)
- Sociology (1)
- Systems theory (1)

A further 12 papers were linked to the broader discipline of computing through use of a technology or tool. We conclude that computing education research does indeed make wide use of works from other disciplines, with the key disciplines being computing, education and psychology.

4.4 Research Purpose

For each paper we identified one or more research purposes. In 35 papers we found one research purpose, in 33 papers, two research purposes, and in four papers, three purposes. In some papers where we had identified a secondary purpose, it was not clear whether it was sufficiently substantial to warrant inclusion. For example, a large quantitative study might include a brief qualitative mention of student feedback.

Table 5: Distribution of Research Purposes (72 papers)

<table>
<thead>
<tr>
<th>Research Purpose</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive-information/human system (DI)</td>
<td>10 (14%)</td>
</tr>
<tr>
<td>Descriptive-technical (DT)</td>
<td>2 (3%)</td>
</tr>
<tr>
<td>Descriptive-other (DO)</td>
<td>4 (6%)</td>
</tr>
<tr>
<td>Evaluative-positivist (EP)</td>
<td>36 (50%)</td>
</tr>
<tr>
<td>Evaluative-interpretive (EI)</td>
<td>29 (40%)</td>
</tr>
<tr>
<td>Evaluative-critical (EC)</td>
<td>1 (1%)</td>
</tr>
<tr>
<td>Evaluative-other (EO)</td>
<td>5 (7%)</td>
</tr>
<tr>
<td>Formulative-model (FM)</td>
<td>15 (21%)</td>
</tr>
<tr>
<td>Formulative-process/method/algorithm (FP)</td>
<td>8 (11%)</td>
</tr>
<tr>
<td>Formulative-standards (FS)</td>
<td>3 (4%)</td>
</tr>
</tbody>
</table>

The decision was therefore necessarily subjective.

In Table 5 we list the distribution of all research purposes found in the papers. Evaluative purposes are found in a clear majority of the papers: 62 papers (86%) involve some evaluative aspect. More than a third of the papers (36%) had a formulative aspect, and 22% of papers had some descriptive purpose. Considering these more closely, we note that all DI papers (describing information / human systems) also had a secondary purpose. In seven cases this was evaluating the system (EI/EP), and in three cases formulating a process (FP). Using the classification scheme developed by Simon [35], two of the latter were considered position papers or proposals rather than research papers. There was one pure literature summary [35]. Of the two technical descriptive papers (DT), one also formulated a process and one was a pure technical research work.

We conclude that ICER papers have a clear focus on evaluative work.

4.5 Research Framework

We identified a research framework in 57 of the papers (79%). Of these, 47 had a single framework, nine had two, and one had three frameworks. Fifteen papers did not use a framework. The distribution of frameworks was very uneven, as we expected. In greatest numbers were survey frameworks (28 papers, 39% of all papers), experimental frameworks (11, 15%), constructive research (10, 14%) and grounded theory (9, 13%). Smaller counts included phenomenography (4 papers), case studies (3), Delphi method (1), ethnography (1) and phenomenology (1).

The papers with no clear framework tended to use some qualitative analysis method or content analysis.

4.6 Data Source

As might be expected, the majority of papers (79%) had ‘research’ as one data source, indicating that the data was specifically collected for the research being reported. Twelve papers (16%) used naturally occurring data in combination with research-specific data, while six used natural data alone. Eight papers had literature as their data source, and one paper used reflection. In one technical paper we could not identify any data source, because the paper simply described a tool.

4.7 Analysis Method

With the exception of one technical paper, all papers applied some method of analysis. In 24 papers we identified the use of two different analysis methods; in four papers, three methods; and in one paper, four methods.

The most common method used was some kind of statistical analysis (SA) with tests of significance (30 papers, 42%). Exploratory statistical analysis (ESA) was used in 12 papers (17%), and purely descriptive statistics (DS) in eight papers (11%).

Qualitative methods were widely used. In 25 papers (35%) we identified interpretive qualitative analysis (IQA), meaning some data-driven categorization with no predefined categories. Some form of interpretive classification or content analysis (IC) was used in 19 papers (26%). Here some predefined categorization scheme was either applied as such or refined on the basis of the data. Finally, in 12 papers we considered that argumentation (Arg) was an analysis method. We recognize that all papers have some form of argumentation, but we listed Arg only in cases where argumentation clearly played a strong role in reaching the conclusion, perhaps supported by data and/or theory.
4.8 Validity and Reliability Issues in the Classification

Several of the authors had conducted analysis and classification of literature in earlier studies [23, 24, 28, 32-36]. However, during the process of developing this classification scheme we soon realized the complexity of the task, as we were looking more deeply into the research process than previous studies had done. There were several difficulties. We wanted to identify which TMFI had really been used in a paper, rather than just referred to as related work. Many papers did not report this clearly, and we had to read the papers carefully to decide whether something should be included. In a number of cases the decision had to be negotiated between the classifiers before consensus was found. Working in pairs was a clear advantage in this regard. We recognize that we may still find individual papers for which we have wrongly decided whether to include some TMFI, but we are confident that the big picture is reasonably accurate, and this is our goal. In this kind of research, the scale of numbers is more important than absolute precision.

The same observation holds to some extent for the other dimensions. Reference disciplines were deduced from where the TMFI was published, and in some cases we were unable to check the background of the originators of a TMFI. The main research purpose was clear in most cases, but for some papers the decision to include additional research purposes had to be negotiated. Research framework was quite clear in most cases, as were data source and analysis method. The inclusion of argumentation as a research method was invariably a matter for negotiation. But notwithstanding these difficult issues, we are confident that the big picture is correct, and that our classification scheme is a useful and usable tool.

5. DISCUSSION

Previous literature studies, as mentioned above, have mapped the CER field in some detail, answering questions such as \textit{what has been done and what kind of work has been carried out in CER}. Our goal in this study was to delve deeper, exploring \textit{how research in CER is being carried out}. Huge numbers of papers have been published in computing education conferences and journals in the past 40 years, so we would expect much to have been learnt about teaching and learning in computing. However, after decades of research, we still have only a vague understanding of why it is so difficult for many students to learn programming, the basis of the discipline, and consequently of how it should be taught. Sheard et al. [32] analyzed a wide pool of papers on programming education and expressed a “… need to explore the process of learning in relation to theories and models of learning. It was a concern that we didn’t find many studies that investigated learning within such a theoretical framework.” Therefore it is natural to ask whether there might be unknown problems in the research itself that could explain our slow progress. Does it have enough focus on evaluation, or is the dominating purpose describing or formulating new teaching practices and tools? Does it apply knowledge from related fields, which have a much longer history of research, rather than reinventing the wheel?

In their wide analysis of computing research, Glass, Ramesh and Vessey [14] found that in Computer Science (CS) 79% of papers had a formulative research purpose (which they called research approach), 11% had an evaluative purpose and 10% had a descriptive purpose. In Software Engineering (SE) and Information Systems (IS) the numbers were 55%, 14%, 28% and 24%, 67%, 9% respectively. As many CER researchers have a background in computer science, we might expect the strong tradition of formulative work to have some influence on their work in CER. However, our analysis of ICER papers fails to support this expectation. The corresponding results were 18% formulative, 71% evaluative, and 11% descriptive work. We conclude that the CER presented at ICER is much closer in purpose to IS research than to CS/SE research.

Does CER build on previous work by disciplines such as Education and Psychology? We examined the reference disciplines on which the research was based. Glass et al [14] found that only 11% of CS papers and 2% of SE papers built on theories from other disciplines, whereas in IS 73% of papers used theories outside computing. In our findings, 44% of CER papers in our data pool built their research on TMFI from other disciplines. If we count only disciplines outside Computing, we still have 35% of papers applying TMFI from these disciplines. Thus the CER presented at ICER seems to lie about halfway between CS/SE and IS.

We should be cautious about generalizing these observations. We have analyzed only papers from ICER, a small and young conference, which, according to previous research [35], has a larger percentage of research papers than many other CER conferences. To get a broader picture of CER we will need to look at journals and other conferences, which may give us a different picture.

Casting back to the ICER CFP requirement that papers should have a clear theoretical basis, drawing on existing literature in computing education or related disciplines, we conclude that this requirement has been well met. In a clear majority of papers some TMFI from related disciplines was identified. This is a very good sign, considering that we required the TMFI to be used or extended, not just referred to. Moreover, we were impressed by the diversity of TMFI used, which suggests that the authors really seem to look for TMFI that fit their research contexts. We also investigated papers for which we did not find any TMFI, but we found no clear characteristics to explain this lack. A few of them built on some Tool/Technology, and thus on previous work from computing, while others simply presented novel work. It should be recalled that we did not record as a TMFI a theory, model, framework, or instrument actually developed within the paper we were classifying, so some of these papers could well have been developing TMFI. In addition, it is presumably possible to have a clear theoretical basis, drawing on existing literature, without applying an explicit theory, model, framework, or instrument from that literature.

The other ICER requirement was that papers should have a strong empirical basis, drawing on relevant research methods. Again we could confirm that the goal has been well met. Empirical evaluation has been widely applied, with 86% of papers having at least one evaluative research purpose. The richness of the conference is demonstrated by the observation that both evaluative positivist and evaluative interpretive purposes were widely applied, indicating general acceptance of both quantitative and qualitative evaluation papers.

In summary it seems that ICER does indeed achieve its stated goals. We should now look at other publishing forums to form a broader picture of the CER field.

6. ACKNOWLEDGMENTS

Our thanks to Dr Tony Clear for his advice on classification of critical research papers.
REFERENCES


