A Methodological Review of the Articles Published in the Proceedings of Koli Calling 2001–2004

Justus J. Randolph, Roman Bednarik, and Niko Myller
Department of Computer Science, University of Joensuu
P.O. Box 111,
FI-80101 Joensuu, FINLAND
+358 13 251 7929
firstname.lastname@cs.joensuu.fi

ABSTRACT
Three reviewers conducted a methodological review of the full papers published in the proceedings of Koli Calling 2001 to 2004. They analyzed the papers in terms of methodological characteristics, section proportions, report structure, and region of origin. It was found that the majority of articles were descriptions of programs, projects, or interventions. Of the articles that involved research with human participants, (a) posttest-only designs with controls were used most often, (b) the majority of text was devoted to describing evaluation procedures, and (c) literature reviews were often found to be absent or inadequate. Recommendations and future research directions are presented.

Keywords
Methodological Review, Computer Science Education, Research Methodology

1. INTRODUCTION
In the introduction to last year's *Koli Calling: Finnish/Baltic Sea Conference on Computer Science Education* proceedings, Lauri Malmi wrote that "the primary goal of the conference is to promote the development of computer science (CS) and CS education research" [5]. It is known that methodological reviews promote the development of education research by empirically describing the methodological practices that are used and by making recommendations for improvement (see [3]). Because of the benefits of such reviews, we conducted a methodological review of the papers presented in the previous proceedings of Koli Calling. With hope, this review will help the authors and reviewers of future Koli Calling proceedings to become more aware of the state of the research methodology in the Koli Calling proceedings and thus capitalize on the strengths and be aware of the weakness in the previous research. We feel that now is an especially appropriate time to do a critical review because if, as Fincher & Petre [10] state, "CS education research is an emergent area and is still giving rise to a literature," then CS education research methodology is still in a malleable state where proper, or improper, methodological practices can become recalcitrantly engrained in the traditions, paradigms, and conventions of CS education research in general.

We acknowledge that CS education research is a field made up of many methodological disciplines, perspective, and traditions – each of which makes an important contribution to the field as a whole. In this review, we approach the methodologies in the Koli Calling proceedings from the perspective of the behavioral sciences. Our rationale for applying the methodological standards of the behavioral sciences is based on the assumption that when CS education research involves human participants, the conventions, practices, and ethical and scientific standards of the behavioral sciences should apply.

We invite others who hail from traditions outside of the behavioral sciences to replicate our review from their own perspectives. In that way, a rich and multi-disciplinary view of the state of the research methodology in the Koli Calling proceedings can come to light.

In the next section, we summarize the findings from previous reviews of methodologies of CS education research and evaluation. We follow with a report of the methodologies that were used, the results we found, and end with some answers to the following research questions:

1. What were the methodological characteristics of the papers published in the previous Koli Calling proceedings?
2. What, if any, methodological trends occurred over the first four Koli Calling proceedings?
3. How does this Koli Calling review compare to the findings reported in the other CS-education-related methodological reviews?

2. LITERATURE REVIEW
In this section, we summarize the findings from a methodological review of 20 years of SIGCSE Technical Symposium Proceedings [16] and a review of evaluations of CS education programs [11]. In addition, we report the findings from a methodological review of the full papers published in the 2004 proceedings of the International Conference on Advanced Learning Technologies (ICALT) in order to be able to make cross-discipline comparisons [12]. (Also of interest to readers, but outside the scope of this literature review, is a *Review of Resources for Evaluating K-12 Computer Science Education Programs* [13]).

2.1 Review of SIGCSE Technical Symposium Proceedings
In 2004, Valentine's "CS Educational Research: A Meta-analysis of SIGCSE Technical Symposium Proceedings" was published. Although not technically a meta-analysis, Valentine did take a light-hearted approach to critically analyzing 444 articles published in the SIGCSE proceedings from 1984 to 2003 that dealt with the first year of CS courses. He categorized each of the articles into one of six categories: *Marco Polo, Tools, Experimental, Nifty, Philosophy*, and *John Henry* and looked for trends over the years in terms of the distribution of articles that fell into each of these categories. The category that is most relevant to our review is *experimental*, which he defined as "the author making any attempt at assessing 'the treatment' with some scientific

analysis" (p. 256). The major findings from this review were that

- Of the research articles on first-year courses, 21% were categorized by Valentine as being experimental,
- The proportion of experimental articles had been increasing since the mid-90's,
- The proportion of what he calls *Marco Polo – I went there and I saw this* - types of papers had been declining linearly since 1984, and
- The overall number of papers being presented in the SIGCSE forum had been steadily increasing since 1984.

Valentine's findings, however, should be accepted with a fair amount of skepticism because the methodology used to conduct Valentine's review was lacking (e.g., there were no estimates of reliability about his categorizations.)

2.2 A Review of Program Evaluations in K-12 CS Education

Randolph [11] conducted a meta-analysis and methodological review of the published evaluations of Kindergarten through 12th grade CS education programs. A comprehensive literature search resulted in 29 evaluations that were included in the methodological review and 8 evaluations that were included in the meta-analysis. The main findings were that

- Most of the programs that were evaluated offered direct computer science instruction to general education, high school students in North America;
- In order of frequency, evaluators examined (1) stakeholders attitudes, (2) program enrollment, (3) academic achievement in course courses, and (4) achievement in computer science;
- The most frequently used measures were questionnaires, existing sources of data, standardized tests, and teacher- or researcher-made tests. Only one measure of computer science achievement, which is no longer available, had reliability or validity estimates. The pretest-posttest design with control groups and the posttest-only design without control groups were the most frequently used experimental designs; and
- No interaction between type of program and CS achievement was found.

2.3 A Critical Review of the Research Methodologies Reported in the Full Papers of the Proceedings of ICALT 2004

Randolph et al. [12] conducted a critical review of the proceedings of a major conference in educational technology. (Methodologically, the current review is a replication of Randolph et al. [12] study.) In the Randolph et al. review, 5 reviewers reviewed 126 papers and coded them in terms of *research design; common defects in research design and reporting; the proportions of literature review, description, and evaluation sections; measures;* and *regional affiliation of the first author.* The major findings from that review are summarized below:

- The majority (61%) of full papers in ICALT 2004 do not report research on human participants *[35% would have been classified as what Valentine called experimental]*;
- 71% of the measures concerning human participants are quantitative;
- Experimental designs were used the most – followed by explanatory descriptive designs;
- Of experimental designs, the posttest-only design without controls was used nearly half of the time;
- In terms of the most frequent common defects, many papers lacked congruence between study operations and interpretations, many did not report effect sizes, and many lacked need controls for important study variables;
- 47% of the text in [the ICALT papers] was devoted to program description;
- The most frequently used measures were researcher-made student surveys; and
- Very few studies reported data about reliability or validity of the measures that were used. (p. 13)

In summary, the reviews mentioned above validate Fincher and Petre's [10] assumption that CS education research is an emergent area.

3. METHOD

Three reviewers rated the 59 full papers in the proceedings of Koli Calling 2001 [4], 2002 [5], 2003 [6], and 2004 [7]. (The full papers are called 'sessions'; 'paper sessions'; 'paper presentations'; or 'research paper presentations' and 'discussion paper presentations' in Koli Calling 2001, 2002, 2003, and 2004 respectively.) Each of the reviewers was assigned a unique, random sample from the population of full papers (i.e., each article was rated by one reviewer.) The research method (hereafter a *case)* was the unit of analysis; a paper was the unit of data collection. To calculate estimates of interrater agreement, each reviewer rated the same fourteen randomly-selected articles, which amounted to twenty cases.

3.1 Key variables of the coding sheet

In this section we describe the variables that were included on the coding sheet. (See [14] and [12] for more information about and lessons learned from developing this coding sheet.) The text describing the research categories in Section 3.1.1 is similar to [12], since this review replicates the methodology of the Randolph et al. [12] review.

3.1.1 Research categories

Raters classified each case into one of the seven research categories: *explanatory descriptive, exploratory descriptive, correlational, classification, causal-comparative, experiments or quasi-experiments,* or *other.*

Studies that provided answers to "how" questions by explaining the causal relationships involved in a phenomenon were put into the *explanatory descriptive* category [17]. (Studies using qualitative methods often fall into this category.)

Studies that answered "what" or "how much" questions but did not make any causal claims [17] were put into the *exploratory descriptive* category. (Pure survey research is perhaps the most typical example of the exploratory descriptive category, but certain kinds of case studies might qualify as exploratory descriptive research as well [17].)

A case was categorized as *correlational* if it analyzed how continuous levels of one variable systematically covaried with continuous levels of another variable.

Cases that used some classification technique (e.g., clustering) were put into the *classification* category.

If researchers compared two or more groups on an inherent variable, the case was categorized as *causal-comparative*. For example, if a researcher had compared computer science achievement between boys and girls, that case would have been classified as casual-comparative because gender is a variable that is inherent in the group and can not be naturally manipulated by the researcher.

If the researcher compared experimental and control (or contrast) conditions in the experiment, the case was classified as *experimental or quasi-experimental*. If the case was experimental or quasi-experimental, it was further categorized into one of the following categories [15]:

- Group posttest-only without controls;
- Group posttest-only with controls;
- Group pretest-posttest without controls;
- Group pretest-posttest with controls;
- Group designs with repeated, dependent measures; or
- Single-participant designs.

If the author made causal claims based on a retrospective survey, the case was classified as group posttest-only design without controls. If there were no causal claims, a retrospective survey was classified as exploratory descriptive.

If the article did not report research involving human participants, the reviewers were instructed to classify the article as *other*. Articles marked as *other* were further categorized into:

- Literature reviews or meta-analyses;
- Program descriptions without anecdotal evidence;
- Program descriptions with anecdotal evidence;
- Theoretical , methodological or philosophical papers;
- Technical investigations; or
- Others.

In addition, raters categorized experimental/quasi-experimental cases into (a) those that used random assignment (experimental) and those that did not (quasi-experimental), (b) those that used random selection and those that did not, and (c) those that reported any psychometric information and those that did not.

3.1.2 Proportions of literature review, description, and evaluation sections

The raters divided the sections of the articles into three categories – *literature review, program description,* and *evaluation. Literature review* was defined as sections of text where the previous, relevant literature was presented and analyzed; it excluded text that consisted of the introduction and the problem statement. *Program description* was defined as any text that went beyond explaining the intervention in enough detail to replicate the study. Finally, the text related to reporting of methods and results was considered to be *evaluation*. The reviewers measured the amount of text written for each category and converted the measures into proportions.

3.1.3 Structure of the Paper

The structure of the paper was analyzed based on the presence or absence of the following sections or characteristics:

- Background,
- Purpose statement,
- Research questions are explicitly stated,
- Literature review,
- Methods: Participants,
- Methods: Settings,
- Methods: Instruments,
- Methods: Design,
- Methods: Description of treatment and control procedures,
- Claims in the results section are congruent with methods, and
- The discussion/conclusion section links back to explicit (implicit) research questions.

3.1.4 Regional of affiliation of first author

The geographical region of the affiliation of the first author was recorded for each article. The region categories were *Europe, Asian-Pacific/Eurasia, Africa, Middle East, North America*, and *South America*.

3.1.5 Other categories

We analyzed *measures used, dependent variables, independent variables* and *moderating/mediating factors;* however, we do not report on the results of those categories here because the average levels of interrater agreement were below the lower bound that had been set a priori for (Cohen's) *kappa* (see the next section). For readers interested in the results of these categories, data tables and a possible explanation of the cause for the low reliabilities can be found in the appendix to this paper.

3.2 Analysis

For main results analysis, we calculated the frequencies of levels of nominal variables and calculated descriptive statistics for continuous variables. Fleiss' Multirater *kappa* (see [2]) was used as the interrater reliability statistic for multicategory, nominal variables; a prevalence- and bias-adjusted version of multirater *kappa* was used for binary, nominal variables. A two-way mixed-model, single measure intraclass correlation (consistency definition) was used as the interrater reliability statistic for continuous variables. It was decided a priori that only variables with values of *kappa* above .40 or intraclass correlations above .60 would be included in the analysis. (According to Banerjee et al. [1], .40 is the lower bound of acceptability for *kappa.)* Nichols' [9] SPSS macro – mkappasc.sps – was used to calculate Fleiss' multirater *kappa*. Since *kappa* is known to be sensitive to the prevalence of ratings (see [1]), a slight variation of Nichol's mkappasc.sps, where p(e) was set equal to *1/(number of rating categories)*, was used to calculate prevalence- and bias-adjusted multirater *kappa*.

4. RESULTS

4.1 Interrater Reliability

Reviewers 1, 2, and 3, rated an equal number of articles. Table 1 shows the multirater *kappa* for the key variables reported in this study. The intraclass correlation coefficients, each of which was statistically significant, for the proportion of literature reviews, program description, and evaluation were .94, .97, and .96, respectively

Table 1. Interrater Agreement.

Variable	Kappa
Methodology category	.57
If *other* in methodology category, type of other	.74
Experimental or quasi-experimental design	.68
Random selection	.68
Psychometric information provided	1.0
Experimental design	.60
Structure of the paper (average)	.55

4.2 Methodological characteristics

Of the 74 cases in the population of full articles in the four Koli Calling proceedings, 44 (60%) were categorized in the *other* category. The proportion of cases classified as other compared to all rated cases was 11/18, 10/12, 9/16 and 14/28 for the years 2001, 2002, 2003 and 2004, respectively. There were 30 cases that involved human participants; Table 2 shows the breakdown of those cases.

Table 2. Methodology Category.

Methodology	Frequency	%
Experimental/quasi-experimental	9	30.0
Explanatory description	3	10.0
Exploratory description	9	30.0
Correlational	1	3.3
Causal-comparative	6	20.0
Classification	2	6.7
Total	30	100

Table 3 presents the research design that was used if a case had been categorized as experimental or quasi-experimental. Of these nine cases, none were categorized as experimental and nine were categorized as quasi-experimental. Eight cases (88.9%) did not use random selection. None of the cases reported any psychometric information.

Table 3. Experimental or Quasi-Experimetal Designs.

Design	Frequency	%
Posttest, no controls	0	0.0
Pretest-posttest, no controls	1	11.1
Pretest-posttest with controls	1	11.1
Group repeated measures	3	33.3
Single-participant	0	0.0
Posttest-only, with controls	4	44.5
Total	9	100

Furthermore, the cases that did not involve human subjects (i.e., the *other* category) were classified into six categories. The distributions of those categories are shown in Table 4, below.

Table 4. Proportion of Cases that Did Not Involve Human Participants.

Type of other	Frequency	%
Literature reviews, meta-analyses	4	9.1
Program descriptions without anecdotal evidence	26	59.1
Program descriptions with anecdotal evidence	7	15.9
Theoretical, methodological or philosophical papers	1	2.3
Technical investigations	0	0.0
Other	6	13.6
Total	44	100

4.3 Section proportions

There were seventeen papers that involved research with human participants (n=17 in this and the next section). The proportions of sections (as the percentage of the entire literature review, description, and evaluation sections) devoted to literature review, program description, and evaluation in these papers were approximately distributed around the means, with the standard deviations, presented in Table 5.

Table 5. Proportions of Sections.

Section	M (%)	SD
Literature review	8	9
Program description	27	19
Evaluation	65	19

4.4 Structure of the Paper

The characteristics of the structure of the papers that involved research with human participants is presented in Table 6.

Table 6. Structure of the Papers.

Element	Yes	No	% Yes
Background	14	3	82.4
Purpose	16	1	94.1
Research questions explicitly stated	3	14	17.7
Literature review	8	9	47.1
Methods: Participants	8	9	47.1
Methods: Settings	12	5	70.6
Methods: Instruments	11	6	64.7
Methods: Design	6	10	37.5
Methods: Treatment & control proc.	8	6	57.1
Results aligned with methods	8	9	47.1
Discussion/conclusion relate to explicit (or implied) research questions	10	7	58.8

4.5 Region of first author's affiliation

In the 59 papers, the geographic region of the first author's affiliation was almost always Europe. Only two papers came from outside of Europe; one came from Asian-Pacific/Eurasia and one came from North America. About 90% of all papers came from Finland.

5. DISCUSSION

5.1 Methodological characteristics

Most (44 out of 74, or 60%) of the full papers in Koli Calling proceedings belong to the *other* category (i.e., they do not involve research with human participants.) Of the papers in the *other* category, the vast majority are devoted purely to describing (or marketing) a program, project, or intervention. There are 30 papers (40%) that involve research with human participants. In those 30 papers, the two most common methodologies are experimental/quasi-experimental and exploratory descriptive, with nine cases each. In the experimental/quasi-experimental methodology category, the post-test only with controls design is most commonly used.

When considering the section proportions in the papers that involve research with human participants, most of the text is devoted to descriptions of evaluation. It is striking what a small proportion of space is actually used for literature reviews.

Some alarming results concerning the characteristics of structure of the Koli papers have been found. Critical sections such as *research questions, literature review, participant descriptions* and *research design* are often absent from the papers. Also, the results sections are often not aligned with the methods.

In conclusion, the methodological characteristics that characterize the articles thus far published in the Koli Calling proceedings are listed below:

1. By far, the most frequently published type of paper in the Koli proceedings are pure program (project) descriptions;
2. Of the empirical articles reporting research that involves human participants, exploratory descriptive (e.g., survey research) and quasi-experimental methodologies are common;
3. The structure of the empirical papers that report research involving human participants deviates sharply from structures that are expected in behavioral science papers;
4. Most of the text in empirical papers are devoted to describing evaluation of the program; very little is devoted to literature reviews; and
5. The Koli Calling proceedings do indeed represent mainly the work of Nordic/Baltic, especially Finnish, CS education researchers.

5.2 Trends

Since the Koli Calling Conference has been in operation for a limited number of years, there is not enough longitudinal data to make convincing arguments about methodological trends. (That is why our results are presented as collapsed across the four years that the Koli Conference has been publishing proceedings.)

5.3 Comparison with other reviews

Several commonalities can be seen in the findings of this and other reviews. Compared to Valentine's [16] review of SIGCSE proceedings, where 21% of articles are experimental, and to Randolph et al.'s [12] review of ICALT proceedings, where 40% of articles involve research with human participants; the Koli Calling proceedings (41%) have a only a slightly greater percentage of articles that involve research with human participants. One finding common across all reviews is that psychometric information about the reliability and validity of measures is absent or grossly underreported. This fact gives credence to Randolph's [11] finding that there is a dire need for readily available CS education measures that have been rigorously shown to be reliable and valid. Finally, in both the ICALT 2004 and Koli Calling reviews, the articles are found to be severely lacking in the amount of emphasis given to reviewing the previous literature.

There are several methodological characteristics that are unique to the Koli Calling proceedings that are worth mentioning. Of the articles that involve research with human participants, compared to the 2004 ICALT proceedings, much more text is devoted to evaluation (i.e., there was a larger amount of text devoted to methods and results) than to describing (and marketing) a particular program or project. What's more, the articles in the Koli Calling proceedings more frequently report using designs that are less susceptible to threats to internal validity (such as posttest-only with controls or group repeated measures designs) than the articles in the ICALT 2004 review, which used mainly posttest-only designs *without* controls.

6. RECOMMENDATIONS AND FUTURE WORK

As in the Randolph et al. [12] review, we suggest that the Koli Calling reviewers and authors systematically evaluate their papers based on the elements in the structured abstract model. The elements might be *background, purpose, methods, results*, and *conclusions* for empirical papers and *background, purpose, description of program*, and *implications* for program descriptions. In the next stages of our research agenda, we plan to build a semi-automated methodological review system designed to help reviewers evaluate their papers in a systematic way. We hypothesize that such a systematic methodological review system would be expected to increase the methodological quality of the proceedings in several ways:

- It would provide reviewers with an explicit and common framework for evaluating the papers and for returning suggestions for improvement to authors;
- It would give authors a framework for writing high-quality methodology papers; and
- Finally, it would facilitate a meta-level methodological review of each year's proceedings so that research trends in CSE can be monitored, discussed, analyzed, and – thereby – systematically improved.

We believe that this kind of monitoring will help speed CS education research's transition from an emergent area into an area with a high degree of methodological sophistication and rigor.

In this article we concentrated on behavioral-science grounded CS education research papers, but we want to

extend this work to other types of papers in the fields of CSE and educational technology. We see that this could enhance the quality of research in the areas where CS, CS education, educational technology, and information systems overlap.

7. CONCLUSION

We carried out a methodological review of the full papers in the proceedings of Koli Calling. The papers were analyzed in terms of research methodology and design, section proportions, and structure of the paper. It was found that the majority of articles did not report research involving human participants. Experimental/quasi-experimental and exploratory descriptive were the two most commonly used methodology categories; the posttest-only with controls design was the most commonly used experimental design.

When empirical articles involving research with human participants were published, they were often exploratory descriptions of programs or experiments that were carried out with somewhat rigorous designs. However, the empirical articles in Koli Calling proceedings usually had low proportions of literature reviews or none at all and their structures and elements differed significantly from the structures and elements regularly expected in behavioral science papers. We recommend that reviewers and authors use a model based on the elements that should be included in a structured abstract to evaluate their papers. A semi-automated reviewing system, based on such a framework, is planned for development.

8. ACKNOWLEDGMENTS

This research was supported in part by a special projects grant from Association for Computing Machinery's Special Interest Group on Computer Science Education (SIGCSE).

9. REFERENCES

[1] Banerjee, M., Capozzoli, M., McSweeney, L. & Sinha, D. Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1), 3-23, 1999.

[2] Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382, 1971.

[3] Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. Statistical practices of educational researchers. An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350-386, 1998.

[4] Kuittinen, M. (Ed.) Kolin Kolistelut - Koli Calling 2001: *Proceedings of the First Finnish/Baltic Sea Conference on Computer Science Education.* Report A-2002-1, University of Joensuu, Department of Computer Science, Finland, 2002.

[5] Kuittinen, M. (Ed.) Kolin Kolistelut - Koli Calling 2002: *Proceedings of the Second Finnish/Baltic Sea Conference on Computer Science Education.* Report A-2002-7, University of Joensuu, Department of Computer Science, Finland, 2002.

[6] Kurhila, J. (Ed.) Kolin Kolistelut - Koli Calling 2003: *Proceedings of the Second Finnish/Baltic Sea Conference on Computer Science Education.* Report B-2003-3, University of Helsinki, Department of Computer Science, Helsinki University Printing House, Finland, 2003.

[7] Korhonen, A. & Malmi, L. (Eds.) Kolin Kolistelut - Koli Calling 2004: *Proceedings of the Fourth Finnish/Baltic Sea Conference on Computer Science Education.* Report TKO-A42/04, Helsinki University of Technology, Department of Computer Science and Technology, Finland, 2004

[8] Malmi, L. Foreward. In A. Korhonen & Malmi, L. (Eds.) Kolin Kolistelut - Koli Calling 2004: *Proceedings of the Fourth Finnish/Baltic Sea Conference on Computer Science Education.* Report TKO-A42/04, Helsinki University of Technology, Department of Computer Science and Technology, Finland, 2004, iii.

[9] Nichols, D. *MKAPPASC.SPS/MKAPPASC.TXT (SPSS Macro/SPSS read me), 1997.* Retrieved September 12, 2005 from ftp://ftp.spss.com/pub/spss/statistics/nichols/macros/.

[10] Fincher, S., & Petre, M. *Computer Science Education Research.* Taylor and Francis, 2004.

[11] Randolph, J. J. *Planning and Evaluating Programs in K-12 Computer Science Education.* Utah State University, Dissertation in progress, 2005.

[12] Randolph, J. J., Bednarik, R., Silander, P., Gozanlez, J., Myller, N., & Sutinen, E. A critical analysis of the research methodologies reported in the full papers of the proceedings of ICALT 2004. In *The 5th Annual IEEE International Conference on Advanced Learning Technologies.* Los Almitos, CA: IEEE Press, 2005, 10-14.

[13] Randolph, J. J. & Hartikainen, E. Yhtenäistyvät vai erilaistuvat oppimisen ja koulutuksen polut. In S. Havu-Nuutinen & M. Heiskanen (Eds.) *Kasvatustieteen päivät Joensuussa* 25.-26.11.2004 [Electronic Proceedings of the Finnish Educational Research Days Conference 2004] (pp. 183-193). Finland: University of Joensuu Press. Retrieved August 6th, 2005 from http://joypub.joensuu.fi/publications/other_publications/kasvtied_paivat/

[14] Randolph, J. J., Hartikainen, E. & Kähkönen, E. Lessons learned from developing a procedure for the critical review of educational technology research, Paper presented at *Kasvatustieten Päivät 2004* [Finnish Education Research Days Conference 2004], Joensuu, Finland, November 25th-26th, 2004.

[15] Shadish, W. R., Cook, T. D. & Campbell, D. T. *Experimental and quasi-experimental designs for generalized causal inference,* Boston: Houghton-Mifflin, 2004.

[16] Valentine, D. W. CS educational research: A meta-analysis of SIGCSE technical symposium proceedings. In *Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education.* New York: ACM Press, 2004, 255-259.

[17] Yin, R. K. *Case study research: Designs and methods* (rev. ed), Newbury Park, CA: Sage, 1990.

APPENDIX

In this appendix, the results for the categories that fell below the a priori lower bound of reliability can be found. We hypothesize the prevalence paradox inherent in generalizations of Cohen's kappa (see [1]) played a significant part in the low reliabilities for these categories because we used a *not applicable* category, which turned out to have a very high prevalence. We decided not to use a free-marginal kappa on these categories because we felt that the high agreement and prevalence of the *not applicable* category would artificially inflate kappa to a severe degree.

Table 7 shows how many instances of a particular measure were used in the 30 cases that reported research on human participants. Table 8, 9, and 10 show how many times a certain independent variable, dependent variable, or a moderating or mediating variable were examined, respectively. In the following tables, *frequency* refers to the number of cases in which a measure or variable was used in the 30 cases that involved research with human participants. For example, the second row of Table 7 indicates that grades were used as a measure in 10 out of the 30 cases. (The sums of the frequency columns do not equal 30 because more than one measure or variable could have been used per case.)

Table 7. Measures

Measure	Frequency
Grades	10
Diary	2
Questionnaire	9
Log files	9
Test (teacher/researcher made)	7
Interviews	1
Direct observation	1
Teacher survey	0
Test (standardized)	1
Narrative analysis scheme	0
Number of resubmitted exercises	0
Time on task (electronic)	1
Focus groups	1
Existing attrition data	2

Table 8. Independent Variables.

Variable	Frequency
Student instruction	18
Teacher instruction	1
CS fair/contests	0
Mentoring	0
Speakers at school	0
CS fields trips	0
Other	3

Table 9. Dependent Variables.

Variable	Frequency
Stakeholder attitudes	13
Enrollment (attendance)	4
Achievement in core subjects	1
CS achievement	10
Teaching practice	3
Intention for future CS jobs	0
Program implementation	1
Costs and benefits	0
Socialization	0
Computer use	0
Other	8

Table 10. Moderating or Mediating Variables.

Variable	Frequency
Gender	1
Aptitude	0
Race/ethnic origin	0
Nationality	0
Disability	0
Type of course	5
Other	5