A REVIEW OF RESOURCES FOR EVALUATING K-12 COMPUTER SCIENCE EDUCATION PROGRAMS

Justus J. Randolph, Elina Hartikainen

Education in computer science (CS), defined as "the study of computers and algorithmic processes, including their principles, their hardware and software designs, their applications, and their impact on society" (Tucker et al. 2003, 4), is considered to be a key factor in preparing K-12 students for a socio-technological future. The National Research Council Committee on Information Technology Literacy (1999) [hereafter, *the NRC*] provides strong rationales for teaching students about technology and computer science. On the level of the individual, they argue that people will increasingly need to understand technology to carry out personally meaningful and necessary tasks such as

- using e-mail to stay in touch with family and friends,
- pursuing hobbies,
- helping children with homework and projects, and
- finding medical information or information about political candidates over the World Wide Web.

On a societal level, the NRC makes an argument that an informed citizenry must also be a citizenry that has a high degree of technological fluency because many contemporary public policy debates are associated with information technology. For example, the NRC writes that

> *A person with a basic understanding of database technology can better appreciate the risks to privacy entailed in data-mining based on his or her credit card transactions. A jury that understands the basics of computer animation and image manipulation may have a better understanding of what counts as "photographic truth" in the reconstruction of a crime or accident . . . A person who understands the structure and operation of the World Wide Web is in a better position to evaluate and appreciate the policy issues related to the First Amendment, free expression, and the availability of pornography on the Internet. (4-5)*

Besides the personal and societal needs for people to understand information technology, there are also dire economic needs. For example, The (U.S.) Labor Secretary's Commission on Achieving Necessary Skills, (as cited in "Pros and Cons" 1998), says that "those unable to use technology face a lifetime of menial work."

K-12 computer science education (CSE) is one of the keys to preparing students to meet the social and economic needs of the future (Tucker et al. 1998). K-12 CSE helps meet these needs in two ways. First, the K-12 CSE curriculum, as envisioned by Tucker et al., is aligned with the needs of post-secondary CSE programs so that CS students will be more prepared to get higher degrees, which are necessary for most jobs in computing. Second, the Tucker et al. curriculum for K-12 CSE, which is endorsed by the largest computing organization - the Association for Computing Machinery (ACM), is also designed to prepare students, even those who will not specialize in CS, to participate in a technological future. This is evident from Tucker et al.'s K-8 CSE curricular goals that are described below:

> *In order to live and work successfully in an increasingly information-rich society, K-8 students must learn to use computers effectively and incorporate the idea of algorithmic thinking into their daily problem-solving vocabulary. To ensure these outcomes, schools must provide computing tools that enable students to solve problems and communicate using a variety of media: to access and exchange information; compile, organize, analyze, and synthesize information; draw conclusions and make generalizations from information gathered; understand what they read and locate additional information as needed; become self-directed learners; collaborate and cooperate in team efforts; analyze a problem and develop an algorithmic solution; and interact with others using computers in ethical and appropriate ways. (10) (reprinted with permission)*

Despite the ostensible value of adopting a coherent CSE curriculum, Tucker et al. (1998) report that CSE curricula, in general, have been poorly implemented in U.S. schools. As a side note, Tucker et al.'s findings suggest that the degree of implementation and alignment with a coherent CSE curriculum could be an important program variable in CSE program evaluation.

The Benefits of program evaluation for K-12 CSE

Weiss (1987) describes evaluation mainly as a decision-making and organizational learning tool. Fetterman and Pittman (1986) document the benefits that program evaluation can have for empowering stakeholders. Mark, Henry, & Julnes (2000) describe evaluation as sensemaking as a means for social betterment. Whatever variety of formative evaluation, it is reasonable to say that it can be a moderating variable in program success. Although evaluation is usually thought of as an activity independent of the intervention of a program, evaluation, at least in its formative sense, is an intervention itself.

Our assumption is that the effectiveness of K-12 CSE education in meeting impending economic and social needs is moderated by the quality of program evaluation, which is in turn moderated by the quality of evaluation resources available to K-12 CSE evaluators. If our assumption is correct, then creating high-quality K-12 CSE evaluation resources will lead to better evaluation that will lead to better K-12 CSE. Finally, we assume that K-12 CSE will ultimately lead to the fulfillment of economic and social needs.

The purpose of this paper is to review the state of K-12 CSE evaluation resources so that the next generation of evaluation resources can improve on, or fill the gaps in, the current generation of resources and lead, eventually, to meeting the socio-technological and economic needs of the future. The study question that will be answered in this review is

> To what degree do the current K-12 CSE program evaluation resources have feasibility, propriety, accuracy, and utility?

This rest of this paper describes the criteria used for evaluating the current K-12 CSE program evaluation resources (we use program in the sense of *project,* not in the sense of *software*) and the criteria that were used to make relevancy decisions. It also includes a description of the search strategy, details of study categories, information about the reviewers, data analysis procedures, and results of the literature review. The procedures in this literature review are adaptations of guidelines given in Cooper (1984).

Criteria for valuing the quality of evaluation resources

In this section, we discuss the criteria, which are interpolated from the Program Evaluation Standards (Joint Committee on Program Evaluation Standards [hereafter *Joint Committee*] 1994), that were used to evaluate the quality of K-12 CSE evaluation resources. Each of our criteria for high-quality evaluation resources is based on one of the four categories of program evaluation standards (utility, feasibility, propriety, and accuracy) (Joint Committee 1994).

In this paper, by *K-12 CSE evaluators*, we refer to professional, external evaluators and to K-12 CSE practitioners who are asked to conduct internal evaluations or to take part in participatory evaluations. By *evaluation resource,* we refer to physical or electronic texts that have the primary purpose of helping others conduct K-12 CSE program evaluation.

Feasibility

One of the lessons learned from CSE evaluation practice is that evaluation models and resources are often too tedious or cumbersome for CSE evaluators and CSE practitioners who are also given the job of program evaluation (Billings 1986, Carbone & Kaasbøll 1998, Randolph & Eronen 2004, Randolph, Virnes, & Eronen 2004). Therefore, we argue that high-quality evaluation resources should consist of information that is more than theoretical; the information should lead to evaluation activities that can actually be carried out.

We make a distinction between feasibility and utility in that feasibility refers to the ability for recommendations in the resource to be carried out successfully; utility refers to the ease or perceived inclination for evaluators to use the resource. For example, a checklist of evaluation questions may have ostensible ease to beginning evaluators; however, it is unlikely that a checklist alone would lead to a successful evaluation. The feasibility questions that were used for each resource were

Are the guidelines outlined in this procedure feasible for most evaluation situations?
Are the guidelines practical and keep the amount of disruption to a minimum?

Propriety

As a criterion, we regard propriety in two senses. The first is in the sense of ethical conduct. The second is in the sense of conducting evaluation in the proper context. We suggest that high-quality evaluation resources must address ethics in terms of the treatment of human participants and in terms of the just distribution of social and educational benefits and risks resulting from a program (and from program evaluation itself) (see Thompson-Robinson, Hopson, & SenGupta 2004; Sirotnik 1990). The consideration of social justice is particularly important for K-12 CSE because of the well-established pipelining of female students of computer science (Clark & Teague 1994; Gürer & Camp 2002; Galpin 2002; Teague 1997 2002), because technological advances have both strong positive and negative impacts within and across societies and cultures (Wilson 1998), and because of the fact that an individual's lack of technological acumen will become a factor that increasingly limits participation in the social, political and economic arenas of the future (Tucker et al. 2003; NRC 1999).

Since evaluation begins with the case, it is safe to assume that evaluation resources that are most closely related to the evaluation case will be the most appropriate. For example, one would expect that evaluation resources that deal with the specific theories and practices of K-12 CSE evaluation would have a greater impact on K-12 CSE evaluation than would a general set of evaluation resources. Although general evaluation resources may be helpful for CSE evaluators, in this framework we chose to examine only the resources that were purportedly contextualized to K-12 CSE evaluation.

In summary, we evaluated the propriety of resources based on the following questions:
If the resource is an evaluation model,
(1) Does it address the treatment of human participants?
(2) Does it take into account multiple, stakeholder perspectives?
(3) Does it address the differential effects of the program on participants?
(4) Is it specifically contextualized to the field of K-12 CSE?

Accuracy

Concerning accuracy, we make the assumption that evaluation and evaluation resources both need to be built around accurate data. The accuracy-related question that is asked of each evaluation resource was

Are the guidelines based on solid, current evaluation research?

Utility

Even if evaluation resources meet the first three criteria, there is no benefit to be gathered from evaluation if the evaluation resources are not utilized. Cronbach, (as cited in Torres, Preskill, & Piontek 1976) wrote that

> *The proper function of evaluation is to speed up the learning process by communicating what might otherwise be overlooked or wrongly perceived. . . Success is to be judged by . . . success in communication. Payoff comes from the insight that the evaluators' insight generates in others. (3)*

Just as the payoff in evaluation comes from the insight generated in others, the payoff in evaluation resources comes from the insight generated in others as well. The utility-related question that was asked of each evaluation resource was

> Are the guidelines outlined in this resource easy to use for most types of evaluators?

Procedures for conducting a critical review of K-12 CSE evaluation resources

In this section, we discuss the procedure for conducting a critical review of K-12 CSE evaluation resources. After a comprehensive search for articles that met the criteria for selection, two reviewers used the four criteria described above as the framework for mixed-method valuing, in the spirit of Eisner's connoisseur evaluation (1984), of CSE K-12 evaluation resources. Inter-rater reliability statistics are reported for relevancy decisions and ratings of the quality of resources.

Criteria for inclusion

K-12 CSE evaluation resources, physical or electronic texts on K-12 CSE program evaluation that have a primary purpose of helping others conduct K-12 CSE program evaluation, were included in this review if they met the following conditions:

1. They were written in English, and
2. There explicit purpose was to give guidelines for the summative, formative, and implementation evaluation of K-12 CSE programs or projects. (Examples of CSE programs might include curriculum alignment programs, instructional projects, or teacher training programs.)
3. The resources did not concern the evaluation of particular interventions, teaching strategies, students, or resources that commented on conducting evaluations in general.
4. The resources did not concern the evaluation of computer assisted instruction (CAI) programs, unless CAI was a critical aspect of computing education in that program.
5. Evaluation resources for technology education, which normally subsumes CSE, were not included in this review unless they specifically addressed the evaluation of CSE programs.

For example, Billings' (1986) *An Evaluation Handbook for a Computer Education Program* was included in this review because it was written in English and gave direct guidelines for the evaluation of K-12 computer education programs. NSF's *User-Friendly Handbook for Mixed-Method Evaluations* was not included because it did not specifically address the evaluation of K-12 CSE programs, although it did address the evaluation of programs in general. An article that was on the border, but was not included, was Carbone & Kaasbøll's, *A Survey of Methods Used to Evaluate Computer Science Teaching* (1998) because it addresses the research on CS teaching rather than giving guidelines for evaluating CSE programs. Fincher and Petre's *Computer Science Education Research* (2004), although it would be an indispensable resource for a CSE researcher, was not included because its focus was research rather than program evaluation.

Primary relevancy decisions were made by the author of this proposal. To establish reliability for relevancy decisions, a second rater was asked to make independent relevancy decisions for the set of resources that made it to the final round of the search for resources. The final round included eight resources that met the criteria and four resources that did not. The Kappa statistic was used to indicate the measure of inter-rater reliability of relevancy decisions. Because of the controversies surrounding the use of Kappa as a single rating of inter-rater reliability (Agresti 1996, Crocker & Algina 1986), the *phi* coefficient (i.e., a correlational measure of two binary variables) was also included as a secondary measure of inter-rater reliability.

Search strategy

The search for evaluation resources of K-12 CSE programs was conducted using three strategies. First, a number of leading academic databases via *Ebsco Host* were searched, on July 7th, 2004, using the key words '*program evaluation*' and '*computer science education*' with no limiters placed on year of publication. Second, an Internet search with the Google search engine using the keywords "*computer science education"* and "*program evaluation"* was done. All of the listings that could not be excluded from the link description were searched until a relevancy decision could be made. Third, after a preliminary list of K-12 CSE evaluation resources was made, a message was sent to the Association for Computing Machinery's Computer Science Education Special Interest Group's list serve (ACM SIGCSE n.d.). The message, sent to the 959 members of the list serve on October 31, 2004, asked the members of the list serve to send information about K-12 CSE evaluation resources that met the criteria for inclusion and that were not already on the preliminary list.

Details of study coding categories

Besides demographic information about each resource, the reviewers evaluated the resource in terms of its feasibility, utility, accuracy, propriety, and overall impression. For each applicable criterion, the reviewers gave a rating on a five-point scale. In addition, the reviewers answered the following questions or carried out the following tasks:

What was your overall impression of this evaluation resource?
Would you use it yourself?
Would you recommend it to a colleague?
Please write a short blurb about this resource.

In general, the Evaluation Resources Coding sheet, which was a document that collected all of the review framework questions, and the procedures section of this paper served as the reviewing protocol (Yin 1989). The entire Evaluation Resource Coding Sheet was not used on resources that had a page or less dealing with program evaluation. In those cases, the reviewers only answered the question

What was the overall impression of this evaluation resource?

Reviewer information

The first reviewer, listed as the first author of this paper, is a PhD student in Utah State University's Education Research and Evaluation program who has four years experience in program evaluation. The second reviewer, listed as the second author, is a PhD student in the University of Joensuu's Department of Educational Science. The second reviewer has four years experience in education research.

Data Analysis

For quantitative ratings, only the first reviewer's ratings are reported. The second reviewer's ratings were used to establish inter-rater reliability. Inter-rater reliability of ratings was calculated, with SPSS 11.0, using Pearson's *r* and a two-way mixed-model, single-measure intra-class correlation (consistency definition) using the procedures described in Nichols (1998) and Wuensch (2003 a,b). For qualitative analysis, an emergent coding procedure, as outlined in Merriam (1997) was used to generate categories of findings from each reviewer's ratings.

Results

The search resulted in nine K-12 CSE evaluation resources, which are preceded by an asterisk (*) in the reference section, that met our criteria for inclusion. (For analysis purposes, we grouped Billings 1985 and Billings 1986 together because they are slight variations on the same resource. After grouping the Billing's articles together, there were actually eight resources.) The kappa statistic for the reliability of relevancy decisions was originally 0.57 (10 out of 12 possible agreements); the *phi* coefficient was 0.63. After a calibration process on the criteria for inclusion, relevancy decisions for the Kappa statistic rose to .75 (11 out of 12 agreements) with an asymptotic standard error of .23 and an approximate p value < .01; after calibration the *phi* coefficient was 0.78 (p < 0.01). For the inter-rater reliability on the five resources that had a section dealing with evaluation for more than a page (Almstrum et al. 1996, Anderson 1987, Billings 1985 1986, Ferguson 1985, Torvinen 2004), Pearson's *r*, was 0.70 (p < 0.001, n=25). The intra-class correlation between ratings was 0.69 with a 95% confidence interval upper bound of 0.86 and a lower bound of 0.43.

Table 1 provides a short description of each resource. Table 2 shows ratings for the five resources that were greater than a page in length.

Table 1 Overview of K-12 CSE Evaluation Resources

Evaluation Resource	*Description*
Almstrum et al. 1996	A 16-page article that gives a review of research designs for CSE evaluation. It includes hypothetical examples and discuss the role of technology in evaluation.
Anderson 1987	The evaluation part of this resource is a 7-page evaluation checklist for implementing K-12 CSE programs in Wisconsin, U.S.A., secondary schools.
Billings 1985 1986	These resources are a handbook for evaluating CSE programs and a dissertation that describes the development and improvement of the handbook.
Brady-Ciampa 1983	This resource gives general guidelines, in less than a page, for evaluating the integration of computer literacy into a general education program.
Ferguson 1985	In this resource, Ferguson presents a 22-item checklist for overall, formative, and summative evaluation design of computer literacy programs for K-8 students.
Kurtz et al. 1993	This is an abstract for a panel discussion on evaluating effectiveness of CSE.
Saskatchewan Education 1983	In this resource, two paragraphs are used to discuss the evaluation of CSE programs in secondary education.
Torvinen 2004	This is a multiple-paper format licentiate thesis that concerns the evaluation and implementation of a distance CSE program for secondary school students in eastern Finland.

Note. CSE = computer science education

Table 2 Ratings for Evaluation Resources Longer than One Page

Resource	*Feasibility*	*Propriety*	*Accuracy*	*Utility*	*Overall*
Almstrum et al. 1996	5	3	4	4	5
Anderson 1987	2	3	2	3	2
Billings 1985 1986	5	3	3	4	4
Ferguson 1985	1	1	2	1	1
Torvinen 2004	2	2	3	3	3

Note. Ratings are on a scale from 1 to 5, where 1 is the lowest and 5 is the highest.

Qualitative themes, their evidence, and discussion

This section presents the themes generated from the qualitative data in the Evaluation Resource Coding Sheet and from personal communication between the reviewers. The theme is italicized. Evidence for and discussion of that theme, when it is not apparent, is given after theme.

Feasibility

F1. *The most feasible resources are the ones that presented a variety of choices that could be adapted to different contexts and allow enough specificity for evaluation practices to be carried out.* For example, Alstrum et al. (1996) discussed the context and constraints of an evaluation as a key factor in the choice of research and evaluation designs and was consequently given a high rating of feasibility. Anderson's (1987) checklist of evaluation questions such as

> *Have you developed an evaluation design and questions? . . . Have you allowed for flexibility? . . . Have you evaluated by objectives? (7-8)*

were considered to be too general to result in a beginning evaluator conducting a successful evaluation.

Propriety

P1. *No models discuss the ethical treatment of human participants.*

P2. *Little attention is paid to stakeholder representation in the planning, implementation, and interpretation of evaluations.* Only two resources, (Billings 1985 1986, Torvinen 2004), directly address including stakeholders in the process of the evaluation. Torvinen advocates a participatory action research approach to formative evaluation; however, participatory action research materializes in her evaluation only in the sense that she and her coauthors were involved in the planning and implementation of the program. No stakeholder, other than researchers, were involved in the planning, implementation, or interpretation of the results.

P3. *Differential effects are usually not taken into consideration*, Except for one item in the Ferguson (1985) checklist, - "within the school district, there is equity in the program from school to school" - differential effects are not taken into consideration in the resources in this sample.

P4. *CSE-grounded resources concentrate mainly on research. Evaluation-grounded resources concentrate mainly on the conduct of evaluation.* The major resources (i.e., those longer than a page) fall into two categories: CSE-grounded evaluation sources (Alstrum et al. 1996, Torvinen 2004) or evaluation-grounded resources (Anderson 1987, Billings 1985 1986, Ferguson 1985). (A look at the authors' affiliations supports this assertion.) In the CSE-grounded group, the resources concentrate on research methods, but offer little in the tradition of program evaluation. For example, Almstrum et al. and Torvinen detail methods of research without speaking of methods of evaluation. Both use the term *formative evaluation*, but in a very general sense of using research for intervention improvement. In the evaluation-grounded group, they concentrate on the conduct of program evaluation with little discussion on the contemporary practices, theories, or debates in CSE. Neither group seems to bridge the divide between the fields of CSE and evaluation research well.

Utility

U1. *Resources have utility for beginning evaluators, but they do not have much to offer for experienced evaluators.* This theme was generated on the assumption that beginning evaluators will benefit from resources that explain how to conduct an evaluation and that experienced evaluators, since they already know how to conduct an evaluation, will benefit from resources that explain what are the evaluation practices in a given field. For example, experienced evaluators would benefit from CSE resources that discuss the evaluation debates in CSE evaluation, the types of logic models in CSE programs, the big theories in CSE, the evaluation designs, outcomes, and measured being used. In this sample of resources, the resources are all guidelines that have utility only for beginners.

U2. *As a strength, some resources address the use of technology as tool in evaluation.* Alstrum et al. (1996) anticipated the growing interest in the evaluation community for using technology as a tool for program evaluation (Love 2004) and as a tool for public policy (Mean, Roschelle, Penuel, Sabelli, & Haertel 2003). They discuss the benefits that technology can have for evaluation in addition to the ubiquitous argument about the benefits of evaluation for technology. Practitioners in the cross-disciplinary field of CSE and technology educational evaluation seem to be, logically, the most appropriate authors for this type of resource.

Accuracy

A1. *Few resources refer to recent literature.* Of the five major resources in this sample, only two were published in the last ten years.

A2. *If resources do refer to recent literature, their references are grounded in CSE research or program evaluation almost exclusively.* A cursory examination of the references listed in the resources classified as CSE-grounded or evaluation-grounded show that references in the majority of cases stay within their respective fields. In terms of CSE, this finding is supported by Glass, Ramesh, & Vessey's (2004) findings that in CS and software engineering articles, 89% and 98% of the references are to articles within their own respective fields.

Unexpected outcome: Differing definitions of evaluation

UE1. Because the CSE-grounded evaluation resources concentrate on what program evaluators would call *applied research*, we hypothesize that the term *evaluation* as it is used in the CS literature equivocates between *evaluation* in the sense of *valuing* and in the sense that the term *evaluation* can be used in the computing sciences. The definition of *evaluation* as it is listed in the *Online Computing Dictionary* (n.d.) is shown below:

1. Converting an expression into a value using some reduction strategy.
2. The process of examining a system or system component to determine the extent to which specified properties are present.

The second sense of evaluation in the definition directly above, which is akin to *analysis,* is the one that we found often in the CSE literature.

Below is an example, which was found in one of the CSE-grounded resources in this sample, of *evaluation* being used in the *analysis* sense.

> *"By combining the evaluation of the learning results and the analysis of student feedback. . ." (Meisalo, Suhonen, Sutinen, & Torvinen 2004, 134; included in Torvinen 2004).*

In this case, *evaluation* is clearly used as a synonym of *analysis* rather than as a synonym for *valuing* because the program, not the learning results, is the evaluation object, (i.e., the thing to be given value to.) To dispel an alternate theory that the use of *evaluation* as *analysis* is an artefact of non-native English, another example is provided from an article where all but one of the authors is affiliated with a university in a primarily English-speaking country

> *"… if [an educational technology is] applied with the deliberative study of its use in context and without the evaluation of the technology's impact on this use, 'educational' technology remains a toy" (Alstrum et al. 1996, 201).*

In this sentence, we conclude that the authors use *evaluation* in the sense of *analysis* rather than in the *valuing* sense because it is unlikely from the context of the rest of the passage that the authors mean ". . . . *the valuing of the technology's impact on this use.*" It is likely that they mean, ". . . *the analysis of the technology's impact on this use.*"

In the CSE-grounded resources we saw *evaluation* used both in the *analysis* sense and in the *valuing* sense; however, in the program evaluation-grounded resources we only noticed *evaluation* being used in the sense of *valuing.* We hypothesize that the differences in the uses of the term *evaluation* across the two fields could account for the differences in the content matter of articles written on the subject of evaluation (see P4).

Conclusion

This study has shown that the current resources could be improved in a number of ways. In terms of feasibility, evaluation resources have to be general enough to apply to a number of contexts but specific enough to enable procedures to be carried out. Regarding propriety, there are several issues. Guidelines addressing the treatment of human participants involved in evaluation are entirely absent. Although, there are substantial gender differences in the computing disciplines, none of the resources in this sample directly addressed how to evaluate programs in a way that deals with a program's differential effects. Also, contrary to current evaluation standards, little is mentioned about the inclusion of stakeholders in the design, conduct, and interpretation of evaluations and evaluation results. What's more, the current resources do not bridge the divide between disciplines. In terms of utility, several resources do a very good job at making evaluation practice clear to beginners; however, they offer little for experienced evaluators. The discussion of the use of technology is a niche that cross-disciplinary evaluation practitioners would be well-qualified to fill. Concerning accuracy, it was found that there are few resources that are recent. If they are recent, they are self-referential and do not capitalize on the added value that is expected from combining what is known in evaluation research with what is known in CSE research. Finally, it is hypothesized that there is a difference in the way that CSE-grounded and program-evaluation-grounded researchers use the term *evaluation.*

REFERENCES

ACM Special Interest Group on Computer Science Education. (n.d.). *ACM Special Interest Group on Computer Science Education home.* Retrieved November 1, 2004 from http://www.sigcse.org/

Agresti, A. (1996). *An introduction to categorical data analysis.* New York: John Wiley & Sons.

*Almstrum, V. L., Dale, N., Berglund, A., Granger, M., Little, J. C., Miller, D. M., Petre, M., Schragger, P., Springsteel, F. 1996. Evaluation: Turning technology to tool – report of the Working Group on Evaluation. *Proceedings of the 1st conference on integrating technology into computer science education* (pp. 201-217). New York: ACM Press.

*Anderson, M. E., 1987. *A guide to curriculum planning in computer education.* Madison, WI: Wisconsin State Dept. of Public Instruction. (ERIC Document Reproduction Service No. ED287469)

*Billings, K. J. 1985. *An evaluation handbook for a computer education program.* Eugene, OR: International Council for Computers in Education. (ERIC Document Reproduction Service No. ED291338)

*Billings, K. J. 1986. The development of an evaluation handbook for a computer education program. *Dissertation Abstracts Internationa*l, 47(06), 2131A. (UMI no. AAT 8620322)

*Brady-Campia, B. 1983. *A model for the integration of a computer literacy component into the general education curriculum*. (ERIC Document Reproduction Service No. ED228960)

Carbone, A. & Kaasbøll, J. J. 1998. A survey of methods used to evaluate computer science teaching. *Proceedings of the 6th annual conference on the teaching of computing and the 3rd conference on integrating technology into computer science education: Changing the delivery of computer science education* (pp. 41-45). New York: ACM Press.

Clark, V. A. & Teague G. J. 1994, March. A psychological perspective on gender differences in computing participation. In *Proceedings of the Twenty-Fifth SIGCSE Symposium on Computer Science Education* (pp. 258-262). New York: ACM Press.

Cooper, H.M. 1984. *The integrative research review: A systematic approach.* Beverly Hills, CA: Sage.

Crocker, L. & Algina, J. 1986. *Introduction to classical & modern test theory.* New York: Holt, Rinehart, and Winston.

Eisner, E. W. 1985. *The art of educational evaluation: A personal view*. Philadelphia, PA: Falmer Press.

*Ferguson, B. 1985. *Designing and evaluating a quality computer program*. U.S.A, CA: Falbrook Union Elementary School District, CA. (ERIC Document Reproduction Service No. ED256010)

Fetterman, D. M. & Pittman, M. A. (Eds.). 1986. *Educational evaluation: Ethnography, in theory, practice, and politics*. Beverly Hills, CA: Sage.

Fincher, S. & Petre. *Computer Science Education Research.* London: Routledge Falmer.

Galpin, V. 2002. Women in computing around the world. *ACM SIGCSE Bulletin 34*(2), 94-100.

Glass, R. L., Ramesh, V. & Vessey, I. 2004, June. An analysis of research in computing disciplines. *Communications of the ACM, 47*(6), 89-94.

Gürer, D. & Camp, T. 2002. An ACM-W literature review on women in computing. *ACM SIGCSE Bulletin, 34*(2), 121-127.

Joint Committee on Standards for Educational Evaluation. 1994. *The program evaluation standards* (2nd ed.). Thousand Oaks, CA: Sage.

*Kurtz, B. L., Dale, N., Engel, J., Miller, J., Barker, K., & Taylor, H. 1993. Evaluating effectiveness in computer science education. *Proceedings of the Twenty-fourth SIGCSE Technical Symposium on Computer Science Education.* (p. 293). New York: ACM Press.

Love, A. P. (Ed.). 2004, fall. Harnessing technology for evaluation. *The Evaluation Exchange, 10*(3), 1-27.

Mark, M. M., Henry, G. T., & Julnes, G. 2000. *Evaluation: An integrated framework for understanding, guiding, and improving policies and programs.* San Francisco: Jossey Bass.

Means, B., Roschelle, J., Penuel, W., Sabelli, N. & Haertel. G. 2003. Technology's contribution to teaching and policy: Efficiency, standardization, or transformation? *Review of Educational Research, 27,* 159-181.

Meisalo, V., Suhonen, J., Sutinen, J., & Torvinen, S. 2004). Formative evaluation scheme for a Web-based course design. In *Proceedings of the 7th Annual Conference on Innovation Technology and Technology in Computer Science Education (ITiCSE 2002).* (pp. 130-134). New York: ACM Press.

Merriam, S. B. 1997. *Qualitative research and case study applications in education.* Hoboken, NJ: John Wiley & Sons.

National Research Council Committee on Information Technology Literacy. 1999, May. *Chapter 1: Why know about information technology: Being fluent with information technology*. Washington, DC: National Academy Press. Retrieved October 12, 2004 from http://books.nap.edu/html/beingfluent/ch1.html

Nichols, D. P. 1998. Choosing an intraclass correlation coefficient. *SPSS Keywords, 67.* Retrieved November 19, 2004 from http://www.utexas.edu/its/rc/answers/spss/spss4.html.

Online Computing Dictionary. (n.d.). *Evaluation.* Retrieved November 19, 2004 from http://www.instantweb.com/D/dictionary/index.html

Pros and cons of technology in the classroom: The Pea/Cuban debate. 1998, February 5. Retrieved October 11, 2004, from http://tappedin.org/archive/peacuban

Randolph, J. J., Eronen, P. J. 2004. Program and evaluation planning lite: Planning in the real world. Paper presented at *Kasvatustieteen Päivät 2004* [Educational Research Days Conference 2004], November 25th and 26th, 2004, University of Joensuu, Finland.

Randolph, J. J., Virnes, M., and Eronen P. J. (in press). A model for designing and evaluation teacher training programs in technology education. In Courtiat, J-P., Davarakis, C., & Villemur, T. (Eds.), *Technology enhanced learning: Proceedings of the 18th IFIP World Computer Congress* (pp. 69-79). New York: Kluwer.

*Saskatchewan Education. 1999. *Computer science 20, 30: Curriculum guidelines for the secondary level.* Retrieved October 18, 2004 from www.sasked.gov.sk.ca/docs/cs/2030/

Sirotnik, K. A. (Ed.) 1990, Spring. Evaluation and social justice: Issues in public education. *New Directions for Program Evaluation, no. 45*. San Francisco: Jossey Bass.

Teague, J. 1997, July. A structured review of reasons for the underrepresentation of women in computing. In *Proceedings of the 2nd Australian Conference on Computer Science Education* (pp. 91-98). New York: ACM Press.

Teague, J. 2002, June. Women in computing: What brings them to it, what keeps them in it? *ACM SIGCSE Bulletin, 34*(2), 147-158.

Thompson-Robinson, M., Hopson, R., & SenGupta, S. (Eds.) 2004. In search of cultural competence in evaluation. *New Directions for Program Evaluation, no. 102.* San Francisco: Jossey-Bass.

R. T. Torres, H. Preskill, & M. E. Piontek. 1996. *Evaluation strategies for communicating and reporting: Enhancing learning in organizations.* Thousand Oaks, CA: Sage Publications.

*Torvinen, S. 2004. *Aspects of the evaluation and improvement process in an online programming course. Case: The ViSCoS program*. Licentiate Thesis, University of Joensuu, Finland.

Tucker, A., Deck, F., Jones, J., McCowan, D., Stephenson, C., & Verno, A. (ACM K-12 Education Task Force Curriculum Committee). 2003, October 22. *A model curriculum for K-12 computer science: Final report of the ACM K-12 Education Task Force Curriculum Committee.* Retrieved October 20, 2004, from http://www.acm.org/education/k12/k12final1022.pdf

Weiss, C. H. 1998. *Evaluation: Methods for studying programs and policies* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Wuensch, K. L. 2003a. *The Intraclass correlation coefficient.* Retrieved November 19, 2004 from http://core.ecu.edu/psyc/wuensch/docs30/IntraClassCorrelation.doc

Wuensch, K. L. 2003b. *Inter-rater agreement.* Retrieved November 19, 2004 from http://core.ecu.edu/psyc/wuenschk/docs30/InterRater.doc.

Wilson III, E. J. 1998. *Globalization, information technology, and conflict in second and third worlds: A critical review of the literature*. New York: Rockefeller Brothers Fund. Retrieved September 22, 2004 from http://www.rbf.org/pdf/wilson_info_tech.pdf

Yin, R. K. (1989). *Case study research: Designs and methods,* (Rev. ed.). London: Sage.