The Peer Review Process of Teaching Materials

Report of the ITiCSE'99 Working Group on Validation of the quality of teaching materials

Deborah Knox (co-chair)
The College of New Jersey, USA
knox@tcnj.edu

Don Goelman
Villanova University, USA
goelman@vill.edu

Sally Fincher (co-chair)
University of Kent at Canterbury, UK
S.A.Fincher@ukc.ac.uk

James Hightower
California State University, Fullerton, USA
hightower@acm.org

Nell Dale (co-chair)
University of Texas at Austin, USA
ndale@cs.utexas.edu

Ken Loose
University of Calgary, Canada
loose@cpsc.ucalgary.ca

Elizabeth Adams
James Madison University, USA
adamses@jmu.edu

Fred Springsteel
University of Missouri, USA
csfreds@showme.missouri.edu

ABSTRACT

When an instructor adopts teaching materials, he/she wants some measure of confidence that the resource is effective, correct, and robust. The measurement of the quality of a resource is an open problem. It is our thesis that the traditional evaluative approach to peer review is not appropriate to insure the quality of teaching materials, which are created with different contextual constraints. This Working Group report focuses on the evaluation process by detailing a variety of review models. The evolution of the development and review of teaching materials is outlined and the contexts for creation, assessment, and transfer are discussed. We present an empirical study of evaluation forms conducted at the ITiCSE 99 conference, and recommend at least one new review model for the validation of the quality of teaching resources.

1 INTRODUCTION

1.1 Web-based Resources

Computer science educators are faced with an environment that changes quickly. We are experiencing burgeoning enrollments, a diverse student population, and a need to remain current in our technology knowledge base. At recent SIGCSE Technical Symposium meetings, faculty expressed need to access materials in support of their teaching. These needs include access to traditional material such as syllabi, tests, and projects, as well as to innovative teaching materials. The latter might include interactive software, visualizations, multimedia based units, etc. Immediate access to materials is now technologically feasible, allowing the easy dissemination of such resources. A number of web sites are available in support of the quest to find teaching materials.

Among these web sites there are generalized lists of materials as well as specialized sites devoted to particular areas of computer science. One (of many) such useful resources is Computer Science Education Links, which is a categorized list of links to teaching materials [McCauley 1999]. Another listing of CS related materials, some of which are tools to use in support of teaching, is Computer Science Education Resources [Barnett 1999]. Users don't have time to browse, so the above collections of materials are helpful. Users need ease of use; this suggests that good navigation support (searching versus browsing) is desirable.

Repositories can supply needed materials in a unified framework, providing a guarantee of the quality of the materials. None of the "collection" sites noted above supports a strong review model. One repository site that does provide reviewed materials is the National Engineering Education Delivery System (NEEDS) [Muramatsu 1999]. While this repository focuses on engineering materials, there is some overlap in the disciplines. (A recent announcement indicates that the NEEDS digital library will expand to cover all areas in science, math, engineering, and technology.) Of special note is their premier courseware competition. In each of the past two years, approximately five courseware packages have been awarded premier status. Each of these packages has undergone an extensive review process. This evaluation process is detailed in [Muramatsu 1999] and discussed in Section 2.

The development of the Computer Science Teaching Center (CSTC) is supported by the National Science Foundation and by the ACM Education Board. One focus of the CSTC is on increasing the availability of materials to enhance the teaching and learning of computer science [Knox 1999]. This digital library is being designed to support the access of quality teaching materials, including peer reviewed materials.

1.2 Approaches to evaluation

The question of what makes a laboratory project "good" or what makes a visualization demonstration "worthy" of class or lab time is a question deserving investigation. A second fundamental question is how any resource not developed "in house" enhances the learning experience of our students. The first phase of addressing these questions is to validate the quality of the material, i.e., to provide a level of confidence that the material is sound and well-founded for the topic.

It is important to insure the quality of the materials available for a number of reasons:

- We want to provide an enriching learning experience for our students, at minimal extra cost (in time or effort) to educators.
- We want to gain the confidence of users of the materials so they will revisit the repository and use additional materials.
- We want educators to be encouraged to submit materials for inclusion.

As professionals, we accept a variety of measures of quality. These measures include the peer review of written material to be published in journals or presented at conferences, and established criteria for accreditation of programs of study or institutions.

The evaluation of teaching materials is an open research question. In the area of computer science education, we are accustomed to reviewing papers describing teaching methods or projects, e.g., the SIGCSE Technical Symposium, but in general there is neither resource nor forum for refereeing teaching materials. We need to explore and establish appropriate methodologies for the review of teaching materials.

1.3 Progression of Working Group Contributions

This Working Group builds upon the work of the 1998 Dublin Working Group, who started collecting materials on assessment of teaching materials (http://www.tcnj.edu /~cstc) and made recommendations to utilize an Editorial Board and a formal review process [Grissom 1998, ACM].

At the 1997 ITiCSE Conference, a Working Group convened to discuss the peer review of laboratory materials [Joyce 1997]. This group categorized submissions to the predecessor of the CSTC and identified qualities of a good lab, e.g., portability, completeness, outstanding content, successfully class-tested and subsequent revision prior to review, stimulates learning, stimulates student interest in the topic, and flexibility. These were features recommended for identification during a peer review process. This initial attempt to identify qualities of good lab materials was only a beginning to the process of ensuring quality resources. A more formal approach needs to be established, a problem which this Working Group addressed.

While we frequently discuss the CSTC in this Report, it is our belief that the recommendations of this Working Group Report are applicable to other repositories as well.

1.4 Organization of this Report

The Working Group focused on the mechanisms that could be used to instill confidence in a user about the quality of adopted teaching materials. Peer review of resources was determined as the most appropriate means. The next section of this report considers how reviews are conducted for traditional media, software, and research papers. Section 3 identifies the stakeholders in the review process: submitter, reviewer, editor, and users. In addition, we present a model for the review process that starts in the context of submission (creation), progresses through the context of assessment and concludes with the context of use, which results in the transfer of materials (adoption). In Section 4, we present an empirical study conducted during the ITiCSE 99 conference. Five different styles of review forms are outlined and results of a survey of CS educators are presented. Reliability testing was performed and is reported on in Section 4 as well. The section finishes with a recommendation that a scaled, multiple section review form be applied to teaching materials. Our report concludes with some thoughts for future work in Section 5. A variety of evaluation models are included in the Appendix, as well as tabular results from the empirical studies. Additional materials are housed at www.tcnj.edu/~cstc/krakow/ appendix.html.

2 BACKGROUND

The advent of the web has made the exchange of post-secondary teaching materials easy and convenient. With the ease of distribution come questions about quality. As such questions are relatively new to university-level faculty, we undertook to examine previous work which had concentrated on teaching materials developed for elementary and high school education. Note that our definition of teaching materials in the introduction is very inclusive. Although we focused on the evaluation of materials in a computer-based repository, this does not mean that the materials must be computer-based. Therefore, we first look at evaluating traditional teaching materials.

2.1 Traditional Media

Media and the Curriculum is one of a three volume set entitled *Selecting Materials for Instruction* [Woodbury 1980]. This book contains an in-depth look at different media with suggested evaluation criteria. Although the guidelines are designed for elementary and high-school materials, there are certain suggestions that might be useful for post-secondary materials. There is an emphasis on the use of checklists.

The chapter on evaluating pictorial media outlines instructional objectives for pictures and provides a 4-part scale by which a reviewer can rate the materials against the objectives. Another form asks yes/no questions about the

quality of the pictures themselves. There is a yes/no checklist for evaluating textbook illustrations accompanied by a method for quantifying the results: the number of yes's as a percentage of total number of items less those marked non-applicable.

Traditional criteria for evaluating print materials are listed, including accuracy, authenticity, currency, literary quality, content, organization, and age-level appropriateness. This category includes textbooks, curriculum guides, magazines, and newspapers.

Lists of questions organized under the following categories are suggested as guides for evaluating non-print media.

- authenticity
- technical qualities
- utilization
- overall rating
- content

There is also a list of criteria with the section titles including "appropriate to purpose" and "appropriate to users." A further list includes such categories as aesthetic value and concept development.

The following criteria are suggested as guides for evaluating games and simulations.

- Does it teach or reinforce anything?
- Is it fun?
- Does it create a more positive attitude toward the subject in general?
- Does it encourage more interest and learning of the subject?
- Is it adaptable?

Another list of simulation criteria includes categories such as interest and verisimilitude (are the right things abstracted).

To evaluate television as a learning tool, a 7-point scale for questions including accuracy of content, relevance of content, quantity of material covered, pacing, level of material, organization and planning, and follow up possibilities are suggested.

In summary, two points stand out in all of the evaluation criteria listed for media evaluation. The first is that all of the checklists are directive. They state a principle and ask if it has been met (yes/no/NA) or they have a scale upon which to measure how completely the principle has been met. The second is that "meets objectives" is included in all checklists, either implicitly or explicitly.

2.2 Software

2.2.1 Duchastel

The use of software as a teaching tool began with Plato [Alessi 1985] in the 1960's, but never blossomed until the advent of the microcomputer in the 1980's. Philippe C. Duchastel summarizes the history of the call for evaluation of educational software products [Duchastel 1987]. Duchastel describes three models for educational software evaluation: product review, checklist procedure, and user observation. Product review by an individual is subjective but capitalizes on a person's expertise. Reviewers have a mental set of categories and characteristics, which they use in making a review.

The checklist procedure tries to systematize the evaluation process by requesting the evaluators to rate the product on a delineated set of characteristics representing a number of dimensions. As Duchastel points out, the tricky part of the process is to determine the correct characteristics—the parameters of good educational software.

User observation is a review model that examines the educational software in a laboratory setting. Students are often video taped while interacting with the software, so that the session can be further analyzed later. This model is very rich in data, but rarely performed because of the costs involved.

2.2.2 SYNTHESIS

The SYNTHESIS Coalition (http://www.synthesis.org/), a National Science Foundation sponsored coalition of eight schools, developed an electronic database of engineering educational courseware, called the National Engineering Education Delivery System (NEEDS). The NEEDS database includes three types of materials: non-reviewed courseware, endorsed courseware, and premier courseware. They conducted a literature search into evaluation techniques for educational courseware, from which they created an extensive checklist review form to use to review endorsed courseware and tested it with a large group of engineering educators. The review form was determined to be too long and complicated, so they compromised on a ten-question yes/no form for endorsed courseware, which is included in this Working Group's web materials (see appendix). The premier courseware award is reserved for exceptional courseware determined in competition. The evaluation form for premier awards is an extensive two-page form and is included as part of the web based appendix.

2.2.3 ECALM

Evaluating Computer Assisted Learning Material produced by Durham University provides a different perspective for reviewing software [Harvey 1997]. Rather than view the process from the standpoint of an expert evaluating a submitted resource, they address the issue of how a user should go about evaluating a piece of software for his or her own use. This report recommends that a prospective user think about which aspects of the package are important for his or her particular needs. These aspects then form the basis for a checklist, which can act as a guide during the first stage of evaluation. The four different aspects, which Durham University suggests as a start, are:

- subject content and material structure
- usability

- pedagogy and the quality of the approach adopted by the package and how it encourages quality in learning through assessment, and
- layout and the stylistic presentation of the material within the package.

2.3 Paper Reviews

As outlined below (3.0), we determined that papers and teaching materials are fundamentally different. Papers are written to inform our colleagues; teaching materials are written to enhance learning in our students. Nevertheless, an examination of paper review forms gives us some insight into the types of forms we might use. A selection of paper review forms is given in the web appendix.

2.4 Summary

From this survey of existing mechanisms of evaluation used by different communities we identified two areas for further consideration:

- The use of checklists (whether designed to be used against stated criteria or whether designed to elicit tacit reviewer knowledge) appeared to be a common and successful method for capturing evaluative judgements.
- The evaluation criteria within the categories: *technical soundness*, *appropriateness for the audience*, *meets its stated goals*, and *evaluation of writing style* all seemed to be appropriate, as did the additional *ease of use* category. These categories provided the basis for the draft forms used in the empirical study.

In the next section, we reflect on how to review computer science teaching materials. This broad investigation led us to examine why we want to review and to determine the stakeholders in the process.

3 DEVELOPMENT OF A CONTEXT MODEL

The concept of a repository of peer-reviewed teaching and learning materials (electronic or not – although this report confines itself to electronic) is a relatively new one, with a short history. It has drawn on two existing models: the peer-review of research papers and the notion of a library.

Both of these models have a long history, and the transition of their use to this new endeavor is not entirely fluid.

3.1 Stakeholders in the process

Peer review of research papers is a well-understood mechanism within the academic community. Its purpose is to guarantee the rigor of methodology, originality and acceptability (to the research community) of the work reported. It is a "gatekeeper" mechanism that defines certain threshold standards for given, well-defined disciplinary research areas. This mechanism works because publication (the dissemination of results for the advancement of the discipline) is a public endeavor which academics owe to their geographically distributed research community. One consequence of this is that the dissemination products of the research endeavor (articles, papers and other publications) are all constructed with the specific intention of being submitted to this formal public scrutiny.

Peer-review of teaching materials is a more complex matter. First, teaching occurs almost wholly in private, behind the closed classroom door. There is neither public currency nor consensual standards between pieces of practice or among practitioners. Consequently, it is difficult to understand the process of peer-review in the same way. With research, there are at least two primary stakeholders in the process: the submitters (who seek entry to the community and status within it) and the reviewers who arbitrate on their acceptability. In a teaching-materials review process (and especially in the proposed review process for a repository) we have identified four categories of stakeholder:

- the submitter of the resource
- the reviewer of the submitted resource (as called upon by the repository editor)
- the editor as engaged in the post-review decision of whether to admit the submitted resource or not, and
- the users. This category can be seen to consist of two distinct elements, teachers who incorporate materials into their classes and students who use the resources in the process of their learning.

Consequently, when considering the selection of appropriate of review criteria, all these stakeholders have to be accounted for, as shown in Figure 1.

The expectation of the submitter of course materials is that others will be excited about the material and find it useful. The submitter is primarily interested in the acceptance of the resource. If it fails to be accepted, there needs to be appropriate feedback to the submitter so that a decision can be made regarding revision and resubmission.

The review is concerned with the problem of properly conveying, in a constructive way, any difficulties found in the course materials. The reviewer wants this done as efficiently as possible so that the review can be dispatched easily.

The editor is concerned with the integrity of the collection of accepted course materials, and that the reviewer conveys information regarding the validity of proper classification of the material. The editor relays review information to the submitter to assist in producing an accepted product that is valid in terms of its correctness, usefulness and classification.

The instructor-user wants to find materials for teaching. The information available at the repository must facilitate the instructor's decision regarding a resource's suitability. The instructor has an expectation that this material will work as advertised with as little time as possible invested in obtaining it.

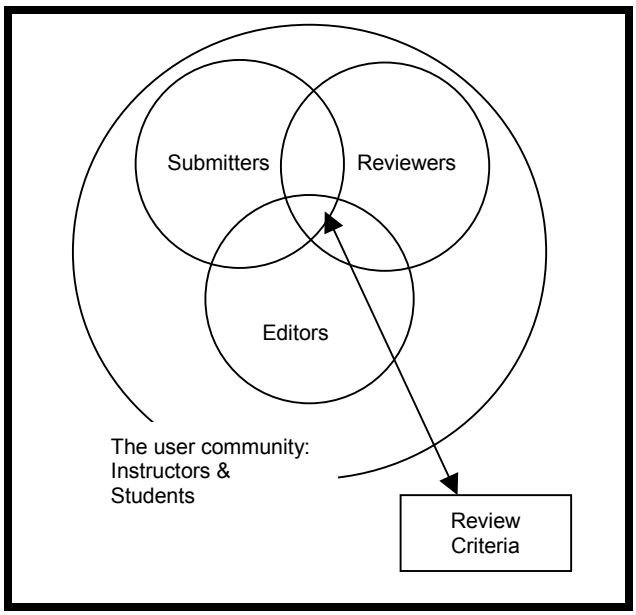

Figure 1: Stakeholders in the Review Process

The student-user needs to have course material with clear and understandable instructions. This material needs to be at a level that is challenging (not too simple nor too complex for the student at the point in the course).

3.2 Modeling the Review Process

When placed against previous work, which examined the process of review and the information flow within it [Joyce 1997, Grissom 1998], it is clear that each of these stakeholder groups is associated primarily with a single stage in the review process. The review cycle involves the stakeholders in a feedback model, as shown in Figure 2.

Submissions are passed from the editor to the reviewer. After review, the results are returned to the editor and feedback is provided for the submitter if revision is required or the resource has been accepted. When the material is accepted it is put into the repository.

This information flow can be divided into four stages, each associated with a stakeholder interest: Pre-evaluation, Evaluation, Editorial Evaluation, and the Afterlife of Evaluation, thus:

Stage	Stakeholder
Pre-evaluation	Submitter
Evaluation	Reviewer
Editor Evaluation	Editor
Afterlife	Users

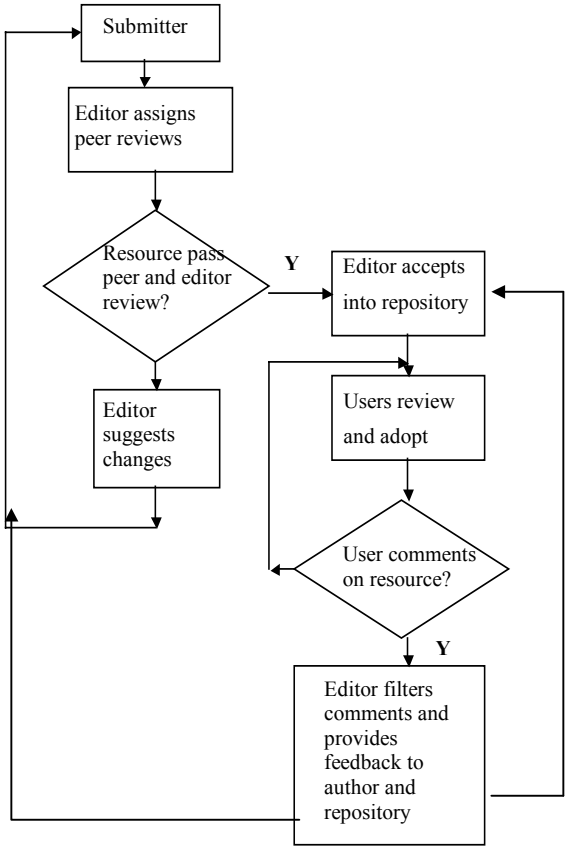

Figure 2: Feedback Model of Review

This model of the review process recognizes the distinctive nature of the materials being reviewed – that teaching materials are not created for insertion into a repository. It views the review process as one slice in the life-cycle of a piece of practice, hence the coinage "afterlife" for the informal review processes that are undertaken by users. (Such a process might be recognized by comments such as "Wouldn't it be nifty if it covered that concept, too" or "Why doesn't it cover this as well as that" or "This is terrible for this purpose, but I can fit it into another course" or "This also worked well in an advanced course by adding the following requirements…"). When looked at in this way, this model can be expanded, thus:

Stage	Stakeholder	Purpose
Pre-evaluation	Submitter	Creation
Evaluation	Reviewer	Assessment
Editor Evaluation	Editor	
Afterlife	Users	Transfer

3.3 Contexts

With peer-review of research papers, all stakeholders are engaged in the same purpose (albeit on different sides of the fence). As identified above, teaching materials are initially constructed for a purpose that is *not* peer-review. They are created for specific use in a single, individual classroom. It is a second (creative) step to re-cast them against given criteria and submit them to a repository. Each material has a history of its life in the classroom before review and, equally, a future in other people's classrooms after it has been through the process of review. Consequently, there are several contexts against which it may be (must be) judged.

- First, each material has to be "packaged" for the repository against specific submitter criteria. (We would not wish to suggest that this would be a particularly lengthy or arduous task. With the existence of submitter criteria, it is to be hoped that academics would create new teaching materials to meet those criteria as a matter of course.)
- Second, each material has to be evaluated with regard to its originating context. That is to say, evaluated with explicit reference to the pedagogic purpose and institutional context it was created for. It would not be productive to evaluate materials created for use in the second year of a course at a community college against criteria anticipating use with final year students at MIT. Materials have to be evaluated initially on their own terms.
- Third, each material must be evaluated for technical presentation and content. The material must be portable to the extent that another teacher with similar set-up should be able to install and use them with few problems.
- Fourth, each material must be worthwhile in the context of the discipline. For example, it would not be useful to submit excellent materials that assisted students in learning long division. That is not appropriate for the teaching and learning of university-level Computer Science; it is not disciplinarily appropriate. Not only do disciplinary criteria define what content is appropriate, but they may also address whether the pedagogic aims are worthwhile and/or significant.
- Fifth, each material must be useful within the context of the repository as a whole.
- Finally, each material will be evaluated in the context of its transfer to other instructors and institutions.

These separate contexts shape and expand the model of review of teaching materials and start to allow us to define sets of evaluation criteria:

Every resource has a history – created for use in someone's classroom

Stage in Review	Stakeholder	Process
Pre-evaluation Before going to review, the submitter re-creates the product against submission criteria.	**Submitter**	**Creation: Context of Submission**
Evaluation The reviewer evaluates the work against review criteria.	**Reviewer**	**Assessment: Value in three contexts** Submissions are evaluated with regard to three contexts: • Context of Original Classroom (evaluated against "Learning Criteria") • Context of Technical presentation and content (evaluated against "Technical Criteria") • Context of other practice within the discipline (evaluated against "Disciplinary Criteria")
Editorial Evaluation The editor evaluates the work against specified criteria.	**Editor**	**Assessment: Repository Context** The editor evaluates only against criteria that are relevant to the context of a repository.
Afterlife of Evaluation Evaluative activity does not finish with the end of the formal evaluation process.	**User**	**Transfer: Context of Use** Users feedback their reactions and comments on the product in use.

Every resource has a future of use – transferred to other people's classrooms

3.4 Summary

The reviewer evaluation received the most attention during the Working Group sessions. In particular, the contexts for assessment evolved into three categories, including learning criteria, technical criteria, and disciplinary criteria; influenced by the categories identified in our survey, particularly "learning criteria" which we believe encompasses *appropriateness for the audience* and "technical criteria" which is clearly based upon *technical soundness*. These categorizations help organize the reviewer form and guide the reviewer through the process. Having generated a conceptual framework for further exploration, we proceeded to concentrate on expanding this framework, again using guidelines from our survey and particularly investigating forms and checklists of criteria.

4 EMPIRICAL STUDY

After careful consideration of the types of information needed by the various stakeholders (submitter, reviewer, editor, and users), and thoughtful discussion of the variety of forms, the Working Group developed a survey to administer to the ITiCSE conference attendees to provide feedback on their preferred model. After these results were analyzed, the Working Group then undertook a small experiment to assess the reliability of the forms that had received the most votes.

4.1 Evolution of Models for Review Forms

After reviewing the materials on evaluation in traditional media, software, and journal articles, the Group identified two general models for evaluation forms: *open-ended* and *directed*.

The open-ended model is one in which the reviewer is asked to give his or her opinion on the worth of the submitted teaching material. Within this category, forms can be further classified as *unguided* or *guided*. Unguided forms give the reviewer one or two very open-ended questions, such as "Do you like this material? Explain why or why not." or "Do you think this material should be in the repository? Justify your answer." Guided forms have open-ended questions, but the questions are chosen to guide the reviewer to look at certain dimensions of the material. Questions such as "Evaluate the writing of the material in terms of style and grammatical correctness" or "Does this material enhance student learning?" fall into this category.

Directed forms contain specific questions such as "Are the concepts accurately described?" or "Is any needed terminology adequately defined?" Directed forms may be further classified by length (short or long) and by type of reply expected, scaled or unscaled. That is, questions may be phrased in a yes/no/not applicable format, or the reviewer may be asked to rate the question on a given scale. The examples are shown in a yes/no form, but they could be rephrased in a scaled form as "How accurately are the concepts described?" or "How adequately is any needed terminology defined?" Examples of five of these types of forms are available at the web appendix.

In the discussion, it became clear that each specific model might offer some advantages to particular stakeholders and disadvantages to others. However, a model may also appeal to specific individuals on a personal level, not related to the community they represent.

4.2 Relative Advantages and Disadvantages of Different Forms

- **Open-ended, unguided (A)** This has the advantage of complete flexibility, but its disadvantages are that it is hard to compare reviews (for the Editor and Submitter) and that dimensions of the resource may be ignored.
- **Open-ended, guided (B)** This also has the advantage of flexibility and, for the Editor, that all the required dimensions are addressed. Its disadvantages are that (for the Editor and Submitter) it is hard to compare reviews and that (for the Editor and User) important dimensions of the resource may be missed
- **Directed, unscaled, short** This has advantages for the Reviewer that it is easy to use and for the Editor that all required dimensions are addressed. Its disadvantages are that it is inflexible and lacks shaded responses.
- **Directed, unscaled, long (D)** This benefits the Reviewer and Editor in that the details are channeled, and it makes it easy for the Editor to compare reviews. For all stakeholders, more information is gathered. Its general disadvantage is that it lacks shaded responses, and specifically burdens the Reviewer by taking longer to fill out.
- **Directed, scaled, short (C)** This has the advantage that it allows shades of gray. It benefits the Editor because it is quantifiable and the Reviewer because it is easy to complete. Its disadvantages are that important dimensions of the resource may be missed, and Reviewers are constrained to the categories listed.
- **Directed, scaled, long (E)** The advantages of this form are perceived to be primarily for the Editor and are that the responses are quantifiable and allow objective comparison of reviews. The disadvantages are perceived to be for the Reviewer in that it takes longer to fill out and constrains responses to the categories listed.

For our experiment, we chose to use five of the six models, feeling that the unscaled short form did not give enough information.

4.3 Survey and Recommendations

Five review forms were constructed, based on the review form models discussed in the previous section. Questions were chosen from the categories outlined in the literature review. Conference attendees were requested to view each of the five forms from the perspective of the four stakeholders. After examining all of the forms, the attendees were asked to choose which form would be their

favorite if they were a Submitter, if they were a Reviewer, if they were an Editor, and if they were a User.

The results are shown in Figure 3. By and large, the results of the survey were consistent with the predictions of Section 4.2 and showed the preferred review instrument to be the directed, scaled, long one. However, the popularity of the open-ended guided model, at least when viewed from the roles of Submitter and Reviewer, was less expected. This result, no doubt, bears out our earlier comment that there is variation among individuals, which is independent of their community.

The Working Group reconsidered the formulations at hand and decided to investigate the two preferred models further, with the addition of another important dimension: the subjective opinion of the individual completing the form. This was done by adding the question, "Would you use this resource in your own classroom?"

4.4 Testing for Reliability

The Working Group took the modified versions of Forms B and E and applied the forms to three resources. Two were traditional laboratory exercises; the third was a software package. The tallies from these two forms are available at the web appendix.

4.4.1 *General Impressions of Form B*

Applying Form B to potential submissions identified problems with this form as outlined below.

- The answers to the reviewer questions lacked the specificity of the other form since it was not based on a scale and all reviewers did not provide a yes/no answer. This could potentially be a problem for some stakeholders and an advantage for others.
- If the submission meets the review requirements for a question posed, or if the question is not appropriate for the submission, requesting clarification proves difficult for the reviewer.
- Different reviewers responded similarly with respect to any one submission, but did so under different categories and for different reasons. It would be necessary to carefully consider the wording of each of the questions so that like responses might occur in the same question rather than in another heading.
- There were no questions regarding the goals of the resource and whether these were met, except in an oblique fashion. These should be included.

4.4.2 General Impressions of Form E

There were a number of problems in Form E, which were identified as the Working Group applied the form. There was an assumption that the "cover sheet" provided background information. As these were hypothetical submissions, they did not include a "cover sheet", which would be expected to include information generated against the Submitter Criteria, posited in section 3, above. Several of the Working Group members marked related questions NA and several marked them Poor. This explains some of the diversity in answers to questions 1 through 4. Other problems are as follows.

- There were problems with the scales used for the responses in that some questions really required a yes/no response rather than a response on a 4-point scale. The decision was that questions on the review sheet that required a yes/no response should be reworded so that the question could be answered using the scale.
- Some of the terms required changing, especially those that referred to 'completeness' which was interpreted differently by different group members.
- There was a problem because of the lack of thematic groupings for the questions resulting in diverting the attention of the referee from questions about content by inserting questions about required resources, for example. Regrouping the questions should avoid the problem.
- Some of the questions were inappropriate because they were specific to one of the stakeholders. Two of these were of importance to the editor, one to the referee The ones needed by the editor could readily be removed, the one for the reviewer could be covered in the information supplied by the editor and/or submitter.

Stakeholders →	Submitter	Reviewer	Editor	User
Form A	0	2	1	7
Form B	10	8	5	5
Form C	1	7	2	0
Form D	5	4	4	1
Form E	11	12	15	12
Totals	27	33	27	25

Figure 3: Poster Session Preference Feedback

Form E was further revised and is in the appendix as E version 2. The Working Group reapplied this new form to the examples previously tested for a "level of comfort" check. Each Working Group member felt comfortable with the revised forms.

4.4.3 Analysis Across Forms B and E

Since the three prototype resource materials were rated using both Forms B and E, it was possible to use the results to examine the reliability of the evaluation of the resources using the different forms. The forms differed substantially, so for the comparison of results the specific questions posed in Form E were regrouped so that they corresponded to the more general questions in Form B. The following comparisons used these categorizations.

An initial finding, following the regrouping, was that there were a substantial number of items in Form E which were not covered in Form B. Form B lacked any question concerning audience and goals.

A more detailed analysis of the data provided some additional generalizations. The discrimination in Form B was poorer. The questions were phrased to elicit a yes/no response, and there were very few 'no' responses. For both the 'yes' and 'no' responses, raters generally qualified the answers in some way to avoid an outright 'no'. For Form E, the use of four choices did not allow the rater to remain neutral and forced an opinion. If we look at the questions on Form E that relate to a question on Form B and assume that two choices on the negative end of a scale are a basis of concern, then from 10 to 30 percent of the questions on Form B, marked 'yes,' are really questionable. This negative reaction is not captured in Form B.

As was expected, there were fewer comments on Form E. This was seen as a disadvantage to the submitters, especially if there was a suggestion that the resource should be resubmitted after revision. It was obvious that some effort would need to be made to point out to reviewers that comments were needed especially in cases where negative scale values were used. On the other hand, the use of a substantial number of specific items to focus reviewer attention and responses was seen as very positive from the editor's point of view. This focus makes it easier for the editor to determine the suitability of the resource and to compare reviews among reviewers. This focus is also beneficial to the submitters because it draws attention to specific shortcomings of the resource.

Given the increased information gleaned from Form E and the substantial increase in discrimination for this form, the Working Group decided to refine Form E on the basis of the findings. The form can be revised again after it has been used for a while.

4.5 Summary

The Working Group identified several important matters for further consideration and research:

The long scaled directed form lends itself nicely to quantification, on the basis of which the editor can proceed further with her or his decision. However, some of the questions are certainly more significant than others. This leads to the question of how one decides the relative weights to assign the different questions. This, of course, is more a concern of the editor than the other stakeholders. The Working Group concluded that if weights were used, such information needed to be shared with the other stakeholders.

The assessment of a teaching resource, even if it has been submitted to a repository and undergone editorial and peer review, does not end with its acceptance. It is critical that a feedback loop from adopters of the resource be incorporated into the evaluation process. This feedback assessment, as noted in Figure 2, provides important information to the creator of the resource and to other potential users, as well as to any review or editorial communities involved. The design and implementation of such "post-evaluation" instruments has been discussed, but selecting the appropriate ones in this context is still open. "In particular, referees are not expected to attest to the correctness of the programs; correctness will be attested through use by readers, with software and test data attaining progressively higher levels of certification through additional reports of satisfactory use by readers in their applications or research projects." [JEA 1995].

This is related as well to the general question of correlating the predictions of the reviewer with the satisfaction of the user. If a review model is to be retained and enhance user confidence, it must have a proven track record of high correlation with user satisfaction.

When correlating the conceptual, contextual work of Section 3 and the empirical work of Section 4, we observed that by combining the relevant *form* of evaluation instrument with the various *categories* of stakeholder a modular structure was suggested. It is our tentative conclusion that a multi-purpose instrument might be developed where appropriate question sets could be "plugged in" depending on the specific given combination of resource, stakeholder and form.

5 CONCLUSIONS

This report presents some motivations to validate the quality of teaching materials, including the need for currency in our field and the desire to access materials to help enhance our students' learning experiences. We outline how a review process differs for a variety of resources, drawing upon the literature. Based on a literature review and group discussions, we delineated six different models of review forms, and developed samples of five of them to use in an empirical study. These five models were used in a survey, and the results allowed the Working Group to refine the models. We present the results of testing for the reliability of two of the forms, and further refine the scaled, long form review model. The recommendation of the Working Group is that the refined

Form E version 2 be used for peer review (by the CSTC and other repositories) to gain further information on its ability to provide user confidence in a teaching resource to be adopted and integrated into a user's course.

5.1 Future Work

The goals of this Working Group were originally proposed as:

1. Evaluation of one or more proposed review models through application to CSTC resources, specifically lab materials.
2. Refinement of the review models.
3. Assessment and enhancement of reviewer training materials.

As we conclude this Working Group experience, each of the first two goals was visited during our sessions. A number of characteristics for the review of teaching materials were defined and tested, both for suitability within specific contexts and also for reliability. The sample size was small and heavily biased, but provided good insight. The assessment of the review model developed will be an on-going project. As more reviews are conducted (for the CSTC), we will gain important feedback, which will be used to further refine the model. In addition, a subgroup will continue work on developing training materials for reviewers.

6 REFERENCES

ACM (1993) ACM SIG Editor Manual: Summarizing the Differences Between Refereeing, Formal Reviewing, and Reviewing. Available: http://www.acm.org sig_volunteer_info/editors_manual/ref_and_review.txt [July 1999]

Alessi, S.M. and Trollip, S.R. (1985) Computer-based instruction: methods and development. Prentice-Hall, Englewood Cliffs, NJ

Barnett, L. (1999) Computer Science Education Resources. Available: http://gum.richmond.edu/~lbarnett/ Informatics/CS_Ed.html [July 1999]

Duchastel, P.C. (1987), "Structures and Methodologies for the Evaluation of Educational Software", *Studies in Educational Evaluation*, Vol. 13, pp. 111-117.

Grissom, S., Knox, D. (joint chairs), Copperman, E., Dann, W., Goldweber, M., Hartman, J., Kuittinen, M., Mutchler, D., Parlante, N. (1998) "Developing a Digital Library of Computer Science Teaching Resources, Report of the ITiCSE'98/ACTC'98 Working Group on The Online Computer Science Teaching Center," ACM ITiCSE 98, Dublin, Ireland, August 1998, SIGCSE Bulletin, December 1998.

Harvey, J. (1997) "Choosing courseware: Some guidelines to first step evaluation." Chapter 7 of "Implementing Learning Technology," Greg Stoner, Editor, as part of the Learning Technology Dissemination Initiative. http://www.icbl.hw.ac.uk/ltdi/implementing-it/cont.htm

JEA (1995) ACM Journal of Experimental Algorithmics, Bernard M. E. Moret, Editor. Available: http://www.jea. acm.org /guidelines_eval.html [July 1999]

Joyce, D., Knox, D. (joint chairs), Gerhardt-Powals, J., Koffman, E., Kreuzer, W., Laxer, C., Loose, K., Sutinen, E., Whitehurst, A. (1997) "Developing Laboratories for the SIGCSE Computing Laboratory Repository: Guidelines, Recommendations, and Sample Labs: Report of the Working Group on Designing Laboratory Materials for Computing Courses," ACM ITiCSE 97, Uppsala, Sweden, June 1997, SIGCSE Special Issue 1997, pp. 1-12.

Knox, D., Grissom, S., Fox, E. (1999) Computer Science Teaching Center, Available: http://www.cstc.org [July 1999]

McCauley, R. (1999) Computer Science Education Links. Available: http://www.cacs.usl.edu/~mccauley/ edlinks/ [July 1999]

Muramatsu, B. (1999) NEEDS: National Engineering Education Delivery System, Available: http:// www.needs.org/ [July 1999]

Woodbury, M., (1980), *Selecting materials for instruction: Media and the Curriculum* Libraries Unlimited, Inc., Littleton, Colorado, 1980

Appendix

Due to space limitations, additional materials are maintained at a web site for this Working Group. The Appendix materials may be found at www.tcnj.edu/~cstc/krakow/appendix.html. These items include

- evaluation forms assessed by the Working Group in our preliminary work, as well as
- the Review Forms A through E we designed and used for the survey discussed in this Report, and
- tallied results from applying the Review Forms B and E to three resources (two labs and one visualization).

Other evaluation forms available at the Appendix web site include:

- the NEEDS project forms (both the 10 question form as well as the premier courseware evaluation form),
- the evaluation form used for the SIGCSE 98 Symposium papers, and
- the IEEE Frontiers in Education 98 paper evaluation form.

The following items are contained in this Report's (paper) Appendix:

- Review Form B (open ended, guided)
- Review Form E (directed, scaled, long)
- Review Form E (revised to version 2)

Review Form B

Submission: _____

Reviewer name: _____

Please answer the following questions:

Will the resource enhance or facilitate student learning?
Why or why not?

Are the concepts correctly and accurately described?
Why or why not?

Is the resource complete and technically sound?
Why or why not?

Are the goals worthwhile, and would the resource be useful to instructors?
Why or why not?

Is the writing clear and grammatically correct?
Why or why not?

Would you use this resource in your own classroom?
Why or why not?

Should the resource be included in the repository?
Why or why not?

Additional Comments:

Review Form E

Submission: _____

Reviewer name: _____

Please check the appropriate box.

Question	NA	Excellent	Good	Fair	Poor
Is the audience specified?					
Is the resource appropriate for its audience?					
Are the goals of the resource specified?					
Is the resource effective at accomplishing its stated goals?					
Will the resource enhance or facilitate student learning?					
Are the concepts correctly and accurately described?					
Is any needed terminology adequately defined?					
Are the concepts used in the resource complete in the context?					
Is the resource technically sound?					
Does the resource operate correctly?					
Is the resource complete?					
Are the technical requirements clearly documented?					
If it is software, is it easy to install?					
Has the resource been correctly categorized?					
Are the goals of the resource worthwhile?					
Does the resource teach something significant?					
Is the resource likely to be useful to instructors?					
Is the resource written clearly and grammatically?					
Are there supporting references in the document as needed?					
Is media used appropriately and not gratuitously?					
Is the resource original?					
Would adding this resource to the repository be of benefit?					
Is this resource free of copyright restrictions?					
Would you use this in your own classroom?					
Overall, should the resource be included in the repository?					

Additional Comments:

Review Form E (V 2.1)

Submission: _____

Reviewer name: _____

Please check the most appropriate box.

Question	Very	Adequately	Somewhat	Poorly	NA
Learning context					
How well specified is the audience?					
How appropriate is the resource for its audience?					
How well are the goals of the resource specified?					
How effective is the resource in accomplishing its stated goals?					
How accurately are the concepts described?					
How thoroughly are the necessary concepts defined within the resource itself?					
How well is any needed terminology defined?					
How well does the resource enhance or facilitate student learning?					
Presentation style					
How accurate is the coversheet documentation?					
How clear and grammatical is the writing?					
If required, how well are supporting references cited?					
How appropriately is non-text media used?					
Technical context					
If the resource is software:					
How well does the software operate?					
How easy is it to install?					
How clearly are any supporting technical resources specified?					
Disciplinary context					
How worthwhile are the goals of the resource?					
How significant are the topics taught in the resource?					
How useful is the resource likely to be to instructors of this topic?					
Repository context					
How beneficial would adding this resource be to the repository?					
How positive do you feel about including this resource in the repository?					
From a personal perspective, how well did you like this resource?					
If appropriate, how likely would you be to use this in your classroom?					

Additional Comments: