Evaluation Guidelines for Multimedia Courseware

William Gibbs, Pat R. Graves, and Ronan S. Bernas
Eastern Illinois University

Abstract

This research sought to identify criteria important to multimedia instructional courseware evaluation and validate them with a panel of instructional technology experts. Data were collected with a Web-based survey using a modified Delphi research technique. During three rounds of questioning, experts judged 91 criteria as important to the evaluation of instructional courseware. The study also examined the effect of conducting panel discussions online. The method of panel discussion presented in the study enabled geographically dispersed discussants to examine criteria collectively. Although limited, participant commentary helped refine the criteria list. In general, participants agreed in their opinions and gave consistent criteria ratings. (Keywords: computer-based instruction, courseware selection criteria, evaluation guidelines, hypermedia, instructional software, multimedia.)

Computer-based instructional courseware is software developed for the purpose of providing instruction. It is generally thought to fall into one of the following categories: drill and practice, tutorial, simulation, instructional game, and problem solving (Hannafin & Peck, 1988). In the early 1980s, the educational market was flooded with diverse courseware applications that lacked instructional quality. Hannafin and Peck pointed out that, although there were examples of effective educational computer programs, much of the available software was terrible. Roblyer (1988) stated that poor-quality courseware affected educators' use of computer-assisted instruction (CAI).

The increased availability of instructional courseware generally, and poor-quality courseware particularly, engendered attempts to develop methods for software assessment (Shuell & Schueckler, 1989) and helped make evaluation more common. The importance of evaluation is recognized, and it is vitally important for educators to partake in software reviews (Dudley-Marling & Owston, 1987; Zahner, Reiser, Dick, & Gill, 1992). However, evaluation can be an intricate process complicated by a variety of intervening factors. Educators frequently are not cognizant of the most suitable methods to use when evaluating and selecting instructional programs. Evaluation entails methodical decision making about the value of a particular object and identification of reliable methods with which to base decisions or judgments (Gros & Spector, 1994). According to Gros and Spector, difficulties encountered in evaluation relate to "not determining reliable methods, which produce generally useful judgments" (p. 37).

A review of the literature yields a variety of instructional courseware evaluation methodologies. Heller (1991), for example, lists several organizations, initiatives, and services that developed evaluation methodologies such as the National Council of Teachers of Mathematics, MicroSIFT, EDUCOM's Software Initiative, or York Educational Software Evaluation Scales (YESES). Each approached the evaluation process somewhat differently. In some cases, students

and educators evaluate courseware based on criteria that generally fall into one of four categories: accuracy, effective instructional strategies, instructional objectives, and ease of use (Morrison, Lowther, & DeMeulle, 1999). By adopting textbook review protocols, professional organizations and associations have conducted software critiques, provided software ratings and recommendations, and distributed the results (Heller; Jolicoeur & Berger, 1986).

On many evaluation instruments, evaluators judge courseware using predetermined criteria (such as ease of use or instructional quality) and rating scales (such as highly favorable to highly unfavorable). Other types of instruments present a checklist format. Roblyer, Edwards, and Havriluk (1997), for instance, presented a minimum criterion checklist comprised of four primary categories: instructional design and pedagogical soundness, content, user flexibility, and technical soundness. Reviewers marked "yes" or "no" on the checklist to indicate whether the program met the criterion.

Criteria used in evaluation instruments have also come under scrutiny because they are sometimes confusing. Likewise, evaluation rating systems have been criticized for not being comprehensive, understandable, and easy to use (Chang & Osguthorpe, 1987). Developing representative and reliable criteria and categorizing them are challenging.

Software adopters and agencies that publish courseware reviews often use comparative evaluations. Comparative evaluations are often useful for initial software screening (Dudley-Marling & Owston, 1987; Zahner et al., 1992), but some researchers question the value of comparative evaluations because they are subjective assessments and do not provide adequate information about program effectiveness (Dudley-Marling & Owston; Heller, 1991; Jolicoeur & Berger, 1986, 1988; Schueckler & Shuell, 1989; Zahner et al.).

Two alternatives to comparative evaluation include a criterion-based evaluation, the YESES and the Zahner et al. (1992) Courseware Review Model. YESES evaluates software on four dimensions: pedagogical content, instructional presentation, documentation, and technical adequacy (Dudley-Marling & Owston, 1987). Software is rated on each dimension using a four-point criterion-based scale (exemplary software, desirable software, minimally acceptable software, and deficient software). The Zahner et al. model is not based on subjective judgments of predetermined criteria, but on learners' pre- and posttesting and survey data.

This study attempts to identify important criteria that characterize well-designed multimedia instructional courseware. These criteria can serve as a basis for constructing evaluative instruments for software screening. Moreover, this study examines the process of conducting panel discussions online and provides a review of participants' commentary.

For the purposes of this study, multimedia instructional courseware is defined as a single instructional unit with one or more lessons designed for self-instruction or small-group instruction using a stand-alone computing system. Whether the courseware is drill, tutorial, simulation, gaming, or problem solving, it would most commonly be available on CD-ROM. In this instance, multimedia is defined in accordance with Tolhurst's (1995) definition, which sug-

gests *multimedia* is a comprehensive term encompassing both *hypermedia* and *hypertext* when it uses the interactive nature of the computer to include nonlinear access to informational content. We chose this definition of multimedia courseware for the following reasons:

1. It narrowed the scope of software selection to a defined media type (specifically CD-ROM).
2. Many instructional courseware programs are developed and delivered on CD-ROM.
3. Although the Internet and Web offer seemingly unprecedented opportunity for instructional delivery through Web sites, the design for such learning environments is notably different, in many ways, than CD-ROM delivery. The presentation of video, audio, and graphics, for example, is often limited because of bandwidth restrictions. Additionally, the Web is an open environment where users can potentially link to sites of varying quality and interface and to sites containing unrelated informational content. The CD-ROM format offers a multimedia designer greater control of the content presentation and interface.

METHOD

This study used a "group of experts" with identified interests and expertise in instructional technology and software development and a modified Delphi survey research method. The *Delphi* technique is a group process approach enabling several individuals to communicate as a whole about complex issues and work toward a problem solution while remaining anonymous (Linstone & Turoff, 1975; Oswalt, 1982; Sapre, 1990; Smeltzer & Leonard, 1994). It provides an alternative means for decision making other than face-to-face meetings, and it allows knowledgeable people to share their opinions while geographically separated.

For this study, the problem was to identify and validate important criteria that characterize well-designed multimedia instructional courseware. Anonymous participants completed an online Web survey on three occasions, or *rounds*, over a two-month period. The Web survey automatically updated response frequencies and commentary. In doing so, it permitted asynchronous electronic communication between the participants and us.

Participants

We identified study participants from three sources: the Association for Educational Communications and Technology (AECT) membership directory, a review of related research literature, and the Who's Who in Instructional Technology Web site. Individuals were selected to participate as experts if:

- they had published articles in the last five years on computer-based courseware design, development, or evaluation;
- they taught courses about these topics; or
- their primary employment responsibilities related to these areas.

Twenty-one individuals completed Round 1, 18 of the original 21 completed Round 2, and 14 completed Round 3. This is not a high percentage of participation in terms of survey research. Because the Delphi technique is an alternative method to face-to-face meetings for problem solving, fewer knowledgeable individuals are typically needed than for research surveys with a sample size representing a large population. Therefore, it may be more appropriate to consider the number of respondents according to the dynamics of small work groups. When the meeting goal is problem solving, planning, or decision making, 6 to 12 people should participate (Bovée & Thill, 2000). When meeting goals are more complex, groups of 12 to 13 people can be more effective (Smeltzer & Leonard, 1994). The number of participants in each round exceeded these numbers. Statistical procedures were used to examine the effects of varying participation in each round; only a few differences between rounds were noted, and they are explained in the Results section of this article.

Participants represented seven countries and were employed in one of the following categories: professor, director of educational technology, instructional technologies specialist, consultant, head of interactive media, lecturer, or research associate. Their areas of expertise included teaching and technology integration, instructional design, software development, courseware materials development, computer conferencing and faculty development, research, educational software evaluation for teachers, and interface design.

Materials

We reviewed more than 50 research articles and textbooks to identify criteria related to the evaluation of multimedia instructional courseware. We developed an original list consisting of 22 categories, 9 subcategories, and more than 230 criteria, which we deemed too lengthy and unmanageable. To consolidate the list, we and two assistants reviewed all statements again to identify redundancies and rearrange categories. Together, we then discussed each statement and made a final decision as to its categorization or removal from the list. The resulting 81 criteria were grouped into 14 categories. The list was meant to serve as a beginning point from which participants could suggest modifications, additions, or deletions.

On three separate rounds, participants rated evaluation criteria using a Web form set up for the study. They also provided comments and suggested new criteria. In all three rounds, the original 14 categories and statements remained unaltered. Maintaining the initial list helped determine whether participants reached consensus through the three rounds of questioning. After each round, however, we reviewed all participant commentary to identify new criteria from their input. All new statements were appended to the original list and labeled as New Items from Round 1 and New Items from Round 2, respectively.

Based on participant commentary, we added 17 statements after Round 1 and 12 statements after Round 2. Because participants did not see the survey after Round 3, there was no additional category for that round. At the completion of the study, the survey consisted of 16 evaluation categories and 110 criteria (Table 1). It was necessary to keep the added criteria separate for analysis pur-

Table 1. Number of Items and Mean Rating for Each Criteria Category

Category	Number of Items	Mean
Evidence of effectiveness (field test results)	1	4.64
Instructional adequacy	6	4.50
Information content	6	4.28
Clear, concise, and unbiased language	6	4.24
Support issues	5	4.18
Interface design and navigation	11	4.17
Program adequacy	4	4.17
Visual adequacy	9	4.10
Feedback and interactivity	8	4.00
Instruction planning	5	3.97
Information reliability	3	3.96
Motivation and attitude	7	3.93
Classroom management issues	5	3.76
Documentation	5	3.55
New items from Round 1	17	n/a
New items from Round 2	12	n/a

poses; however, at the completion of the study, many of these items were interspersed throughout the original 14 categories based on the topic addressed.

Procedure

In Round 1, participants rated their opinion of 81 courseware evaluation criteria. Participants rated each criterion on a 5-point Likert scale from Strongly Agree (5) to Strongly Disagree (1) by selecting the appropriate button on the Web survey form. A Not Applicable (N/A) option and a comment field accompanied each statement. The N/A option allowed participants to respond with other than a 1 to 5 rating in the event, in the participant's opinion, such a rating did not apply to a particular statement. The comment field afforded participants the opportunity to type comments, suggest rewording, or provide alternative criteria. All new criteria suggested in Round 1 were appended to the survey in Round 2.

In Rounds 2 and 3, the survey presented the previous round's response frequencies and commentary for each item. Each individual participant's rating for each item from the previous round was also available to allow participants to consider their responses in relation to those given by others. Participants again rated each criterion using the same scale as in Round 1, except a Comments column presented two text fields for each survey item. The first (top) field displayed the comments made by all participants for a particular item during the previous round(s). In the second (bottom) field, participants typed any additional comments they had about the item or entered their reaction to the comments of other participants. Round 3 was similar to Round 2, except new criteria suggested in Round 2 were added to the survey. Round 3 completed the study.

RESULTS

Expert Ratings of Criteria

Using the assigned values (1 to 5), a weighted overall mean was calculated for each statement across rounds to account for the different number of participants in each round. We categorized criteria into five levels according to mean score and indicated the percentage of the 110 statements within each mean range. Participants strongly agreed with 15% of the 110 statements, agreed with 68% of the statements, and were neutral on approximately 17% of the statements. They did not Disagree or Strongly Disagree with any statements (Table 2). Twenty-nine "write-in" statements were derived from participant commentary, of which 16 were original items and 13 were derivations of criteria previously introduced. Across the three rounds, there were 109 occurrences of N/A ratings given to 60 different statements.

In general, mean ratings of criteria were high. "Does the courseware provide accurate information?" had the highest overall mean rating ($M = 4.91$), and "Does the software manufacturer provide any indications of teacher/trainers' and learners' attitudes to the courseware?" received the lowest rating ($M = 2.55$).

Table 2. Percent of Criteria by Mean Rank

Mean Range	Category	Percent of Statements in Category
4.51–5.00	Strongly agree	15
3.51–4.50	Agree	68
2.51–3.50	Medium/neutral	17
1.51–2.50	Disagree	0
1.00–1.50	Strongly disagree	0

Criteria with a mean greater than 4.50 are displayed in Table 3 in descending order. Participants strongly agreed these criteria are important to the evaluation of multimedia instructional courseware. Of all other statements presented on the survey, these 16 items represented the most important criteria.

In the highest mean range ($M = 4.51$–5.00), Strongly Agree, most criteria related to instructional adequacy (25%) and information content (18%) issues. No criteria from the documentation, program adequacy, classroom management, visual adequacy, motivation and attitude, write-in items from Round 1, and write-in items from Round 2 categories ranked in this mean range.

Most criteria in the second highest mean range ($M = 3.51$–4.50), Agree, were from the write-in items from Round 1 (17%) and write-in items from Round 2 (12%). These statements were obtained from participant commentary. Examples include, "Do pictures, graphic images, and other mediated representations facilitate understanding?" ($M = 4.43$); "Can learners freely explore the courseware?" ($M = 4.41$); and "Will the courseware provide any educational gains for learners?" ($M = 4.41$). In some cases, these statements represented modifications of original survey items.

Table 3. Highest Rated Criteria: Overall Mean Ranking 4.51–5.00

Overall Mean	Category	Criteria
4.91	Information content	Does the courseware provide accurate information?
4.74	Information reliability	Are the answers provided to questions correct?
4.72	Instructional adequacy	Are practice activities provided in the courseware to actively involve the learner?
4.66	Feedback and interactivity	If a test is used, are test questions relevant to courseware objectives?
4.66	Clear, concise, and unbiased language	Are sentences written clearly?
4.64	Evidence of effectiveness	Did learners learn from the courseware?
4.64	Instruction planning	Is a definition of the target audience and prerequisite skills given in the courseware?
4.62	Feedback and interactivity	Is feedback appropriate?
4.60	Instructional adequacy	Are instructional objectives clearly stated?
4.59	Support issues	Are the computer hardware and software requirements for the courseware specified?
4.56	Information content	Are examples, practice exercises, and feedback meaningful and relevant?
4.55	Interface design	Is the courseware screen layout (screen elements—titles, text areas, navigation buttons, etc.) easy to understand?
4.55	Instructional adequacy	Is the purpose of the courseware and what is needed to complete the lesson made explicit?
4.53	Information content	Is the information current?
4.53	Interface design	Do learners understand directions for using the courseware?
4.51	Instructional adequacy	Does the courseware provide adequate support to help learners accomplish the lesson objectives?

Finally, most statements in the third highest mean range (M = 2.51–3.50), Neutral/Medium, were from the write-in items from Round 1 (21%) and write-in items from Round 2 (15%) categories. Fifteen percent of the statements in this level were documentation issues, and 10% were issues of feedback and interactivity.

Criteria Category Rating

For each of the original 14 categories, a mean was calculated by averaging ratings for the relevant items across rounds. Nine of the 14 categories received a mean of 4.00 or greater. Participants rated criteria in these categories as being important to the evaluation of multimedia instructional courseware. All of the mean category ratings tended to be high. Evidence of effectiveness (M = 4.64) and instructional adequacy (M = 4.50) represented the highest rated categories, and classroom management (M = 3.76) and documentation (M = 3.55) the lowest (Table 1, p. 6).

Participant Commentary

Comments participants made about the criteria were tallied for each of the three rounds. A total of 263 comments were received. Participants typed 130 comments in Round 1, 114 comments in Round 2, and 19 comments in Round 3. In Rounds 1 and 2, each participant made an average of six comments. In Round 3, participants, on average, provided only one comment. The amount of commentary diminished considerably in Round 3, possibly because there were fewer participants.

We coded participant commentary using categories adopted from the four steps of the problem-solving model (Newell & Simon, 1972) as discussed by Le Maistre (1998). The steps of the problem-solving model include:

1. recognition of the existence of a problem and establishment of a goal statement;
2. construction or evocation of the appropriate problem space;
3. selection and application of operators, action, or decisions that move the problem solver through the problem space; and
4. evaluation of new knowledge state attained.

It was believed that validating criteria necessitated participants to work through these problem-solving stages; thus, coding categories based on the model proved useful.

We reviewed all commentary independently and marked according to whether a comment was valuable to the purpose of the study. Each comment was segmented based on its meaning. The information contained in the comment segment was then coded (Ericsson & Simon, 1993). A total of 355 codings were compared. Discrepancies in the codings were reconciled at this time.

Of the 263 comments, we judged 240 (91%) as valuable to the study and 23 (9%) as not specifically relevant to the study's purpose. The majority of coded segments (42%) were categorized as problems identified with the materials and statements indicative of participant knowledge or experience (26%) (Table 4).

Table 4. Proportion of Comments by Category

Category	Percent of Comments
Problems identified with the material	42.2
Problems identified in participant commentary	1.1
Statements indicative of participant knowledge or experience	26.0
Proposed revision	14.1
Evaluation of proposed revision	0.0
Agreement/support directed at researchers	5.7
Agreement/support directed at participant comments	4.0
Disagreement directed at researchers	0.0
Disagreement directed at participant comments	2.0
Observations	4.2
Miscellaneous	1.1

Consensus of Participant Opinion

We examined how consistent participant criteria ratings were within a specific round and how consistent the item ratings were across rounds. For each round, a mean of the standard deviation of criteria ratings was obtained. The standard deviation decreased from Rounds 1 to 3 (Round 1 $SD = 0.997$, Round 2 $SD = 0.976$, Round 3 $SD = 0.861$). As participants were exposed to the criteria, they appeared to gain more agreement in their opinions about the statements.

We examined the extent to which criteria ratings differed from one round to the next by comparing the mean of each round. This analysis provided an indicator of how consistent criteria ratings were across rounds. A one-way ANOVA for dependent means was conducted on the ratings across rounds for each of the 81 original items, and there were no significant differences in the ratings for 76 of the items. However, there were significant differences in the ratings across rounds for the following three items:

1. "Is the content and vocabulary level for intended users appropriate?" $F(2, 26) = 4.98$, $p = .015$. A subsequent Tukey's test indicated that the rating in Round 1 ($M = 4.57$) was significantly higher than the rating in Round 2 ($M = 4.21$) and Round 3 ($M = 4.14$), $p < .05$. Ratings in Rounds 2 and 3, however, were not found to be significantly different from each other.
2. "Will the courseware accept mouse and keyboard input?" $F(2, 26) = 5.35$, $p = .011$. The rating in Round 1 ($M = 3.36$) was found to be significantly higher than the rating in Round 2 ($M = 2.79$) when a Tukey's test was performed, $p < .05$. The rating in Round 3 ($M = 3.00$) was not significantly different from the ratings in Round 1 or 2.
3. "Do learners have control over navigation through courseware?" $F(2, 26) = 4.98$, $p = .015$. Specifically, the rating in Round 2 ($M = 4.50$) was significantly higher than the rating in Round 1 ($M = 4.07$), $p < .05$. The rating in Round 3 ($M = 4.43$), however, did not significantly differ from Round 1 & 2 ratings.

Two other items exhibited marginal significance, $p = .05$: "Can the information in the instructor's guide/manual be easily incorporated into class activities?" and "Is the courseware compatible with the hardware you have at your school?"

The mean ratings of the 17 write-in statements from Round 1 were compared on Rounds 2 and 3 using a *t*-test for dependent means. Results indicate that the ratings given to the 17 write-in statements in Round 2 did not significantly differ from those given in Round 3.

DISCUSSION

Expert Ratings and Possible Uses of the Criteria

Participants rated 91 of the 110 criteria as $M = 3.51$ or higher. They judged criteria as not applicable on 109 occasions, but the N/A ratings had marginal effect. Based on their review, analysis, and commentary, a final list was developed. A selection of the revised criteria, which received a mean rating of 3.5 or higher, is presented in Appendix A (www.iste.org/jrte/, select this article on the table of contents).

The findings suggest that participants perceived most of the criteria as valuable to the evaluation of multimedia instructional courseware. The relatively high criteria ratings should be interpreted cautiously. Participants had similar backgrounds and conceivably found the statements, the premises being evaluated, and the language used identifiable and familiar and, thus, gave favorable ratings. Other participant groups (such as K–12 educators) may yield a different distribution of ratings. It is reasonable to suggest that most of the criteria presented were important for this group of participants.

During the three rounds of questioning, participants sought to validate or invalidate the importance of each criterion. Although further study is needed about the criteria presented in Appendix A, they offer practitioners and courseware designers a foundation from which to evaluate and design courseware. A discussion detailing methods for using the criteria list is beyond the scope of this article, and more inquiry about them is needed. However, based on the expert reviewers' responses, it seems plausible, at least initially, to suggest the following uses.

- *Courseware trial reviews:* Students use courseware for a specific time period, and the instructor observes their use. The instructor may assign a rating scale (e.g., 1 to 5) for some criteria and "check-off" categories (e.g., yes or no) to others to indicate whether the courseware meets the criterion. By selecting the appropriate rating or category, the instructor indicates the extent to which the courseware met the criterion and, in doing so, evaluates its appropriateness.
- *Pretrial reviews:* A single instructor uses a courseware program for a specified time period. The instructor may assign a rating scale (e.g., 1 to 5) for some criteria and "check-off" categories (e.g., yes or no) to others to indicate whether the courseware meets the criterion. It is important to note that a single instructor cannot evaluate all the criteria presented in Appendix A. For

example, the criterion, "Do learners seem satisfied with the learning experiences provided by the courseware?" can only be determined after students use the courseware. However, using the criteria in the pretrial reviews can help focus the instructor's attention on key evaluative elements of courseware that may otherwise go unnoticed.

- *Criteria guide:* For individuals unable to conduct courseware trials but who are reviewing information (e.g., courseware reviews, specifications, etc.) about a program, the criteria list can help focus attention on particular criteria that they may not otherwise consider. Thus, when reviewing information about a program, the criteria list may confirm that specific features are present and call attention to important missing elements.
- *Courseware designer guide:* Many criteria have implications for courseware designers. Designers could use the list as a guide of important criteria to consider when developing instructional courseware.

Participant Commentary

Participants made 263 comments, many of which we found to be valuable to the purpose of the study. As expected, most comments (42.2%) identified problems with criteria related to clarity, specificity, relevance or importance, accessibility, duplication, and incompleteness (the extent to which items or concepts were missing from a criterion). The majority of problems pertained to issues of clarity, specificity, and relevance. For example, in several instances, participants identified unclear or imprecise statements. In response to the criterion, "If a test is used, do test questions measure learner understanding?" a participant remarked, "Understanding is interpreted differently from performance in my response." Another participant suggested that the criterion, "Is a definition of the target audience and prerequisite skills given in the courseware?" lacked specificity by stating, "these are two very separate issues." Participants found one or more problems with 64 statements. We reworded, consolidated, expanded, and deleted criteria to address each concern regarding clarity, specificity, relevance, accessibility, duplication, and completeness. Twenty-seven statements were removed from the criteria list, and all but 7 of the 64 problem criteria were modified.

Statements reflective of participant knowledge or experience were the second most frequently occurring types of comment. For example, in response to the criterion "Do illustrated concepts use familiar analogies or metaphors?" one participant remarked "I feel these are useful for novel concepts, but analogies and metaphors can create very 'fat' instruction in places where they are not needed to understand the material. And they can become annoying." The remark does not specifically identify a problem or suggest a revision to the criterion statement, but it reflects the participant's experiences and perception about the topic. Approximately 26% of the comments were coded as statements indicative of participants' knowledge or experience.

Although 42.2% of the commentary related to problematical issues, only 14% of it proposed revisions to the original criteria. Overall, there were fewer revision-type comments than anticipated. Revisions given by participants fell

into one of the following categories:

- combining items—recommendations to combine criteria;
- removing items—recommendations for removing duplicate criteria;
- rewriting—recommendations for rewording criteria (the premise remained unchanged and only the wording was altered); and
- adding on—recommendations for adding a word, concept, or new criterion (the respondent perceived that a particular concept was not addressed in the original statement).

Participants proposed revisions for 27 statements, the majority of which fell within the Add-on category. For example, responding to the statement, "Will the courseware accept mouse and keyboard input?" a participant commented, "Where are your questions about touch windows and augmented keyboards and switches and scanning input. We have a lot of individuals with disabilities who use assistive technology and I see no questions asking about this!" As a result, the following statement was added to the criteria list, "Will the courseware accept alternative input devices to accommodate individuals with disabilities?" (M = 3.91). In addressing each revision comment, we reworded, consolidated, added, and deleted criteria. Participants removed 10 statements from the criteria list and modified all but 4 of the 27 statements for which they proposed revisions. The following four statements were not revised:

1. "Do learners have a sense of position within the courseware?" (M = 4.38)
2. "Is the courseware screen layout (screen elements—titles, text areas, navigation buttons, etc.) easy to understand?" (M = 4.55)
3. "Does the user guide or online documentation provide sufficient detail and complete indexes of the information available in the courseware?" (M = 3.68)
4. "Are instructions for software installation and operation explicit, accurate, and complete?" (M = 4.36)

These statements received high mean ratings, which indicated that participants found them to be important. Each statement received only one comment and, while valid, the comments did not, in our view, provide adequate justification for altering the original statements, particularly given their high mean ratings. For example, a participant suggested combining the statements, "Does the user guide or online documentation provide sufficient detail and complete indexes of the information available in the courseware?" (M = 3.68) and "Is a user/technical guide or online documentation provided with the courseware?" (M = 3.86). In our view, the statements evaluate two distinct concepts; given their high mean ratings, they were not altered.

Across the three rounds, there were 109 occurrences of N/A ratings given to 60 different statements, 24% of which were assigned to write-in criteria. Participants assigned 33 N/A ratings in Round 1, 45 in Round 2, and 31 in Round 3. Most (57%) of the 60 statements received only one N/A rating. Three statements each received six N/A ratings, the most given to a single criterion. All three statements were removed from the criteria list. We eventually removed 21

of the 60 statements given N/A ratings. The participants modified 17 of these statements and left 22 unchanged. Although we examined N/A ratings, we also based our decisions to remove or alter items on participant commentary and mean rating. For example, the criterion, "Are the computer hardware and software requirements for the courseware specified?" (M = 4.59) received one N/A rating, yet it remained on the criteria list because of its high mean rating and lack of comments proposing it be removed.

Although the online mode of communication offered many advantages in terms of data collection, access, and convenience, it appeared to limit dialogue among participants. The 263 comments given by participants consisted of 2,953 words. If one considers that the average person speaks 100 to 150 words per minute (Heinich, Molenda, Russell, & Smaldino, 1996), it is reasonable to assume that much more commentary would have been exchanged had the panel of experts met face-to-face. All communications were conducted asynchronously online during two-week spans of time for each round. For some, communicating asynchronously may not have been conducive for in-depth discussion of the criteria. Participants transformed their thoughts into text messages typed into a Web page. The process of encoding thoughts into text form is often more laborious and time consuming than verbalization and may have affected the amount of commentary received. Additionally, the absence of verbal and non-verbal cues might have contributed to less dialogue. Verbal and non-verbal expressions often indicate, for example, that a listener is confused causing the discussant to expound on a topic and provide more detail.

It should be noted that, in efforts not to influence the discussions, we made no attempt to moderate the panel of experts. Participants contributed to the extent they thought appropriate. In retrospect, more prompting and questioning from us may have resulted in increased dialogue. Although an analysis of online informational exchange methods is beyond the scope of this article, it was an important factor in data collection. More research is needed to examine methods of conducting expert panel discussions online, which yield an amount of data comparable to face-to-face discussions. Moreover, the amount of commentary is not indicative of its quality. We judged 91% of the comments as valuable to the study's purpose. Additional inquiry must also investigate the quality of commentary received online relative to face-to-face panel discussions.

Consistency of Participant Opinion

Within rounds, the consistency of criteria ratings increased as the study progressed. This suggests that, among participants, a degree of agreement about the criteria existed. Exposure to the criteria and other participant ratings and commentary may have fostered more congruous ratings in later rounds.

Across rounds, it appears participant attitudes about the criteria remained constant overall. Of the 98 criteria analyzed, only three statements showed significant mean differences from one round to another. The statements "Is the content and vocabulary level for intended users appropriate?" (M = 4.26), "Will the courseware accept mouse and keyboard input?" (M = 3.23), and "Do learners have control over navigation through courseware?" (M = 4.30) each exhibited differ-

ences in rating across rounds. In these cases, individuals who adjusted their ratings appeared to be influenced by the majority. For example, most participants gave the statement, "Is the content and vocabulary level for intended users appropriate?" a 4 (agree). Every instance in which a rating was adjusted, the participant changed it from 5 (strongly agree) on Round 1 to 4 (agree) on Rounds 2 and 3.

These findings indicate that online expert panel discussions such as the one used in this study enable geographically dispersed discussants to examine criteria collectively. It also appears that opinions about criteria will form as discussants are exposed to criteria and review each other's ratings and commentary. However, the extent to which the online medium aided or impeded the panel discussions is unknown and is an area of further inquiry.

SUMMARY

This study set out to identify important criteria that characterize well-designed multimedia instructional courseware and to examine a process of conducting panel discussions online. During three rounds of questioning, a panel of experts rated the importance of 81 criteria, recommended modifications to the criteria list, and proposed additional items. Participants rated most criteria highly; in other words, they strongly agreed or agreed the criteria were important. Many criteria judged to be important related to issues of instructional adequacy, informational content, and evidence of effectiveness. Participants proposed 29 additional write-in statements, of which only 16 were original. The final list of criteria consisted of 97 items grouped in 14 categories. Many participant comments identified problems with criteria, proposed revisions, and expressed participant perceptions about or experiences with the premises being evaluated. Most problems related to unclear wording, preciseness, and the criterion's relevance to courseware evaluation. The majority of revisions proposed by participants consisted of adding new criteria or adding to an existing criterion. Participants suggested these types of revisions because they perceived them as not being addressed in the original list.

The final criteria list offers educators and courseware designers a set of evaluation guidelines subjected to review and analysis by experts in the instructional technology field. In this regard, it provides educators an initial framework from which to review the capability and appropriateness of courseware. It also provides courseware designers an indication of key features and issues to consider in the design of courseware applications.

We made three observations about the online panel discussion. First, participant review and analysis of criteria and their commentary was beneficial for refining the original list of items. Second, commentary was limited, and it diminished in Round 3 of the study. This raises additional questions about the amount of dialogue that would have transpired if discussants met face-to-face and conveyed their thoughts verbally and non-verbally. The effort needed to encode one's thoughts into text form relative to verbalization, the asynchronous nature of the discussion format, the lack of verbal and non-verbal communication, and the absence of a moderator to guide dialogue, among other things, may have restrained participant commentary. Third, participants appeared to

agree in their opinions concerning the criteria and, overall, gave consistent criteria ratings for the duration of the study. These observations suggest that online panel discussions are possible; however, the nature and dynamics of such discussions and the factors influencing them are areas for additional research. ■

Contributors

William Gibbs is an associate professor and the head of the Department of Media at Eastern Illinois University. He received his PhD in instructional systems from Pennsylvania State University. His research interests include knowledge acquisition, technology-based learning environments, and methods for effective instructional software evaluation. Pat R. Graves is a professor in the School of Business at Eastern Illinois University. She teaches business communications and computer-related courses. Her research interests include the visual display of information and computer-mediated communication. Dr. Graves is co-editor of the *Business Education Index* published by Delta Pi Epsilon, a graduate research association for business teacher educators. Ronan S. Bernas is an assistant professor at the Department of Psychology of Eastern Illinois University. He received his PhD in psychology (Committee on Human Development) from the University of Chicago in 1995. His research is on argumentative and explanatory discourse. He examines the learning and conceptual changes that occur during argumentative and explanatory discourse. (Address: Dr. William J. Gibbs, Department of Media, Eastern Illinois University, 600 Lincoln Ave., Charleston, IL 61920; cfwjg1@ux1.cts.eiu.edu.)

References

Bovée, C. L., & Thill, J. V. (2000). *Business communication today.* Upper Saddle River, NJ: Prentice Hall.

Chang, L. L., & Osguthorpe, R. T. (1987). An evaluation system for educational software: A self-instructional approach. *Educational Technology, 27*(6), 15–19.

Dudley-Marling, C., & Owston, R. D. (1987). The state of educational software: A criterion-based evaluation. *Educational Technology, 27*(3), 25–29.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis verbal reports as data.* Cambridge, MA: The MIT Press.

Gros, B., & Spector, M. (1994). Evaluating automated instructional design systems: A complex problem. *Educational Technology, 34*(5), 37–46.

Hannafin, M. J., & Peck, K. L. (1988). *The design, development, and evaluation of instructional software.* New York: Macmillan.

Heinich, R., Molenda, M., Russell, J. D., & Smaldino, S. E. (1996). *Instructional media and technologies for learning.* Columbus, OH: Merrill.

Heller, R. S. (1991). Evaluating software: A review of the options. *Computer Educator, 17*(4), 285–291.

Jolicoeur, K., & Berger, D. E. (1986). Do we really know what makes educational software effective? A call for empirical research on effectiveness. *Educational Technology, 26*(12), 7–11.

Jolicoeur, K., & Berger, D. E. (1988). Implementing educational software and evaluating its academic effectiveness: Part I. *Educational Technology, 28*(9), 7–13.

Linstone, H. A., & Turoff, M. (1975). *The Delphi method: Techniques and applications.* Reading, MA: Addison-Wesley Publishing Co.

Le Maistre, C. (1998). What is an expert instructional designer? Evidence of expert performance during formative evaluation. *Educational Technology Research and Development, 46*(3), 21–36.

Morrison, G. D., Lowther, D. L., & DeMeulle, L. (1999). *Integrating computer technology into the classroom.* Columbus, OH: Prentice Hall.

Newell, A., & Simon, H. A. (1972). *Human problem-solving.* Englewood Cliffs, NJ: Prentice-Hall Inc

Oswalt, B. J. (1982). *Identification of competencies necessary for computer literacy and determination of emphasis placed on each competency in introduction to data processing courses offered at the high school level.* Unpublished doctoral dissertation, University of Memphis, TN.

Roblyer, M. D. (1988). Fundamental problems and principle of designing effective courseware. In D. J. Jonassen (Ed.), *Instructional design of microcomputer courseware* (pp. 7–33). Hillsdale, NJ: Lawrence Erlbaum Associates.

Roblyer, M. D., Edwards, J., & Havriluk, M. A. (1997). *Integrating educational technology into teaching.* Columbus, OH: Merrill.

Sapre, P. M. (1990). *Research methods in business education.* Little Rock, AR: Delta Pi Epsilon.

Schueckler, L. M., & Shuell, T. J. (1989). A comparison of software evaluation forms and reviews. *Journal of Educational Computer Research, 5*(1), 17–33.

Shuell, T. J., & Schueckler, L. M. (1989). Toward evaluating software according to principles of learning and teaching. *Journal of Educational Computer Research, 5*(2), 135–149.

Smeltzer, L. R., & Leonard, D. J. (1994). *Managerial communication.* Boston: Irwin McGraw-Hill.

Tolhurst, D. (1995). Hypertext, hypermedia, multimedia defined? *Educational Technology, 35*(2), 21–26.

Zahner, J. E., Reiser, R. A., Dick, W., & Gill, B. (1992). Evaluating instructional software: A simplified model. *Educational Technology Research and Development, 40*(3), 55–62.