Software for the Automated Evaluation of Web-Delivered Instruction

George Nickles, Amy Pritchett

School of Industrial and Systems Engineering
Georgia Institute of Technology

Introduction

Many forms of technology have been used to mediate education between instructor and student, ranging from simple chalkboard drawings to complex intelligent tutoring systems. Recently, the advantages of the Internet, including speed of communication and use of a variety of media, have made it the focus of much educational research.

The effectiveness of Internet mediated education must be proven through evaluation. Evaluation in the context of educational systems is briefly defined as examining the effectiveness of an educational system (or component of that system) in meeting learning and teaching goals. Bloom, Hastings, and Madaus[1] give a classic, more detailed definition.

There are many measurement issues to consider when preparing an educational evaluation. One is to understand what form of evaluation is being conducted. The three major forms of evaluation are planning, formative, and summative,[2, 3] corresponding to the system's life cycle. Planning evaluation takes place early in the design phase to ensure the system is consistent with known educational theories. Formative evaluation takes place during development and implementation and is intended to drive improvement. It has been compared to quality control, as it continuously searches for weaknesses in the technology and opportunities for improvement.[1] Summative evaluation judges the effectiveness of an educational system after completing one or more cycles of operation. While all are necessary, formative evaluation is the type most-closely tied to instructors' efforts to improve the educational system.

A recent survey finds 74% of engineering instructors use the Internet to provide instructional material to students,[4] adding a component to their educational systems that needs evaluating. However, the same survey also notes that only 41% of instructors who use the Internet report evaluating the Internet components of their courses. The survey results suggest some reasons for this low number, particularly noting a lack of time and evaluation support resources for instructors. There are few off-the-shelf tools for evaluating learning over the Internet, requiring instructors (or departments) to build their own to suit their needs. Instructors may not receive release time to develop such tools, may simply not have time due to other constraints, or may not have the required evaluation knowledge and programming skills to develop evaluation tools.

The survey also suggests that where evaluation tools are made available to instructors, they are likely to use them only within their time constraints; therefore the tools must be quick and easy to use for formative evaluation.

Two important measures of effectiveness of a formative evaluation are latency and resolution. The main purpose of formative evaluation is to drive improvement, so decreasing the latency between administering an evaluation and generating results and feedback means that improvements can be implemented sooner. The survey mentioned above[4] found that the most widely used evaluation tool (88% of engineering instructors) is end-of-term surveys provided by the university or department. However, such surveys are generally intended to be summative for the entire academic term and may only drive improvement for the next term. Also, with respect to the parameter of resolution, these surveys typically are very broad and attempt to cover education across the whole term. An instructor may wish to evaluate one day's lesson or one section of a course website in detail. Thus, the most widely used evaluation tool in engineering education does not provide low latency (i.e., shorter time) or tightly focused resolution, and is not specifically directed at Internet mediated education.

In the following sections, evaluation measurements and analysis methodologies will be briefly reviewed with respect to how each can be implemented for Internet mediated education. Following this, a recently-developed evaluation system will be described that integrates the various measurements and encourages the use of formative evaluation to improve instruction.

Evaluation Measurements & Methodology

Process Measures

Process measures examine the process by which a task is completed. In the context of an educational system, the task is the behavior exhibited during teaching by the instructor and learning by the student. One popular way to collect process data from an Internet web site is through the web server log. When analyzed, this can reveal how many visitors came to the site, how often certain parts of the site are accessed, where visitors are from through their IP address, approximately how long individual pages are viewed, and patterns of use over time.[5] Many programs exist to extract and report this type of information from a server's log file.

However, a difficulty in using a web server log is how to predict learning from the behavior recorded in the log. For example, a large number of hits on an individual page may be interpreted as either that page being very useful to students (and thereby referenced often) or very difficult to understand (thereby requiring multiple viewings to fully grasp its material). There are very few tools designed to examine a server log for an educational system, and they are typically more focused on the usability of the course web site rather than learning.

In addition, tools for tracking student activities on a course website are often not easily applied to formative evaluation. For example, the student tracking tool on WebCT allows faculty to view the lecture material most recently viewed by students and the length of time of each access. However, this view is on a student-by-student basis, requiring faculty who desire an overall assessment of lecture material access to step through each student's latest activities, rather than

being provided with a summary of the use of specific pieces of instructional material by all students.

The literature is sparse on predicting learning from behavior while using a web site. However, some recent work indicates that more access of relevant Internet material correlates with better performance. A study of server logs by McNulty, Halama, Dauzvardis, and Espiritu[6] show a positive correlation between use of the educational system and in-class exam scores. Although not raw web server logs, Lu, et al[7] using WebCT logs also found that more visits to relevant content coincided with higher exam scores. In addition, in a more interactive Internet-based system, Bruckman, Jensen, and DeBonte[8] found that higher "time on task" measured by number of commands entered (collected from specially generated logs) were positively correlated to the performance measures used.

Even from this one trend, some use can be made of web server logs in evaluation. It is difficult to quantify in an absolute sense how much use of a website equates to a certain letter grade of performance. However, they can be used to make relative comparisons, such as to compare sections of an educational website to each other. Sections of an educational website that motivate students to use them frequently would show greater use in the logs, and it can be predicted that these components will produce more learning. Sections with very low use should be investigated further with other metrics.

Performance Measures

Performance measures examine the outcomes of a task, typically grading them against a criterion. In the context of education systems, performance measures are referred to as assessments of learning outcomes, which may or may not be recorded as a grade for the class. Internet-based performance measures may take many different forms, including quizzes implemented as a standard HTML form on a web page or dynamically generated pop-up windows with interactive questions. Each assessment of student performance, or particular questions in an assessment, may be identified as relating to a specific topic in the course, hereby also enabling evaluation of that topic. In this case, the assessment may be presented immediately after the student views the topic, or the next time the student logs in, or even the next time the student logs in after a preset time delay. Each performance measure need not be attached to specific learning material but could simply be presented the next time each student logs in after a preset time to examine certain goals over the entire semester.

Many Internet-delivered assessment tools exist. Typical features include automation of the process of assessment distribution and student submission, scoring of the assessment, and automatic feedback to student and instructor.[9-11] Some are designed mainly for students as a non-graded self-check on understanding of the material.[12] While many of these are stand-alone systems, course delivery packages, such as WebCT, have an assessment module built-in. Two of the greatest benefits of an automated assessment system are the time saved by instructors in distribution and grading[13] and the immediate feedback available to students.[14]

However, the types of Internet-based assessment that can be scored by machine are limited. Any free response question still requires scoring by a human. Even if only single words or phrases are required to answer the question, issues such as capitalization, misspelling, homonyms, and

synonyms must be considered if the computer is the sole judge of the correct answer. Likewise, the response is a numerical value, a rounding error or typing error can lead to an incorrect response. In one study, students stepped through a mathematical procedure, entering intermediate values that led to the solution during the process.[15] The assessment program had to be tested and revised multiple times as students would introduce small errors due to rounding in an early step (which the system could tolerate) that would affect the final answer (which the system may or may not tolerate). The only perfectly reliable type of machine-scored question has a set, discrete number of potential responses, only one of which is correct and can be selected from the whole set.

In addition to problems with automatic scoring, cheating becomes an issue when exams are proctored electronically. Generally, strict security measures are used including a human supervisor to ensure no dishonesty occurs when students' grades are significantly affected by the assessment.[16] Therefore, online assessment may be better suited for evaluation of the website.

Unlike web server logs that passively collect information, assessments require active engagement from students to be useful. Although some students may be self-motivated to assess their own learning, other incentives may be required to get the majority to complete assessments. Only using assessments that are graded for evaluation is one incentive, albeit problematic for the reasons noted above. Using sample questions similar to those that may appear on a graded exam for non-graded assessments may be another incentive.

Even with these limitations, performance measures can demonstrate learning, the main outcome desired of any educational system. When well designed, they can examine learning of a very specific topic, may be able to classify any errors made, and reduce the time burden on the instructor.

Subjective Opinion Measures

Subjective opinion measures examine attitudes of users of the system. There are many difficulties in determining how to interpret subjective opinion data, as by their nature they measure opinion rather than an observable physical characteristic. Opinions may be shaped by motivation, difficulty, aesthetics, a sense of having learned, or a combination of these and other factors that may be difficult to tease apart. Student's opinions of system components may reveal a strong like or dislike, both of which may be cause for further investigation or may directly show highly effective or ineffective educational methods. Surveys and interviews are two frequently used subjective opinion measures.

Surveys for educational evaluation are typically administered on paper, although some universities are now using Internet administered midterm and end-of-term surveys. Returning to the issue of resolution, these standard surveys typically examine the course as a whole without providing sufficient detail to improve specific parts of the course website. Additionally, these surveys are likely to be more focused on traditional course activities such as lectures and instructor-student interactions than the Internet components of the course.

Many examples of on-line surveys exist, especially for consumer or product marketing research. There are some on-line survey tools available for educational evaluation, such as those built in to

course delivery systems. Similar to performance measures, surveys must use questions designed to have discrete responses, such as a rating scale, in order to be machine scored. Still, the advantages of machine scoring are great in terms of saving faculty time.

Methodology

A serious measurement issue is the difficulty of collecting valid educational measurements, as we cannot directly access the cognitive process of learning.[17] Thus, measurements such as those mentioned above must be used to indirectly measure the cognitive process of learning. Process measures are collected unobtrusively and reveal the behavior of the students, which has shown some promise for predicting learning. However, behavior captured by a web server log can only capture interactions with the web site, not any work students do off-line such as working example problems on paper. Performance measures examine whether expected outcomes of learning are present and can be focused very tightly on specific topics of interest. Yet, these are obtrusive measures and may direct learning only toward expected questions rather than the broader body of knowledge on the web site. Also, machine scoring is a limiting factor when implementing performance measures on the Internet for the reasons previously mentioned. Subjective opinion measures may also be tightly focused and can be used to determine specific features of a web site that are viewed positively or negatively by students. However, interpretation may be difficult due to the subjective nature of the data.

One methodology that has been proposed to circumvent this problem is measurement triangulation.[5, 7] The concept of measurement triangulation in evaluation is that the more diverse types of measurements that can be taken on an item of interest, the more precise and accurate the result. The measurements used in triangulation typically fall into the three categories mentioned above.

Another way to perform triangulation is to examine the correlation between each pair of measurements collected. This is difficult for several reasons. First, how the data sets from each measurement should be paired is not clear as some are taken over time (behavior) while others are not (an individual application of a survey or assessment). Also, data with categorical and ordinal scales is typically collected from surveys and must be aligned with interval (and possibly ratio) scaled data for analysis. One method to use correlation is to give assessments and surveys at set intervals, such as every one or two weeks. Summary statistics for each students' use of the website over that period can be paired with their assessment and survey responses. Correlations such as these can show effects of changes over time and the general trend of the website.

Another triangulation method is to compare each measurement to a criterion to create an indicator of positive or negative effects of learning. Then, the set of all indicators are examined together to interpret the results. While assessment measures have a means of indicating positive or negative effects by using percent of questions correct, a criterion for a rating scale is more difficult to determine. Also, server logs have no fixed criterion, simply that greater use suggests (but does not prove) greater learning. A reasonable criterion can be chosen for the survey (e.g., the middle value of the scale) and the server logs (e.g., the per student averages for the entire website) so that an analysis of this type can be conducted. This method has the advantage of being able to aggregate all sources of data at once and examine them all at least in the same

positive/negative indicator sense. The meaning provided by each set of indicators is not fully understood.

As an example of triangulation, Lu, Zhu, and Stokes[7] use one measurement from each category - process measures (surveys) performance measures (exams), and subjective opinion (web server logs) - to examine the Internet component of a course. However, they examined the Internet component over the entire semester and had a large number of measurements to draw together from electronic log files and hardcopy exams and surveys, which likely took a great deal of time. The effort required to perform this evaluation was high, the resolution was low, and the latency was fairly high. This illustrates the significant amount of time and resources that must be spent to make triangulation an effective formative evaluation strategy.

Integrated Evaluation System

As seen above, the three categories of measurement used in evaluation can be implemented on the Internet. However, there has been no single tool that integrates all these measures in a comprehensive analysis and provides feedback to the instructor suitable for formative evaluation with low latency and high resolution. Some course delivery packages may have all these tools available for instructors, but they are more focused on assessing students rather than evaluating the delivery system. Also, each measurement type is presented separately, making triangulation on a single component of the delivery system more difficult.

Recognizing the advantages of triangulating with each category of evaluation measurement, all three have been integrated into a single evaluation component of an Internet-based course delivery system. This system, named IT Web, is in beta testing in the School of Industrial and Systems Engineering at Georgia Tech. The evaluation component of IT Web includes a log file tool, an assessment tool, a survey tool, and an analysis tool. Instructors can access the evaluation component from their course management webpage. In the evaluation component, instructors select a section of their course website to evaluate, typically an individual lecture topic. The evaluation component then prompts the instructor for the necessary parameters, such as survey questions and responses. Each tool in the evaluation component is described below.

Log File Analysis Tool

The log file analysis tool consists of Perl scripts designed to examine the web server log file with results presented with PHP. The log file is automatically generated by the web server and is a record of every electronic transaction with the server. It records information such as the IP address that requested a file, the time the file was requested, and the name of the file requested.

Basic statistics on access of individual files are available. In addition, access of an individual file can be compared to a baseline, typically the average access rate (per student) of all files, to determine if it is used more or less frequently. Also, another baseline for comparison can be specified, such as all other files in the same course. Values such as hits and visits are presented in units per student so comparisons can be made across courses or all of IT Web. This tool provides the ability to show relative use of individual website sections and to infer greater or less learning relative to the baseline.

Assessment Tool

The assessment tool is implemented in PHP with a MySQL database. The instructor is prompted to enter a set of questions for each assessment related to the section of the website being studied. These questions may be true/false, multiple choice, or fill in the blank (with a single entry). The instructor must supply the questions and all answers. The instructor can also select various options, such as what feedback will be provided. Students will see a pop-up window after viewing the section of the website being evaluated. They will then fill out the assessment and submit it.

Statistics on assessment scores are generated upon request of the instructor, such as overall score on an assessment and responses to individual questions. If the questions are well designed, with each incorrect answer revealing incorrect learning by the students, the system can diagnose what parts of the section of the website need modification.

Another ongoing development is a feature to link evaluation of a single topic covered by the website and students' graded assignments. Individual homework assignments, quizzes, and exam questions will be matched to their respective topics from the website, and scores on these questions will be used in the assessment tool.

Survey Tool

The survey tool is similar in design to the assessment tool and is also implemented in PHP with a MySQL database. In addition to supplying their own questions, instructors are able to select from a set of survey questions (Figure 1). Three types of questions are available: Likert scale (for rating individual components on certain criteria), multiple choice (selecting the statement that most closely matches the student's opinion), and short answer (for free form response). Instructors may also select options such as when the surveys will be presented to students. In addition to the specified questions, students are given the opportunity to submit anonymous comments to the instructor concerning the website. Although these cannot be machine scored, they are an opportunity for students to share urgent needs or suggestions with the instructor.

Although subjective opinion may not directly reveal learning, it may reveal what sections of the website are difficult to learn from. Subjective questions asking about specific components of the web site will reveal student preferences and therefore whether they had difficulty learning from those components. In addition to basic statistics on the items of the surveys, instructors are presented with the sections of the website that rate poorly and specific questions for those sites that rate poorly in order to diagnose the problems.

Analysis Tool

In the current version, the analysis tool draws the results from the measurements together to allow the instructor to generate reports on each separately and together for triangulation. The instructor can select an individual topic from the course to analyze and generate a report showing use of this topic in terms of page hits and visits and the results of assessments and surveys administered on that topic. The measurements are also converted into positive/negative learning

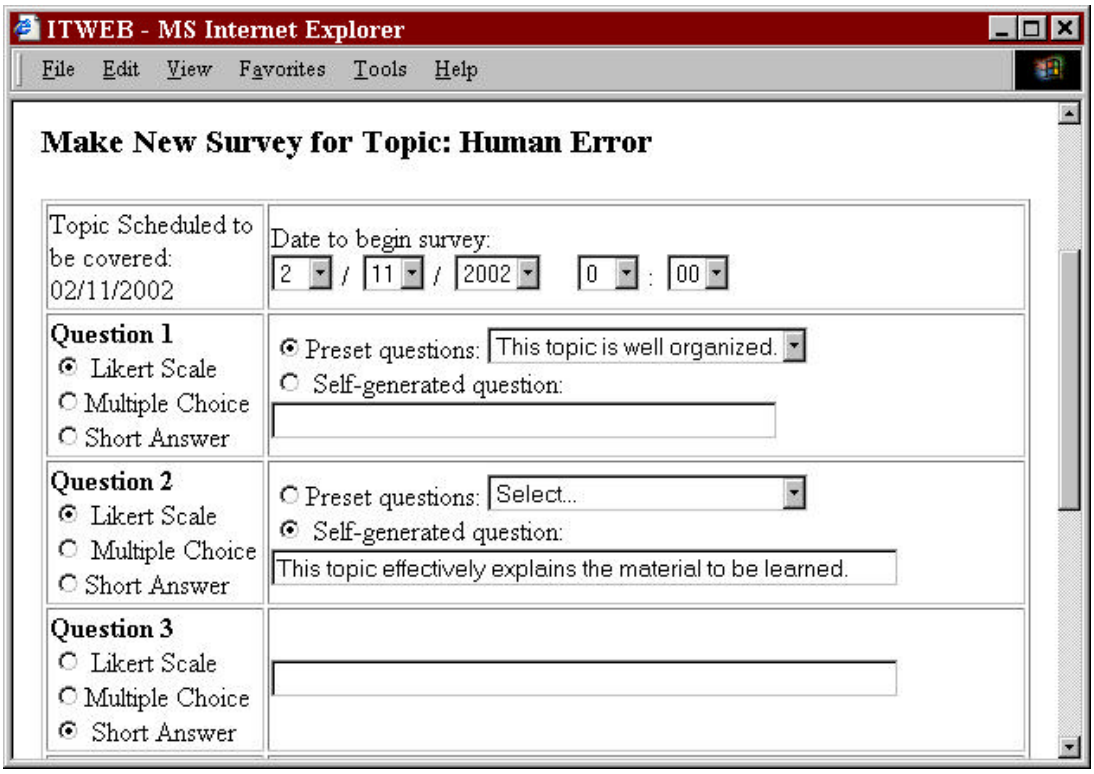

Figure 1: Creating a new survey for a topic in the IT Web evaluation component.

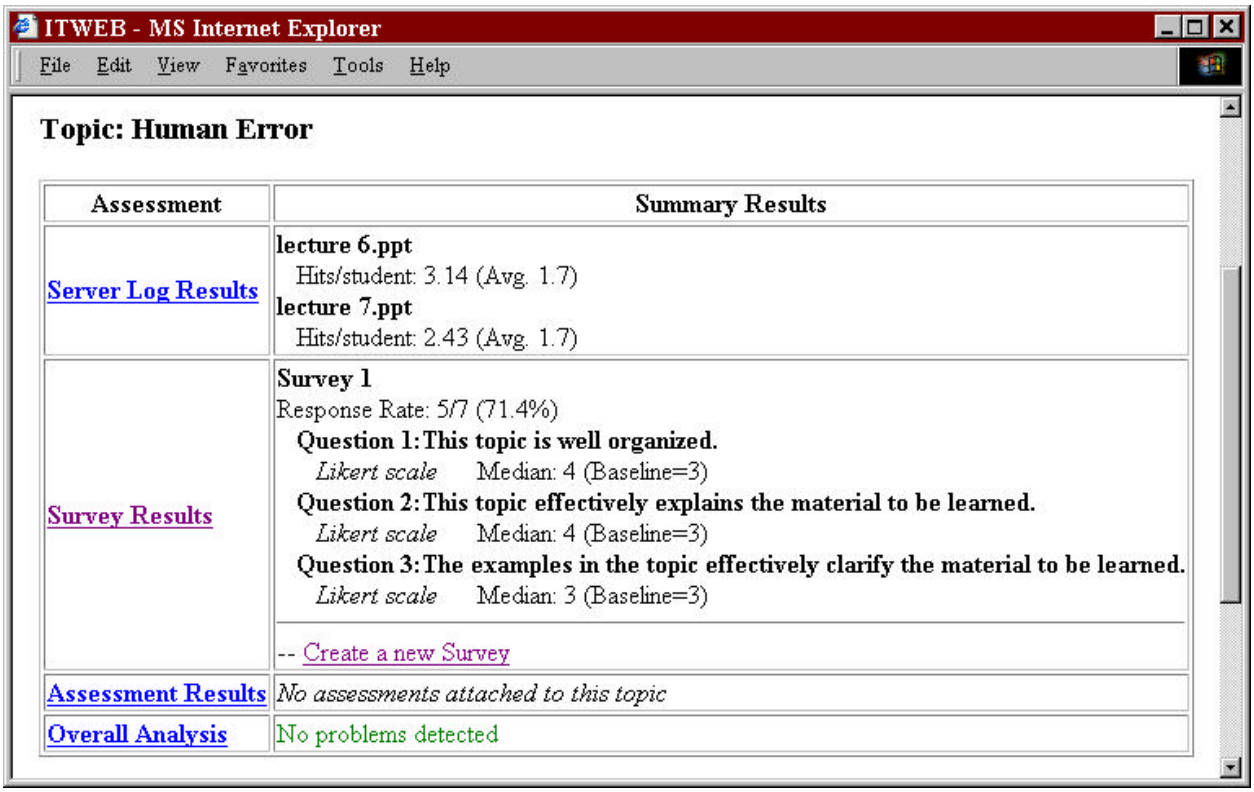

Figure 2: Overall evaluation results for a single topic in the IT Web evaluation component.

indicators and analyzed by the analysis tool to interpret the performance of the section in light of all the measurements combined (Figure 2). For example, if all the measurements tend to be positive (high usage rates, high assessment scores and grades, and high survey ratings) it can be inferred that this section of the website is effective for learning. However, if assessment scores and survey ratings for a section are low but usage rates are high, it may be inferred that students made many efforts to learn from the website, but it was not effective for learning. The specific points of the website that proved difficult may be determined from examining scores on individual assessment questions and survey results. In both of these cases, the tool can perform these analyses and report them to the instructor.

It should be emphasized that the analysis tool is not simply a reporting device, but also provides recommendations based on the analysis. If built-in questions were used on the survey that can be recognized by the analysis tool, results from these will also be analyzed individually and incorporated into the report. Also, the analysis tool examines individual questions on each assessment and the specific part of the website they are associated with. In the previous example where problems are indicated by overall low assessment scores and survey ratings, individual assessment scores may indicate a particular topic that reflects poor performance. In addition, the survey may show especially low ratings on the built in question "Examples are clear and effective" indicating the examples in that topic need revision. All this information from each measurement will be available to the instructor, but triangulation readily reveals a problem and even this first implementation of the analysis tool can provide specific feedback for improvement.

The next version, which is currently under development, will include many more evaluation functions. The assessment tool will be integrated with the assignments and grading section of the website, allowing graded assignments to be included in the evaluation. This will be done to the level of individual questions on assignments and exams to provide high resolution on individual topics on the website. In addition, the evaluation component will have a correlation triangulation tool, which will guide the instructor through creating the necessary assessments and surveys to properly administer and analyze this type of triangulation.

Instructors can select multiple ways to receive feedback from the analysis tools. Instructors can request the analysis tool to e-mail them a weekly evaluation report. This report provides a summary of results from each measurement for the week and a very basic triangulation analysis of the website's performance, alerting the instructor if there appears to be a problem. The instructor may also go directly to the analysis tool in the evaluation component from their course website management tools. This provides the summary from the weekly e-mail, and allows more detailed analysis and perusal of all data collected for the evaluation. Instructors are able to view an evaluation summary for any time period of their course up to the entire term. Also, more detailed reports can be generated on individual assessments, surveys, and the server log. These will show individual questions with scores or ratings and more detailed web site use statistics.

In future versions, instructors can have an evaluation summary, similar to the e-mail summary, displayed when they log in the website. Also, analysis of assessments, surveys, and the server log will include dynamically generated graphs and charts for easier examination.

Conclusions

The evaluation component of IT Web offers great benefits to instructors for the investment of time necessary to set up the assessments and surveys. Evaluation support is built in to the system, guiding the instructor in preparing and analyzing the measurements. Also, the time that would be spent distributing, processing, and analyzing the results of each measurement and integrating results from each into a cohesive report is saved, not only by reducing latency but also making it feasible for instructors to conduct an evaluation within their time constraints. Feedback is provided quickly to the instructor so changes can be made during the term, rather than waiting until the end of the term to identify and correct problems. Also, the resolution of the evaluation can be adjusted to focus on a large or small section of the system.

Acknowledgements

This research is sponsored by grant #EEC-0080315 from the National Science Foundation.

Bibliography

1. B. S. Bloom, J. T. Hastings, and G. F. Madaus, *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill, (1971).

2. N. Walker, "A Primer on Evaluating Your Engineering Education Research Project," accessible from http://www.succeed.ufl.edu/pubs/papers/evaluation_primer/index.html: SUCCEED Engineering Education Coalition, (1997).

3. F. Stevens, F. Lawrenz, and L. Sharp, "User-Friendly Handbook for Project Evaluation: Science, Mathematics, Engineering, and Technology Education," accessible from http://www.ehr.nsf.gov/EHR/RED/EVAL/handbook/handbook.htm: National Science Foundation Division of Research, Evaluation, and Communications, Directorate for Education and Human Resources, (1993).

4. G. Nickles, A. Pritchett, and L. Trotti, "Methods of Measuring Teaching Effectiveness in the Classroom and on the Internet: A Survey of Engineering Instructors," presentation at American Society of Engineering Education, Albuquerque, NM, (2001).

5. A. L. Ingram, "Using Web server logs in evaluating instructional Web sites," *Journal of Educational Technology Systems*, vol. 28, pp. 137-157, (1999-2000).

6. J. A. McNulty, J. Halama, M. F. Dauzvardis, and B. Espiritu, "Evaluation of Web-based Computer-aided Instruction in a Basic Science Course," *Academic Medicine*, vol. 75, pp. 59-65, (2000).

7. A. X. Y. Lu, J. J. H. Zhu, and M. Stokes, "The Use and Effects of Web-Based Instruction: Evidence From a Single-Source Study," *Journal of Interactive Learning Research*, vol. 11, pp. 197-218, (2000).

8. A. Bruckman, C. Jensen, and A. DeBonte, "Gender and Programming Achievement in a CSCL Environment," presented at Computer Supported Collaborative Learning 2002: Foundations for a CSCL Community, Boulder, CO, (2002).

9. J. E. Bartlett, K. A. Reynolds, and M. W. Alexander, "inQuisit(c): A tool for online learning," *Journal of Online Learning*, vol. 11, pp. 22-24, (2000).

10. C. E. Brawner, "Practical Tips for Using Web-based Assessment Systems," *T H E Journal (Technological Horizons in Education)*, vol. 28, pp. 38, (2000).

11. B. Armstrong, D. Dingsdag, and D. Neil, "Electonric assessment software for distance education," presented at Proceedings International Workshop on Advanced Learning Technologies. IWALT 2000. Advanced Learning Technology: Design and Development Issues, Palmerston North, New Zealand, (2000).

12. R. J. White and C. A. Hammer, "Quiz-o-Matic: A free Web-based tool for construction of self-scoring on-line quizzes.," *Behavior Research Methods, Instruments & Computers*, vol. 32, pp. 250-253, (2000).

13. R. W. Hall, L. G. Butler, N. K. Kestner, and P. A. Limbach, "Combining feedback and assessment via Web-based homework," *Campus-Wide Information Systems*, vol. 16, pp. 24-26, (1999).

14. M. Chetty, "Scheme for on-line Web-based assessment," *Engineering science and education journal*, vol. 9, pp. 27-32, (2000).

15. H. D. Him, G. M. Nickles, M. S. Leonard, D. L. Kimbler, and A. K. Gramopadhye, "Asynchronous Learning Applied to Classroom Instruction: Lessons Learned," presented at Industrial Engineering Research Conference '99, Phoenix, AZ, (1999).

16. L. Carswell, P. Thomas, M. Petre, B. Price, and M. Richards, "Understanding the 'Electronic' Student: Analysis of Functional Requirements for Distributed Education," *Journal of Asynchronous Learning Networks*, vol. 3, (1999).

17. A. L. Brown, "Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings," *The Journal of the Learning Sciences*, vol. 2, pp. 119-156, (1992).

Biographical Information

GEORGE NICKLES
George Nickles is a PhD student in the School of Industrial and Systems Engineering at Georgia Tech with an emphasis in Human-Integrated Systems. He received his B.S and M.S. in Industrial Engineering from Clemson University. His research interests include cognitive engineering, educational technology, evaluation, and training.

AMY PRITCHETT
Amy Pritchett is an Assistant Professor in the School of Industrial and Systems Engineering, and a Joint Assistant Professor in the School of Aerospace Engineering, at Georgia Tech. She received her S.B., S.M. and Sci.D. from MIT's Department of Aeronautics and Astronautics. Her research specializes in the design of complex human integrated systems, including cockpit design and educational technology.