Student Assessment in Online Courses: Research and Practice, 1993–2004

John A. Ross & Maura Ross
Ontario Institute for Studies in Education (OISE)
University of Toronto

Abstract

Research on student assessment in online environments has not been extensive, although manuals for instructors provide broad guidelines and specific procedures. In this article we review the most frequently reported approaches to online assessment in postsecondary settings, giving particular attention to systems for assessing the quality of student participation. We also extrapolate from research on assessment in face-to-face courses to identify strategies that could be usefully adapted to online assessment. Research on the reliability and validity of online assessment methods is mixed and there is not much of it. We suggest to online instructors that there appears to be a disjunction between assessment

Résumé

Bien qu'on trouve les grandes lignes directrices et des procédures spécifiques dans les manuels des instructeurs, il n'y a pas de recherche poussée sur l'évaluation des étudiants dans un environnement en direct. Dans cet article, nous revoyons les approches les plus fréquemment signalées en milieu postsecondaire sur l'évaluation en direct, tout en portant une attention particulière aux systèmes évaluant la qualité de la participation étudiante. Aussi extrapolons-nous de la recherche sur l'évaluation faite dans des cours face-à-face afin d'identifier des stratégies pouvant être adaptées utilement à l'évaluation en direct. La recherche sur les questions de fiabilité et de validité des méthodes d'évaluation en direct est contradic-

Canadian Journal of University Continuing Education
Vol. 31, No. 2, Fall 2005
pp. 37–57

methods and instructional ideologies and suggest to researchers that there is an urgent need to investigate the consequential validity of online assessment.

toire, aussi n'est-elle pas abondante. Nous suggérons aux instructeurs en ligne, qu'il semble avoir une disjonction entre les méthodes d'évaluation et les idéologies pédagogiques, et nous suggérons aux chercheurs qu'il importe au plus haut point d'entreprendre une enquête sur la validité consécutive de l'évaluation en direct.

Introduction

Manuals for online instructors provide meager advice about student assessment, and there is scant research evidence about the effects of such guidance. In this article, we begin by reviewing the literature on student assessment in online courses in post-secondary institutions. When appropriate, we compare the findings to research on student assessment in face-to-face courses. After a brief rehearsal of our search procedures, we describe how online assessment differs from face-to-face assessment, identify the most frequently reported online student-assessment strategies (and some less frequently discussed alternatives), outline the results of research on their reliability and validity, and, finally, suggest directions for research and practice.

Literature Search

Fink's (1998) procedures were followed to create a narrative review that was systematic, reproducible, and explicit. To locate studies, multiple online searches of the ERIC database for reports published between 1993 and 2004 were conducted, using the keywords *online courses, distance education, distributed learning, computer mediated communication,* or *virtual classroom*, combined with *student evaluation, evaluation methods, evaluation,* or *evaluation criteria,* and repeated with *validity* or *reliability.* Our searches produced 428 unique documents, which were culled to 88 by reviewing their abstracts and then to 33 articles and reports by applying four criteria: peer-reviewed journals were selected over non-refereed sources, recent documents over less-recent documents, reports containing empirical evidence over accounts of personal practice, and reports containing greater relevance to the research questions that guided the search over those of lesser relevance.[1] We also limited our review to studies of online instruction in post-secondary settings.[2]

Additional materials were located by examining references in the "hits" and by drawing upon research on student assessment in face-to-face courses to provide context for our interpretations. The studies were coded on the basis of our research questions. We repeated the search with the keywords noted above for JSTOR (www.jstor.org) for 1995 to 2005, which generated 177 hits, and for AACE (Association for the Advancement of Computing in Education at aace.org), which generated 131 hits. Sixteen of these hits were new and relevant to our review. Finally, in conducting our review, we were mindful of the sources of error and bias identified by qualitative (Ogawa & Malen, 1991) and quantitative researchers (Matt & Cook, 1994). For example, we tested our claims with an explicit search for contrary evidence.

Online Assessment Versus Face-to-Face Assessment

Online assessment differs from face-to-face assessment. First, assessment opportunities are more frequent online because students can simultaneously communicate with each other and with the instructor and because the duration of an online class is longer (potentially 168 hours per week). Second, online courseware provides instructors with tracking tools, which tally the number of times students login to a course, post contributions, and download the posts of others, making these actions easy to observe and simple to count. Third, nonverbal behaviour and status indicators that distort assessments are stripped away in online settings. Fourth, the decentralized nature of computer-mediated communication (CMC) classrooms shifts the focus from the teacher to the student by demanding greater interaction and participation (Berge, 1997). The multiple levels of discourse enabled by CMC not only facilitate collaboration and active learner involvement, but can also transform fragmented surface knowledge to deeper levels of understanding (Koschmann, 1996).

Students' ability to participate in multiple ways makes the instructor's job more laborious (Hsu, Marques, Hamza & Alhalabi, 1999; Levin, Levin, & Waddoups, 1999; Macdonald, 2001), while creating opportunities for innovative assessment. Research to date indicates that the assessment strategies used in online courses are similar to those used in face-to-face courses, even though online instructors encounter challenges not experienced in face-to-face situations.

Online Student Assessment Strategies

Most reports contained only a paragraph or two on the assessment of online learners, and many (e.g., McCarthy-McGee, 2000; Mellon, 1999; Palloff & Pratt, 1999) did not report the evidence that supported their assertions (an exception is Liang & Creasy, 2004). Some authors described the theory behind their recommendations (e.g., Hsu et al., 1999; Wade, 1999), but most did not. Teacher-directed procedures were discussed much more frequently than assessment methods that provided opportunities for learner involvement in assessment decisions.

Examinations, Tests, and Essays

Many state-supported institutions must include formal exams, regardless of the mode of course delivery. Discussions of grading issues in the online literature tend to address administrative issues, such as how to evaluate and return tests and examinations electronically (e.g., automated marking of multiple-choice questions using .cgi scripts). Palloff and Pratt (1999) discussed the convenience of electronic tests and quizzes that provide instructors and students with immediate feedback. Other advantages of computer-assisted testing include repeatability, reliability, equitability, timeliness, and flexibility (Brown, Race, & Bull, 1999). Rovai (2000) recommended that computer-assisted testing tools be used for lower-level cognitive tasks in low-stakes assessments.

Examinations are recommended, especially when the stakes are high, as a defence against cheating. Ensuring that the students who are registered for an online course are the ones actually contributing to course discussions and completing course assignments is a key concern in the literature (Lieblein, 2001; Rowntree, 1998). Scheduling and proctoring real-time examinations to combat online cheating were common suggestions. However, such advice fails to consider the unique needs of many online distance learners (Palloff & Pratt, 1999; Rovai, 2000). These assessment techniques are more suited to the measurement of recall-type objectives and less sensitive to the assessment of deeper understanding. Reliance on individual accountability measures, such as tests, fails to recognize that many online courses emphasize collaborative productions.

Essays have the potential to elicit complex intellectual behaviour and for that reason are recommended to online instructors (see, e.g., Macdonald, 2001). End-of-course essays in online settings are indistinguishable from similar assignments in face-to-face courses. Rubrics developed to guide markers and to communicate expectations to students are similar in both forms of course delivery. The approach of simply modifying conventional assessment tools to the online environment is common (Levin et al., 1999).

Performance Assessments: Counting and Rating Online Participation

The simplest form of performance assessment is measuring the frequency of participation in online discussions, for example, using Web logs (Chen, Liu, Ou, & Liu, 2000; Pappas, Lederman, & Broadbent, 2001; Rovai & Barnum, 2003). The assumption is that more frequent participation indicates higher productivity.

At a more sophisticated level, Wade (1999) recommended that transcripts of online sessions be assessed using standards for written communication, that is, grammatical and organizational structure, as well as content. Reeves (2000) proposed that concept mapping be used to measure conceptual understanding, and Brown et al. (1999) suggested that instructors assess Information Technology skills by examining either the product or the process used to generate it. Witmer (1998) created a three-pronged approach that examined students' conceptual understanding, computer-mediated conferencing skills, and utilization of the medium.

Theory-driven approaches to assessing online interaction offer fruitful possibilities. For example, Henri's (1992) online discourse-analysis model distinguished four content domains: social, interactive, cognitive applications, and metacognitive skills. Classifying samples of discourse for each student could provide frequency counts or ratings of quality within each domain. Although such schemes have formative evaluation utility, it is not clear how such frequencies and ratings can be translated into grades.

Garrison, Anderson, and Archer (2001), Gunawardena, Lowe, and Anderson (1997), and Salmon (2000) have each proposed multi-stage schemes for representing knowledge construction within online courses. Their models could be used as rubrics for assessing the progress of individuals or groups. However, if these phases are not linear (Gunawardena et al. suggested they may be recursive), distinguishing high from low performance becomes problematic. Stage theory may have greater use as a modifier of schemes for interpreting online discourse; in other words, the criteria for judging the quality of a student's contribution to a discussion may vary with the stage of the course.

Several researchers have developed schemes for analyzing student constructions that could enable instructors to distinguish the quality of online performance. Woodruff (1995) viewed online learning environments as micro communities that use argument as a form of inquiry and develop shared knowledge through constructive conflict and proposed an argumentation hierarchy for assessing contributions to a shared interpretation of a text. Level 1 arguments build a set of collectively valid statements, for example, students expressing unelaborated agreement with a proposition. Level 2 arguments elaborate an idea by suggesting warrants, evidence, or ways to test the idea. Level 3 arguments identify discrepancies between a proposed idea and conventional belief, for example, identifying misconceptions. Level 4 contributions challenge an idea by presenting contrary evidence, thereby suggesting an alternative hypothesis. Similar schemes based on the co-construction of argument in online courses have been developed by Ross (1996), Hmelo-Silver (2003), Luppicini (2002), and Pena-Shaff and Nicholls (2004). These categories could be converted to an assessment rubric.

To date, taxonomies of knowledge construction in online courses have been generic; that is, they are applicable regardless of course content. A subject-specific approach to measuring contributions to group productivity would focus on conversational turns that contribute to student learning within a specific scholarly community. Discipline-based explanations are built around a discipline's conventions for building new knowledge. Leinhardt (2001) presented a model of instructional explanations, consisting of a query, the use of examples, the role of intermediate representations, such as analogies and models, and devices that limit or bound explanation (errors, principles, and conditions of use). Leinhardt illustrated the model with examples from face-to-face classes in mathematics and history, showing how each of these elements was manifest in quite different forms in the two disciplines. Leinhardt (1993) demonstrated that the same categories could be used to classify the quality of individual student contributions to group understanding. Cobb, Wood, Yackel, and McNeal (1992) and Ross (1995) constructed coding schemes for interpreting student conversations as they jointly solved mathematics problems, providing examples of challenges, justifications, and explanations of solutions. To date, none of the coding schemes that build upon the unique characteristics of subject-specific reasoning has been applied to online student assessment.

Finally, discourse-analysis studies in face-to-face courses identify patterns associated with student achievement. These studies indicate that it is sequences of dialogue (requests and responses) that need to be measured, not isolated utterances. For example, asking for simple information contributes to learning if an answer is received; if it is not, the effect is negative. Asking for an explanation and receiving one is potentially the most powerful learning strategy, but the results have been mixed. Some researchers have found a positive effect; others have not. However, when an explanation-seeker is given no response or an unelaborated factual reply, the effects are invariably negative. (For reviews of the evidence, see Ross & Cousins, 1995b, and Webb, 1989.) All forms of academic help-giving contribute to the help-givers' learning, including giving unelaborated information (facts and procedures), evaluations, and, especially, explanations. The discourse-analysis approach has the advantage of being simpler to code than schemes for assessing contributions to arguments. In addition, the categories identified as salient are founded on a highly consistent set of findings that is derived from process-product studies on the impact of particular discourse patterns on the achievement of speakers and listeners. We found no reports of online student assessment that drew upon this literature.

Rating the quality of student interactions is an especially suitable technique for online assessment. These strategies are founded on well-developed theories of learning in collaborative, constructivist environments, and there is substantial evidence of their face validity as outcome measures (in the case of contribution to argument) and their construct validity (in the case of conversational moves that predict student achievement). These strategies are particularly attractive for online courses because no new set of tasks is required to generate the assessment data—the same activities serve instructional and assessment purposes. Rating the quality of student interactions is considerably easier in online than in face-to-face courses because what students contribute is automatically recorded. In contrast, audio-recording of student conversation in face-to-face courses is highly intrusive and susceptible to distortions from social desirability, and transcribing the talk is hugely expensive.

Self- and Peer Assessment

Macdonald (2001) argued that online distance learning lends itself to peer assessment on a one-to-one and a one-to-many bases. Macdonald found that posting exemplars for peer review resulted in an iteractive evaluation process that was beneficial to the instructor, tutors, and students. Asynchronous computer-mediated conferences promote peer assessment when users read, reflect on, and post feedback to contributions; Pena-Shaff and Nicholls (2004) found that 89% of messages posted to a course bulletin board were responses to earlier messages. O'Reilly and Morgan (1999) reported that peer assessment motivates learners and helps build a sense of community. Yet we found little evidence of peer assessment in use.

The literature we reviewed gave little attention to self-assessment. Levin et al. (1999) recommended that instructors model exemplary work as a strategy for promoting self-assessment in asynchronous forums. Self-assessment is particularly important in online courses because it can provide information about affective states that influence achievement, such as students' goal orientations and beliefs about their ability to master the content of the course (Ross, Rolheiser, & Hogaboam-Gray, 1999). The visual cues that inform instructors about these states in face-to-face settings are missing in online courses.

Portfolio Assessment

We found no discussion of portfolio assessment in online courses, even though portfolios are increasingly used in university and college courses to measure student achievement in face-to-face settings. For example, Slater, Ryan, and Samson (1997) asked students to provide evidence in three portfolio assignments that they had mastered each objective in a face-to-face course. The evidence could consist of journals, tests, homework assignments, and other artifacts. Each entry had to include a reflection that indicated how the evidence related to the objective. The application of portfolio assessment to online courses would entail working through the same set of issues confronting portfolio assessment in face-to-face courses: defining learning objectives to be demonstrated, developing and distributing assessment criteria (e.g., through rubrics), and identifying types of evidence acceptable to the assessors. The only difference is that students would likely submit their portfolios electronically. However, we did not find penetrating discussion of these issues in the online literature, despite surfacing 404 hits for *portfolio* in the AACE database.

Summary

The literature we reviewed identified a small group of traditional measures (tests, examinations, and essays) that dominate online assessment. Currently, great attention is being given to performance tasks in the form of a range of strategies for assessing online participation. The most promising of these involve theory-driven approaches generated by conceptualizing how knowledge is constructed, either generically or in the context of particular disciplines. Feasible procedures for operationalizing these ideas remain to be worked out. Other alternate assessment strategies (peer, self-, and portfolio assessment) have received less attention but offer fruitful possibilities based on experiences of student assessment in face-to-face courses.

Psychometric Properties of Online Assessment

Reliability

Studies examining consistency over time indicate that on some measures, particularly frequency of interaction, there are duration effects. For example, student collaboration increases over the duration of an online course (Pena-Shaff & Nicholls, 2004; Thorpe, 1988) as learners move through a series of learning stages (Garrison et al., 2001; Henri, 1992; Salmon, 2000). This secular trend may be problematic if the norms for assessing frequency of online discourse are static.

Inter-rater reliability (i.e., whether different raters assign the same categories or qualities to transcripts of interaction) was high (80–90% agreement)[3] for the transcript-analysis schemes examined by Fahy, Crawford, Ally, Cookson, and Keller (2000). In contrast, other researchers (Macdonald, 2001; Smith & Coombe, 2000) found qualitative evidence to suggest that students tended to disagree with the assessments of student interactions made by external markers, such as tutors. This finding might be attributed to tutors' misunderstanding of assessment criteria and their application; tutors meet with instructors infrequently and receive limited direction (McCulloch, 1997).

Some researchers have warned that peer assessments in online courses may be unreliable. Macdonald (2001) observed that students' lack of content knowledge could limit agreement between instructor and peer evaluations. Durham (1990) argued that peer evaluations are vague and unreliable, even when the assessment is based on the average assigned by several students (multiple raters usually increase reliability). However, Topping's (1998) review of 67 quantitative studies of peer assessment in face-to-face, post-secondary classes found that 72% of the studies reported high reliability (typically reporting correlation coefficients that were in the 80s). High reliability was more likely to occur when expectations were clarified for students and a climate of trust was created—conditions that could as easily be created in online as in face-to-face courses. However, lower reliability was observed when peers assessed contributions to group projects. Topping noted that a few studies have included computer-assisted procedures to support peer assessment. For example, Downing and Brown (1997, cited by Topping, 1998) had students in a Web-enhanced, face-to-face course post drafts of their essays to a Web site and critique each other's products by e-mail. However, the reliability of these and other computer-assisted peer evaluation procedures has not been reported.

The reliability of online assessment in online courses may be impeded by a lack of transparency about how online courses are graded (Purnell, Cuskelly, & Danaher, 1996; Reeves, 2000; Schrum & Berge, 1997). Student confusion over assessment criteria impacts negatively on reliability when students have assessment roles (i.e., in peer and self-evaluation).

Validity

Even when assessment is tightly controlled by the instructor, opaque criteria may impact on the validity of student assessments. Students who are unclear about the standards for appraising their work may fail to provide the kind of evidence that instructors need to make evaluative judgments.

As in face-to-face courses, the most serious validity threat in online courses is a mismatch between measurement tools and learning objectives. Many online instructors design their assessments to increase the frequency of student interaction (see, e.g., Palloff & Pratt, 1999; Schrum & Berge, 1997), but the connection between participation frequency and achievement has not been demonstrated. For example, a student reading posts is not necessarily actively engaged in the content presented, even if we were to assume, as many users of tracking tools do, that opening or downloading messages means they are read. "Scanning" is often employed to lessen the cognitive overload associated with multiple online postings. As with offline reading, scanning employs surface processing in order to determine the sections or posts that merit a more in-depth examination. Increased reading time is not necessarily indicative of increased comprehension, higher-level thinking skills, and/or deep learning (Reinking, 1988). CMC tools that track the amount of time a user spends online are not accurate measures of student learning. There are firmer grounds for claims about the validity of discourse-analysis schemes, although the evidence of their validity comes from studies of student interaction in face-to-face settings. The most promising approach—schemes based on conceptualizing how knowledge is constructed online—comes from qualitative studies. We were not able to find any studies that correlated scores on these process variables with outcome measures.

Rourke and Anderson (2004), in their review of quantitative content-analysis research, identified several shortcomings of such schemes that have implications for their use as student-assessment measures. For example, they noted that since only a few of the coding schemes have been tested in more than a single study, few norms are available for interpreting the frequency of the codes, and evidence is lacking to correlate them with achievement. Rourke and Anderson concluded that these schemes may work well for low-inference categories, but the claims behind high-inference use (e.g., assuming that messages with particular characteristics show that higher-level thinking was required to generate them) need to be demonstrated. Another factor influencing the validity of using message-content analysis to assess the quality of student performance is the presence of systemic bias in response patterns. Hewitt (2003) found evidence of "recency" effects, that is, students tend to respond to the most recent notes in a conference and ignore earlier messages that might have greater relevance to the conceptual issues in the course.

Other threats to the validity of the interpretations of assessment measures have been identified. Particularly important are personalogical variables, which are defined as students' characteristics that interact with assessment procedures to distort appraisal. The *Principles for Fair Student Assessment Practices for Education in Canada* (Joint Advisory Committee, 1993) require that "assessment methods should be free from bias brought about by student factors extraneous to the purpose of the assessment" (p. 4). Although this document identifies developmental stages and special circumstances as factors for evaluators to consider, it fails to highlight the skills and knowledge outlined in the literature as being important to online learners, which might affect the interpretability of the evidence. Wade (1999) described a familiar list of characteristics shared by successful students in online courses: "strong self-starter, being self-disciplined, being knowledgeable of the technology requirements of the specific format, and being able to meet other students and the faculty in a virtual environment" (p. 96). Dasher-Alston and Patton (1998) recommended that instructors assess a student's ability to succeed in an online learning environment, although they did not set out any procedures for doing so. McCarthy-McGee (2000) took this assertion further by suggesting that instructors identify different student populations in online courses and tailor their instruction accordingly. Berge and Myers (2000) found that 30% of online instructors engaged in some kind of data collection to measure student knowledge, skills, and attitudes that might affect online learning. Student features that affect online learning also influence the assessment of that learning.

A major contributor to the validity of online assessment is that the removal of personal identifiers from online contributions (although never complete) reduces instructor and student bias by eliminating peer pressure and the halo effect. However, Brown et al. (1999) reported that students were uncomfortable with peer and self-assessment, believing each to be influenced by pre-existing social networks or by self-delusion.

As noted earlier, counting and classifying participation are key assessment techniques. Gruber (1995) found that the number of times a post is read is influenced by social structures, prejudice, time constraints, and gender. The frequency and quality of student contributions depend in part on comfort level with the medium. Ross (1998) found the gender composition of student groups influenced the students' experience in an online course. Women exercised less procedural leadership, had reduced influence on group products, contributed less to the advancement of their group's argument, and overall had fewer productive contributions—all behaviours that have been proposed as indicators of online participation. Students who are uncomfortable expressing their thoughts in writing may be concerned that their posts will be ridiculed because of their content or mechanics. This is particularly inhibiting for students in different programs and/or second-language learners. Cultural differences may lead to systematic over- or underestimates of student engagement in the course, and cultural differences in argumentation (e.g., whether it is permissible to overtly disagree with the instructor or a peer) may interfere with assessment based on schemes for rating contributions to knowledge construction. Rovai (2000) discussed the implications of different online discourse patterns by contrasting the linear deductive nature of English with the circular approach of Japanese. These differences in computer-usage patterns (Panero, Lane, & Napier, 1997, identified four computer-usage "dimensions") differ from the online developmental stages identified earlier in that they are non-hierarchical. However, the former could distort classification of students into the latter.

In an online environment, a student's demonstration of conceptual understanding may be depressed by a lack of proficiency in using the technology. Messages that arrive incomplete, garbled, or not at all speak volumes about the competence of the sender. Readers (and assessors) may extrapolate from the form in which the message arrives to its content, ascribing lower ability to the sender. Despite these concerns, Ross (1996) found that lack of computer skills had a negligible effect on participation in an online graduate course. Students with high and low computer skills scored equally on knowledge-construction measures. In contrast, students with weaker prior knowledge of the course content scored substantially below those with stronger prior knowledge on the same measures. Conversely, Purnell et al. (1996) found that a lack of access to resources required for task completion was a major concern for geographically isolated students. The ability to search and retrieve relevant online documents mitigates the effects of distance, as do university library support systems, but the inequality in opportunity to learn likely distorts online assessment in ways not encountered in face-to-face courses.

We found little evidence of the consequential validity of online assessment, that is, evidence that the assessment procedures contributed to student learning, with one exception: the timing of feedback to students affected student satisfaction and learning. Purnell et al. (1996) found the timeliness of instructor feedback was a major concern for a group of geographically isolated students engaged in an online course. Similarly, McCarthy-McGee (2000) reported that frequent interaction enabled the instructor in a graduate-level course to be more proactive, which resulted in a better student experience. Macdonald (2001) found that university students particularly valued feedback in the early stages of a course, as it helped to confirm they were on the right track or it provided redirection. These studies highlighted the importance of a feedback loop in learning. Thorpe (1998) suggested that the asynchronous nature of CMC creates immediacy in the feedback loop, which increases the learning pace, as well as reducing the control students have over timing. We found no research on the effects of particular methods on students or teachers, compared to the extensive research on the consequential validity of specific assessment methods in face-to-face courses (e.g., Moss, 1998).

Summary

Evidence of the reliability of online assessment procedures is scanty and relatively weak. Inter-rater reliability is inconsistent, although the factors that contribute to higher reliability in online courses (e.g., untrained tutors) have been identified. There is some evidence to suggest that assessments are not consistent over time. No internal-consistency measures (e.g., inter-item correlations) have been reported for online assessment. In contrast, the literature on the reliability of student-assessment strategies in face-to-face courses is extensive.

Research on the validity of online assessments is in its early days. We found no studies that correlated measures of the quality of online participation with achievement. Although many of the schemes in use or in development are founded on substantial evidence, the claims have been extrapolated from research in face-to-face settings. Researchers have also identified powerful threats to the validity of assessment judgments, particularly those related to student cultures and opportunities to learn, which might distort online assessments.

Directions for Research and Practice

The literature reviewed for this article suggested that advances in online teaching in post-secondary courses have outpaced progress in student assessment. The following directions for research and practice emerged from our review.

Issues for Online Instructors

1. *Unify theories of instruction and assessment.* Online instruction lends itself to a constructivist orientation to learning. Yet the assessment strategies most frequently reported were tests, examinations, and essays more suited to the measurement of declarative knowledge in transmission-oriented settings. The most fruitful way of aligning assessment with instruction is building student-assessment procedures from a theory of how knowledge is collaboratively constructed in online settings.
2. *Make transparency an explicit goal of assessment.* Students are more likely to achieve course goals when they know what they are expected to know and do. Involving students in developing the criteria used to assess their work is a powerful communication strategy, one that was not mentioned in the literature we examined, despite the commitment to student control of their learning.
3. *Link assessment criteria directly to outcomes.* Some of the assessment criteria in the studies we reviewed were only distantly related, if at all, to course outcomes. For example, there was little evidence to suggest that frequency of participation is linked to achievement. In contrast, research in online and face-to-face environments has identified a few universally applicable categories of interaction that either correlate with or credibly represent deep processing of course concepts. The most promising of these categories, not yet explored in online settings, is student contributions to the development of a scholarly community, uniquely defined by the discipline in which the course is housed.

4. *Explore approaches to online assessment that share assessment responsibility with students.* Learner involvement in online assessment is highly compatible with the online instructor's role as a facilitator of student knowledge construction. Assessment strategies that provide a decision-making role for learners (peer, self-, and portfolio assessment) appear to be underused. Assessment-training activities, such as co-development of rubrics, shared selection of exemplars, and instructor modeling of assessment practice, enable students to share assessment tasks.

Issues for Researchers

1. *Examine the reliability of online assessment.* Little research has been conducted on the reliability of online assessment. Instead, the research has focused on inter-rater indicators to the neglect of other procedures, and the evidence on the reliability of online assessment has been mixed. Studies of internal consistency, consistency over time, within-rater consistency, and between-rater consistency would be helpful, especially if researchers identify the conditions under which particular assessment strategies have high and low reliability.
2. *Examine the validity of online assessment procedures.* A number of studies have incidentally examined the validity of online assessment, identifying a number of sources of validity threats. This useful theme needs to be continued, drawing upon the extensive knowledge about assessment validity developed in face-to-face environments. Especially useful would be studies of alternate or authentic assessment in online contexts. Research that manipulates online-assessment procedures to reduce sources of invalidity has high theoretical and practical value.
3. *Develop new assessment procedures suitable for constructivist forums.* Knowledge construction, a central goal of many approaches to online instruction, occurs within a disciplinary context. Research that develops models of subject-specific knowledge construction and translates these models into practical assessment tools is needed. Testing the validity of these tools against existing online-assessment strategies is a critical part of the research agenda.
4. *Incorporate stage theories into online assessment*. Researchers have found that interactions within an online course go through predictable patterns as a scholarly community develops. The next step is to adapt assessment strategies to this development, highlighting particular criteria for particular phases or redefining standards in response to changing opportunities to learn.
5. *Conduct studies of the consequential validity of online assessment*. Current research on online assessment is focused on psychometric character-

istics for accountability purposes. In contrast, research on face-to-face assessment emphasizes consequential validity, in other words, the extent to which particular assessment strategies have beneficial effects on student achievement and/or instructional practice. A key question is whether the findings from face-to-face studies generalize to online contexts.

A final concern that is relevant to both groups is the issue of standards for student assessment. None of the well-known statements of standards (e.g., American Educational Research Association, 2000; Joint Advisory Committee, 1993) addresses the specific assessment issues that confront online instructors.

Conclusion

The literature on student assessment in online courses has made a promising start, but is still in its infancy. In the immediate future, we anticipate that innovations will flow from studies of assessment in face-to-face contexts to online environments. However, online courses have intrinsic advantages for working out the assessment implications of constructivist pedagogy. As the field of online assessment matures, we anticipate that the flow of innovations will reverse direction.

Endnotes

1. The research questions were:
   - How does student assessment differ from student assessment in face-to-face courses?
   - What online student assessment practices are recommended?
   - What are the psychometric properties of online student assessment practices?
   - What are the effects of different online student assessment practices on students and/or on course processes?
   - To what extent are online student assessment practices rigorous, transparent, and fair?
2. We omitted assessment practices in elementary and secondary schools because few studies of online courses in these environments have been reported and assessment practices differ substantially between K-12 and post-secondary settings.
3. Percentage agreement is very sensitive to chance, especially when the raters use fewer of the categories than the scheme provides. A chance-adjusted procedure like Cohen's Kappa is more appropriate than percentage agreement.

References

American Educational Research Association. (2000). Position statement of the American Educational Research Association concerning high-stakes testing in preK-12 education. *Educational Researcher, 29*(8), 24–25.

Berge, Z. L. (1997). Characteristics of online teaching in post-secondary, formal education. *Educational Technology*, *37*(3), 35–47.

Berge, Z. L., & Myers, B. (2000). Evaluating computer mediated communication courses in higher education. *Journal of Educational Computing Research*, *23*(4), 431–450.

Brown, S., Race, P., & Bull, J. (Eds.). (1999). *Computer-assisted assessment in higher education*. London: Kogan Page.

Chen, G., Liu, C., Ou, K., & Liu, B. (2000). Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology. *Journal of Educational Computing Research*, *23*(3), 305–332.

Chou, C. C. (2001). Formative evaluation of synchronous CMC systems for a learner-centered online course. *Journal of Interactive Learning Research*, *12*(2/3), 173–192.

Cobb, P., Wood, T., Yackel, E., & McNeal, B. (1992). Characteristics of classroom mathematics traditions: An interactional analysis. *American Educational Research Journal, 29*(3), 517–544.

Dasher-Alston, R. M., & Patton, G. W. (1998). Evaluation criteria for distance learning. *Planning for Higher Education*, *27*(1), 11–17.

Durham, M. (1990). Computer conferencing, students' rhetorical stance and the demands of academic discourse. *Journal of Computer Learning*, *69*(4), 265–272.

Fahy, P. J., Crawford, G., Ally, M., Cookson, P., & Keller, V. (2000). The development and testing of a tool for analysis of computer mediated conferencing transcripts. *The Alberta Journal of Educational Research*, *36*(1), 85–88.

Fink, A. (1998). *Conducting research literature reviews: From paper to the internet*. Thousand Oaks, CA: Sage.

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education*, *15*(1), 7–23.

Gruber, S. (1995). Re: ways we contribute: Students, instructors, and pedagogies in the computer mediated writing classroom. *Computers & Composition, 12*(1), 61–78.

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research*, *17*(4), 397–431.

Henri, F. (1992). Computer conferencing and content analysis. In A. Kaye (Ed.), *Collaborative learning through computer conferencing* (pp. 117–136). Berlin: Springer-Verlag.

Hewitt, J. (2003). How habitual online practices affect the development of asynchronous discussion threads. *Journal of Educational Computing Research*, *28*(1), 31–45.

Hmelo-Silver, C. E. (2003). Analyzing collaborative knowledge construction: Multiple methods for integrated understanding. *Computers and Education, 41*, 397–420.

Hsu, S., Marques, O., Hamza, M. K., & Alhalabi, B. (1999). How to design a virtual classroom: 10 easy steps to follow. *T.H.E. Journal*, *27*(2), 96–109.

Joint Advisory Committee. (1993). *Principles for fair student assessment practices for education in Canada*. Retrieved March 31, 2004, from http://www.education.ualberta.ca/educ/psych/crame/files/eng_prin.pdf.

Koschmann, T. (Ed.). (1996). *CSCL: Theory and practice of an emerging paradigm*. Mahwah, NJ: Lawrence Erlbaum Associates.

Leinhardt, G. (2001). Instructional explanation: A commonplace for teaching and location for contrast. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 333–357). Washington, DC: American Educational Research Association.

Leinhardt, G. (1993). Weaving instructional explanations in history. *British Journal of Educational Psychology, 63*, 46–74.

Levin, J., Levin, S. R., & Waddoups, G. (1999). Multiplicity in learning and teaching: A framework for developing innovative online education. *Journal of Research on Computing in Education*, *32*(2), 256–269.

Liang, X., & Creasy, K. (2004). Classroom assessment in web-based instructional environment: Instructors' experience. *Practical Assessment, Research & Evaluation, 9*(7). Retrieved March 4, 2004, from http://PAREonline.net/getvn.asp?v=9&n=7

Lieblein, E. (2001). Critical factors for successful delivery of online programs. *Internet and Higher Education*, *3*(3), 161–174.

Luppicini, R. J. (2002). Toward a conversation system modeling research methodology for studying computer-mediated learning communities. *Journal of Distance Education, 17*(2), 87–101.

Macdonald, J. (2001). Exploiting online interactivity to enhance assignment development and feedback in distance education. *Open Learning*, *16*(2), 179–189.

Matt, G. E., & Cook, T. D. (1994). Threats to validity of research synthesis. In H. Cooper & L. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York: Russell Sage Foundation.

McCarthy-McGee, A. F. (2000). Addressing evaluation and assessment while delivering online learning for the army. *Internet and Higher Education*, *3*(3), 175–181.

McConnell, D. (2002). The experience of collaborative assessment in e-learning. *Studies in Continuing Education*, *24*(1), 73–92.

McCulloch, K. H. (1997). Participatory evaluation in distance learning. *Open Learning, 12*(1), 24–30.

Mellon, C. A. (1999). Digital storytelling: Effective learning through the internet. *Educational Technology*, *39*(2), 46–50.

Moss, P. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6–12.

Ogawa, R., & Malen, B. (1991). Towards rigor in reviews of multivocal literatures: Applying the exploratory case study method. *Review of Educational Research, 61*(3), 265–286.

O'Reilly, M., & Morgan, C. (1999). Online assessment: Creating communities and opportunities. In S. Brown, P. Race, & J. Bull (Eds.), *Computer-assisted assessment in higher education* (pp. 149–161). London: Kogan Page.

Palloff, R. M., & Pratt, K. (1999). *Building learning communities in cyberspace: Effective strategies for the online classroom*. San Francisco: Jossey-Bass.

Panero, J. C., Lane, D. M., & Napier, H. A. (1997). The computer use scale: Four dimensions of how people use computers. *Journal of Educational Computing Research*, *16*(4), 297–315.

Pappas, G., Lederman, E., & Broadbent, B. (2001). Monitoring student performance in online courses: New game—new rules. *Journal of Distance Education, 16*(2). Retrieved November 3, 2005, from http://cade.athabascau.ca/vol16.2/pappasetal.html.

Pena-Shaff, J. B., & Nicholls, C. (2004). Analyzing student interactions and meaning construction in computer bulletin board discussions. *Computers and Education, 42*, 243–265.

Purnell, K., Cuskelly, E., & Danaher, P. (1996). Improving distance education for university students: Issues and experiences of students in cities and rural areas. *Journal of Distance Education*, *11*(2), 75–101.

Reeves, T. C. (2000). Alternative assessment approaches for online learning environments in higher education. *Journal of Educational Computing Research*, *23*(1), 101–111.

Reinking, D. (1988). Computer mediated text and computing differences: The role of reading time, reader preference, and the estimation of learning. *Reading Research Quarterly*, *23*(4), 484–498.

Ross, J. A. (1998). Differential participation of males and females in a computer-mediated communications course. *Canadian Journal of University Continuing Education, 24*(1), 83–100.

Ross, J. A. (1996). The influence of computer communication skills on participation in a computer conferencing course. *Journal of Educational Computing Research*, *15*(1), 37–52.

Ross, J. A. (1995). Students explaining solutions in student-directed groups: Cooperative learning and reform in mathematics education. *School Science and Mathematics, 95*(8), 411–416.

Ross, J. A., & Cousins, J. B. (1995). Impact of explanation seeking on student achievement and attitudes. *Journal of Educational Research, 89*(2), 109–117.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing, 6*(1), 107–132.

Rourke, L., & Anderson, T. (2004). Validity in quantitative content analysis. *Educational Technology Research and Development, 52*(1), 5-18.

Rovai, A. P. (2000). Online and traditional assessments: What is the difference? *Internet and Higher Education*, *3*(3), 141–151.

Rovai, A. P., & Barnum, K. T. (2003). On-line course effectiveness: An analysis of student interactions and perceptions of learning. *Journal of Distance Education, 18*(1), 57-73.

Rowntree, D. (1998). Assessing the quality of materials-based teaching and learning. *Open Learning, 13*(2), 12–22.

Salmon, G. (2000). *E-moderating: The key to teaching and learning online.* London: Kogan Page.

Schrum, L., & Berge, Z. L. (1997). Creating student interaction within the educational experience: A challenge for online teachers. *Canadian Journal of Educational Communication*, *26*(3), 133–144.

Slater, T., Ryan, J., & Samson, S. (1997). Impact and dynamics of portfolio assessment and traditional assessment in a college physics course. *Journal of Research in Science Teaching, 34*(3), 255–271.

Smith, E., & Coombe, K. (2000, September). *Distance education at arm's length: Outsourcing of distance education marking.* Paper presented at the "Distance Education: An Open Question?" conference, Adelaide, Australia.

Thorpe, M. (1998). Assessment and third-generation distance education. *Distance Education*, *19*(2), 265–286.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Wade, William. (1999). Assessment in distance learning: What do students know and how do we know they know it? *T.H.E. Journal*, *27*(3), 94–100.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International Journal of Educational Research, 13*(1), 21–39.

Witmer, D. F. (1998). Introduction to computer-mediated communication: A master syllabus for teaching communication technology. *Communication Education*, *47*(2), 162–173.

Wolf, A. (1995). *Competence-based assessment*. Buckingham, UK: Open University.

Woodruff, E. (1995). *Investigating collaborative maieutics: An examination of the effects of face-to-face and computer networked communication mediums on peer-assisted knowledge-building*. Unpublished doctoral dissertation, University of Toronto, Toronto, Ontario, Canada.

BIOGRAPHIES

John A. Ross is a professor of curriculum, teaching, and learning and head of the Ontario Institute for Studies in Education (OISE), University of Toronto, Field Centre in Peterborough, Ontario. He has taught online courses for eight years. His research focuses on program evaluation, mathematics education, computer-mediated conferencing (CMC) instruction, and student assessment.

John A. Ross est professeur de programmes d'études, d'enseignement et d'apprentissage, et il est chef du Centre local de l'IÉPO / UT à Peterborough en Ontario. Pendant huit ans, il a donné des cours sur Internet. Sa recherche vise l'évaluation de programmes, l'enseignement des mathématiques, la formation par téléconférence informatisée, et l'évaluation des étudiants.

Maura Ross is a doctoral candidate in the Curriculum, Teaching and Learning Department at the Ontario Institute for Studies in Education (OISE), University of Toronto. She has experience teaching students from elementary to postsecondary. Her research interests include online learning, student assessment, teacher education, and the practical application of information and communications technology (ICT) in the classroom.

Maura Ross est candidate au doctorat dans le département des programmes d'études, d'enseignement et d'apprentissage à OISE. Son expérience en enseignement s'étend des étudiants de l'école élémentaire aux étudiants du postsecondaire. Ses intérêts de recherche comprennent l'apprentissage en direct, l'évaluation des étudiants, la formation des enseignants, et l'application pratique des technologies de l'information et des communications dans la salle de classe.