Using Rubrics to Assess Information Literacy: An Examination of Methodology and Interrater Reliability

Megan Oakleaf
326 Hinds Hall, Syracuse, NY 13244. E-mail: moakleaf@syr.edu

Academic librarians seeking to assess information literacy skills often focus on testing as a primary means of evaluation. Educators have long recognized the limitations of tests, and these limitations cause many educators to prefer rubric assessment to test-based approaches to evaluation. In contrast, many academic librarians are unfamiliar with the benefits of rubrics. Those librarians who have explored the use of information literacy rubrics have not taken a rigorous approach to methodology and interrater reliability. This article seeks to remedy these omissions by describing the benefits of a rubric-based approach to information literacy assessment, identifying a methodology for using rubrics to assess information literacy skills, and analyzing the interrater reliability of information literacy rubrics in the hands of university librarians, faculty, and students. Study results demonstrate that Cohen's κ can be effectively employed to check interrater reliability. The study also indicates that rubric training sessions improve interrater reliability among librarians, faculty, and students.

Introduction

Academic librarians seeking to assess information literacy skills often focus on testing as a primary means of evaluation. Educators have long recognized the limitations of traditional tests that include fixed-choice question types (e.g., multiple choice, matching, and true/false). In fact, test limitations cause many K–16 educators to prefer rubric assessment to fixed-choice testing; however, most academic librarians have not adopted a rubric-based approach to information literacy assessment and therefore are unable to take advantage of the instructional benefits offered by rubric assessments. This article seeks to rectify this missed opportunity by describing the benefits of a rubric-based approach to information literacy assessment, identifying a methodology for using rubrics to assess information literacy skills, and analyzing the interrater reliability of information literacy rubrics in the hands of university librarians, faculty, and students.

Received January 30, 2008; revised November 28, 2008; accepted November 28, 2008

 DOI: 10.1002/asi.21030

Rubrics Defined

Based on assessment for learning (Oakleaf, 2009), motivation, and constructivist educational theory (Oakleaf, 2008, p. 244), rubrics are "descriptive scoring schemes" created by educators to guide analysis of student work (Moskal, 2000). Rubrics describe the parts and levels of performance of a particular task, product, or service (Hafner, 2003, p. 1509). Rubrics are often employed to judge quality (Popham, 2003, p. 95), and they can be used across a broad range of subjects (Moskal, 2000). Full-model rubrics, like the one used in this study, are formatted on a grid or table. They include criteria or target indicators down the left-hand side of the grid and list levels of performance across the top (Callison, 2000, p. 34). Criteria are the essential tasks or hallmarks that comprise a successful performance (Wiggins, 1996, p. V-6:2). Performance-level descriptors "spell out what is needed, with respect to each evaluative criterion . . . [for] a high rating versus a low rating" (Popham, 2003, p. 96).

Rubrics can be described as holistic or analytic. Holistic rubrics provide one score for a whole product or performance based on an overall impression. Analytic rubrics, like the one employed in this study, "divide . . . a product or performance into essential traits or dimensions so that they can be judged separately—one analyzes a product or performance for essential traits. A separate score is provided for each trait" (Arter & McTighe, 2000, p. 18). To obtain a holistic score from an analytic rubric, individual scores can be summed to form a total score (Nitko, 2004, p. 226).

Benefits of Rubric Assessment

Rubric assessment of information literacy skills results in a number of benefits to students, librarians, and faculty (Oakleaf, 2008, p. 245). For students, "a properly fashioned rubric can . . . help students learn much more effectively" (Popham, 2003, p. 95). Rubrics allow students to understand the expectations of their instructors. By making instructor-expectations clear, rubrics make rankings, ratings, and grades more meaningful (Bresciani, Zelna, & Anderson, 2004, p. 31). They provide direct feedback to students about what they have learned and what they have yet to learn. Second,

students can use rubrics for self-evaluation. Finally, rubrics emphasize "understanding rather than memorization, 'deep' learning rather than 'surface' learning" (Pausch & Popp, 1997).

University librarians and faculty also can benefit from rubric assessment in two important ways (Oakleaf, 2009). First, the rubric creation process provides an opportunity to discuss and determine agreed-upon values of student learning. Callison (2000) wrote that "Rubrics are texts that are visible signs of agreed upon values. They cannot contain all the nuances of the evaluation community's values, but they do contain the central expressions of those values" (p. 36). Stevens and Levi (2005) listed the facilitation of communication with students, educators, and other stakeholders as a key reason to use rubrics (p. 23). Bresciani et al. (2004) confirmed that rubrics "make public key criteria that students can use in developing, revising, and judging their own work" (p. 30). They also noted that once rubrics are developed, they can be used to norm educators' expectations and bring them in line with the vision for student learning (p. 31).

Second, rubric assessment offers university librarians and faculty assessment data full of rich description. Rubrics provide "detailed descriptions of what is being learned and what is not" (Bresciani et al., 2004, p. 30). This descriptive data can be used to document how to improve instruction (Bernier, 2004, p. 25). Furthermore, rubric assessment data are so detailed and well-defined that they "combat accusations that evaluators do not know what they are looking for in learning and development" (Bresciani et al., 2004, p. 30). Because rubrics are easy to use and to explain, they generate data that are easy to understand, defend, and convey (Andrade, 2000, p. 14). Finally, the level of detail found in rubrics helps prevent inaccurate (Popham, 2003, p. 95) or biased assessment data (Bresciani et al., 2004, p. 31). Because rubrics clarify schemes for assessment ahead of time, they reduce subjectivity in grading (Moskal, 2000). According to Callison, rubric assessment "is more likely to be reasonably objective and consistent from lesson to lesson and from student to student, especially useful in team teaching situations that involve collaboration among library media specialists and other teachers" (2000, p. 35).

Limitations of Rubric Assessment

Like other assessment tools, there are limitations associated with rubric assessment (Oakleaf, 2008, p. 247). Many limitations of a rubric approach to assessment are rooted in poor rubric construction. Not all rubrics are well-written (Popham, 2003, p. 95), and crafting a good rubric requires time, practice, and revision (Callison, 2000, p. 35). Tierney and Simon (2004) cautioned that unfortunately, "the most accessible rubrics, particularly those available on the Internet, contain design flaws that not only affect their instructional usefulness, but also the validity of their results."

Another limitation of rubric assessment is time. While creating rubrics is inexpensive monetarily, some assessors find the process time-consuming (Tierney & Simon, 2004). Part of that perception might be due to a lack of familiarity or expertise (Bernier, 2004, p. 25); librarians do not always know how to make a rubric and thus believe the process will take too much time. Prus and Johnson (1994) acknowledged the potential cost of time required to create a rubric, but felt that the advantages outweigh the costs. They wrote: "As in virtually all other domains of human assessment, there is a consistently inverse correlation between the quality of measurement methods and their expediency; the best methods usually take longer and cost more faculty time, student effort, and money" (p. 25). Stevens and Levi (2005) argued that rubrics actually make grading easier and faster by "establishing performance anchors, providing detailed, formative feedback, ... supporting individualized, flexible, formative feedback, ... and conveying summative feedback" (p. 73).

Reliability

There is "nearly universal" agreement that reliability is an important property in educational measurement (Colton, 1997, p. 3). Reliability is a measure of consistency (Moskal, 2000); however, in performance assessment, reliability is more than getting the same score twice. For performance assessors, two forms of reliability are considered significant. The first form is interrater reliability, which refers to the consistency of scores assigned by multiple raters (Moskal, 2000). The second is intrarater reliability, which refers to the consistency of scores assigned by one rater at different points of time (Moskal, 2000). Because this study investigates the use of rubrics by multiple rater groups, interrater reliability is of more concern than is intrarater reliability.

Interrater Reliability

Many assessment methods require raters to judge or quantify some aspect of student behavior. For example, raters are often used to "empirically test the viability of a new scoring rubric" (Stemler, 2004). In such cases, interrater reliability is a very useful measure. Interrater reliability refers to "the level of agreement between a particular set of judges on a particular instrument at a particular time" and "provide[s] a statistical estimate of the extent to which two or more judges are applying their ratings in a manner that is predictable and reliable" (Stemler, 2004). Raters, or judges, are used when student products or performances cannot be scored objectively as right or wrong but require a rating of degree (Stemler, 2004). This use of raters results in the subjectivity that comes hand in hand with a rater's interpretation of the product or performance (Stemler, 2004). To combat potential subjectivity and unfairness, many assessors develop rubrics to improve the interrater reliability of constructed-response and performance assessments. Moskal and Leydens (2000) stated that rubrics respond to concerns of subjectivity and unfairness by formalizing the criteria for scoring a student product or performance. They wrote that "The descriptions of the score levels are used to guide the evaluation process. Although scoring rubrics do not completely

DOI: 10.1002/asi

eliminate variations between raters, a well-designed scoring rubric can reduce the occurrence of these discrepancies."

There are three general categories of interrater reliability: consensus estimates, consistency estimates, and measurement estimates. Consensus estimates are based on the belief that "reasonable observers should be able to come to exact agreement about how to apply the various levels of a scoring rubric to the observed behaviors" (Stemler, 2004). In contrast, consistency estimates are based on the assumption that "it is not really necessary for two judges to share a common meaning of the rating scale, so long as each judge is consistent in classifying the phenomenon according to his or her own definition of the scale" (Stemler, 2004). Finally, measurement estimates are based on the belief that "one should use all of the information available from all judges (including discrepant ratings) when attempting to create a summary score for each respondent" (Stemler, 2004). For the focus of this study, consensus estimates are the most relevant form of interrater reliability.

Consensus Estimates

Consensus estimates of interrater reliability assume that independent raters should be able to agree on how to use a rubric to score student products or performances. If two raters can agree exactly on a rubric score to assign to a student's work, then the two raters "may be said to share a common interpretation of the construct" (Stemler, 2004). This type of estimate is most useful when data are "nominal in nature and different levels of the rating scale represent qualitatively different ideas" (Stemler, 2004). Consensus estimates also are useful when "different levels of the rating scale are assumed to represent a linear continuum of the construct, but are ordinal in nature (e.g., a Likert scale). In that case, the judges must come to exact agreement about the quantitative levels of the construct under investigation, rather than attempting to evaluate qualitative differences in scoring categories" (Stemler, 2004).

There are three main ways of calculating consensus estimates of interrater reliability. The most popular method is the simple percent-agreement figure. This figure is calculated by "adding up the number of cases that received the same rating by both judges and dividing that number by the total number of cases rated by the two judges" (Stemler, 2004). Three advantages of the simple percent-agreement statistic are that it has "strong intuitive appeal," it is a simple calculation process, and it is easy to explain (Stemler, 2004). There also are two disadvantages to this statistic. First, this calculation is used to compare two raters, and the present study includes 25 raters plus the researcher. Second, the percent-agreement statistic does not correct for chance. In other words, the statistic does not consider the random probability of a rater assigning a particular score. In rubric assessment, the limited amount of criteria and levels of performance description increase the probability of a rater assigning a particular score by chance rather than intention. As a result, the percent-agreement statistic is likely to be artificially inflated. To correct for chance, there is a procedure to modify the percent-agreement statistic. The modification involves requiring not only exact agreement but also adjacent scoring categories on the rating scale. This relaxes the need for exact agreement among raters, but it has one disadvantage. If the rating scale has only a limited number of categories (e.g., a 1–4 scale), the estimate may be inflated (Stemler, 2004). As Stemler (2004) noted,

> If the rating scale has a limited number of points, then nearly all points will be adjacent, and it would be surprising to find agreement lower than 90%. The technique of using adjacent categories results in a situation where the percent agreement at the extreme ends of the rating scale is almost always lower than the middle.

Because the rubric used in this study had only three levels of performance description, this method was not used to analyze interrater reliability.

A second method of calculating a consensus estimate of interrater reliability is Kendall's coefficient of concordance. Kendall's coefficient is used to estimate agreement among multiple raters, corrects for chance, and is appropriate for ordinal responses that are numerically coded (SAS, 2006). Because the rubric used in this study yields responses that are both ordinal and numerically coded, Kendall's coefficient seems a good match for this study. However, one major disadvantage of this statistic is that it offers no agreed-upon index for interpreting results. That is, there are no cutoffs for levels of acceptable or unacceptable reliability estimates. As a result, this statistic was not used to estimate interrater reliability.

The third method of calculating a consensus estimate of interrater reliability, and the method used in this study, is Cohen's κ statistic. This statistic estimates the degree of consensus among multiple raters on nominal data after correcting for the "amount of agreement that could be expected by chance alone based on the values of the marginal distributions" (Stemler, 2004). Therefore, Cohen's κ indicates whether the agreement among raters is better than chance would predict. Stemler (2004) explained:

> The interpretation of the kappa statistic is slightly different than the interpretation of the percent-agreement figure. A value of zero on kappa does not indicate that the two judges did not agree at all; rather, it indicates that the two judges did not agree with each other any more than would be predicted by chance alone. Consequently, it is possible to have negative values of kappa if judges agree less often than chance would predict.

Furthermore, this statistic offers the advantage of an index that allows researchers to easily interpret results. Landis and Koch (1977, p. 65) assigned labels (see Figure 5) to corresponding ranges of Cohen's κ. Statistical support documentation points to this as the definitive index for κ (SAS, 2006). As a final advantage, "kappa is a highly useful statistic when one is concerned that the percent-agreement statistic may be artificially inflated due to the fact that most observations fall into a single category" (Stemler, 2004).

DOI: 10.1002/asi

There are two limitations of Cohen's κ statistic. First, "kappa values for different items or from different studies cannot be meaningfully compared unless the base rates are identical" (Stemler, 2004). Therefore, it is difficult to compare κ statistics over different assessment situations; however, this is not a disadvantage that is significant in this study. Second, κ requires a greater number of observances to achieve an acceptable standard error. This requirement is not significant for this study because it includes a sufficient number of student responses.

There are several advantages to using consensus estimates. For instance, consensus estimates are well suited to working with "nominal variables whose levels on the rating scale represent qualitatively different categories" (Stemler, 2004). Consensus estimates also can help determine how judges might misinterpret how to apply a rubric. Stemler (2004) stated that "A visual analysis of the output allows the researcher to go back to the data and clarify the discrepancy or retain the judges." Another advantage of consensus estimates is that they identify raters who have been trained enough to agree on how to interpret a rating scale. When that occurs, two raters may be treated as equivalent, and both raters need not score all student products or performances. Stemler (2004) confirmed that:

> When judges exhibit a high level of consensus, it implies that both judges are essentially providing the same information. One implication of a high consensus estimate of interrater reliability is that both judges need not score all remaining items . . . because the two judges have empirically demonstrated that they share a similar meaning for the scoring rubric. In practice, however, it is usually a good idea to build in a 30% overlap between judges even after they have been trained in order to provide evidence that the judges are not drifting from their consensus as they read more items.

When raters are trained to a level of agreement, summary scores can be figured by taking the score of one rater or averaging the scores given by all raters (Stemler, 2004). Although this advantage is not explored in this study, it has practical implications for future applications of rubrics.

Purpose of the Study

Because rubrics offer numerous potential benefits to librarians and faculty seeking to assess information literacy skills, a number of information literacy rubrics have appeared in the library and information science literature. The following authors recorded the use of rubrics to assess information literacy in higher education: D'Angelo (2001), Merz and Mark (2002), Rockman (2002), Emmons and Martin (2002), Buchanan (2003), Choinski, Mark, and Murphey (2003), Franks (2003), Gauss and Kinkema (2003), Hutchins (2003), Kivel (2003), Kobritz (2003), Warmkessel (2003), Smalley (2003), and Knight (2006). While these authors reported the use of information literacy rubrics, none have adequately examined the methods for training raters nor explored interrater reliability. As a result, an investigation of the use of rubrics by university librarians, faculty, and students is merited. This study investigates the viability of a rubric approach to information literacy assessment and details a methodology for both using rubrics and analyzing interrater reliability. It addresses the following research question: *To what degree can different groups of raters (librarians, English faculty, students) provide consistent scoring of artifacts of student learning using a rubric?* This central research question can be divided into two smaller areas of investigation:

- Can raters provide scores that are consistent with others in their rater group?
- Can raters provide scores that are consistent across groups?

Background

At North Carolina State University (NCSU), first-year students complete an online information literacy tutorial called "Library Online Basic Orientation" (LOBO; www.lib.ncsu.edu/lobo2) during English 101, a required writing course. As students progress through the tutorial, they are prompted to answer open-ended questions that reinforce or extend concepts taught in the tutorial. In the web evaluation section of the tutorial, students type the URL of a web site they have chosen as a possible resource for their research paper assignment. In subsequent screens, they respond to questions about the web site. Figure 1 depicts the prompt that focuses on web site authority. Student responses to the prompt are collected in a secure database within LOBO and offer a rich dataset for assessing the achievement of learning outcomes.

NCSU Libraries

This study focuses on student answers to the web site authority prompt. During the study period, more than 800 students responded to this open-ended question. To assess student responses, the researcher (also the NCSU instruction librarian and author) designed a full-model, analytic rubric (see Figure 2) based on the Association of College and Research Libraries (ACRL) standards to assess student ability to evaluate web sites for authority. The rubric included four criteria and three levels of performance. The criteria listed in the rubric were "Articulates Criteria," "Cites Indicators of Criteria," "Links Indicators to Examples from Source," and "Judges Whether or Not to Use Source." The rubric also described student behavior at three levels: Beginning, Developing, and Exemplary. The instruction librarian and 25 other raters used this rubric to score the 75 student responses to the study question.

Methodology

This study employed a survey design methodology. The data for the study came from student responses to open-ended questions embedded in the LOBO online tutorial. These textual data were translated into quantitative terms through the use of a rubric. Using a rubric, raters coded student answers into preset categories, and these categories were assigned point values. The point values assigned to student

DOI: 10.1002/asi

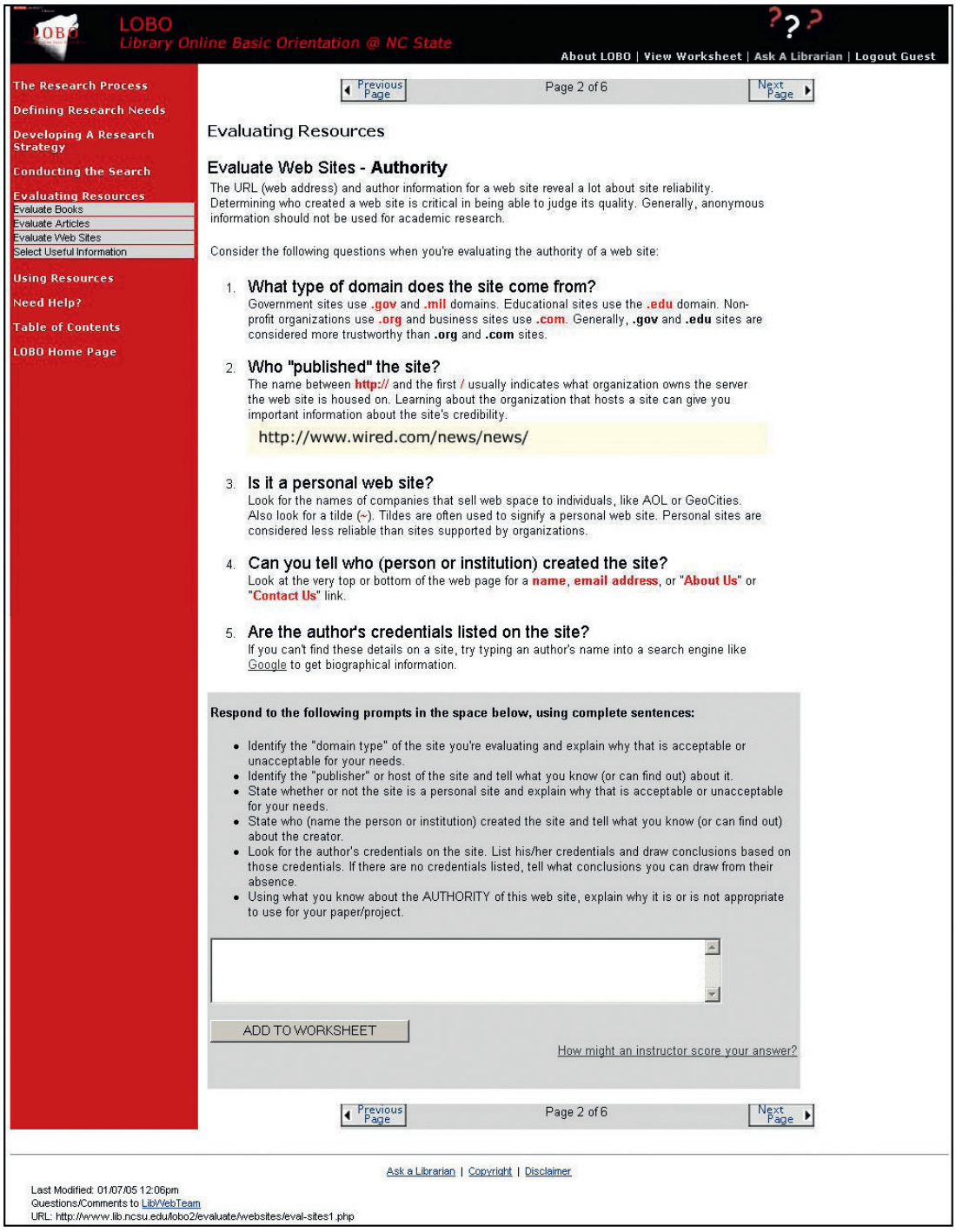

FIG. 1. LOBO tutorial prompt.

responses were subjected to quantitative analysis to test for interrater reliability. According to Y.S. Lincoln (personal communication, June 25, 2005), this approach is called "discovery phase" or preliminary experimental design, and it is commonly employed in the development of new rubrics.

Raters

Twenty-five raters participated in this study. The raters were evenly divided into five groups: NCSU librarians, NCSU English 101 instructors, NCSU English 101 students, instruction librarians from other Association of Research Libraries (ARL) libraries, and reference librarians who had limited instruction responsibilities from other ARL libraries.

NCSU librarians were selected as a rater group for this study because they are responsible for the instruction of information literacy skills at NCSU, and they are interested in ways to assess student learning. The individual NCSU librarians involved in this study represent a cross section

DOI: 10.1002/asi

Evaluation Criteria	Beginning	Developing	Exemplary
Articulates Criteria	0 - Student does not address authority issues. ◯	1 - Student addresses authority issues, but does not use criteria terminology. ◯	2 - Student addresses authority issues and uses criteria terminology such as: author, authority, authorship, or sponsorship. ◯
Cites Indicators of Criteria	0 - Student does not address authority indicators. ◯	1 - Student refers vaguely or broadly to authority indicators, but does not cite specific indicators. ◯	2 - Student cites specific authority indicators such as: domain, server/publisher/host, or ~ in URL; presence of personal or corporate author name, email, "About Us" or "Contact Us" links; or author credentials. ◯
Links Indicators to Examples from Source	0 - Student does not address examples of authority indicators from the site. ◯	1 - Student refers vaguely or broadly to examples of authority indicators from the site under consideration, but does not cite specific examples. ◯	2 - Student cites specific examples of authority indicators from the site under consideration. ◯
Judges Whether or Not To Use Source	0 - Student does not indicate whether or not the site is appropriate to use for the purpose at hand. ◯	1 - Student indicates whether or not the site is appropriate to use for the purpose at hand, but does not provide a rationale for that decision that cites authority issues and/or indicators. ◯	2 - Student indicates whether or not the site is appropriate to use for the purpose at hand and provides a rationale for that decision citing authority issues and/or indicators. ◯

RESEARCHER USE ONLY: Total Score ___/8

FIG. 2. Study rubric.

of reference and instruction librarians in subject specialties, instructional experience, assessment knowledge, gender, and race. English 101 instructors were selected as another rater group because they use the tutorial in their teaching. Instructors were selected to reflect the composition of English 101 faculty in teaching experience, race, and gender. Five students were selected as raters to ensure student input into the rubric assessment process. The students were enrolled in English 101 during the previous semester and were selected among many volunteers to represent the intended major, GPA, race, and gender make-up of incoming students. Finally, 10 librarians at other campuses were included in two additional rater groups to explore the performance of raters outside the NCSU campus context. All 10 librarians worked at five ARL library systems. Each of the five library systems was represented by one instruction librarian and one reference librarian.

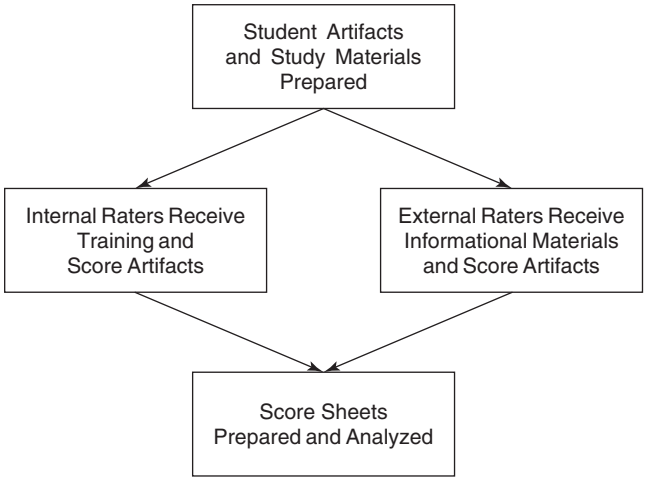

FIG. 3. Study procedure.

Study Procedure

The procedures followed in this study can be divided into four parts (see Figure 3). First, the researcher prepared the artifacts of student learning and study materials. Then, the three NCSU (i.e., internal) rater groups met in person for a 6-hr training and scoring session. Third, the two non-NCSU (i.e., external) rater groups received, scored, and returned their study materials. Finally, the score sheets were prepared for statistical analysis.

Preparation of Study Materials

Student responses to the LOBO tutorial writing prompt were prepared using a multistep process. First, the researcher retrieved all student responses from the LOBO answer database and separated them from personally identifying information. After null and unscorable responses were removed, the remaining 800 responses were numbered consecutively. Using a random number table, 75 student responses were selected for the study—an amount sufficient for delivering statistically significant results. Each of the 75 responses was placed on a score sheet. Score sheets included the student response, the scoring rubric, and three

DOI: 10.1002/asi

code numbers: the number of the response, the position of the response among the 75 to be scored, and the rater's number.

Next, the researcher scored each of the 75 responses three times using the study rubric. Afterwards, the researcher reviewed the scores and reconciled any divergent scores. After each student response was assigned a score, the researcher sorted the student responses into three large groups of 25 student responses to ensure an equal number of high, medium, and low scoring responses in each group of 25. Finally, within each group of 25 responses, individual responses were arranged in their original random order. This process resulted in three separate groups of student responses with an equal number of high, medium, and low scoring responses. The three separate groups were numbered 1–25, 26–50, and 51–75 and distributed to raters in this order. This preparation process enabled the researcher to later examine the reliability with which raters scored the first third, middle third, and final third of the student responses.

Fifteen additional student responses were retrieved from the LOBO database for a separate purpose. These 15 responses were selected as model or "anchor" papers to be used in the training session for internal raters. The 15 anchor responses were not chosen randomly. Rather, they were selected because they represented the wide range of student responses included in the study sample.

In this study, the preparation of materials for internal and external raters differed because the research design imitates the realities of assessment in academic libraries. In academic libraries, information literacy assessments are typically either created on campus where training is available, or they are imported from a separate institution and only written materials are available for consultation. In this study, the internal raters participated in a training session, a likely experience for librarians using a "home-grown" assessment tool. Materials prepared for the internal rater training included a meeting agenda, consent forms, poster-sized versions of the rubric, a Power Point presentation, copies of the ACRL *Information Literacy Competency Standards for Higher Education* (2000), copies of *LOBO Information Literacy Objectives and Outcomes* (Oakleaf, 2006, p. 384–388), screenshots of LOBO, and open-ended comment sheets to be completed by raters at the close of the scoring session.

External raters, like academic librarians attempting to use an assessment tool from another institution, were provided with a substantial amount of background material, directions, and examples. Materials prepared for the external rater mailing included several handouts: an inventory of materials, the context of the study, consent forms, directions for scoring the 75 study responses, open-ended comment sheets to be completed by raters at the close of their study participation, return-mail checklists, and postage-paid return-mail envelopes.

The Internal Rater Experience

The internal rater portion of this study was conducted in one 6-hr session during which the researcher met with 15 NCSU librarians, English 101 instructors, and students. As raters entered the training session, the researcher divided them into five small groups. Groups consisted of 1 librarian, 1 English 101 instructor, and 1 English 101 student to elicit diversity of opinion during the training session. The researcher began the session by explaining the purpose of the study, defining information literacy, and describing the need for tools to assess information literacy skills. Next, the researcher introduced rubrics by providing a definition, describing the component parts (criteria and performance levels), and providing brief examples. She also reviewed the relevant sections of LOBO, including the tutorial content and the open-ended student writing prompts. The researcher described the origins of the outcomes assessed by the study rubric and explained the relationship between the outcomes and the rubric criteria and performance levels. This part of the internal rater training took 45 minutes.

After a short break, the researcher followed a multistep process to familiarize the raters with the task of scoring student responses. This "norming" process was modeled on recommendations made by Maki (2004, p. 127). Maki referred to this process as "calibration" and described calibration as the process of "establishing interrater reliability in scoring students texts" (p. 126). She noted that calibration is "developed over successive applications of a scoring rubric to student work . . . to assure that rater responses are consistent" (p. 126). Maki outlined the six steps in this process:

1. Ask raters to independently score a set of student samples that reflects the range of texts students produce in response to a direct method.
2. Bring raters together to review their responses to identify patterns of consistent and inconsistent responses.
3. Discuss and then reconcile inconsistent responses.
4. Repeat the process of independent scoring on a new set of student samples.
5. Again, bring all scorers together to review their responses to identify patterns of consistent and inconsistent responses.
6. Discuss and then reconcile inconsistent responses. This process is repeated until raters reach consensus about applying the scoring rubric.

Ordinarily, two to three of these sessions calibrate raters' responses (p. 127).

In this study, the researcher added an initial step to the process. She began by sharing five "anchor" responses with the raters to demonstrate the range of student responses and then modeled the scoring process by "thinking aloud." Next, the raters independently scored five other anchor responses and discussed the scores they assigned in their small groups. In discussions, raters were asked to focus on inconsistent scores and attempt to reconcile them. Next, the small groups reported their scores to the full group, and the full group discussed the remaining inconsistencies and attempted to reconcile them. This part of the training lasted 75 minutes. After a second short break, raters independently scored five more anchor responses, discussed them in

DOI: 10.1002/asi

small groups, and finally worked as a full group to eliminate inconsistencies in scoring. This time, the process took about 60 minutes.

Raters felt at this point that they were ready to score student responses on their own, and they began to score the 75 study responses. They received three packets of 25 student responses; as they turned in each packet, they received a new one for scoring. Most raters required 45 to 75 minutes to score all responses; 1 rater took 95 minutes. After raters finished scoring study responses, they completed an open-ended comment sheet and left the scoring session.

The External Rater Experience

The 10 external (non-NCSU) raters did not participate in a training session. Instead, they were provided with study materials, background information, and directions via the mail. When raters opened their study packets, they encountered several documents. The first document inventoried the contents of the study packets. The second document included the purpose of the study, the major research questions, information explaining raters' role in the study, and directions for participating in the study. Raters were provided with the URL for LOBO and directions to login as a guest, screenshots from LOBO, the full and student versions of the rubric, and handouts including the ACRL *Information Literacy Competency Standards for Higher Education* and *LOBO Information Literacy Objectives and Outcomes*. The packet also included 75 LOBO student responses and an open-ended comment sheet for raters. Finally, raters were directed to place completed study materials in the postage-paid envelope and return them to the researcher.

Preparation for Statistical Analysis

At the close of the study, all raters returned their rubric score sheets and open-ended comment sheets. The open-ended comment sheets were transcribed for later use as a source for raters' perceptions and anecdotal comments. The rubric score sheets were organized for data entry and analysis. Data from the rubric score sheets were entered into an Excel spreadsheet. The number of the response, the position of the response among the 75 study responses, and the rater's number were included in the spreadsheet. For each response, each rater's score for the four criteria were recorded, along with the total score (0–8).

Statistical Analysis

The major purpose of this study was to determine to what degree different groups of raters can provide consistent scoring of artifacts of student learning using a rubric. The interrater reliability of rubric scores was examined both within groups and across groups using Cohen's κ. This statistic was calculated for each of the four criteria included in the rubric. It also was calculated for the total score assigned to the student response. Because of limitations of the κ statistic, total scores (0–8) were converted to letter grades (A, B, C, U) according to recommendations in the literature (see Figure 4) (Mertler, 2001). After this conversion, Cohen's κ was run on the "grade" assigned to the total student response.

Total rubric score	Letter grade
7–8	A
5–6	B
3–4	C
0–2	U

FIG. 4. Total scores and their associated letter grades.

κ	Strength of agreement
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

FIG. 5. Level of agreement indicated by κ score.

To illustrate the reliability within each group of raters, charts were generated that showed the κ for each rater group. To clarify the meaning of each κ statistic, the level of agreement indicated by the κ scores was labeled (see Figure 5) using the index provided by Landis and Koch (1977, p. 165). For example, rater groups that produced a κ of .41 to .60 are labeled "Moderate."

Results

Interrater Reliability Within Groups

To determine the reliability of rubric scores provided by raters, each rater group was examined separately. The groups defined by the study are NCSU librarians, English 101 instructors, English 101 students, external (non-NCSU) instruction librarians, and external (non-NCSU) reference librarians. The raters' scores also were examined in two large groups: internal (NCSU) raters and external (non-NCSU) raters.

The first rater group, the NCSU librarians, provided a moderate level of agreement when scoring student responses to the LOBO tutorial (see Figure 6). In three of the four criteria listed on the study rubric, NCSU librarians provided moderately reliable scores. NCSU librarians' ratings of the first rubric criterion, "Articulates Criteria," yielded a moderate κ of.54 ($SE = .03$). On the second rubric criterion, "Cites Indicators of Criteria," their scores produced a moderate κ of .54 as well ($SE = .03$). For the third rubric criterion, "Links Indicators to Example from Source," the ratings provided by NCSU librarians showed a fair κ of only .24 ($SE = .03$). Still, the librarians produced a moderate κ of .59 ($SE = .03$) for the final criterion, "Judges Whether or Not to Use Source." After the total rubric numerical scores were calculated and transformed into letter grades, a κ of .41 ($SE = .02$) shows that

DOI: 10.1002/asi

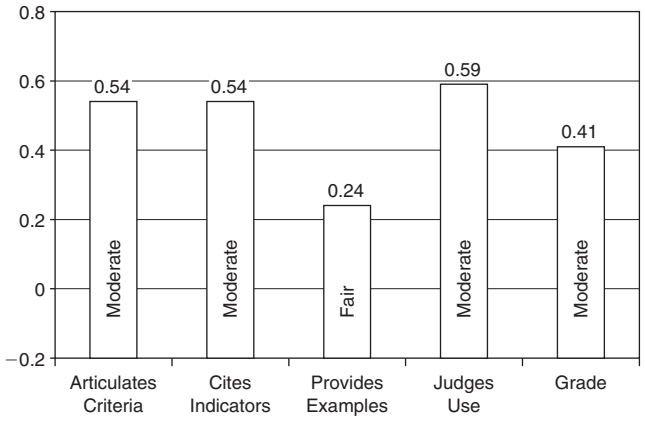

FIG. 6. NCSU librarian agreement levels.

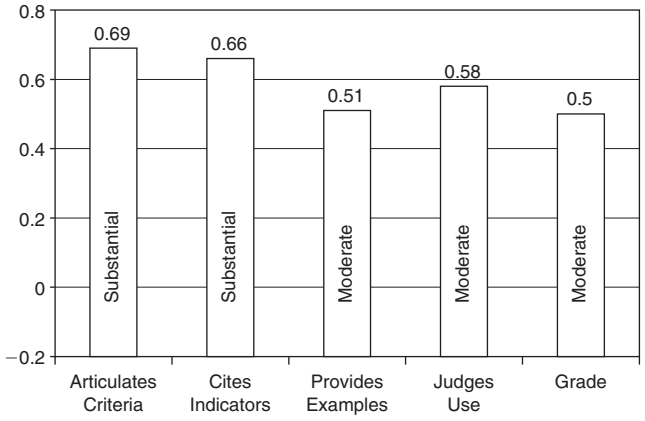

FIG. 7. English 101 instructor agreement levels.

NCSU librarians produced moderate agreement on the grade they assigned to student responses. These κ statistics indicate that within their group, NCSU librarians provided moderately reliable scores, but had difficulty coming to consensus on the third criterion of the rubric, "Links Indicators to Examples from Source."

As the second rater group, the English 101 instructors achieved moderately and substantially reliable results when scoring student responses to the LOBO tutorial (see Figure 7). For two of the criteria included in the study rubric, instructors produced substantially reliable scores. For the first criterion, "Articulates Criteria," instructors' ratings yielded a κ of .69 ($SE = .03$), and for the second criterion, "Cites Indicators of Criteria," they yielded a κ of .66 ($SE = .03$). Both κ scores correspond to a substantial level of agreement. For the third and fourth criteria on the study rubric, "Links Indicators to Examples from Source" and "Judges Whether or Not to Use Source," the instructors provided rankings with κs of .51 ($SE = .03$) and .58 ($SE = .03$), respectively, showing moderate levels of agreement. For the total grade assigned to student responses, English 101 instructors' ratings produced a κ of .50 ($SE = .03$), indicating a moderate level of agreement. These κs demonstrate that within their rater group, English 101 instructors were able to provide moderately to substantially reliable scores in all areas of the rubric and the grade assigned to student responses. In fact, English 101 instructors produced the greatest within-group reliability of all rater groups studied.

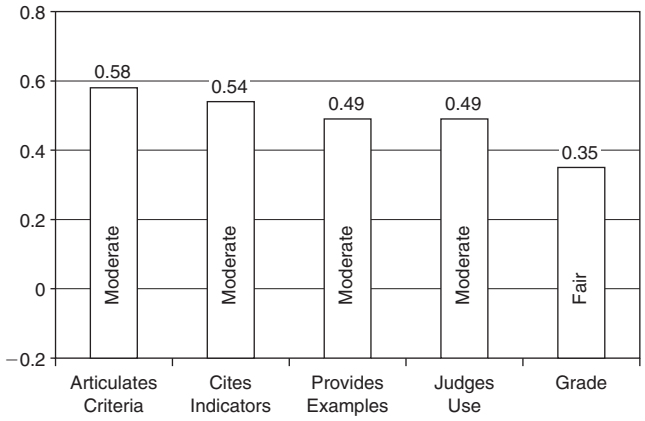

FIG. 8. English 101 student agreement levels.

The third rater group was comprised of English 101 students. This group produced a fair to moderate level of agreement with their scores of responses to the LOBO tutorial (see Figure 8). Across the four criteria included in the rubric, the students provided moderately reliable scores. For the first criterion, "Articulates Criteria," the students' scoring yielded a κ of .58 ($SE = .03$). For the second criterion, "Cites Indicators of Criteria," the κ for student scores is .54 ($SE = .03$). For both the third and fourth criteria, "Links Indicators to Examples from Source" and "Judges Whether or Not to Use Source," the students' ratings showed a κ of .49 ($SE = .03$). Across all the criteria in the rubric, these κ statistics indicate a moderate level of agreement; however, English 101 students' ratings of the total response, when converted to a letter score, indicate only a fair level of agreement, with a κ of .35 ($SE = .03$). These statistics indicate that within their rater group, English 101 students were able to achieve moderate levels of reliability for each criterion included in the rubric, but only achieved a fair level of agreement on the grade assigned to student responses.

The last two rater groups included external instruction and external reference librarians. These two groups provided scores that demonstrate slight and fair agreement (see Figures 9 and 10, respectively). While external instruction librarians achieved moderate agreement for the fourth criterion of the study rubric, "Judges Whether or Not to Use a Source," with a κ of .47 ($SE = .03$), their ratings for the other three criteria showed only slight agreement. κs for the first three criteria were .12, .18, and.19 ($SE = .03$), respectively. For the grade assigned to student responses, external instruction librarians demonstrated a fair level of agreement, with a κ of .23 ($SE = .02$). Overall, the levels of agreement produced by external instruction librarians were lower than would be acceptable.

External reference librarians also produced slight and fair levels of agreement. Like the external instruction librarians, external reference librarians came to moderate agreement on the fourth criterion of the rubric, "Judges Whether or Not

DOI: 10.1002/asi

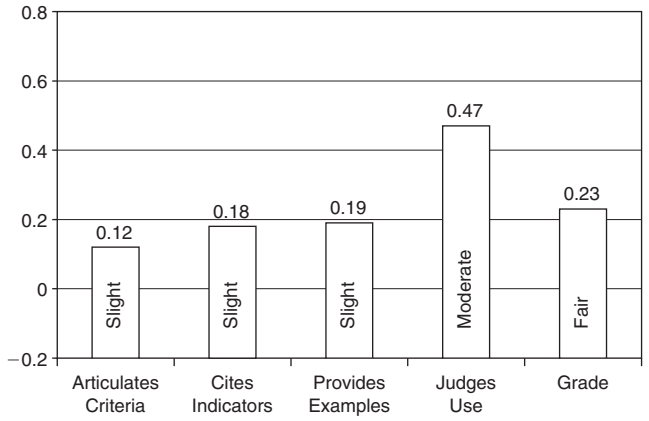

FIG. 9. External instruction librarian agreement levels.

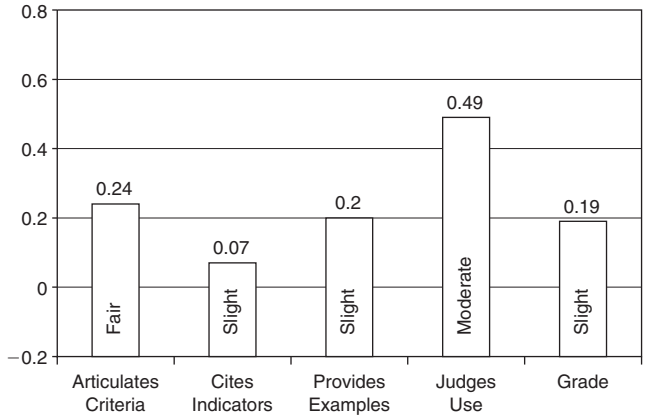

FIG. 10. External reference librarian agreement levels.

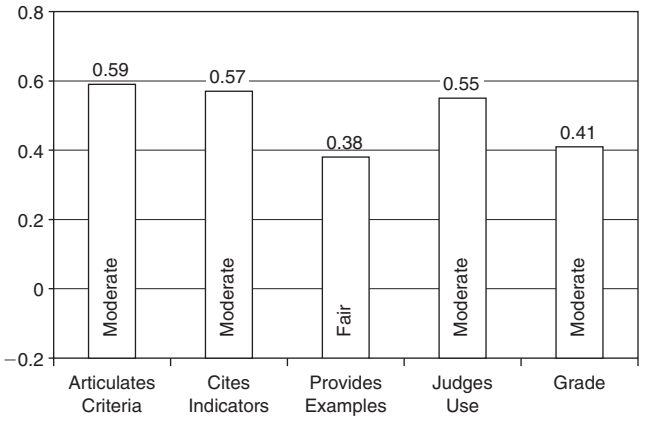

FIG. 11. Internal rater agreement levels.

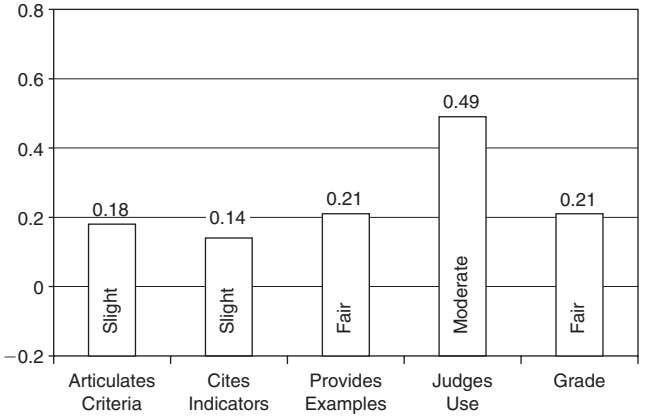

FIG. 12. External rater agreement levels.

to Use Source," with a κ of .49 ($SE = .03$). For the second and third rubric criteria, "Cites Indicators of Criteria" and "Links Indicators of Criteria," external reference librarians produced slight levels of agreement, with κs of .07 and .20 ($SE = .03$), respectively. For the grade assigned to the overall student responses, a κ of .19 ($SE = .02$) shows that external reference librarians' ratings demonstrated only a slight level of agreement. These κ statistics show that within their two groups, external reference and instruction librarians were unable to achieve greater than slight to fair levels of agreement in all areas except the fourth criterion of the rubric, "Judges Whether or Not to Use Source." For this criterion alone, both groups of external librarians achieved a moderate level of agreement.

The scores assigned to student responses by the 25 raters in this study can be grouped into two larger categories: internal raters and external raters. κs for these two larger categories (see Figures 11 and 12). Overall, internal raters yielded moderate levels of agreement ranging from .55 to .59 ($SE = .01$) for the first, second, and fourth rubric criteria. For the third rubric criteria, the internal raters' scores produced a κ of .38, showing a fair level of agreement on this criterion. Internal raters produced a moderate level of agreement for the grades assigned to student responses. The κ for this measure was .41 ($SE = .01$). These levels demonstrate a generally moderate level of agreement within all internal raters. The third rubric criterion, "Links Indicators to Examples from Source," produced only a fair κ statistic.

In contrast, external raters produced only slight to fair levels of agreement. For the first and second criteria included in the rubric, external raters' scores showed only slight agreement, with κs of .18 and .14 ($SE = .01$), respectively. The third rubric criterion, "Links Indicators to Examples from Source," shows a κ of .21 ($SE = .01$), which indicates a fair level of agreement. The fourth criterion shows moderate agreement, with a κ of .49 ($SE = .01$), but the grade assigned by external raters indicates a fair level of agreement, with a κ of .21 ($SE = .01$).

Summary of Interrater Reliability Within-Groups Results

This study sought to answer two questions regarding the reliability of a rubric approach to the assessment of information literacy skills. These are the answers to the first major research question posed by the study: Can raters provide scores that are consistent with others in their rater group? The answer to this question varies by rater group:

- Within their rater group, NCSU librarians provided moderately reliable scores for the first, second, and fourth criteria

DOI: 10.1002/asi

on the rubric and the total grade assigned to the student responses, but they had difficulty coming to consensus on the third criterion of the rubric, "Links Indicators to Examples from Source."

- Within their rater group, English 101 instructors were able to produce moderately to substantially reliable scores in all areas of the rubric and in the total grade assigned to student responses. English 101 instructors had the greatest within-group reliability of all rater groups studied.
- Within their rater group, English 101 students were able to achieve moderate levels of reliability for each criterion included in the rubric, but only achieved a fair level of agreement on the total grade assigned to student responses.
- Within their rater group, internal raters demonstrated a generally moderate level of agreement. Only the third rubric criterion, "Links Indicators to Examples from Source," produced a fair κ statistic.
- Within their rater groups, external instruction and reference librarians were unable to achieve greater than slight to fair levels of agreement in the first, second, and third criteria of the rubric. The exception was the fourth criterion of the rubric, "Judges Whether or Not to Use Source." For this criterion, both groups of external librarians achieved a moderate level of agreement. The external instruction librarians demonstrated a fair level of agreement on the total grade assigned to student responses. The external reference librarians demonstrated slight agreement in the same area.

In summary, the internal rater groups (NCSU librarians, English 101 instructors, and English 101 students) provided moderately consistent scores with others in their rater groups. In contrast, external librarians could not achieve acceptable levels of agreement.

Significant Differences Among Rater Groups

A number of statistically significant differences in reliability were revealed by analyzing the κ statistics of rater groups using two-sided t tests with an α level of .05. At the 95% confidence level, a t statistic over 1.96 is deemed significant. Statistically significant differences were found when comparing NCSU librarians with instructors, instructors with students, NCSU librarians with students, external reference librarians with external instruction librarians, NCSU librarians with external librarians, and all internal raters with all external raters.

Comparing the reliabilities of NCSU librarians and English 101 instructors revealed four significant differences between these two rater groups. The first statistically significant difference involved the first three criteria on the study rubric. For the first criterion, "Articulates Criteria," the instructors' substantial κ level of .69 was significantly greater than the NCSU librarians' moderate κ of .54 ($t = 3.5$), indicating that the English 101 instructors produced scores showing a greater degree of reliability for this rubric criterion. The second criterion of the rubric, "Cites Indicators of Criteria," was scored with significantly greater reliability by instructors with a substantial κ of .66 than the NCSU librarians with a moderate κ of .54 ($t = 2.8$), and the third criterion, "Links Indicators to Examples from Source," also was scored with significantly greater reliability by instructors with a moderate κ of .51 than the librarians with a fair κ of .24 ($t = 6.4$). Finally, the reliability of the grades assigned by instructors with a moderate κ of .50 was significantly greater than κs produced by librarians, a .41 ($t = 2.5$). There was no statistically significant difference between the reliability of the scores produced by NCSU librarians and English 101 instructors for the fourth criterion of the study rubric, "Judges Whether or Not to Use Source." Taken as a whole, these significant differences indicate that the English 101 instructors produced more reliable scores of student responses than did NCSU librarians.

There were four statistically significant differences in the reliabilities of the scores produced by English 101 instructors and those produced by English 101 students. In three areas of the study rubric, the scores supplied by instructors had significantly greater reliability than those of the students. The first criterion of the rubric, "Articulates Criteria," was more reliably scored by instructors with a substantial κ of .69 than by students with a moderate κ of .58 ($t = 2.6$). The second criterion also revealed a statistically significant difference. For "Cites Indicators of Criteria," instructors produced more reliable results, showing a substantial κ of .66, than the students ($t = 2.8$). For this criterion, the students' ratings yielded only a moderate κ of .54. Instructors also were shown to produce more reliable scores for the fourth criterion, "Judges Whether or Not to Use Source," as well. This was indicated by a statistically significant difference ($t = 2.1$) between the instructor's κ for this criterion at .58 and the students' κ at .49. The fourth statistically significant difference ($t = 3.5$) appeared when the moderate reliability of the grades assigned by the English 101 instructors ($k = .50$) was compared to the fair reliability of the grades assigned by the students ($k = .35$). There was no statistically significant difference between the reliability of the scores produced by English 101 instructors and English 101 students for the third criterion of the rubric, "Links Indicators to Examples from Source." Overall, these significant differences signify that English 101 instructors scored student responses more reliably than did English 101 students.

The reliability of the scores assigned by NCSU librarians and English 101 students differed in only two statistically significant ways. For the third criterion of the study rubric, "Links Indicators to Examples from Source," the moderate κ statistic of .49 provided by the students' ratings was significantly greater than the fair κ for the librarians' ratings, which was only .24 ($t = 5.9$). While both librarians and students showed moderately reliable scoring for the fourth rubric criterion, "Judges Whether or Not to Use Source," the librarians' scores ($k = .59$) were significantly greater than the student's scores ($k = .49$). The t statistic for this difference was 2.1. There was no significant difference between the reliability of librarians' scores and the reliability of the students' scores for the first and second rubric criteria and the grade assigned to student responses. The lack of significant differences between the NCSU librarians and English 101 students indicated that neither group produced more reliable scores than the other. The fact that the two significant differences that appeared in the data were split (i.e., one shows

DOI: 10.1002/asi

the librarians with greater reliability, the other shows the students with more reliable results) underscores the lack of substantial differences between the scores of these two groups.

There were two statistically significant differences between the reliability of scores provided by external instruction librarians and by external reference librarians. For the first criterion of the study rubric, external reference librarians showed significantly greater reliability ($t = 2.8$), although the reliability of the external reference librarians on this criterion was still only fair. For the second criterion of the rubric, external instruction librarians demonstrated significantly greater reliability ($t = 2.59$), but the reliability of both groups showed only a slight agreement. There were no significant differences between the reliability of the scores provided by external instruction librarians and by external reference librarians for the third and fourth rubric criteria and the final grades assigned to the student responses. Because only two significant differences were identified between these two groups and these significant differences indicated greater reliability in opposing directions (i.e., one showing the greater reliability of external reference librarians and the other showing the greater reliability of external instruction librarians), these two groups did not appear to provide substantially different scores for student responses.

The reliability of the scores assigned by NCSU librarians differed significantly from the reliability of scores provided by external librarians in four ways. First, the κ for the NCSU librarians' scores for the first criterion in the study rubric showed a moderate level of agreement ($k = .54$). The κ for external librarians was significantly lower at .18, showing only slight agreement ($t = 11.4$). Similarly, the κ for NCSU librarians' scores for the second rubric criterion was a moderate .54 while the external librarians showed only slight agreement with a κ of .14. These κ statistics were significantly different ($t = 12.6$). The t test for the fourth rubric criterion showed a statistically significant difference ($t = 3.2$) between the reliability of NCSU librarians' ratings ($k = .59$) and those of the external librarians ($k = .49$). The reliability of the grade assigned by the raters also was significantly different ($t = 6.4$), with the NCSU librarians showing moderate reliability with a κ of .41 and the external librarians demonstrating only fair reliability with a κ of .21. Although there was no significant difference between these two groups for the third criterion of the study rubric, NCSU librarians clearly produced more reliable scores of student responses than did their non-NCSU counterparts.

In the same way, the reliability of all ratings for internal raters was significantly greater than that of all ratings for external raters. For the first criterion, "Articulates Criteria," the internal raters demonstrated a moderate κ of .59 while the external raters provided only slight agreement ($t = 29.0$). For the second rubric criterion, "Cites Indicators of Criteria," the κ for internal raters was .57. This κ also is significantly different ($t = 30.4$) from the external raters ($k = .14$). Likewise, internal raters produced greater reliability for the third ($t = 12.0$) and fourth criteria ($t = 4.2$). Finally, internal raters yielded moderate levels of agreement ($k = .41$) when assigning grades to student response, a significant difference ($t = 14.1$) from the fair level of agreement provided by external raters ($k = .21$). The significantly greater level of agreement among internal raters over external raters was clearly demonstrated.

Summary of Significant Differences Among Rater Groups Results

This study sought to answer two questions regarding the reliability of a rubric approach to the assessment of information literacy skills. These are the answers to the second major research question posed by the study: Can raters provide scores that are consistent across groups?

Several statistically significant differences were discovered by comparing rater groups:

- In nearly all areas, English 101 instructors achieved higher levels of reliability than did NCSU librarians. The exception to this is the fourth criterion of the study rubric, "Judges Whether or Not to Use Source." In this area, there was no statistically significant difference between the reliability of scores assigned by the English 101 instructors and those by the NCSU librarians.
- In nearly all areas, the English 101 instructors produced higher levels of reliability than did the English 101 students. However, for the third rubric criterion, "Links Indicators to Examples from Source," there was no statistically significant difference in the reliability of their scores.
- In most areas, there was no statistically significant difference between the reliabilities of NCSU librarians and those of the English 101 students. However, English 101 students showed higher reliability for the third criterion of the rubric, "Links Indicators to Examples from Source." On the other hand, NCSU librarians achieved higher levels of reliability for the fourth criterion on the rubric, "Judges Whether or Not to Use Source."
- In all areas of the rubric and in the total grade assigned to student responses, internal raters demonstrated higher levels of reliability than did external raters.
- In most areas, there was no statistically significant difference between the reliabilities of external instruction librarians and those of external reference librarians.
- In nearly all areas, NCSU librarians achieved higher levels of reliability than did external librarians. The one exception to this is the third rubric criterion, "Links Indicators to Examples from Source." In this area, there was no statistically significant difference between the reliability of scores assigned by the NCSU librarians and those by the external librarians.

In summary, English 101 instructors produced significantly higher reliabilities than did NCSU librarians and English 101 students. In fact, few significant differences were discovered between the levels of agreement exhibited by the NCSU librarians and English 101 students. However, NCSU librarians produced much higher levels of agreement than did external instruction librarians and external reference librarians. Overall, internal raters produced significantly higher levels of agreement than did external raters.

DOI: 10.1002/asi

Intrarater Reliability Differences Throughout the Scoring Process

In this study, the intrarater reliability of all raters increased as they scored the 75 student responses. When comparing the reliabilities of the first third, middle third, and last third of student responses using a Bonferroni adjustment, a t test over 2.5 indicates a significant difference. In this study, the scores raters assigned to Responses 26 to 50 were more reliable than the scores they assigned for Responses 1 to 25 ($t = 3.03$). Additionally, the scores assigned to Responses 51 to 75 were more reliable than those assigned to Responses 26 to 50 ($t = 3.03$). Note that because the groups of responses compared in this test were scored by the same raters, these t-test scores are statistically conservative. As a result, a greater difference might actually exist than what these t scores indicate. Because of *SE* increases when only 25 responses are examined (rather than the full 75), analysis of the smaller rater groups was not statistically feasible.

Discussion

In this study, different rater groups arrived at varying levels of agreement within their groups. For example, the *English 101 instructors achieved the greatest levels of agreement* within a group, levels that were significantly higher than those of any other five original rater groups. The English 101 instructors were the only one of five original rater groups to attain moderate or substantial levels of agreement across all areas of the study rubric and the grade assigned to each of the student responses. Although definitive reasons for this group's success must await future research, it seems likely that the English 101 instructors' familiarity with rubrics used to assess writing may have increased their ability to come to agreement using rubrics to assess information literacy. It also is possible that the English 101 instructors, through educational background or personal experience, were familiar with outcomes-based assessment. As teachers, they also are likely to value the ability to produce consistent scores for complex artifacts of student learning.

English 101 students also produced moderate levels of agreement across all areas of the research rubric. This level of consistency might be attributed to previous experiences with rubrics. Although it is probable that the study rubric was the students' first experience with a rubric designed to assess information literacy skills, they may have transferred skills acquired from earlier rubric experiences in other subject areas to their activities in this study.

As a group, *NCSU librarians produced moderately consistent scores* in most areas of the assessment. Interestingly, there was little significant difference between the levels of reliability achieved by NCSU librarians and by English 101 students; however, in two areas, students and NCSU librarians differed significantly. For the fourth criterion of the rubric, "Judges Whether or Not to Use the Source," librarians produced greater levels of agreement than did the students. For the third criterion of the rubric, "Links Indicators to Examples from Source," NCSU librarians achieved less consistency than did the students. The reason that NCSU librarians came to only a fair level of agreement on this criterion is unclear and should be investigated in future research. Aside from this weak area, the NCSU librarians demonstrated far greater reliabilities within their rater group than did the external instruction librarians and the external reference librarians.

External librarians (both instruction librarians and reference librarians) *could not produce consistent scoring* of student responses in this study, and there was no statistically significant difference between the overall performance of external reference librarians and external instruction librarians. These raters' professional experiences as librarians did not appear to be as important as the fact that they were external to the assessment environment and, as a result, did not receive training. Taken as a group, external librarians were unable to achieve more than fair levels of agreement on all areas of the assessment, with the exception of the fourth rubric criterion, "Judges Whether or Not to Use Source" (For this fourth criterion, external librarians achieved moderate levels of agreement, but this does not indicate a particular level of expertise in this area. All rater groups were able to achieve a moderate level of agreement for this criterion.) Overall, the external librarians achieved significantly lower levels of reliability than did NCSU librarians and internal raters as a whole.

Conclusion

Although instructors and students exhibit a level of proficiency in the use of rubrics, this study demonstrates that librarians may be less proficient; however, this study indicates that internal librarians can be trained to become moderately consistent raters. It may be that external librarians, provided with additional training, could become consistent raters as well. Additional rubric training for external raters should teach (a) major concepts underlying outcomes-based assessment, (b) differences between analytic and holistic approaches to assessment, (c) strategies for comprehending rubric content, (d) techniques for reconciling one's personal beliefs with rubric assumptions, (e) methods for tolerating difficulties in student learning artifacts, and (f) ways to understand library context and campus culture (Oakleaf, 2007, p. 38).

This study also demonstrates that Maki's (2004, p. 126) model of calibration is a useful training tool for preparing librarians to become proficient rubric raters. In addition, the study highlights two suggestions for using the six-step Maki calibration model. First, the researcher in this study modeled rubric use by employing "think aloud" techniques. The researcher began the rubric training session by talking through the application of the study rubric to assess five anchor responses. This prompted discussion among raters and brought norming issues out early in the calibration process. This initial step in the rater calibration process appeared to expedite rater readiness. Second, Maki's model calls for two to three rounds of rater practice scoring and discussion (p. 127). In this study, two rounds were used before raters felt

DOI: 10.1002/asi

confident about independently scoring student responses. In hindsight, it is possible that a third or even fourth round may have been advisable. It also may be that multiple calibration sessions should be required for librarians who have spent little or no time using rubrics in the past.

Although this study indicates that librarians require training to consistently and accurately use rubrics, the benefits associated with rubric assessment far outweigh the time spent in training. Indeed, 96% of the raters stated that they believe rubrics have great instructional value. All internal raters stated that they could envision using rubrics to improve information literacy instruction, and all but 1 external rater agreed (Oakleaf, 2006, p. 377).

Recommendations for Future Research

Because this study is the first of its kind in the area of information literacy instruction, the findings described in this study await testing and confirmation by future researchers. Besides replicating this research in other environments, several areas of research should be explored. First, future research could determine what characteristics made English 101 instructors the most successful raters in this study. Characteristics might include previous experience with rubrics, familiarity with outcomes-based assessment, or tolerance of minor errors in student work. Future studies could examine these characteristics by investigating other faculty populations, including library and information science educators. In addition, future researchers could explore the effects of different uses of rater training. For instance, the study design could be altered to compare the reliability of scores provided by internal librarians who participated in training and those who did not. In a similar vein, future research could investigate the effects of different types and levels of rater training in external librarian populations. Finally, future investigations could include evaluations of a wide variety of performance assessments, including student bibliographies, research journals, and portfolios. All these areas of additional research will help build a strong foundation for future uses of information literacy assessment rubrics.

References

Andrade, H.G. (2000). Using rubrics to promote thinking and learning. Educational Leadership, 57(5), 13–18.

Arter, J., & McTighe, J. (2000). Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance. Thousand Oaks, CA: Corwin Press.

Association of College and Research Libraries. (2000). Information literacy competency standards for higher education. Retrieved January 1, 2009, from http://www.ala.org/ala/mgrps/div/acr/standards/informationliteracycompetency.cfm

Bernier, R. (2004). Making yourself indispensible by helping teachers create rubrics. CSLA Journal, 27(2), 24–25.

Bresciani, M.J., Zelna, C.L., & Anderson, J.A. (2004). Assessing student learning and development: A handbook for practitioners. National Association of Student Personnel Administrators.

Buchanan, L.E. (2003). Assessing liberal arts classes. In E.F. Avery (Ed.), Assessing student learning outcomes for information literacy instruction in academic libraries (pp. 68–73). Chicago: Association of College and Research Libraries.

Callison, D. (2000). Rubrics. School Library Media Activities Monthly, 17(2), 34–36, 42.

Choinski, E., Mark, A.E., & Murphey, M. (2003). Assessment with rubrics: An efficient and objective means of assessing student outcomes in an information resources class. Portal: Libraries and the Academy, 3(4), 563–575.

Colton, D.A. (1997). Reliability issues with performance assessments: A collection of papers. Iowa City, IA: ACT.

D'Angelo, B.J. (2001). Integrating and assessing information competencies in a gateway course. Reference Services Review, 29(4), 282–293.

Emmons, M., & Martin, W. (2002). Engaging conversation: Evaluating the contribution of library instruction to the quality of student research. College and Research Libraries, 63(6), 545–560.

Franks, D. (2003). Using rubrics to assess information literacy attainment in a community college education class. In E.F. Avery (Ed.), Assessing student learning outcomes for information literacy instruction in academic libraries (pp. 132–147). Chicago: Association of College and Research Libraries.

Gauss, N., & Kinkema, K. (2003). Webliography assignment for a lifetime wellness class. In E.F. Avery (Ed.), Assessing student learning outcomes for information literacy instruction in academic libraries (pp. 161–171). Chicago: Association of College and Research Libraries.

Hafner, J.C. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. International Journal of Science Education, 25(12), 1509–1528.

Hutchins, E.O. (2003). Assessing student learning outcomes in political science classes. In E.F. Avery (Ed.), Assessing student learning outcomes for information literacy instruction in academic libraries (pp. 172–184). Chicago: Association of College and Research Libraries.

Kivel, A. (2003). Institutionalizing a graduation requirement. In E.F. Avery (Ed.), Assessing student learning outcomes for information literacy instruction in academic libraries (pp. 192–200). Chicago: Association of College and Research Libraries.

Knight, L.A. (2006). Using rubrics to assess information literacy. Reference Services Review, 34(1), 43–55.

Kobritz, B. (2003). Information literacy in community college communications courses. In E.F. Avery (Ed.), Assessing student learning outcomes for information literacy instruction in academic libraries (pp. 207–215). Chicago: Association of College and Research Libraries.

Landis, J.R., & Koch, G.G. (1977). The measure of observer agreement for categorical data. Biometrics, 33, 159–174.

Maki, P.L. (2004). Assessing for learning: Building a sustainable commitment across the institution. Sterling, VA: Stylus.

Mertler, C.A. (2001). Designing scoring rubrics for your classroom. Practical Assessment Research and Evaluation, 7(25). Retrieved January 1, 2009, from http://pareonline.net/getvn.asp?v=7&n=25

Merz, L.H., & Mark, B.L. (2002). Clip note #32: Assessment in college library instruction programs. Chicago: Association of College and Research Libraries.

Moskal, B.M. (2000). Scoring rubrics: What, when, and how? Practical Assessment Research and Evaluation, 7(3). Retrieved January 1, 2009, from http:// pareonline.net/getvn.asp?v=7&n=3

Moskal, B.M., & Leydens, J.A. (2000). Scoring rubric development: Validity and reliability. Practical Assessment Research and Evaluation, 7(10). Retrieved January 1, 2009, from http://pareonline.net/getvn.asp?v=7&n=10

Nitko, A.J. (2004). Educational assessment of students. Upper Saddle River, NJ: Pearson Education.

Oakleaf, M.J. (2006). Assessing information literacy instruction: A rubric approach (Doctoral dissertation, University of North Carolina at Chapel Hill, 2006). Dissertation Abstracts International, Proquest No. 1095444541.

Oakleaf, M.J. (2007). Using rubrics to collect evidence for decision-making: What do librarians need to know? Evidence Based Library and Information Practice, 2(3), 27–42.

Oakleaf, M.J. (2008). Dangers and opportunities: A conceptual map of information literacy assessment approaches. Portal: Libraries and the Academy, 8(3), 233–253.

DOI: 10.1002/asi

Oakleaf, M.J. (2009). The information literacy instruction assessment cycle: A conceptual framework. Journal of Documentation, 65(4).

Pausch, L.M., & Popp, M.P. (1997, April). Assessment of information literacy: Lessons from the higher education assessment movement. Paper presented at the meeting of the Association of College and Research Libraries, Nashville, TN.

Popham, W.J. (2003). Test better, teach better: The instructional role of assessment. Alexandria, VA: Association for Supervision and Curriculum Development.

Prus, J., & Johnson, R. (1994). A critical review of student assessment options. New Directions for Community Colleges, 88, 69–83.

Rockman, I.F. (2002). Rubrics for assessing information competence in the California State University. Retrieved April 10, 2005, from http://www.calstate.edu/LS/1_rubric.doc

SAS. (2006). Compute estimates and tests of agreement among multiple raters. Retrieved September 1, 2008, from http://support.sas.com/kb/25/006.html

Smalley, T.N. (2003). Bay Area Community Colleges information competency assessment project. Retrieved November 17, 2003, from http://www.topsy.org/ICAP/ICAProject.html

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research, and Evaluation, 9(4). Retrieved January 1, 2009, from http://pareonline. net/getvn.asp?v=9&n=4

Stevens, D.D., & Levi, A. (2005). Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning. Sterling, VA: Stylus.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scales levels. Practical Assessment Research and Evaluation, 9(2). Retrieved January 1, 2009, from http://pareonline.net/getvn.asp?v=9&n=2

Warmkessel, M.M. (2003). Assessing abilities of freshmen to reconcile new knowledge with prior knowledge. In E.F. Avery (Ed.), Assessing student learning outcomes for information literacy instruction in academic libraries (pp. 249–256). Chicago: Association of College and Research Libraries.

Wiggins, G. (1996). Creating tests worth taking. In R.E. Blum & J. A. Arter (Eds.), A handbook for student performance in an era of restructuring (pp. V-6:1–V-6:9). Alexandria, VA: Association for Supervision and Curriculum Development.

DOI: 10.1002/asi