Potholes and Pitfalls on the Road to Authentic Assessment

Sam Lohmann, Karen R. Diller, and Sue F. Phelps

**abstract:** This case study discusses an assessment project in which a rubric was used to evaluate information literacy (IL) skills as reflected in undergraduate students’ research papers. Subsequent analysis sought relationships between the students’ IL skills and their contact with the library through various channels. The project proved far longer and more complex than expected and yielded inconclusive results. We reflect on what went wrong and highlight lessons learned in the process. Special attention is paid to issues of project management and statistical analysis, which proved crucial stumbling blocks in the effort to conduct a meaningful authentic assessment.

**Introduction**

In 2013, a team of librarians at Washington State University (WSU) Vancouver began an authentic assessment project, a form of assessment that looks for evidence of knowledge and skills in the performance of meaningful real-world tasks. The project was intended to determine whether students’ contact with the library had an impact on their information literacy (IL) skills as reflected in their writing. What began as a seemingly simple one-semester pilot project grew into a multiyear study. Ultimately, due to a series of unforeseen issues, it became a long, time-consuming study with few conclusive or actionable results. In the end, the project was most interesting for what it taught the researchers about the complexity and importance of research design and project management. This article attempts to provide an account of these issues, which include norming and interrater reliability, planning, and iterative design.

To tell this story, we will depart somewhat from the conventions of a primary research article, which typically focuses on reporting and interpreting findings after concisely outlining the research methods. Instead, we will focus on our methods and process, concluding with the lessons we learned from this largely unsuccessful project.
Our statistical results are included in Appendix B, since they are ancillary to the article but may help the reader to understand the study. Though suggestive and generally encouraging, the assessment results are constrained by methodological issues. However, a detailed account of the research process itself not only offers an opportunity for reflection but also may save future researchers time and frustration.

**Background**

**Authentic Assessment**

Over the past decade, librarians at many institutions have sought new methods for assessing the impact and value of library programs, often moving away from traditional usage indicators such as circulation statistics toward quantitative and qualitative measures that attempt to assess the library’s impact on the mission and goals of its parent institution. The need to measure the impact of library instruction has become especially urgent. Because many libraries provide IL instruction, librarians have sought ways to assess the effectiveness of that teaching, usually by assessing students’ IL skills in some way. Since their publication in 2000, the Association of College and Research Libraries (ACRL) Information Literacy Competency Standards for Higher Education, the Standards, were the most influential model for articulating and measuring IL learning outcomes in the United States, at least until the adoption of the ACRL Framework for Information Literacy for Higher Education in 2016. Starting from this common reference point, researchers have used a wide variety of methods to assess IL. These techniques can be broadly divided between indirect methods, which use an instrument such as a standardized test or survey to evaluate IL as an isolated skill set, and direct or authentic assessment methods, which seek evidence of students’ learning within the context of their regular coursework.

Kathleen Montgomery uses the term *authentic assessment* to describe practices that include “the holistic performance of meaningful, complex tasks in challenging environments that involve contextualized problems.” Such practices have given strong support to the use of rubrics for assessment of IL and other skills, as advocated by Montgomery, by the team of Karen Diller and Sue Phelps, and by Megan Oakleaf. Articles published after the completion of data collection for the present study have extended and affirmed the use of rubrics for authentic IL assessment.

**Assessment at WSU Vancouver Library**

WSU Vancouver is an urban campus of Washington state’s only land-grant university and a large Research I institution, which offers a full range of programs and engages in extensive research activity. Its original campus is in Pullman, Washington. WSU Vancouver, across the Columbia River from Portland, Oregon, serves about 3,500 students annually, the majority of them transfer students. IL instruction is a major part of the library’s mission, and librarians frequently engage in classroom teaching. This teaching typically takes the form of one- or two-hour sessions tailored to the instructors’ requests, often combining a basic but essential procedural demonstration with more interactive, critical engagement in IL issues. In addition to one-shot bibliographic instruction ses-
sions, librarians teach a one-credit online IL elective, Accessing Information for Research, each semester. In addition to library instruction, reference interactions, content on the library website, and instruction from nonlibrary faculty may also contribute to students’ IL learning.

WSU Vancouver’s General Education Learning Goals include an IL goal based on the ACRL Standards. Part of the library’s mission is to support this goal. To determine whether the library did this effectively and how it might better support the goal, instruction librarians met and developed an assessment plan. After reviewing previous studies using both indirect and authentic assessment methods, the librarians determined that the most relevant, actionable results would be obtained through authentic IL assessment methods, specifically the use of a rubric to score student research papers.

After securing university funds to help cover the anticipated work hours and receiving an Institutional Review Board waiver, we moved forward with the project.

Our chosen method had the advantage of being flexible, applicable to a wide range of skill levels and assignments, and directly tied to student coursework. If the results could be analyzed in relation to student library use data and library instruction records, they would have the potential to demonstrate the value of existing library interventions and to inform new plans. However, weaknesses became apparent as we moved forward. For instance, there was no way to account for IL learning that occurred outside the library as part of students’ regular coursework, employment, or extracurricular activities. In addition, data collection and analysis proved surprisingly time-consuming, and our methods took some unforeseen directions in the process.

Methods

Data Collection

Since this article’s main purpose is to recount and reflect on methodological issues that arose during our study, we will discuss the research methods chronologically and in detail, beginning with our preparations prior to data collection. Following an explanation of our two rounds of data collection and the complications that arose, we will describe our research questions and the data analysis methods by which we sought to answer them.

Preparation and Planning

In preparation, we substantially revised a rubric that had been used for a previous IL assessment project on campus. Because WSU’s stated undergraduate learning goal closely follows the language of the ACRL Standards, the latter were used to guide both the previous and the revised rubrics. The present assessment focused more narrowly on individual student research papers rather than on summative and reflective portfolios, so we chose to limit the rubric to the goals and outcomes that could be directly
observed in a typical research paper, leaving out those that could only be assessed through observation and documentation of students’ research practices prior to writing. Eight learning outcomes were identified, based on aspects of ACRL Standards One (“. . . determines the nature and extent of the information needed”), Four (“. . . uses information effectively to accomplish a specific purpose”), and Five (“. . . understands many of the economic, legal, and social issues surrounding the use of information and accesses and uses information ethically and legally”); and on corresponding university-wide undergraduate learning goals. These outcomes were described in the rubric in terms of three qualitative rankings, “Emerging,” “Developing,” and “Integrating,” each subdivided into two possible scores, resulting in a 6-point ordinal scale—that is, a scale that allowed for ranking of the data. Standards Two and Three were omitted because the researchers did not believe they could be objectively assessed using student writing. The full text of the rubric appears as Table 1.

We also created a survey (see Appendix A) to gather data on students’ contact with the library through various channels, as well as other potentially relevant information, such as the students’ major, gender, number of semesters completed, and transfer status (that is, whether the student transferred or began at WSU Vancouver as a first-year student). A staff member obtained participants’ enrollment records from WSU’s Office of Institutional Research and compared them with the bibliographic instruction records routinely kept by library staff. Once the number of bibliographic instruction sessions attended by each student had been recorded, the staff member removed all identifying information from the demographic surveys and data spreadsheets to ensure anonymity before providing them to the researchers.

**Initial Sample**

With this process in place, we selected a sample of courses for the first iteration of the assessment, in the fall of 2013. For this initial sample, we sought courses that included students at various stages in their college careers (both first-year and transfer students), involved a substantial writing assignment with a research component, and had high enrollment. Given these criteria and the size of our institution, we did not believe that a true random sample of courses could be achieved. Instead, we tried to make a representative selection. Three English course sections at the 100, 200, and 300 levels were chosen, along with a 400-level History course section. By arrangement with the instructors, we visited the four classes to solicit participation and distributed demographic surveys and consent forms to those who chose to participate. We gave participants a brief, intentionally vague description of the research project, explaining that we would look at their research papers and that their anonymity would be protected. We avoided references to IL or to the specific purpose of the study. Near the end of the semester, 46 usable papers were collected, anonymized, and prepared for scoring.

A team of four librarians participated in scoring the student papers using the rubric. In preparation, they held a norming session in which they discussed the rubric, scored a set of sample papers individually, compared the results, and further discussed points of ambiguity and disparity that arose. The goal was not to produce identical scores but to produce scores with a difference of no more than one point for each rubric facet and no
Table 1.
Rubric for assessment of information literacy

<table>
<thead>
<tr>
<th></th>
<th>Emerging 1</th>
<th>Developing 2</th>
<th>Developing 3</th>
<th>Developing 4</th>
<th>Integrating 5</th>
<th>Integrating 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Determines the extent and type of information needed*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Understands the multiple viewpoints relevant to need even if assignment does not state this as a requirement.</td>
<td>Draws primarily from anecdotal or personal experience.</td>
<td>Limited but more than own viewpoints</td>
<td>Balanced. Good representation of multiple viewpoints.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. Uses multiple source types (i.e., journal article, book, newspaper, etc.)</td>
<td>Sources all from one source type.</td>
<td>Mix of two source types. Partially satisfies the information type of the assignment with some variety of source material.</td>
<td>Mix of more than two source types. Satisfies or exceeds the expected information type of the assignment.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c. Extent of source materials is appropriate to assignment</td>
<td>Amount of source material is below requirements of assignment.</td>
<td>Amount of source material partially satisfies the requirements of the assignment.</td>
<td>Amount of source material satisfies or exceeds the requirements of the assignment.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Assess credibility and applicability of information sources.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accepts information without question. For example: quotes sources without comment or evaluation; sources are not timely for topic; sources are inappropriate for project.</td>
<td>Articulates and/or applies basic evaluation criteria to information sources in a partial or limited way. For example: mix of sources which are timely and not timely; mix of authoritative and nonauthoritative sources; mentions one aspect of credibility but ignores others.</td>
<td>Clearly articulates and applies evaluation criteria to the information. For example: all sources are timely for the topic; all sources are authoritative; mentions more than one evaluative criterion per source.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5. Uses information effectively to accomplish a purpose (successful completion of assignment).

a. Demonstrates understanding of the importance of putting sources into context and maintaining the contextual meaning of sources.

- Uses sources out of context.
  - For example: distorts opposing viewpoints.
- Demonstrates some understanding of how context is important when using sources to support arguments.

b. Successfully integrates own knowledge with the knowledge of others.

- Relies heavily on quotations.
  - For example: quotations do not serve a purpose; quotations are not integrated into an overall argument or thesis. For annotated bibliographies: sources are not relevant to topic, and student doesn’t recognize this.
- Uses more paraphrasing than quotations.
  - For example: quotations or references serve a purpose but are not well used for that purpose; some quotations or paraphrases are effectively integrated into an overall argument or thesis. For annotated bibliographies: majority of sources are relevant to topic.

Citing/documenting sources. 

- Makes multiple errors when citing sources in text and in reference list.
- Makes minimal errors when citing sources in text and in reference list.
- Makes no errors when citing sources in text and in reference list.

6. Access and use information ethically and legally.

a. Recognizes plagiarism and need for documentation.

- Uses information without referencing the sources of that information.
- Acknowledges the source of information most of the time. Somewhat clear as to which work is the student’s and which is from a source.
- Always acknowledges the source of information.

Table 1, cont.

<table>
<thead>
<tr>
<th></th>
<th>Emerging</th>
<th>Developing</th>
<th>Integrating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>5. Uses information effectively to accomplish a purpose (successful completion of assignment).</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Demonstrates understanding of the importance of putting sources into context and maintaining the contextual meaning of sources.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Uses sources out of context.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- For example: distorts opposing viewpoints.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Demonstrates some understanding of how context is important when using sources to support arguments.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b. Successfully integrates own knowledge with the knowledge of others.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Relies heavily on quotations.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- For example: quotations do not serve a purpose; quotations are not integrated into an overall argument or thesis. For annotated bibliographies: sources are not relevant to topic, and student doesn’t recognize this.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Uses more paraphrasing than quotations.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- For example: quotations or references serve a purpose but are not well used for that purpose; some quotations or paraphrases are effectively integrated into an overall argument or thesis. For annotated bibliographies: majority of sources are relevant to topic.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citing/documenting sources.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Makes multiple errors when citing sources in text and in reference list.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Makes minimal errors when citing sources in text and in reference list.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Makes no errors when citing sources in text and in reference list.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>6. Access and use information ethically and legally.</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. Recognizes plagiarism and need for documentation.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Uses information without referencing the sources of that information.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Acknowledges the source of information most of the time. Somewhat clear as to which work is the student’s and which is from a source.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Always acknowledges the source of information.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Rubric item numbers follow the numbering of standards and sample outcomes in the Association of College and Research Libraries Information Literacy Competency Standards for Higher Education (2000).
more than eight points in the aggregate score that resulted from adding the eight facet scores together. Through the discussion process, the four raters agreed on sufficiently similar scores and developed a consensus as to how the rubric would be applied in practice.

To minimize the effect of individual raters, each paper was scored separately by two randomly assigned raters. In cases where their total scores for a given paper differed by more than eight points (that is, by more than one point per rubric facet on average), a third rater with no knowledge of the other scores would also evaluate the paper. Because the rubric scale was ordinal, subsequent analysis would use the median of the two or three rater scores. Using this method, an initial sample of 46 papers from fall 2013 was scored by four raters.

In addition to simply assessing the IL skills of our sample, we wanted to identify any significant relationship between IL and contact with the library through various channels. We also hoped to detect any significant differences related to the demographic variables collected in our survey. In the statistical analysis process, we sought significant relationships between resolved median IL rubric scores as dependent variables—that is, the variables we wanted to measure—and 12 independent variables, indicating library contact and demographic features, factors that might cause some change in the rubric scores. Tables 2, 3, and 4 summarize these variables and our approach to quantifying them, as well as central tendencies and distributions for our sample. Table 2 lists the nine rubric variables, including eight individual facets reflecting distinct skills, as well as an aggregate total score. The demographic factors listed in Table 3 include gender, transfer status, age, and number of semesters at WSU Vancouver. Table 4 lists the library contact variables, including an aggregate library contact score used to estimate overall exposure to library interventions.

Because the contact variables shown in Table 4 (C1–C5) are measured on a variety of scales, it was necessary to recode these measures before adding them together to produce a weighted aggregate library contact variable (C6). As shown in Table 5, variables for time in library, website use, library assistance, and attendance at a bibliographic instruction session (C1, C2, C3, and C5, respectively) were divided into categories for lower contact (one point per variable) and higher contact (two points per variable). In that way, each would have equivalent weight in the aggregate score. Because we assumed that an individual appointment with a librarian would have higher impact than other contact methods, the appointment variable (C4) received greater weight, with three points added in cases where an appointment had been made. By adding the results of these five transformed variables, an aggregate score between 5 and 11 was obtained as an estimate of overall library contact.8

First Complication: Expanding the Sample

Analysis of the initial sample suggested positive correlations between IL and library contact, but these correlations were not statistically significant ($p > .05$). We decided to extend the study and seek a larger, more representative sample of courses and students. Because it would be difficult to reach a large sample on our small campus, we elected to retain the initial sample, add a second sample using identical methods, and analyze
# Table 2.
Rubric variables with medians from two rounds of scoring

<table>
<thead>
<tr>
<th>Variable</th>
<th>Rubric item*</th>
<th>Resolved median in first round (N = 77), with interquartile range (IQR)†</th>
<th>Resolved median in second round (N = 77) with IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1: Multiple viewpoints</td>
<td>1a. Understands the multiple viewpoints relevant to need even if the assignment does not state this requirement.</td>
<td>4.0 (IQR = 1.25)‡</td>
<td>3.0 (IQR = 1)</td>
</tr>
<tr>
<td>R2: Source types</td>
<td>1b. Uses multiple source types (i.e., journal article, book, newspaper, etc.)</td>
<td>4.5 (IQR = 1)</td>
<td>4.0 (IQR = 1.5)</td>
</tr>
<tr>
<td>R3: Extent of sources</td>
<td>1c. Extent of source materials is appropriate to assignment.</td>
<td>5.0 (IQR = 1.5)</td>
<td>4.0 (IQR = 1.5)</td>
</tr>
<tr>
<td>R4: Assessment of sources</td>
<td>2. Assess credibility and applicability of information resources.</td>
<td>3.0 (IQR = 1.25)</td>
<td>2.5 (IQR = 1)</td>
</tr>
<tr>
<td>R5: Context</td>
<td>5a. Demonstrates understanding of the importance of context and maintaining the contextual meaning of sources.</td>
<td>4.0 (IQR = 1.5)</td>
<td>3.0 (IQR = 1)</td>
</tr>
<tr>
<td>R6: Integration</td>
<td>5b. Successfully integrates own knowledge with the knowledge of others.</td>
<td>3.5 (IQR = 1.5)</td>
<td>3.0 (IQR = 1.25)</td>
</tr>
<tr>
<td>R7: Citation format</td>
<td>5c. Citing/documenting sources. FORMAT ONLY.</td>
<td>4.0 (IQR = 1.75)</td>
<td>3.5 (IQR = 1.5)</td>
</tr>
<tr>
<td>R8: Documentation</td>
<td>6. Recognizes plagiarism and need for documentation.</td>
<td>4.5 (IQR = 1.75)</td>
<td>3.5 (IQR = 4.5)</td>
</tr>
<tr>
<td>R9: Aggregate rubric score§</td>
<td></td>
<td>31 (IQR = 8)</td>
<td>28 (IQR = 10)</td>
</tr>
</tbody>
</table>

* Rubric item numbers follow the numbering of standards and sample outcomes in the Association of College and Research Libraries Information Literacy Competency Standards for Higher Education (2000).
† Interquartile range (IQR) refers to the range of the middle 50 percent between the lower quartile and the upper quartile of the sample.
‡ Variables R1–R8 are measured on a six-point ordinal rubric scale from 1 (“emerging”) to 6 (“integrating”).
§ Variable R9 represents the sum of variables R1–R8, yielding a score between 8 and 48.
# Table 3.
## Demographic variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unit or scale of measurement</th>
<th>Result for our sample (N = 77)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1: Gender</td>
<td>Nominal: male or female</td>
<td>57% male, 43% female</td>
</tr>
<tr>
<td>D2: Transfer status (whether the student started Washington State University (WSU) as a first-year college student or as a transfer student)</td>
<td>Nominal: started as first-year student at WSU Vancouver or transferred with more than 30 credits from another institution</td>
<td>16% started as first-year students; 86% started as transfer students</td>
</tr>
<tr>
<td>D3: Age</td>
<td>Ratio (number of years); divided into two age groups for analysis (under 25 or 25 and over)</td>
<td>Mean age is 25 (standard deviation = 8.17);* 68% under 25, 32% 25 and over</td>
</tr>
<tr>
<td>D4: Is this your first semester at WSU Vancouver?</td>
<td>Nominal: yes or no.</td>
<td>51% yes, 49% no</td>
</tr>
<tr>
<td>D5: How many semesters have you taken classes at WSU Vancouver?</td>
<td>Ratio: number of semesters.</td>
<td>Mean number of semesters is 2.3 (standard deviation = 1.61)</td>
</tr>
<tr>
<td>D6: What department is your academic major in?</td>
<td>Nominal: Choice of 22 departments</td>
<td>Ranked by frequency: History (16); Business (10); Social Sciences (8); Human Development (7); English (5); Biology, Computer Science, Psychology (4 each); Engineering, Environmental Science (3 each); Creative Media and Digital Culture, Education, Public Affairs, Mathematics, Neuroscience, (2 each); Anthropology, Political Science, Sociology (1 each).*</td>
</tr>
</tbody>
</table>

Standard deviation is a measure of how tightly the numbers cluster around the mean.
### Table 4.
Library contact (C) variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Survey question</th>
<th>Unit or scale of measurement</th>
<th>Result for our sample (N = 77)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1: Time in library</td>
<td>How much time do you estimate you spend at the library?</td>
<td>5-point ordinal scale: Never (1), Once or twice a semester (2), Once a month (3), Once a week (4), At least once a day (5).</td>
<td>Median = 4 (interquartile range [IQR] = 2)*</td>
</tr>
<tr>
<td>C2: Website use</td>
<td>How often do you use the Washington State University (WSU) Vancouver Library website to do research?</td>
<td>5-point ordinal scale: Never (1), Once or twice a semester (2), Once a month (3), Once a week (4), At least once a day (5).</td>
<td>Median = 3 (IQR = 2)</td>
</tr>
<tr>
<td>C3: Library assistance</td>
<td>How many times have you asked for library assistance (talking to someone at the reference desk, e-mailing the library, etc.) while doing research?</td>
<td>4-point ordinal scale: Never (1), Once or twice (2), 3–5 times (3), More than 5 times (4).</td>
<td>Median = 2 (IQR = 2)</td>
</tr>
<tr>
<td>C4: Appointment with librarian</td>
<td>Have you ever made an appointment to meet with a librarian one-on-one at WSU Vancouver Library?</td>
<td>Nominal: no/yes.</td>
<td>96% no, 4% yes</td>
</tr>
<tr>
<td>C5: Number of bibliographic instruction sessions attended</td>
<td>N/A (data from enrollment records)</td>
<td>Ratio (number of bibliographic instructions ranged from 0 to 4).</td>
<td>Ranked by frequency: 1 (42%), 0 (36%), 2 (17%), 3 (4%), 4 (1%). Mean is 0.92 (standard deviation = 0.9)*</td>
</tr>
<tr>
<td>C6: Aggregate library contact score</td>
<td>N/A</td>
<td>Ordinal: Weighted sum of values as shown in Table 5. Minimum possible score is 5, maximum is 11.</td>
<td>Median = 7 (IQR = 1)</td>
</tr>
<tr>
<td>C7: Communication method</td>
<td>If you ask for library assistance while doing research, your method of communication is (check all that apply):</td>
<td>Inclusive nominal scale with four categories (in person, phone, e-mail, or instant messaging [IM]), plus an open-ended “Other” category.</td>
<td>Ranked by frequency: in person (46), no response (21), e-mail (14), phone (4), IM (4).</td>
</tr>
</tbody>
</table>

*Interquartile range (IQR) refers to the range of the middle 50 percent between the lower quartile and the upper quartile of the sample.
† Standard deviation is a measure of how tightly the numbers cluster around the mean.
the combined results. To better approximate the makeup of the campus, we looked for courses that might include a greater proportion of transfer students. A second sample was gathered in fall 2014, consisting of 31 papers from three course sections: a 300-level Human Development course, a 300-level History course (required for transfer students in all majors), and a 400-level History course. (A fourth course section was initially included, but the instructor withdrew after deciding that no research assignment would be required.) Participants in both samples completed the same demographic survey, and the same rubric was used to score all papers. A new group of raters was convened to score the new papers, with the intention of combining the data from both groups into a total sample of 77 papers.

**Further Complication: Re-Norming and Rescoring**

At this point, human error complicated our plans. Due to changes in staffing, one person left and another person joined the rater group. Perhaps because this seemed like a minor
change (the new rater was also a member of the research team), we neglected to perform a second norming session before scoring the papers. Considerable time passed before we realized the significance of this oversight. It only came to light when we reviewed our preliminary results and noted the high level of disagreement between raters on a rubric facet that should have been among the easiest on which to agree, variable R7, which dealt strictly with the formatting of citations and references. Once we realized that our norming had been inconsistent, we attempted to correct the problem by convening a new group of raters, holding a new norming session, and then repeating the scoring process for all 77 papers. Since three of the five raters had been involved in the previous rounds of rating, the papers were systematically assigned to avoid repetition for those three raters. That is, papers were distributed randomly but with the constraint that no one would rate a paper they had already scored in the previous sessions. We planned to base our statistical analysis on the new score results but retained the earlier results for comparison.

Final Complication: Reliability and Rater Disagreement

Surprisingly, the new scores resulted in lower rater agreement than the previous set of scores, even though the initial set was produced by two different groups of raters without consistent norming. In the new round of scoring, 25 papers (32 percent) received total scores that varied by more than eight points, necessitating a third rating by another reader to resolve the disagreement. This is above the maximum acceptable level of rater disagreement, 30 percent, recommended by Steven Stemler. In comparison, the initial round of scores resulted in only about half as many needing a third rating; that is, 13 papers (17 percent) received total scores from two raters that differed by more than eight points. While this level of disagreement might be acceptable under other circumstances, we cannot fully credit these scores because they were obtained without a consistent norming process. Thus, neither set of scores can be considered a reliable measure of IL. When considering these scores in relation to such factors as library contact and demographics, correlations cannot be convincing even when they are statistically significant. Under these conditions, the results of our statistical analysis (summarized in Appendix B) are suggestive at best.

It is unclear why we saw such high levels of disagreement even after norming. Comparison between pairs of raters did not indicate that any rater was at variance with the group, nor have we identified any major ambiguities in the rubric. Our best guess is that several raters, perhaps all, had difficulty in applying the rubric consistently to a wide variety of subject areas and course levels. Perhaps this problem could have been remedied through more extensive norming discussions, a more diverse sample of papers for norming, or use of a more sensitive reliability measure such as intraclass correlation, which describes the extent to which two or more raters agree when scoring the same set of things (discussed under Question 2, “How reliable and consistent were the rubric scores given by different raters?”).
Statistical Analysis Process

Because of these setbacks, the following account of our data analysis will focus on the process of analysis—including secondary research and decision-making prior to running the numbers. We will address the results of our analysis only in passing and only when they pertain to an understanding of the process. The primary value of our study lies in the methodological questions it raises, rather than in the results, many of which may not be valid for our sample population.

Although not all the goals of the investigation were clearly articulated when we initially planned our study, five research questions had emerged by the time we began analyzing the data. Two of the five dealt with the validity of the method itself, while the other three sought significant differences and evidence of impact in the resulting IL scores. We will address these questions one at a time to provide a sufficiently detailed context for the recommendations in the “Conclusion” section and the results presented in Appendix B. Table 6 lists the research questions and summarizes our statistical methods. Repeated scoring resulted in two sets of scores that could not be combined into a single meaningful data set. For each question, therefore, we applied statistical tests to both sets of scores and compared the results.

Resolving Discrepant Scores

To minimize the influence of individual raters, each student paper was read by two randomly assigned raters, and the median score was used for analysis purposes, as described later. To resolve discrepant scores and arrive at an accurate score for analysis, we adapted James Penny and Robert Johnson’s recommendation of the “parity method,” which was found to provide greater validity than other resolution methods in a study of rubric scoring. We used the median between two scores if they differed by eight points or less (or one point per rubric facet) and the median between three scores if the difference was greater than eight points, necessitating resolution by a third rater. In the analyses described later, we therefore used the resolved scores, derived from the medians of either two or three raters’ scores, as our IL rubric variables (R1–R9).

Question 1: Is the rubric scoring internally consistent in a way that suggests that all its parts contribute to IL as a single unifying construct?

This question involves two types of relationships among the rubric facets (variables R1–R8): internal consistency and significant difference. Consistency would indicate a shared construct (that is, IL) underlying all eight variables, each of which would then contribute to a meaningful aggregate score (variable R9). Significant difference would indicate that each variable measured a separate, independent skill or ability. If there was a lack of internal consistency, it would make sense to discard the aggregate score, but the scores for individual IL skills could still be analyzed separately. A statistical test called Cronbach’s alpha was used to test internal consistency among resolved median scores for variables R1–R8. This test also identifies variables that could be removed to
Table 6.
Research questions and statistical methods

<table>
<thead>
<tr>
<th>Research question</th>
<th>Statistical methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question 1: Is the rubric scoring internally consistent in a way that suggests that all its parts contribute to information literacy as a single unifying construct?</td>
<td>Cronbach’s alpha and Wilcoxon signed-rank test*</td>
</tr>
<tr>
<td>Question 2: How reliable and consistent were the rubric scores given by different raters?</td>
<td>Intraclass correlation coefficient (ICC);† modified percent agreement</td>
</tr>
<tr>
<td>Question 3: Do correlations exist between information literacy and contact with the library?</td>
<td>Spearman’s rho and Kendall’s tau‡</td>
</tr>
<tr>
<td>Question 4: In what information literacy skills were the participants strongest and weakest?</td>
<td>Comparison of resolved medians</td>
</tr>
<tr>
<td>Question 5: Did any of the demographic factors have a significant relationship to information literacy?</td>
<td>Mann-Whitney U and Kruskal-Wallis H tests§ for differences between groups; Spearman’s rho and Kendall’s tau for correlations</td>
</tr>
</tbody>
</table>

*Cronbach’s alpha is a statistical test used to measure internal consistency. The Wilcoxon signed-rank test determines significant differences between each possible pair of rubric variables.
†Intraclass correlation describes the extent to which two or more raters agree when rating the same set of things.
‡Spearman’s rho and Kendall’s tau indicate the closeness of the relationship between two variables.
§The Mann-Whitney U test and Kruskal-Wallis H test are used to assess differences between groups.

Improve consistency. To test for significant differences between each possible pair of rubric variables, we conducted Wilcoxon signed-rank tests of significance, a standard nonparametric test—that is, one in which the data are not assumed to fit a normal distribution—used in repeated-measures studies.

Question 2: How reliable and consistent were the rubric scores given by different raters?

Question 2 proved the most complex and required us to seek models outside the library literature. It pertains to two related issues, inter-rater reliability and methods of
resolving rater disagreement. In planning our study, we initially assumed that we could rely on a single method for both purposes and chose the “modified” or “broadened” percent agreement measure described by Stemler, which involves calculating the percentage of agreement between two raters.12 The modified approach measures agreement within an acceptable range (for example, within one point) rather than exact agreement. This measure provided a convenient means of resolving rater disagreement and organizing our norming process, but we eventually realized it was not appropriate as a measure of inter-rater reliability.

Initially, we measured agreement as follows: During norming, raters attempted to reach agreement within one point for each rubric facet, or within eight points for the total score. This process allowed us to identify rubric facets with especially low agreement, which became the focus of the raters’ norming discussion. Following norming, and once each paper in our sample had been scored by two raters, this measure was also used to identify instances of disagreement that required a third rater to provide an accurate median score.

Although this was a convenient, straightforward way to identify and resolve rater disagreement, the percent agreement method is not considered a meaningful measure of inter-rater reliability because it has the potential to overestimate the level of consensus.13 After our second round of rating was completed, we came to understand this problem and reviewed the literature on inter-rater reliability in search of an appropriate method.

Stemler identifies three types of inter-rater reliability estimates, the consensus, consistency, and measurement approaches.14 A measurement estimate is most appropriate in this scenario because the rubric scale is ordinal and because each paper was rated by more than one rater, with the intention of using the central tendency (median) as a final score for analysis purposes. Various methods are used for this type of estimate, depending on the research design. Following Kevin Hallgren’s discussion, we determined that the most appropriate method would be to calculate intraclass correlation, how much two or more raters agree when rating the same set of things, since in our study all subjects were rated by multiple, randomly assigned raters, and we were concerned with the magnitude of disagreement rather than with a simple test of exact agreement.15 (In Stemler’s terminology, we were interested in measurement rather than consensus or consistency.) Specifically, we needed a one-way mixed, absolute-agreement, average-measures intraclass correlation, for the following reasons: randomly assigned pairs rated each paper (indicating a one-way model); we were interested in absolute agreement rather than rank consistency because we intended to use the mean between two raters’ scores for each paper; all papers were rated by at least two raters (indicating the average-measures unit); and we did not seek to generalize to a larger population of raters (mixed effect).16 We applied this measure to ratings by randomly assigned pairs of raters across the entire sample. To evaluate our score resolution method, we also calculated intraclass correlation for the subsets of cases that required a third reader for resolution and for those that did not. Domenic
Cicchetti provides a widely cited standard for evaluating levels of reliability, whereby reliability is described as poor for intraclass correlation values below 0.40, fair for values between 0.40 and 0.59, good for values between 0.60 and 0.74, and excellent for values between 0.75 and 1.\(^7\)

As a measure of inter-rater reliability, intraclass correlation is more meaningful and more sensitive than agreement rate. While agreement rate only detects cases of concurrence within a specific predetermined range, intraclass correlation measures correlation between raters for a given variable across the entire set of scores. In other words, intraclass correlation can detect patterns of consistency between raters even if one rater tends to score higher than another. For this reason, higher rates of agreement do not always coincide with higher intraclass correlation scores.

We became aware of the intraclass correlation approach late in the project, after the second round of norming and scoring. Had we understood it earlier, we could have applied an appropriate inter-rater reliability test to a sample of papers, each evaluated by all raters, in the norming phase and perhaps avoided the need to test inter-rater reliability for our entire sample afterward. This would have saved time and allowed us to move into the rating process with greater confidence in the reliability of our ratings.

**Question 3: Do correlations exist between IL and contact with the library?**

For Question 3, we tested for significant correlations between rubric scores and library contact (variables R1–R9 and C1–C6 in Tables 2 and 4) using Spearman’s rho, a statistic that indicates the closeness of the relationship between two variables. Spearman’s rho is a standard nonparametric test appropriate for ordinal-level variables, in which the possible values can be ranked higher or lower than another. Because the narrow range of rubric scores resulted in many tied ranks, we also used another nonparametric correlation test, Kendall’s tau. As this was the central question for our study, we attempted to be as thorough as possible.

We tested the relationship of each form of library contact and the aggregate score for all forms to each rubric facet score and to the aggregate score for an overall IL construct. One of the contact variables, the method of communication used to contact the library (C7), was not analyzed because it involved overlapping, nonexclusive groups and did not impact the degree of contact between the student and the library.

**Question 4: In what IL skills were the participants strongest and weakest?**

Having established the significance of differences among rubric variables using Wilcoxon tests as described under Question 1 (“Is the rubric scoring internally consistent in a way that suggests that all its parts contribute to information literacy as a single unifying construct?”), Question 4 could be answered by comparing and ranking the resolved median scores for each rubric facet (R1–R8 in Table 2).

**Question 5: Did any of the demographic factors have a significant relationship to IL?**
We conducted Mann-Whitney U tests to assess the differences in rubric scores (variables R1–R9) between groups based on demographic variables (D1–D4). This is a standard nonparametric test used to compare independent groups. Spearman’s rho and Kendall’s tau were used to test for significant correlations between rubric scores and either age (variable D3) or number of semesters at WSU (variable D4). For variable D6, academic major, a Kruskal-Wallis H test was used to assess differences among 18 majors. Kruskal-Wallis is a nonparametric test like Mann-Whitney but appropriate for comparing more than two groups.

Conclusion: Lessons Learned

Our statistical results (summarized in Appendix B) remain problematic. However, the long and frustrating process yielded valuable lessons of its own. Authentic IL assessment is a new field of study, and we believe that by providing an honest account of methodological issues in a failed assessment, we make a valuable contribution to the existing literature. No universal method of assessment fits the needs of every institution in every situation, so practitioners will benefit from sharing their mistakes as well as their successes and considering points of ambiguity or uncertainty alongside strong statistical findings. In addition to describing the pitfalls of our assessment experience, we would like to offer five general recommendations that may be of use in future assessment projects, in IL and other areas. Some may seem obvious, but our experience shows how complicated they can become. When undertaking authentic assessment, researchers should plan for norming, find models for data analysis, take a large sample, dedicate plenty of time to the project, and, above all, stay flexible.

Plan for Norming

As many researchers note, norming is a crucial part of the process for any assessment that involves a rubric or similar tool requiring individual raters to assess aspects of student work. In a norming session, all raters should be present, discuss the instrument and criteria together, and separately score the same sample materials. The goal is to reach a predetermined acceptable level of agreement on scores for each sample, even if this requires multiple rounds of discussion and scoring. Norming serves four important purposes: to establish usable consensus standards for the application of the rubric; to test the

When undertaking authentic assessment, researchers should plan for norming, find models for data analysis, take a large sample, dedicate plenty of time to the project, and, above all, stay flexible.

In a norming session, all raters should be present, discuss the instrument and criteria together, and separately score the same sample materials. The goal is to reach a predetermined acceptable level of agreement on scores for each sample, even if this requires multiple rounds of discussion and scoring.
rubric and identify areas of vagueness or ambiguity; to train raters through discussion and trial and error; and to evaluate the group’s inter-rater reliability (in one or more of the distinct senses identified by Stemler). If we had performed norming in a consistent manner in the first round of scoring, we could have saved time and effort by avoiding the second round. If we had spent more time on norming and held raters to a higher standard of agreement, we might have reached acceptable levels of rater agreement in our final set of scores. In addition, if we had incorporated an appropriate measure of inter-rater reliability, we might have avoided the difficulty of attempting to assess inter-rater reliability from a final score group. We could have obtained a robust estimate of reliability using fully crossed samples from norming and from a segment of the research sample, rather than calculating from randomly assigned pairs for the entire sample.

Our norming process, as planned, was brief, used only two sample papers, and was based on the goal of rater agreement within eight points. Although we reached this goal, we had poor rater agreement and inter-rater reliability afterward. Instead, we should have planned norming with a larger selection of sample papers and allowed time for discussion and rescoring to resolve disagreements. In the process, we could have identified the most problematic rubric facets—those with the highest rate of disagreement—and spent additional time discussing them to arrive at a consistent approach. If necessary, once an approach was agreed upon, we could have revised the rubric for clarity or added examples as a reminder to raters. After achieving agreement within eight points on several papers, we should have calculated two-way mixed, absolute-agreement, average-measures intraclass correlation to ensure not only approximate agreement but also inter-rater reliability regarding measurement. A two-way approach would be used in this case because, in norming, we used a fully crossed design with all raters scoring all papers. A one-way approach was used for the actual sample because the papers were rated by randomly assigned pairs of raters.

Plan for Data Analysis

It is imperative to review the literature when planning an authentic assessment, but even the most relevant studies may not translate directly into a sufficiently detailed model for new research. Two of us had conducted an earlier study using rubrics for authentic assessment of students’ research portfolios, and all three had participated in reviewing and discussing the literature while planning the present study, but we still ran into problems at the analysis stage. Our initial literature review concentrated on methods of approaching and gathering data on IL and library impact, rather than on methods of statistical analysis. Thus, our plans for analysis remained somewhat vague until after we had collected our first sample of papers and conducted our first round of scoring. Perhaps this previous experience gave us an unwarranted confidence at the beginning of the project, so that we neglected to clearly connect our research questions to specific testable hypotheses. While it is impossible to anticipate every contingency, we could have saved a great deal of trouble by more carefully considering the statistical analyses in the library literature and by perusing statistical literature from other disciplines to identify specific methods appropriate to our questions before collecting our data. Because we had limited knowledge of statistical analysis and of SPSS, a software package used to perform such analysis, we had to arrive at a process through extensive
trial and error. For example, we spent a long time considering Fleiss’s kappa and Cohen’s kappa as measures of inter-rater reliability before we found a clear explanation in the literature indicating that we should instead use a type of intraclass correlation since we were comparing ordinal rather than nominal variables. There were many false starts and hours of tedious recoding in SPSS.

We cannot overstate the importance of writing clearly defined research questions and connecting them to statistically measurable hypotheses; of including or consulting a researcher with statistical expertise; and of looking beyond the library literature to identify appropriate methodologies. Our search yielded several studies that match our general approach—rubric assessment of IL using authentic artifacts of coursework—such as those by Mark Emmons and Wanda Martin, Lorrie Knight, and Sue Samson. However, none provided an account of methodology that was both sufficiently detailed and appropriate for our purposes. A study by Megan Oakleaf explicitly focused on methodology, including norming practices and measures of inter-rater reliability. Its methods were not applicable to our study, however, because it treated its three-level rubric classifications as nominal rather than ordinal variables—that is, variables for which the values cannot be organized in a logical sequence rather than variables for which the values can be logically ordered or ranked—and because it used a consensus approach to inter-rater reliability. The former precludes the use of an aggregate score to measure IL as a generalized construct, and the latter precludes the use of central tendency or adjudication to resolve rater disputes.

In our initial literature review, we did not look at methodological literature outside of library and information science journals. If we had, we might have identified useful starting points such as Stemler and Hallgren on computing inter-rater reliability, and Penny and Johnson on methods of resolving rater disagreement. Had we discovered these early in the process, we might have attempted a rigorous norming process using an inter-rater reliability test for measurement such as intraclass correlation and saved ourselves the trouble of testing inter-rater reliability on our final set of scores. Although similar calculations are commonplace in the social sciences, the literature on library assessment rarely addresses statistical methods in such detail. Had we begun with preliminary statistical research, alongside our initial research on authentic assessment of IL, we might have started data collection with a clear, detailed plan to address each of our questions and thus have saved significant time.

Take a Large and Representative Sample

Our initial sample from three courses in fall 2013 did not yield statistically significant results. It was small and did not adequately represent the range of undergraduate research courses on our campus. Due to delays in data entry and analysis, we could not take a new
sample until fall 2014, when we collected additional data from three other courses in an attempt not only to increase our sample size but also to diversify our sample in terms of course level, assignment requirements, and subject area. After combining samples, we could detect statistical significance, but in the process, we created two new problems: due to staffing changes, we needed a slightly different group of raters (and neglected to redo our norming), and we extended the duration of the project well beyond its intended scope. Had we gathered a larger sample in the first round, we might have repeated the project (or a more streamlined version of it) with a new sample a year or two later and ended up with meaningful longitudinal data, repeated observations taken from a larger sample over time, useful for measuring changes.

There are three main reasons why it was difficult to achieve a larger sample. First, WSU Vancouver is a small campus, so we had a limited pool of students and courses from which to pull. In an earlier assessment project, library staff benefited from a systematic campus-wide capstone portfolio collection, a compilation of student work that offered a large, ready-made pool of data. For the present study, that option was not available and we needed to solicit participation individually from faculty and from students in their courses. This led to the second problem: since we needed to solicit faculty participation individually, we had to guess which courses would most likely include a substantial research assignment. This information was not readily available for all courses, so we relied on librarians’ personal knowledge of courses and relationships with faculty in their subject liaison areas. The third problem was that, for the courses we chose, not all students agreed to participate in the study. Initially, participation was higher than we expected for most courses, but in both semesters, we lost participants along the way for reasons that included changing assignment requirements, late or missing assignments, and students dropping courses.

To counteract these difficulties in the future, we can begin planning earlier and reach out to faculty in a more systematic manner. It may help to collect information on research assignments over an academic year before data collection and to contact all faculty in relevant disciplines, either through e-mail or in person at faculty meetings, to draw from the widest possible pool of courses with some research component. Given a large enough pool of participants, a truly random sample might be achieved. It might also help to offer some type of incentive to students for participation in the study.

**Authentic assessment is a time-consuming activity, especially when methods and processes are being established for the first time. It requires the participation of multiple team members over a period that may range from several weeks to several years.**

**Dedicate Plenty of Time**

Authentic assessment is a time-consuming activity, especially when methods and processes are being established for the first time. It requires the participation of multiple team members over a period that may range from several weeks to several years. In
our case, a study that we intended to complete in six months took more than four years, including the time needed for additional data collection, repeated scoring, and repeated data analysis described here. Although we realized that it would take time to develop the rubric, to work with faculty and with the Office of Institutional Research, and to anonymize student papers, demographic surveys, and enrollment data, we did not consider the time it would take to determine appropriate statistical methods and conduct statistical analysis, especially for a lead researcher who was unfamiliar with SPSS and had limited experience with this type of research. And, of course, we did not anticipate the need to collect additional data and eventually to re-norm and rescore with a new group of raters.

Critical reflection and discussion are necessary at each stage of an assessment project, not only during data collection and analysis but also before and after them. Reflection and discussion require time and planning but can also save time in the long run. In our case, a discussion of rubrics and standards led us to focus on those IL skills that could truly be assessed through student writing. Later, in our analysis, we had to examine and think about our statistical results, which led us to recognize and attempt to correct a major flaw in our scoring process, inconsistent norming. If we had built these critical exchanges into our process in a more organized way, we might have avoided the missteps that ultimately invalidated much of our data. Nevertheless, we culled some valuable information from our failed assessment by taking time to reflect on what worked, what did not, and what lessons might prove useful to others in the field.

Each of these stages extended the duration of the project, but the problem was exacerbated by the somewhat sporadic way in which we could devote time to it. Like many academic librarians, we needed to balance a wide variety of duties throughout the year and found it difficult to set aside large blocks of time for research. While there was perhaps no way to truly anticipate the time needed, it would have helped for each researcher to schedule portions of uninterrupted work time for each phase of the project—while retaining flexibility, of course.

Stay Flexible

The final takeaway for this project may seem trivial, but it is important, especially for a long-term project like this. From planning to data collection to analysis, flexibility is vital when dealing with the varied, contingent, and unpredictable factors of authentic assessment. While the study failed in many ways, it was successful as a learning experience for the researchers and, we hope, as a case study.

From planning to data collection to analysis, flexibility is vital when dealing with the varied, contingent, and unpredictable factors of authentic assessment.
for future researchers. These successes were largely the result of open-ended, iterative planning and of flexibility in adapting to circumstances that did not always fit the plan.

For example, rather than trying to force our assessment of student work to address each of the ACRL Standards, we focused our rubric on those areas of the document that could reasonably be assessed by reading students’ writing without observing their research process firsthand. We carefully revised our rubric to make it specific and clear while still applicable to a wide variety of student work at all academic levels, potentially including annotated bibliographies as well as expository or argumentative essays.

Other examples of flexibility include our willingness to research, test, and adapt statistical methods appropriate to our research questions and sample—in some cases, methods not specifically described in the existing IL literature—and to gather additional data and redo norming, scoring, and analysis when necessary. Finally, when it became clear that our results were compromised, we chose to reframe our thinking about the project to learn from our mistakes, clarify and record the methodological issues involved, and identify key strategies that we and other researchers could use to avoid such problems in the future.

Sam Lohmann is a reference librarian at Washington State University Vancouver; he may be reached by e-mail at: sam.lohmann@wsu.edu.

Karen R. Diller is the library director at Washington State University Vancouver; she may be reached by e-mail at: diller@wsu.edu.

Sue F. Phelps is the health sciences and outreach librarian at Washington State University; she may be reached by e-mail at: asphelps@wsu.edu.

Appendix A

Survey of Participants

The following demographic and behavior survey was distributed to students. Identification numbers were arbitrarily assigned and used to identify participants without using their names or student ID numbers. For Question 1, regrettably, only two options are given for gender. In retrospect, we failed to make the survey inclusive and welcoming to students of all gender identities. Were we to redo this study in the future, we would include an open-ended third option for this question.

Demographic Survey

Please answer the following questions.

1. I am: [Male; Female]

2. When you first started at Washington State University (WSU) Vancouver, you: [started as a first-year (freshman) student; transferred in with more than 30 credits from another institution]
3. Age?
4. Is this your first semester at WSU Vancouver? [Yes; No]
   a. If no, how many semesters (NOT counting the current semester) have you taken classes at WSU Vancouver?
5. What department is your academic major in? [Twenty-one options including Undeclared]
6. How much time do you estimate you spend at the library? [At least once a day; Once a week; Once a month; Once or twice a semester; Never]
7. How often do you use the WSU Vancouver library website to do research? [At least once a day; Once a week; Once a month; Once or twice a semester; Never]
8. How many times have you asked for library assistance (talking to someone at the reference desk, e-mailing the library, etc.) while doing research? [More than 5 times; 3–5 times; Once or twice; Never]
9. If you ask for library assistance while doing research, your main method of communication is: [In person (ask the reference desk in the library); Phone; E-mail; Instant messaging (IM); Other (please specify)]
10. Have you ever made an appointment to meet with a librarian one-on-one at the WSU Vancouver Library? [Yes; No]

Appendix B

Results of Statistical Analysis

The following is a summary of the results of our statistical analysis, reported in the order of our research questions (see Table 6). It is important to note that the low level of agreement and inter-rater reliability, as discussed under Question 2 (“How reliable and consistent were the rubric scores given by different raters?”), casts doubt on the validity of the remaining questions, which all to some degree assume the validity of the rubric scores (variables R1–R9).

In addition, the fact that we rescored all papers resulted in two sets of scores that could not meaningfully be combined into a single data set. For all five questions, therefore, we analyzed both sets of scores and compared the results. In reporting our results, we rely primarily on the second set of results because they were generated after a consistent norming process and have similar inter-rater reliability to the first set based on intraclass correlation. We have only noted values from the first set when there is an important difference from the second. To save space, only significant \( p < .05 \) statistics are reported here, with highly significant \( p < .01 \) statistics marked as such. Complete data tables are available from the authors upon request.

**Question 1:** Is the rubric scoring internally consistent in a way that suggests that all its parts contribute to IL as a single unifying construct?

This question addresses two separate aspects of the instrument: internal consistency and distinctness. Cronbach’s alpha showed good internal consistency between the re-
solved medians for the eight rubric facets (variables R1–R8), alpha = 0.88 (and alpha = 0.83 for the first round of scores). The removal of any variable would result in a lower alpha score, suggesting that all variables contributed to the shared IL construct and should be retained when calculating the aggregate IL variable (R9).

Paired Wilcoxon signed-rank tests found significant differences between 22 of the 23 pairs of variables where the resolved medians were not equal, as well as between two of the five pairs of variables with equal resolved medians (see Table 7). Similar results were found in the first round of scoring, with 23 out of 28 pairs of variables showing a significant difference. This supports our assumption that the eight rubric variables measure distinct, independent qualities of student work.

**Question 2: How reliable and consistent were the rubric scores given by different raters?**

We used one-way mixed, absolute-agreement, average-measures intraclass correlation to estimate inter-rater reliability of measurement. We also calculated the rate of agreement. Table 8 summarizes the results for both measures and for the first (not consistently normed) and second (normed) sets of scores. As the table shows, the first round of scores generally showed higher agreement between raters, but there was no clear difference in inter-rater reliability as measured by intraclass correlation.

In the second round of scoring, intraclass correlation was in the fair range as defined by Domenic Cicchetti—that is, intraclass correlation = 0.51—for the aggregate rubric score (variable R9). Intraclass correlations for individual rubric facets (variables R1–R8) ranged from –0.23 (poor) for R1 (multiple viewpoints) to 0.71 (good) for R2 (multiple source types). When we separated out the group of cases (n = 52) that had rater agreement within eight points for R9 and therefore did not require a third rater for resolution, there was, predictably, a marked improvement, with intraclass correlation = 0.86 (excellent) for R9 and correlations ranging from fair to excellent for individual facets. In the group of cases (n = 25) that required a third reading, again unsurprisingly, the correlations for total score and most rubric facets were lower than for the group, with intraclass correlation = 0.49 (poor) for R9.

A simpler but less sensitive measure, rate of agreement within one point (or within eight points for the aggregate score), was also used, since it was the basis for our method of resolving disparities in score. The rate of agreement between raters on the overall score (within a predetermined parameter of eight points) was 68 percent, slightly below our agreed-upon goal of at least 70 percent agreement. For the individual rubric facets, agreement within one point was lowest for R3 (extent of sources, 60 percent) and highest for R2 (multiple source types, 79 percent). As noted in the section on “Statistical Analysis Process,” rate of agreement and intraclass correlation measure distinct aspects of inter-rater reliability and agreement, and higher rates of agreement do not always coincide with higher intraclass correlation scores.

**Question 3: Do correlations exist between IL and contact with the library?**

To test for relationships between rubric scores and library contact, nonparametric correlation tests were run between the individual and aggregate contact variables (C1–C6) and the individual and aggregate rubric variables (R1–R9). The results of these tests are summarized in Table 9.
Table 7.
Difference in scores between pairs of rubric variables*

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
<th>R6</th>
<th>R7</th>
<th>R8</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td></td>
<td>Z = 5.41†</td>
<td>Z = 6.41†</td>
<td>Z = 5.38†</td>
<td>Z = 4.04†</td>
<td>Z = 0.88</td>
<td>Z = 1.88</td>
<td>Z = 2.13‡</td>
</tr>
<tr>
<td>R2</td>
<td>Z = 5.41†</td>
<td></td>
<td>Z = 0.29</td>
<td>Z = 6.66†</td>
<td>Z = 6.14†</td>
<td>Z = 5.59†</td>
<td>Z = 3.44†</td>
<td>Z = 3.42‡</td>
</tr>
<tr>
<td>R3</td>
<td>Z = 5.44†</td>
<td>Z = 0.26</td>
<td></td>
<td>Z = 6.89†</td>
<td>Z = 6.28†</td>
<td>Z = 5.73†</td>
<td>Z = 3.46†</td>
<td>Z = 3.69‡</td>
</tr>
<tr>
<td>R4</td>
<td>Z = 5.38†</td>
<td>Z = 6.66†</td>
<td>Z = 6.89†</td>
<td></td>
<td>Z = 2.95†</td>
<td>Z = 4.28†</td>
<td>Z = 5.29†</td>
<td>Z = 5.20‡</td>
</tr>
<tr>
<td>R5</td>
<td>Z = 4.04†</td>
<td>Z = 6.14†</td>
<td>Z = 6.28†</td>
<td>Z = 2.95†</td>
<td></td>
<td>Z = 2.70†</td>
<td>Z = 4.35†</td>
<td>Z = 4.48‡</td>
</tr>
<tr>
<td>R6</td>
<td>Z = 0.88</td>
<td>Z = 5.59†</td>
<td>Z = 5.73†</td>
<td>Z = 4.28†</td>
<td>Z = 2.70†</td>
<td></td>
<td>Z = 2.69†</td>
<td>Z = 2.83‡</td>
</tr>
<tr>
<td>R7</td>
<td>Z = 1.88</td>
<td>Z = 3.44†</td>
<td>Z = 3.46†</td>
<td>Z = 5.29†</td>
<td>Z = 4.35†</td>
<td>Z = 2.69†</td>
<td></td>
<td>Z = 0.19</td>
</tr>
<tr>
<td>R8</td>
<td>Z = 2.13‡</td>
<td>Z = 3.42‡</td>
<td>Z = 3.69†</td>
<td>Z = 5.20†</td>
<td>Z = 4.48‡</td>
<td>Z = 2.83†</td>
<td>Z = 0.19</td>
<td></td>
</tr>
</tbody>
</table>

*This table shows the results of paired Wilcoxon signed-rank tests, which determine significant differences between each pair of rubric variables in the second set of resolved median scores.
†Score is statistically significant, p < .05.
‡Score is highly statistically significant, p < .01.
Both Spearman’s rho and Kendall’s tau found a low, positive, highly significant correlation between aggregate library contact (variable C6) and aggregate rubric score (variable R9). This suggests that, overall, increased library contact had some positive impact on IL skills. Aggregate library contact (variable C6) also showed weak, positive correlations with all the individual rubric items (variables R1–R8). Of these, only R5 (understanding and using sources in context) showed highly significant correlations for both Spearman’s and Kendall’s tests. This result may indicate that various forms of library contact had a cumulative positive impact on students’ ability to consider and communicate the context of passages cited in their papers. Moderately significant correlations were found in at least one of the two tests for all but one of the remaining rubric variables.

The individual library contact interventions (variables C1–C5) were also tested for correlation with aggregate rubric score (variable R9) and with individual rubric facets (variables R1–R8). Variable C2 (frequency of website use) showed low, positive, highly significant correlations with R9 and with three individual rubric variables, R5 (under-
Table 9.
Correlations between library contact (C) and rubric (R) scores

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spearman’s rho (75)*</td>
<td>Kendall’s tau (75)*</td>
<td>Spearman’s rho (75)</td>
<td>Kendall’s tau (75)</td>
<td>Spearman’s rho (75)</td>
<td>Kendall’s tau (75)</td>
</tr>
<tr>
<td>R1</td>
<td>0.16</td>
<td>0.13</td>
<td>0.28†</td>
<td>0.23†</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td>R2</td>
<td>0.21</td>
<td>0.23</td>
<td>0.22</td>
<td>0.18</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>R3</td>
<td>0.21</td>
<td>0.15</td>
<td>0.14</td>
<td>0.11</td>
<td>-0.12</td>
<td>-0.11</td>
</tr>
<tr>
<td>R4</td>
<td>0.18</td>
<td>0.15</td>
<td>0.26†</td>
<td>0.22†</td>
<td>-0.07</td>
<td>-0.06</td>
</tr>
<tr>
<td>R5</td>
<td>0.19</td>
<td>0.15</td>
<td>0.37†</td>
<td>0.31†</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>R6</td>
<td>0.03</td>
<td>0.02</td>
<td>0.28†</td>
<td>0.22†</td>
<td>-0.06</td>
<td>-0.05</td>
</tr>
<tr>
<td>R7</td>
<td>0.08</td>
<td>0.07</td>
<td>0.51†</td>
<td>0.42†</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>R8</td>
<td>0.09</td>
<td>0.06</td>
<td>0.35†</td>
<td>0.29†</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>R9</td>
<td>0.15</td>
<td>0.10</td>
<td>0.39†</td>
<td>0.30†</td>
<td>0.08</td>
<td>0.06</td>
</tr>
</tbody>
</table>

*Spearman’s rho and Kendall’s tau indicate the closeness of the relationship between two variables.
†Score is statistically significant, p < .05.
‡Score is highly statistically significant, p < .01.
Potholes and Pitfalls on the Road to Authentic Assessment

standing of context), R7 (citation format), and R8 (documentation of sources). Moderately significant positive correlations were found between C2 and R1 (multiple viewpoints), R4 (assessment of sources), and R6 (integration of sources). This suggests that frequent use of the library website, even in the absence of other forms of library contact, resulted in increased levels of IL skills in general and of certain specific abilities.

None of the other individual contact variables showed a highly significant correlation with the rubric variables, although there were moderately significant, low, positive correlations between the number of bibliographic instruction sessions attended (variable C5) and two of the rubric facets, R5 (understanding of context) and R6 (integration of sources). (These correlations were not found in the first set of scores.) In the case of variable C4 (appointment with a librarian), we cannot say whether a relationship exists, since only 3 of the 77 participants in our sample had made such an appointment.

The general patterns of relationship between library contact and rubric score are similar between the first and second sets of scores. Although this similarity does not compensate for the reliability issues that throw both sets of scores into question, it does reassure us that our assessment reflects a real underlying relationship between library contact and IL.

Question 4: In what IL skills were the participants strongest and weakest?

Comparing the median resolved scores for individual IL variables (R1–R8), we found that our sample scored highest on variety of source types and extent of sources (variables R2, multiple source types, and R3, extent of sources, respectively; median = 4 for both), followed by citation format and documentation (R7 and R8, median = 3.5 for both), then by use of multiple viewpoints, understanding of context, and integration of sources (R1, multiple viewpoints; R5, understanding of context; and R6, integration of sources; median = 3 for both). Assessment of sources (R4) received the lowest scores (median = 2.5). The Wilcoxon tests, which we conducted as a test of independence of variables (see Question 1), appear to confirm the significance of these differences.

Scores from the first round of scoring showed the same general pattern. In both rounds, our sample tended to score highest on extent of sources and lowest on assessment of sources, with the remaining skills occurring in a similar order between them. While it is difficult to make a statistically meaningful comparison between the two sets, this similarity suggests general agreement between the two groups of raters. Medians and interquartile ranges (IQRs) for both sets of scores are shown in Table 2. Despite the methodological issues, this ranking of variables will provide a useful guideline for the library’s instructional efforts, since it helps identify areas where students struggle and may need additional or alternative outreach.

Question 5: Did any of the demographic factors have a significant relationship to IL?

We tested for significant differences in IL learning between groups based on demographic factors. Using the appropriate tests (summarized in Table 6), we found no correlations or significant differences in score between groups based on gender, age, number of semesters at Washington State University (WSU), or college major (variables D1, D4 and D5, and D6, respectively). The only significant differences found were related to transfer status.
For variable D2, transfer status, when a Mann-Whitney U test was used to compare rubric scores of transfer students (n = 65) and non-transfer students—that is, students who started at WSU in their first year of college (n = 12)—a highly significant difference was found for variable R3 (extent of sources) and moderately significant differences in variables R1 (multiple viewpoints), R2 (multiple source types), R5 (understanding of context), and R9 (aggregate rubric score). Students who started college at WSU scored the higher median scores for all rubric facets. Statistical values appear in Table 10. These findings may indicate a need for additional or improved library outreach to transfer students, especially involving use of source materials.

As Table 10 shows, our findings on transfer status were similar but not identical for the first set of scores. In the first set, differences were also significant between transfer groups for variables R2 (multiple source types), R5 (understanding of context), and R9, but not for R1 (multiple viewpoints) or R3 (extent of sources). Non-transfer students received the higher median scores on all IL skills in both rounds of scoring, except in the case of R3 (extent of sources) for the first round of scores, where the medians were equal.

These findings raised a question about the relationship between contact with the library and the time a student has spent at WSU. Given significant relationships between aggregate library contact (variable C6) and most of the rubric variables (see Table 9 and Question 3, “Do correlations exist between IL and contact with the library?”), we wondered why the time spent at WSU (variables D4 and D5) had no effect on rubric scores. We believe that this is explained by the fact that aggregate library contact is based on five contact variables, only three of which might increase over time. The only individual contact variable that appeared to have a significant relationship to rubric variables was library website use, which has no clear relationship to a student’s time at WSU. The time spent in the library weekly is also unrelated to time at WSU. The relevant survey questions asked about the frequency of current contact, not about cumulative contact over time. Thus, while the time-related contact variables may have had some combined effect on rubric scores, it is not surprising that no difference could be found when groups were compared based on time at WSU. A student who frequently used the library space and the website might score high on library contact even in the first semester. While the results are understandable, the lack of a clear distinction between routine library contact and cumulative contact over time is a limitation of our study.
Table 10.
Comparison of rubric scores for transfer and non-transfer students

<table>
<thead>
<tr>
<th></th>
<th>First set of scores</th>
<th>Second set of scores</th>
<th>Mann-Whitney U†</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median for non-transfer students (n = 12)</td>
<td>Median for transfer students (n = 65)</td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>4.5 (IQR = 1)*</td>
<td>4 (IQR = 1.5)</td>
<td>324.0</td>
</tr>
<tr>
<td>R2</td>
<td>5 (IQR = 0.5)</td>
<td>4.5 (IQR = 1.25)</td>
<td>218.0†</td>
</tr>
<tr>
<td>R3</td>
<td>5 (IQR = 0.75)</td>
<td>5 (IQR = 1.5)</td>
<td>263.0</td>
</tr>
<tr>
<td>R4</td>
<td>3.75 (IQR = 0.875)</td>
<td>3 (IQR = 1)</td>
<td>324.0</td>
</tr>
<tr>
<td>R5</td>
<td>4.5 (IQR = 1.375)</td>
<td>3.5 (IQR = 1.5)</td>
<td>212.0†</td>
</tr>
<tr>
<td>R6</td>
<td>4 (IQR = 1.375)</td>
<td>3.5 (IQR = 1.5)</td>
<td>229.5‡</td>
</tr>
<tr>
<td>R7</td>
<td>4.25 (IQR = 2.25)</td>
<td>4 (IQR = 1.75)</td>
<td>283.5</td>
</tr>
<tr>
<td>R8</td>
<td>4.75 (IQR = 1.875)</td>
<td>4.5 (IQR = 1.5)</td>
<td>308.5</td>
</tr>
<tr>
<td>R9</td>
<td>36.75 (IQR = 8.75)</td>
<td>30 (IQR = 7)</td>
<td>188.0§</td>
</tr>
</tbody>
</table>

*R Interquartile range (IQR) refers to the range of the middle 50% between the lower quartile and the upper quartile of the sample.
†The Mann-Whitney U test is used to assess differences between groups.
‡Score is statistically significant, p < .05.
§Score is highly statistically significant, p < .01.
Rows in boldface are at least moderately significant (p < .05) for both sets of scores.
Notes


8. In retrospect, it might have been preferable to use a standardized score such as a z-score to calculate aggregate library contact, thus avoiding the use of arbitrary break points selected by the researchers. Such a method would lend equal weight to all contact variables.


15. Hallgren, “Computing Inter-Rater Reliability for Observational Data.”


22. Hallgren, “Computing Inter-Rater Reliability for Observational Data.”


24. Oakleaf, “Using Rubrics to Assess Information Literacy.”


27. Diller and Phelps, “Learning Outcomes, Portfolios, and Rubrics, Oh My!”
