THE ONLINE RESEARCH AND COMPREHENSION ASSESSMENT (ORCA) PROJECT

PROJECT REPORT #10

Hand Versus Auto Scoring In The Online Research and Comprehension Assessments

Cheryl Maykel, Elena Forzani, & Donald J. Leu, University of Connecticut

Julie Coiro, University of Rhode Island

and Jonna Kulikowich, Pennsylvania State University

Portions of this material are based on work supported by the U. S. Department of Education under Award No. R305G050154 and R305A090608. Opinions expressed herein are solely those of the authors and do not necessarily represent the position of the U. S. Department of Education, Institute of Educational Sciences.

We wish to express our appreciation to previous members of the New Literacies Research Lab for work with developing assessments: Ian O’Byrne, Heidi Everett-Cacopardo, Greg McVerry and Lisa Zawilinski. We also wish to thank Jonna Kulikowich, and Nell Sedransk for help with developing the assessments and the scoring system. We especially wish to thank Mike Hillinger and Mark Lorah from Lexicon Systems for their help with conceptualizing the assessments and with the important programming required to develop the auto-scoring system.
Hand Versus Auto Scoring In The Online Research and Comprehension Assessments

The goal of the ORCA (Online Research and Comprehension) Project (Leu, Kulikowich, Sedransk, & Coiro, 2009-14) was to develop valid, reliable, and practical assessments of online research and comprehension in two formats: ORCA-Multiple Choice and ORCA-Simulated. To meet the design criteria of practicality, all of the ORCA-Multiple Choice items were automatically scored in both the Pilot and Validation years of the project. During the Validation year we sought to increase the practicality of the ORCA assessments by expanding the extent of auto-scoring. This expansion included some of the constructed response (CR) items from the ORCA-Simulated, a more performance-based assessment format. The purpose of this technical report is to evaluate the accuracy of the automatically scored CR items used as part of the ORCA Project (Leu, Kulikowich, Sedransk, & Coiro, 2009-14).

Williamson et al. (2010) recommend two criteria to evaluate automated scoring systems: 1) “Automated scores are consistent with the scores from expert human graders,” and 2) “The way automated scores are produced is understandable and substantively meaningful.” This report attempts to address each of these two criteria in its evaluation of the ORCA auto-score program.

Rationale and Research Review

Generally speaking, multiple choice (MC) assessments allow educators to assess a wide range of factual knowledge in a short amount of time, while CR assessments allow for a greater depth of inquiry into a student’s knowledge in a specific area (Livingston, 2009). We may be able to gain a richer understanding of student knowledge through analyzing CR items than we can from interpreting the results of multiple choice assessments alone, since students show more of their thinking through the longer responses required of CR items, and because students have to
generate the correct answer without a list of potential responses from which to choose (Livingston, 2009). For these reasons, some argue that CR items are better suited for assessing higher-level thinking skills (Hancock, 1994; Martinez, 1999), such as critical evaluation and synthesis (Stiggins, 2005).

The ability to automatically score CR items has important potential in education. As more assessments, including those that measure the Common Core standards, become available in digital formats, more automated scoring programs can easily be used (Streeter, Bernstein, Foltz, & DeLand, 2011). However, scoring CR items, as opposed to scoring MC items, can take a significant amount of time from teachers’ busy days (Ventouras et al., 2010). CR items often require human, hand scoring as opposed to automated, computerized scoring (McClellan, 2010). As a result of the time required to both administer and score CR assessments, many educators now rely more on MC items (Bleske-Rechek, et al., 2007).

An assessment that allows students to provide open-ended responses while scoring those responses automatically and reliably may save teachers time while still providing a richer analysis of student thinking. Additionally, automated scoring systems are advantageous since they can produce scores that are more reliable and cost-effective than human scoring (Topol, Olson, & Roeber, 2011), and research supports the idea that automated scoring systems can be both accurate and efficient (Livingston, 2009).

**The Present Study**

The current investigation involved two graduate student expert scorers (herein referred to as the student experts) who were responsible for the initial and ongoing training and reliability checking of the undergraduate student hand-scorers throughout the Pilot and Validation years of the ORCA Project. A third individual, who is a researcher in the area of online reading and
research (herein referred to as the professional expert), was also involved in this investigation. The professional expert allowed for a blinded investigation of the accuracy of the auto-scoring system that was used, since the student experts completed all of the score comparisons. The professional expert executed a systematic method for selecting students from the larger ORCA sample for the present investigation, which, simply put, involved comparing the hand-scored scores of the two expert scorers to those generated by the auto-score function.

Method

Participants

The participants in the present study were selected from a larger sample of approximately 1,300 seventh grade students from two Northeastern states who participated in the Validation year of the ORCA Project (see Forzani & Maykel, 2013 for more information on the full sample of students). Eighty students from this sample were used in the present investigation.

Procedures

Online Reading Comprehension Assessments (ORCAs).

Four research topics that required students to read and conduct research using the Internet were developed in two different assessment formats (ORCA-Multiple Choice, or ORCA-MC, and ORCA-Simulated) as part of The ORCA Project (Leu, Kulikowich, Sedransk, & Coiro, 2009-2014). These assessments have previously demonstrated high levels of reliability and validity (see Kulikowich, J. M., Cui, W., Liu, C., & Leu, D. J., 2014). The students that were selected for the present analysis completed two assessments, one on each of two days. Students completed one ORCA-Multiple Choice assessment on one day, and one ORCA-Simulated assessment on the other day, in varying order.

Both formats were intended to be performance-based since they required students to
engage in an authentic Internet research scenario that was described to students. However, students only engaged in the research task in limited ways in the ORCA-MC, whereas students were more fully immersed in the research task in the ORCA-Simulated format. For example, in the ORCA-MC format, students merely had to select correct Internet search keywords from a set of four possible choices, whereas in the ORCA-Simulated, students had to actually conduct a search by typing appropriate keywords into a search engine. The present study investigated only scores from the ORCA-Simulated assessments.

The ORCA-Simulated assessments were created in an online environment and were intended to mimic the actual, open Internet, though all of the content provided within the online environment was controlled. These assessments required students to search for, select, and use websites from the project’s own search engine, called “Gloogle,” which was populated with a predetermined set of sites. Some of the sites were relevant to the research scenario, while others were intended to distract students, just as adds and irrelevant sites would in actual search engines. Students were guided through the research tasks by two programmed student avatars: “Brianna” and “Jordan.” Students were directed to read the information at different sites, take notes on what they learned, and share that information using various social media tools.

In all topics, students were presented with science research problems that focused on health and human body systems, an area common to many seventh grade science curricula. Two of the topics focused on the eyes, and two focused on the heart. The topics included energy drinks and heart health, snacks and heart health, video games and eye health, and cosmetic contact lenses and eye health. The assessments were framed around two types of research: “Learn More About (LMA)” and “Investigate Conflicting Claims (ICC).” Half of the topics presented the research problem to students via an email message from a school board president
(LMA scenarios) and half via a class wiki with a message from a teacher (ICC scenarios). The LMA scenarios asked students to write a short report about what they learned from their research. The ICC scenarios asked students to investigate both sides of the issue and take a position in a short report of their findings. One final distinction between the tasks is that half were considered restricted tasks, and asked students to search for a specific article, whereas the other half were classified as unrestricted, and simply asked students to find an article about the given topic. Table 1 summarizes the nature of the four research tasks used in this study.

Table 1

Overview of Research Tasks

<table>
<thead>
<tr>
<th>Topic</th>
<th>Body System</th>
<th>Type of Research Scenario</th>
<th>Type of Search Task</th>
<th>Final Communication Task Tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy Drinks</td>
<td>Heart</td>
<td>Learn More About</td>
<td>Restricted</td>
<td>Email</td>
</tr>
<tr>
<td>Snacks</td>
<td>Heart</td>
<td>Learn More About</td>
<td>Unrestricted</td>
<td>Email</td>
</tr>
<tr>
<td>Video Games</td>
<td>Eyes</td>
<td>Investigate Conflicting Claims</td>
<td>Unrestricted</td>
<td>Wiki</td>
</tr>
<tr>
<td>Contacts</td>
<td>Eyes</td>
<td>Investigate Conflicting Claims</td>
<td>Restricted</td>
<td>Wiki</td>
</tr>
</tbody>
</table>

Each assessment scenario included items that evaluated students’ ability to locate, evaluate, and synthesize information found during the research process, as well as items assessing students’ ability to communicate the results of their research via either email or wiki. Four items were devoted to each of the four skill areas; thus, students could earn up to 16 score points for each assessment (see Table 2). The Locate and Synthesize questions did not all appear in a linear sequence according to skill area. Instead, a more natural and logical sequence was used according to what made sense within the research task. For example, students were asked to first locate a web page, then synthesize the information at that webpage, then locate a second
web page, and then synthesize the information they found across both Websites 1 and 2 before locating a third web page. In this way, the Locate and Synthesize skills were interwoven throughout the assessment.

The Evaluate items, however, all appeared in sequence. This was a more logical arrangement, as these four items were all in reference to a single web page. The same was true of the four Communicate items, which were all drawn from a student’s single email message or wiki posting.

Table 2

ORCA Items by Skill Area

<table>
<thead>
<tr>
<th><strong>Reading to Locate Online Information</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Can the student locate the correct email message in an inbox or the correct section of a wiki?</td>
</tr>
<tr>
<td>2. Can the student use appropriate key words in a search engine?</td>
</tr>
<tr>
<td>3. Can the student locate the correct site in a set of search engine results?</td>
</tr>
<tr>
<td>4. Can the student identify correct website addresses in two different search tasks?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Reading to Evaluate Online Information</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Can the student identify the author of a website?</td>
</tr>
<tr>
<td>6. Can the student evaluate an author’s level of expertise?</td>
</tr>
<tr>
<td>7. Can the student identify an author’s point of view?</td>
</tr>
<tr>
<td>8. Can the student evaluate the reliability of a website?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Reading to Synthesize Online Information</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Can the student summarize an important element from one website?</td>
</tr>
<tr>
<td>10. Can the student summarize important elements from two websites?</td>
</tr>
<tr>
<td>11. Can the student summarize important elements from a second set of two websites?</td>
</tr>
<tr>
<td>12. Can the student summarize important elements from the websites in the research task to develop an argument?</td>
</tr>
</tbody>
</table>
Writing to Communicate Online Information

13. **Email Task:** Can the student include the correct address line in an email message?  
**Wiki Task:** Can the student make a wiki entry in the correct location?

14. **Email Task:** Can the student include an appropriate subject line in an email message?  
**Wiki Task:** Can the student use descriptive voice in an informational wiki?

15. **Email Task:** Can the student include an appropriate greeting in an email message to an important, unfamiliar person?  
**Wiki Task:** Can the student include an appropriate heading for a new wiki entry?

16. **Email Task:** Can the student compose and send a well-structured, short report of their research, including sources, in an email?  
**Wiki Task:** Can the student compose and post a well-structured, short report of their research, including sources, in a wiki?

Of the 16 score points in each assessment, five of the ORCA-Simulated items in the wiki topics (Video Games and Contacts) and seven of the items in the ORCA-Simulated email topics (Energy Drinks and Heart Healthy Snacks) were automatically scored (see Table 3) using Microsoft Excel formulas. Items that were auto-scored were those that required a particular response (or set of responses) in order to be scored as correct (e.g., sending a specific URL or clicking on a particular link). For each of these items, an Excel formula was created to determine if a student’s response was correct or incorrect. The formulas allowed for certain, pre-specified variations in spelling where appropriate and, in some cases, required students to include multiple parts of a correct response to receive credit. In contrast, items that did not require a specific response, but instead could have many possible correct answers, were hand scored.
### Auto-scored Items Evaluated in This Study

<table>
<thead>
<tr>
<th>Items from Email ORCAs</th>
<th>Question the Assessment Item Sought to Answer</th>
<th>Scoring Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Locate 1</strong></td>
<td>Can the student locate the correct email message in an inbox?</td>
<td>For <strong>Energy Drinks</strong>, the student must click on the email from Ms. Torres with the subject line: Energy Drinks and the Heart.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>For <strong>Snacks</strong>, the student must click on the email from Ms. Ortiz with the subject line: Healthy School Snacks.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>For <strong>Video Games</strong>, the student must click on the “Video Gaming” wiki page link from the wiki homepage.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>For <strong>Contacts</strong>, the student must click on the “Contacts” wiki page link from the wiki homepage.</td>
</tr>
<tr>
<td><strong>Locate 2</strong></td>
<td>Can the student use appropriate keywords in a search engine?</td>
<td>For <strong>Energy Drinks</strong>, the student must search for the specific article “Energy Drinks Jolt the Heart” in the search engine.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>For <strong>Snacks</strong>, the student must search for &quot;snacks&quot; and either &quot;heart&quot; or &quot;healthy.&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>For <strong>Video Games</strong>, the student must search for &quot;video games”, and either &quot;help” or &quot;helpful&quot;, or &quot;eyes&quot; or &quot;eyesight&quot;.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>For <strong>Contacts</strong>, the student must search for the specific article “Colored Contacts Can Trick, Not Treat for Halloween” in the search engine.</td>
</tr>
<tr>
<td><strong>Locate 3</strong></td>
<td>Can the student locate the correct site in a set of search engine results?</td>
<td>For <strong>Energy Drinks</strong>, the student must click on this specific article from the search results list: <a href="http://www.orcaproject.net/ce_web/sites/WebMD/L1-66_filename.html">http://www.orcaproject.net/ce_web/sites/WebMD/L1-66_filename.html</a>.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>For <strong>Snacks</strong>, the student must click on one of these three links from the search results list: <a href="http://www.orcaproject.net/ce_web/sites/healthy-eating-made-easy/L5-11_heart-healthy-snacks.html">http://www.orcaproject.net/ce_web/sites/healthy-eating-made-easy/L5-11_heart-healthy-snacks.html</a>.</td>
</tr>
</tbody>
</table>
http://www.orcaproject.net/ce_web/sites/HealthySchoolSna
cks/L5-59_filename.html, or
http://www.orcaproject.net/ce_web/sites/healthcentral/L5-
30_filename.html.

For Video Games, the student must click on one of these
links from the search results list:
http://www.orcaproject.net/ce_web/sites/LiveScience/L3-
6_filename.html,
http://www.orcaproject.net/ce_web/sites/Punchng/L3-
72_filename.html, or
http://www.orcaproject.net/ce_web/sites/Methods_Of_Heali
ng/L3-34_filename.html.

For Contacts, the student must click on this specific
article from the search results list:
http://www.orcaproject.net/ce_web/sites/Consumer_Affairs/
L7-1_filename.html.

Can the student identify correct website
addresses in two different search
tasks?

For Energy Drinks, the student must share these two links:
http://www.orcaproject.net/ce_web/sites/WebMD/L1-
66_filename.html and
http://www.orcaproject.net/ce_web/sites/Bostondotcom/L1-
28_filename.html.

For Snacks, the student must share any two of these three
links:
http://www.orcaproject.net/ce_web/sites/healthy-eating-
made-easy/L5-11_heart-healthy-snacks.html,
http://www.orcaproject.net/ce_web/sites/HealthySchoolSna
cks/L5-59_filename.html, or
http://www.orcaproject.net/ce_web/sites/healthcentral/L5-
30_filename.html.

For Video Games, the student must send one of these three
links for the first site:
http://www.orcaproject.net/ce_web/sites/LiveScience/L3-
Punchng/L3-72_filename.html, or
http://www.orcaproject.net/ce_web/sites/Methods_Of_Heali
ng/L3-34_filename.html, and one of these three links for the
second site: http://www.ABCNews.com/L3-
15_filename.html, http://www.HubPages.com/L3-
1_filename.html, or http://www.About.com/L3-
22_filename.html.
For **Contacts**, the student must send these two links:  
[http://www.orcaproject.net/ce_web/sites/Consumer_Affairs/L7-1_filename.html](http://www.orcaproject.net/ce_web/sites/Consumer_Affairs/L7-1_filename.html) and  

**Evaluate 1**  
Can the student identify the author of a website?  

For **Energy Drinks**, the student must indicate Tim Maxey as the author. Variations accepted by the auto-score system include Tim, Mr. Maxey, Tim Maxie, Tim Maxxie, and Tim Maxxey.

For **Snacks**, the student must indicate Melanie Thomassian as the author. Variations accepted by the auto-score system include Melanie and Ms. Thomassian.

For **Video Games**, the student must indicate Troy Bedinghaus as the author. Variations accepted by the auto-score system include Troy, Mr. Bedinghaus, and Troy Bedinghaus, O.D.

For **Contacts**, the student must indicate Julia Dilday as the author. Variations accepted by the auto-score system include Julia and Ms. Dilday.

---

**Email:** Can the student include the correct address line in an email message?  

For **Energy Drinks**, the student must send the email to  
[kmarin@AlbroSchools.edu](mailto:kmarin@AlbroSchools.edu).

For **Snacks**, the student must send the email to  
[mjefferson@AlbroSchools.edu](mailto:mjefferson@AlbroSchools.edu).

**Wiki:** Can the student make a wiki entry in the correct location?  

For **Wiki ORCAs:**  
The student must type a wiki entry either 1) below the other sections if no heading or 2) anywhere in the wiki below the Mr. Henry message and intro if he/she does have a heading.

---

**Email:** Can the student include an appropriate subject line in an email message?  

For **Energy Drinks**, the student must include "energy drinks" and either "heart" or "health" in the email subject line.

For **Snacks**, the student must include “snacks” and either “heart” or “health” in the email subject line.
For Wiki ORCAs, the student must compose an informational wiki entry in a descriptive voice (i.e., without the use of “I” or other reference to personal experiences).

In both the wiki and the email ORCAs, the automatically scored items included all four of the Locate items and the first Evaluate item (determining the author of a web page). In addition to these, the first two email Communicate items (email recipient and email subject line) were also automatically scored (see Table 3). However, the first two wiki Communicate items were not automatically scored. This was a result of the nature of the responses for these items. The first two wiki Communicate items address the questions: “Does the student make a wiki entry in the correct location?” and “Does the student use descriptive voice in an informational wiki?” These items required human judgment to be scored accurately.

**ORCA Scoring.**

An auto-capture system recorded students’ performance in online reports that could be accessed through the ORCA administration site. All of the ORCA-MC responses were automatically scored, as were some of the items from the ORCA-Simulated. The remaining items from the ORCA-Simulated format were hand-scored by a team of four undergraduate student scorers (one for each research scenario) who were trained by two graduate student experts prior to beginning scoring. This scoring was done as part of the larger ORCA research project. The student experts checked the work of the scorers at regular intervals throughout the scoring process. Any scorers who did not meet 90% inter-rater agreement with the student experts were retrained and retested before continuing scoring.
**Training of the Student Experts.**

Although the two expert scorers already had significant experience with the ORCA items, they participated in a training and testing regimen as part of this investigation that was similar to that expected of the undergraduate scorers before they began scoring responses. After undergoing two training sessions that involved scoring five reports from each of the four ORCA-Simulated topics (a total of 20 reports in each session), the two student experts completed a third, test set of 20 reports (equivalent to another set of five student reports from each topic). The student experts met 90% reliability or greater with one another on all score points and were, therefore, deemed able to continue with the present study of comparing their scoring to that of the auto-score program.

**Data Selection and Analysis.** The professional expert first divided all of the ORCA-Simulated Validation year data into four files, one for each of the four research scenarios. Then, within each file, she arranged the students by their ORCA total scores into quintiles. Next, in order to include a representative group of students in the analyses, the professional expert chose 20 students from each file by selecting two students from each state, within each quintile. Finally, the professional expert prepared a list of students’ reports for each student expert to score. She did this by compiling a representative set of 40 students’ reports from the larger selection for each student expert to score (see Figure 1 below). Therefore, a total of 80 students’ reports were used for the comparison of hand- to auto-scoring in this study. The email assessments in these reports contained seven items that were autoscored; the wiki assessments in these reports contained five items that were autoscored. In the 80 reports, a total of 480 items were autoscored and included in this evaluation.

After completing the scoring, the student experts then checked each other’s scoring
against that of the auto-score feature by item (e.g., Locate 2, Evaluate 1), within each of the four scenarios. Given that 10 reports were chosen for this analysis from each of the four scenarios for each student expert to score, the percent agreement between hand- and auto- scoring was simply calculated by dividing the number of correct responses by 10 for each item, within each of the topics, for each hand scorer.

Figure 1. Example sample distribution

![Example for Energy Drinks](image)

**Example for Video Games:**

<table>
<thead>
<tr>
<th>Quintile</th>
<th>State 1</th>
<th>State 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 student</td>
<td>1 student</td>
</tr>
<tr>
<td>2</td>
<td>1 student</td>
<td>1 student</td>
</tr>
<tr>
<td>3</td>
<td>1 student</td>
<td>1 student</td>
</tr>
<tr>
<td>4</td>
<td>1 student</td>
<td>1 student</td>
</tr>
<tr>
<td>5</td>
<td>1 student</td>
<td>1 student</td>
</tr>
</tbody>
</table>

A representative 10 of the students chosen for re-scoring from each topic were assigned to each scorer.

**Results**

Of the total of 480 items from 80 student reports that were reviewed as part of this analysis, there were just five instances of disagreements between the student experts who hand-scored the items and the auto-score program. Thus, the total percentage of agreement between autoscoring and hand scoring for these items was nearly 99% (475/480 x 100 = 98.96). The tables below (Tables 4, 5, 6, & 7) indicate the percentage of agreement between the student expert/hand-scorers, and the auto-score program. Again, these percentages were calculated for each item, within each topic, by dividing the number of items in agreement between the two scores by 10 (the total number of items), and then multiplying that number by 100 to turn it into
a percent. Therefore, the percentages represent the inter-rater agreement between hand and auto
scores.

Table 4

Percent Inter-rater Reliability for Energy Drinks

<table>
<thead>
<tr>
<th>Error Made By</th>
<th>Locate 1</th>
<th>Locate 2</th>
<th>Locate 3</th>
<th>Locate 4</th>
<th>Evaluate 1</th>
<th>Communicate 1</th>
<th>Communicate 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student Expert 1 v Auto-Score Program</td>
<td>100%</td>
<td>90%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Student Expert 2 v Auto-Score Program</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>NA</td>
<td>Student Expert 1/Hand-Scorer</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

The single disagreement in Energy Drinks was a result of a student entering “Energy
drink jolt for the heart” into the search engine (see Table 4 above). In this case, the hand scorer
mistakenly marked the item as correct, and the program accurately marked it as incorrect. For
this item, Locate 2 in this topic, students must enter the title of the article exactly in order to earn
credit, and misspellings are not accepted. Given that this is classified as a restricted task, the
actual title of the article, “Energy drinks jolt the heart” must be entered word for word into the
search engine in order for the student to receive credit. Therefore, the student’s insertion of the
word “for” is what made their response incorrect given the scoring criteria. Overall, with seven
automatically scored items in 20 reports, 140 items were evaluated for reliability, with
disagreement between hand and auto scoring on just one item. Therefore, the total reliability for
energy drinks was over 99% (139/140 x 100 = 99.3).

Table 5

Percent Inter-rater Reliability for Snacks

<table>
<thead>
<tr>
<th></th>
<th>Student Expert 1 v Auto-Score Program</th>
<th>Student Expert 2 v Auto-Score Program</th>
<th>Error Made By</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locate 1</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Locate 2</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Locate 3</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Locate 4</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Evaluate 1</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Communicate 1</td>
<td>100%</td>
<td>90%</td>
<td>Student Expert 2/Hand-Scorer</td>
</tr>
<tr>
<td>Communicate 2</td>
<td>100%</td>
<td>80%</td>
<td>Student Expert 2/Hand-Scorer (2) &amp; Auto-Score Program (1)</td>
</tr>
</tbody>
</table>

In Snacks, there were three items of disagreement between one of the student experts and the program. In two of these cases the student expert scored incorrectly, and in the other the program scored incorrectly (see Table 5 above). For the two items that the student expert marked incorrectly, Communicate 1 and Communicate 2, the student taking the ORCA had not given a response for the item, though the student expert accidentally scored the two items as correct. The program, however, accurately scored these items as incorrect. These errors demonstrate an important concept: very well trained hand-scorers are still human beings, and all human beings are capable of making simple mistakes. A computer scoring system is less likely to make these types of mistakes.

The third item of disagreement, Communicate 2 from a different student report, was an
instance of the program not picking up on a typo within an item that typos are considered acceptable. In this case, the student typed “Heart Healthy sancks,” and the misspelling of the word “snacks” is what caused the program to mark that item as incorrect. The auto score feature only allows for a predetermined set of spelling variations that were entered into the formula, even though any spelling variation that is still understandable to the scorer is technically considered correct in this unrestricted task.

Overall, with seven automatically scored items in 20 reports, 140 items were evaluated for reliability in snacks. We found disagreements between the hand and auto scoring on three items. Therefore, the total reliability for snacks was over 97% (137/140 x 100 = 97.9).

Table 6

*Percent Inter-rater Reliability for Video Games*

<table>
<thead>
<tr>
<th></th>
<th>Student Expert 1 v Auto-Score Program</th>
<th>Student Expert 2 v Auto-Score Program</th>
<th>Error Made By:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locate 1</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Locate 2</td>
<td>90%</td>
<td>100%</td>
<td>Auto-Score Program</td>
</tr>
<tr>
<td>Locate 3</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Locate 4</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Evaluate 1</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
</tbody>
</table>

The single disagreement in Video Games was another example of the program not picking up on a misspelling for a response to the item Locate 2. The student expert marked it as correct, as it should be according to the scoring criteria, and the program marked it as incorrect (see Table 5 above). For this item in Video Games, an unrestricted task, spelling variations are acceptable. The student was required to enter the terms “video games” and either “help” or
“helpful and “eyes” or “eyesight.” Instead, the student actually entered “vidoe games that help your eyes.” As you can see, the student had all of the necessary components to the response; however, the misspelling of the word “video” caused this response to be marked as incorrect by the program.

Overall, with five automatically scored items in 20 reports, 100 items were evaluated for reliability, with disagreement between hand and auto scoring occurring on just one item. Therefore, the total reliability for video games was exactly 99% (99/100 x 100 = 99).

Table 7

*Percent Inter-rater Reliability for Contacts*

<table>
<thead>
<tr>
<th></th>
<th>Student Expert 1 v Auto-Score Program</th>
<th>Student Expert 2 v Auto-Score Program</th>
<th>Error Made By:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locate 1</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Locate 2</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Locate 3</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Locate 4</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
<tr>
<td>Evaluate 1</td>
<td>100%</td>
<td>100%</td>
<td>NA</td>
</tr>
</tbody>
</table>

As indicated in Table 6 above, there were no disagreements between the student experts and the program in the Contacts topic reports. Therefore, among the 100 items evaluated from contacts, the overall agreement between hand and auto-scoring was 100%.

**Discussion**

Recall that Williamson and colleagues (2010) pose two criteria to evaluate automated scoring systems: 1) “Automated scores are consistent with the scores from expert human graders,” and 2) “The way automated scores are produced is understandable and substantively
meaningful.” In this report, we have attempted to provide the information needed to evaluate the reliability of the ORCA scoring system using these two criteria. The information in this report, suggests that the auto-scoring system for the five items in wiki tasks and the seven items in email tasks for the ORCA-Simulated yielded scores that are both reliable and appropriate for scoring the ORCAs. Table 3 shows which items were automatically scored, and all of the responses that were accepted by the scoring system. Though the final ORCA scoring system, in its current form, no longer uses Excel, a similar coding method using the same determinants is employed by the online system.

As the results of this report show, the consensus among scores given by the student experts and those that were automatically assigned is very high. After the two student experts were trained to 90% reliability with one another, they each scored 10 representative student reports from each of the four topic areas. Then, they each compared one another’s scores to those generated by the auto-score feature. These results show that, overall, there were three instances of human error and only two instances of error on the part of the auto-score program. Both of the program errors were the result of misspellings by the students taking the assessments. Although the Excel formulas that were developed for the auto-score program were designed to accept multiple spelling variations for several items, it is simply not possible to conceive of and account for every potential misspelling in the formulas. Overall, 480 items across 80 student reports were included in these analyses, and there were only five total instances where hand scoring was not consistent with auto-scoring, which resulted in an overall inter-rater reliability of nearly 99%. Total agreement for each of the four topic areas ranged from 97% to 100%.

Therefore, the auto-score feature that was used in the Validation year of the ORCA project is considered to be a relatively reliable scoring method for the CR items that it was used
to score. Furthermore, this analysis is promising to others who may wish to use a similar method in future investigations. Although not all of the CR items in the ORCAs can be automatically scored at this time, the ability to automatically score the five to seven items that are currently available for auto-scoring in each topic area can save teachers a significant amount of time, especially when considering that teachers are likely to have several student reports to score.

This report describes an investigation that assessed the reliability of an automated scoring feature that was used with several CR items as part of a large assessment project. Performance-based tasks, which typically require constructed responses from students, can be beneficial when they model good instruction and learning (Darling-Hammond & Pecheone, 2010). Also, the ability to quickly and accurately score student-generated responses can be beneficial to teachers who wish to take advantage of the convenience of a multiple choice assessment without losing information that can only be gained from open-ended responses (Livingston, 2008).
References


