Using The Psychometric Characteristics Of Multiple-Choice, Open Internet, And Closed (Simulated) Internet Formats To Refine The Development of Online Research and Comprehension Assessments In Science: Year Three of The ORCA Project

Donald J. Leu, University of Connecticut
Julie Coiro, University of Rhode Island
Jonna Kulikowich, Pennsylvania State University
Weiwei Cui, National Institute of Statistical Sciences

Please do not quote without permission of the authors.

Portions of this material are based upon work supported by the U.S. Department of Education under Award No. R305G050154 and No. R305A090608. Opinions expressed herein are solely those of the author and do not necessarily represent the position of the U.S. Department of Education.

Paper presented at the Literacy Research Association Conference. San Diego, CA
November 30, 2012
Abstract
The Online Research and Comprehension Assessment (ORCA) Project (Leu, Kulikowich, Sedransk, & Coiro, 2009) is taking place over five years with over 80 schools in several states along the East Coast of the United States. Its goal is to develop and evaluate reliable, valid, and practical assessments of online research and comprehension ability among seventh grade students in science.

This paper summarizes the developmental results, to date, of a large pilot study. The pilot study took place across two states and was conducted to evaluate the psychometric properties of 24 different LESC (Locate, Evaluate, Synthesize, Communicate) assessments, developed over the past several years. Using unidimensional scaling of first LESC scores, it evaluated the psychometric properties of these assessments within three different assessment formats: multiple-choice, open Internet, and closed (simulated) Internet assessments. Participants included representative samples of 1,129 seventh grade students from each of two states in the northeastern U.S. Students were selected using a stratified random sampling plan designed to represent all seventh-grade students in each state.

The evaluation was conducted in order to select and revise a smaller sample of assessments for another round of studies in Year Four. Specifically three research questions were addressed:

1. What are the reliability estimates of scores for 24 LESCs that vary by item format (i.e., Multiple-Choice, Closed, and Open) and by science topic?
2. What are the reliability estimates of scores for the three separate formats?
3. How much variance is explained by each of the 24 LESC scales?

Unidimensional results show that the overall set of LESCs are generally reliable and valid though substantial variation existed between the LESCs, though less so between the formats. We used these results to narrow next year’s LESCs to only two formats (multiple choice and closed) and four of the best performing science topics:

1. How do energy drinks affect heart health? (heart)
2. How can snacks be heart healthy? (heart)
3. Do cosmetic contact lenses harm your eyes? (eyes)
4. Do video games harm your eyes? (eyes)

We also used our analyses of individual item performance to select individual items within these eight LESCs for revision. We are now beginning an analysis of multidimensional scaling to inform our work in preparation for the Year Four assessments.
Using The Psychometric Characteristics Of Multiple-Choice, Open Internet, And Closed (Simulated) Internet Formats To Refine The Development of Online Research and Comprehension Assessments In Science: Year Three of The ORCA Project

An important question today is: How might we best assess students’ ability to conduct online research, comprehend information, and write a short report of their research? This question is important since success in college and career settings now requires much more than the offline reading skills required in the past for both literacy and learning (International Reading Association, 2009; Partnership for 21st Century Skills, 2006). Important to college and career success are the additional online research and comprehension skills required to learn about a topic, locate information, evaluate information, and communicate what you have learned to others on the Internet. These skills are now included in national standards (e.g., Common Core State Standards Initiative [CCSS], 2012), national curricula (e.g. Australian Curriculum, Assessment and Reporting Authority, [ACARA], n. d.) and international assessments of adult literacy and numeracy (e.g. the PIAAC Expert Group on Problem Solving in Technology-Rich Environments, 2009).

Despite their importance, we are only beginning work to develop valid and reliable assessments of online research and comprehension that might be used to inform instruction in the additional literacy and learning skills the Internet requires (Organization for Economic Co-operation and Development [OECD], 2011). The purpose of this study is to report on the performance characteristics of assessments developed to estimate seventh grade students’ ability to conduct online research and comprehend information in science. It presents results from the third year of the Online Research and Comprehension Assessment (ORCA) project (Leu, Kulikowich, Sedransk, & Coiro, 2009), an IES-funded, developmental, assessment grant. The
goal of this project is to develop valid, reliable, and practical assessments of students’ ability to conduct research online and to write a short report of their results. The analyses presented in this paper were used to inform the further development of these assessments, in preparation for a final round that is taking place this year with representative samples of seventh graders in two northeastern states in the U.S.

**Theoretical Framing and Research Perspectives**

**New Literacies Theory**

This study is informed by New Literacies theory (Coiro, Knobel, Lankshear, & Leu, 2008). Central to this notion is that literacy has become deictic; it is in a period of rapid and continuous change (Leu, 2000). Literacy is not just new today; it is new every day as new technologies, with new social practices and new affordances for their use, continuously appear.

How can we develop adequate theory to direct research when the object we seek to study is itself ephemeral, continuously being redefined by a changing technological context? Recently, a dual-level theory of New Literacies has been proposed (Leu, Kinzer, Coiro, Castek, & Henry, in press), suggesting it may be productive to conceive of New Literacies theory on two levels: uppercase (New Literacies) and lowercase (new literacies). *New Literacies*, as the broader concept, benefits from work taking place in the multiple, lowercase dimensions of *new literacies*. Lowercase theories explore a rapidly changing and specific area of new literacies, such as the social communicative transactions occurring with text messaging (e.g., Lewis & Fabos, 2005), or a focused disciplinary base, such as the semiotics of multimodality in online media (e.g., Kress, 2003). Common findings across multiple perspectives may then be integrated into the broader New Literacies theory that is likely, as a result, to be more stable. The greater stability of an
upper-case New Literacies theory may provide theoretical direction to inform the more rapidly changing contexts at lowercase levels.

A common finding across many areas of lowercase, new literacies research, and thus of upper-case New Literacies theory, is that the Internet makes possible new online social practices with new technologies such as text messaging or chat, wikis, email, search engines, and social networks (Greenhow, Robelia, & Hughes, 2009; Lankshear & Knobel, 2006). We used this idea by situating our assessments within a social network that included new social practices associated with each of these technologies.

The New Literacies of Online Research and Comprehension

The new literacies of online research and comprehension (Coiro, 2003; Leu, et al., 2011; Leu, Kinzer, Coiro, Castek, & Henry, in press) is one of many lowercase theories of new literacies. It seeks to describe what happens when we conduct research and read online. It suggests that at least five processing practices occur during online research and comprehension: 1) reading to define important questions or problems (Leu, Kinzer, Coiro, & Cammack, 2004); 2) reading to locate information (Bilal, 2000; Guinee, Eagleton, & Hall, 2003); 3) reading to evaluate information (Sanchez, Wiley, & Goldman, 2006); 4) reading to synthesize information (Goldman, Wiley, Goldman, Graeser, Sanchez, Ash, & Hemmerich, 2009; Leu, et al., in press; Jenkins, 2006); and 5) reading and writing to communicate information (Greenhow, Robelia, & Hughes, 2009). Within these five areas reside the skills, strategies, and social practices that are distinctive to online research and comprehension as well as others that are also important for offline reading in a complex layering of both old and new that we are still seeking to fully understand.
**Reading to define important questions or problems.** We read on the Internet to solve problems and answer questions. How a problem is framed or how a question is understood is a central aspect of online reading comprehension. Work by Taboada and Guthrie (2006) within traditional texts suggests that reading initiated by a question differs in important ways from reading that does not. In this study, we varied questions by using two types: “learn more about” and “take a position” questions.

**Reading to locate online information.** A critical component of successful Internet reading is the ability to read and locate information that meets one’s needs (Guinee, Eagleton, & Hall, 2002; Eagleton, Guinee, & Langlais, 2003; Educational Testing Service, 2007; Sutherland-Smith, 2002). Additional online reading skills and strategies may be required, for example, to generate effective keyword search strategies (Bilal, 2000; Eagleton & Guinee, 2002; Kuiper & Volman, 2008); to read and infer which link may be most useful within a set of search engine results (Henry, 2006); and to efficiently scan for relevant information within websites (Rouet, 2006). In this study, we included items designed to capture students’ ability to query search engines, navigate within their various interfaces, and interpret the relevancy of search results in relation to their particular reading needs.

**Reading to critically evaluate online information.** A third component is the ability to critically evaluate information encountered on the Internet. Critically evaluating online information includes the ability to read and evaluate the level of accuracy, reliability, and bias of information (Goldman, Braasch, Wiley, Graesser, & Brodowinska, 2012; Sanchez, Wiley, & Goldman, 2006). This presents challenges that are different from traditional print and media sources. The content of online information is even more diverse (Tillman, 2003) and commercially biased (Fabos, 2008) and new strategies are required to locate sources of website
information and evaluate expertise, bias, and point of view. In this study, we evaluated students’ ability to: 1) identify the author; 2) evaluate the author’s expertise; 3) determine the point of view; and 4) evaluate the reliability of a website.

**Reading to synthesize online information.** The Internet introduces additional challenges to coordinate and synthesize vast amounts of information presented in multiple media formats, from a nearly unlimited and disparate set of sources (Glister, 2000; Jenkins, 2006; Rouet, 2006). Bulger (2006) highlights the ability to manage, process, and filter multiple electronic documents as a key component of online literacy. The assessments presented here seek to capture students’ ability to synthesize within and across multiple texts using a notepad as students conduct their research and prompts from a system driven avatar.

**Reading to communicate online information.** A final component of successful Internet use is the ability to read and respond (or communicate) via the Internet while interacting with others to seek more information or share what you have learned (Britt & Gabrys, 2001). Each specific communication tool on the Internet is constituted differently and presents a range of new skills, strategies, and social practices to use them effectively (Greenhow, Robelia, & Hughes, 2009). The assessments in this study were designed to measure how readers navigated and communicated the results of their research within either a wiki or an email environment.

**Summary.** This study used both levels of New Literacies theory to frame the investigation. An uppercase theory of New Literacies suggests that new, online social practices become important. Thus, we evaluated students’ ability to read online and conduct Internet research in science within a context that required them to engage in several social practices using chat or text messaging, wikis, email, search engines, and a social network. Online reading comprehension, one of several lower case theories of new literacies, suggests that locating,
evaluating, synthesizing, and communicating information are important areas to evaluate when we conduct research and read online. We looked specifically at students’ ability in each of these areas.

**Recent Assessments of Digital Literacies**

**PISA and PIAAC Assessments**

International work around a broadly defined term, digital literacies is taking place under the leadership of OECD. Two large-scale projects have provided pioneering efforts in this area: the Programme for International Student Assessment (PISA) Digital Reading Assessment (OECD, 2011) and the Problem Solving in Technology-rich Environments portion of the Program for the International Assessment of Adult Competencies, or PIAAC (OECD, 2012). Each, however, appears to contain important limitations that may limit their ability to represent some of the essential qualities of online research and comprehension.

These limitations exist in two respects. First, each assessment tightly restricted the nature of the information space students used. This happened in both the restricted nature of the response types that were used as well as in the number of available websites that were used for items. The large majority of items in these assessment appeared in a multiple-choice format with five answer choices. Far fewer responses required a constructed response. Even these “open-constructed response” items often limited the information space in important ways and did not require the actual use of the technology tool. Email message items, for example, provided much of the information to students and only required them to complete the message. Rather than navigating an email system to set up a message, students were provided with an open email window with a completed “To:” address, a completed subject line, and the stem of the message in the message window. In the PISA Digital Reading Assessment (OECD, 2011) for example,
students only had to provide the ending to the email message containing evidence for the claim. Surprisingly, the PISA Digital Reading Assessment contained a smaller percentage of these types of “constructed response” items (8/29 = 28%) than did the PISA assessment of print reading (69/131 = 53%). Thus, while traditional print reading typically takes place in a far more restricted information space than online research and comprehension, a far greater percentage of multiple choice items, restricting the information space for responses, appeared in the PISA Digital Reading Assessment (72%) than in the PISA assessment of print reading (62/131 = 47%).

The information space was also limited in the PISA Digital Reading Assessment in a second way. Problems in this assessment typically took place within a single website rather than within the extensive set of sites that defines the Internet. Moreover, if students were required to navigate to find a site, directions provided an exact literal match to the correct search result that should be selected. Thus, the challenges presented by locating information in complex information contexts was profoundly minimized.

By restricting the information spaces in these ways, the PISA Digital Reading Assessment may have limited the more complex informational demands of actual online research and comprehension. Using a large percentage of multiple choice response items and limiting the websites used to solve a problem may have resulted in a digital reading assessment that shared many commonalities with traditional print reading. In fact, correlations between students scores on this assessment and a measure of offline reading suggest this possibility.

A second limitation of these assessments is that they represented the nature of the online research process as discrete and unrelated tasks. Instead of constructing a complete research task online, assessments presented a number of discrete, unrelated sets of tasks. Rather than
evaluating the process in its entirety, different elements were evaluated separately. This, too, typifies offline reading assessments more than online research and comprehension.

While both the PISA Digital Reading Assessment and the Problem Solving in Technology-rich Environments portion of PIAAC suffer from inevitable compromises designed to ease the scoring in large-scale assessments of online reading and problem solving, both are further advanced in nature than any other national assessments. In the U.S., for example, not a single state currently includes online research and comprehension tasks in their state reading or writing assessments (Leu, et. al, 2008).

**Online Research and Comprehension Assessments (ORCA)**

Some initial work has taken place to develop assessments of online research and comprehension that use more open and less restricted informational spaces while also containing a complete online research and problem solving task including locating, evaluating, synthesizing, and communicating information. These Online Research and Comprehension Assessments (ORCA) have been designed to capture ongoing and “real-time” research and comprehension processes while individuals read for information on the Internet. Each ORCA is designed to assess how readers navigate and use at least one communication interface (e.g., blog, wiki, instant message, email) while using the Internet to ask questions, seek and evaluate answers, and communicate what they have learned related to their information needs.

A variety of ORCA formats have been used to estimate the online reading comprehension abilities of over 2,600 seventh-graders in language arts and science classrooms. Scores on each of these previous assessments have demonstrated adequate validity and reliability. These have included an *ORCA-IM* (Leu et al, 2005), an *ORCA-Blog* (Leu et al, 2005), *ORCA-Scenario I & II*
measures (Coiro, 2011), two ORCA-Iditarod measures (Leu & Reinking, 2009, and an ORCA-Elementary (Castek, 2008).

Each used the actual Internet. As a result, each also suffered from an important problem -- the stability of the assessment context. The Internet, of course, continuously changes. When target websites disappear or are altered, it makes it difficult to maintain the assessment context over time. The rapidly changing nature of online information presents important stability of assessment issues for measuring online research and comprehension ability.

In the ORCA Project, we sought to develop assessments that included a complete online research and problem-solving task along with a more open information space while also solving the problem of stability. Thus, we have been developing and evaluating three different formats for the assessment of online research and comprehension within the discipline of science, specifically four human body systems that are included in the study of science at the seventh grade level: eyes, ears, heart, and lungs. The assessment formats include: ORCA-Open, ORCA-Closed, and ORCA-Multiple Choice.

**Research Questions**

Three research questions were addressed in the present investigation. All research questions were related to the study of reliability and validity of online reading comprehension scores by format and by research topic. Reliability and validity analyses focus on the 16-score points that were assigned for each topic on the first day of administration of each ORCA. The 16-score points were assigned given four online skill areas: location, evaluation, synthesis, and communication. Collectively, each 16-point scale was called a LESC, the acronym derived from the first letter of each of the four skill sets. Since the data collected to date represent a pilot phase of the research project, the focus on each LESC allowed for inspection of problematic
items that were in need of replacement or revision based on item characteristics. The research questions of the study were as follows:

1. What are the reliability estimates of scores for 24 LESC{s} that vary item format (i.e., Multiple-Choice, Closed, and Open) and by science topic?
2. What are the reliability estimates of scores for the three formats?
3. How much variance is explained by each of the 24 LESC scales?

Methods and Procedures

Participants

Participants included representative samples of seventh grade students from each of two states in the northeastern U.S. The seventh grade level (typically ages 12-13) was selected since this is when students often begin intensive study of content areas and begin to do more extensive online research. The two states exhibited key differences in two regards: the state’s economy and the policy regarding laptops for students. According to data from the U.S. Census Bureau (2012), State 1 ranked fourth in median family income from among all states in the U.S. ($65,753); State 2 ranked thirty-third ($46,033). In State 1 laptop use in classrooms is very limited. There is no statewide policy, providing access to laptops and individual districts have yet to implement a policy, except for very limited, small-scale, pilot studies. In State 2, state policy provides each student with a personal laptop for use at school and at home in grade six and above. One to one laptop use is ubiquitous in every 7th grade classroom.

Nineteen districts (a single participating school in each) were included in the final sample from State 1 and 23 districts (a single participating school in each) from State 2. Because State 2 had a high number of districts with low population, additional districts were included, matched to the very small districts that did not enroll the sample target of 40 students in the seventh grade.
Districts and schools from each state were selected using a stratified random sampling plan designed to ensure a representative sample of all seventh-grade students in that state. Specifically, for each state, the sampling plan stratified districts according to three factors: 1) performance on the state reading comprehension assessment, 2) seventh grade enrollment size and 3) socio-economic factors (using % free/reduced price lunch) and/or state classification. For each factor, the selected schools covered the full range for that state, from the highest to the lowest 10th percentile. Only a single school was selected per district, which enabled geographic distribution of participating schools.

At selected schools, the principal identified one teacher whose students best represented the entire school population and who was willing to participate. For each teacher, two English Language Arts classes were sampled with approximately 20 students in each class. Permissions were returned from a total of 1,384 seventh graders, 725 from State 1 and 706 from State 2. Most students completed two separate online research activities on different days; but because of absences and computer failures, 15.6% of the sample did not complete one of the activities. Table 1 indicates the number of students from each state who completed the online activities on each separate day (Columns 1-3) and the number of students who completed the online activities on both days (Column 4). Thus, the total sample size from which data for this study was analyzed included only the 1,129 seventh grade students who completed the online activities on both days (591 from State 1 and 538 from State 2). See Table 1.
ORCA Assessments

**Development.** Final versions of the 24 LESCs used in the current study were developed in numerous planning and development sessions and multiple rounds of cognitive lab sessions over a span of three years between 2009-2011. Participants included a total of approximately 300 seventh grade students representing students with a range of reading levels and socioeconomic backgrounds from several schools. Cognitive labs proceeded in five phases, as more sophisticated interfaces and scoring functions were designed. Data sources included video recordings of online screen reading, concurrent think-aloud comments, retrospective post-item interview questions, and researcher notes and observations of participants’ behaviors during and after the cognitive lab experience. Weekly team meetings used think-aloud data, performance data, and researcher notes to revise the subsequent week’s items for upcoming cognitive labs. In the end, three formats with eight research topics for each format in science (human body systems), were developed. This totaled 24 separate LESCS (three formats x eight topics).

**Three Formats.** The ORCAs in this project were performance-based assessments, constructed in three different formats: ORCA-Open, ORCA-Closed, and ORCA-Multiple Choice. These provided an important dimension on which the information space for the entire research task was either most restricted (ORCA-Multiple Choice), somewhat restricted (ORCA-Closed), or least restricted (ORCA-Open). All three formats followed a common scenario structure with a common setting (a middle school) and a common sequence of activities that required students to locate, evaluate, synthesize, and communicate a report of the research project in either an email or wiki.

*ORCA-Open and ORCA-Closed.* The ORCA-Open and ORCA-Closed formats took place in a social network created for this study. The assessment program used two system avatars to
direct the sequence of research tasks in each LESC by communicating with each student via chat. For example, Brianna (one of the system avatars) introduced each research problem and then directed each student through the rest of the research project. The introduction of the research problem for the activity “Are energy drinks heart healthy?” may be seen in Figure 1.

FIGURE 1 ABOUT HERE

At various points, students were directed by the avatars to use a search engine to locate various website with information related to the research problem. In the ORCA-Open format, students used the search engine Google and could go anywhere on the open Internet. In the ORCA-Closed format, students used a search engine built for this study “Gloogle” and searched from among approximately 400 websites imported into our closed system. Figure 2 shows the keyword entry and search results from one search. Keyword entry and “first click” selections in search results were captured by the system for scoring students’ locating ability.

FIGURE 2 ABOUT HERE

At other points, students were asked to synthesize the most important information related to the research question within a website or across several websites. The system avatar directed students to write their summary in a notepad within the system. These summaries were captured and used to score students’ synthesis ability.

One of the websites in each research activity was used to evaluate the critical evaluation of source information. At this site, another system avatar appeared and, in a series of chat messages,
asked students to: 1) identify the author/creator of the website; 2) evaluate the author’s level of expertise; 3) determine the author’s point of view; and 4) evaluate the reliability of the information at the website. Student responses were captured by the system and used to score students’ ability to evaluate online information. See Figure 3.

Finally, students were asked by the system avatar to write a report of their research in either an email message to the school board president or in a classroom wiki created by a teacher. In the email task, for example, students needed to reread the initial email containing the research task and then construct and send an email message with the appropriate information, using their notes. A student’s entries in the “To:” and “Subject” windows were captured by the system as was the message in the message window for later scoring of the student’s ability to communicate online. See Figure 4.

**ORCA-Multiple-Choice.** The ORCA-Multiple Choice format presented a similar scenario context as the other formats with similar research problems. It paired screenshots of the same key decision points for each skill assessed in ORCA-Open and ORCA-Closed using prompts and multiple choice item responses. This was designed to be the most restricted information environment. An example item from an ORCA-Multiple Choice appears in Figure 5.
Eight Topics in Science. The discipline of science was selected since this area appears to be especially important to the advancement of knowledge and the well being of nations (National Research Council, 2007; The President’s Council of Advisors on Science and Technology, 2010; Scientific American, 2010). Within science, human body systems (those related to eyes, ears, heart, lungs) were selected as an area of study since this is a curricular area at this age level common to many nations.

Eight research topics (two from each of the selected aspects of the human body) were selected after extensive work in cognitive labs to be of interest to students at this level but not especially familiar. They included:

1. How do energy drinks affect heart health? (heart)
2. How can snacks be heart healthy? (heart)
3. How does the volume level of MP3 players affect hearing? (ears)
4. How well can adults hear mosquito ring tones? (ears)
5. Does third-hand smoke harm your health? (lungs)
6. Can Chihuahua dogs cure asthma? (lungs)
7. Do cosmetic contact lenses harm your eyes? (eyes)
8. Do video games harm your eyes? (eyes)

Research questions 1-4 were developed as “learn more about” research questions. They asked students to use email to compose and communicate their short research report about what they learned to the President of the School Board. Research questions #5-8 were developed as
“investigate conflicting claims” questions. They asked students to use a science teacher’s classroom wiki to take a position and communicate their short research report to the class. Each student received one “take a position” and one “learn more about” task.

Each of the eight topics appeared in each of the three formats. Thus, there were 24 different LESC (three formats x eight topics). Students were randomly assigned to two LESC in the same format with two different topics. Two LESC comprised one ORCA. Each student completed one email and one wiki task in an ORCA.

**Data Capture.** An online data capture system categorized all responses by the system and by the student. This information was used for scoring student performance. Specific elements were selected for output to an excel spreadsheet for scoring.

**Score Points.** Each ORCA-Open and ORCA-Closed LESC was scored using a common 16 score point system. (See Table 2.) The score point system was developed through a number of cognitive labs, data review meetings over three years, and a pre-pilot study. Each of four areas (locate, evaluate, synthesize, communicate) contained four score points, scored on a binary basis (0-1). The score points were scaled in terms of anticipated difficulty by a team of trained researchers from easiest to most difficult. The first three score points in each area were consider “process score points,” essential to successful completion of the final “product score point.”

<table>
<thead>
<tr>
<th>TABLE 2 ABOUT HERE</th>
</tr>
</thead>
</table>

**Scoring.** Spreadsheet data of relevant performance data was output for each score point in the ORCA-Closed and ORCA-Open formats. Since a number items required constructed responses, ORCA-Closed and ORCA-Open formats were scored by trained scorers. In addition,
screen capture videos were used to score several items of the ORCA-Open that took place outside the system and were unable to be captured except by video (mainly outside websites that were located and used in the open Internet). The system automatically scored all responses for the ORCA-Multiple Choice format.

Ten scorers were trained to a minimum inter-rater reliability level of 90% accuracy for each score point on every student’s assessment with expert scorers. Then pairs of trained scorers scored each topic, making certain that inter-rater accuracy levels remained above 90% for each item on three different sets of ten assessments during the scoring of each format. If the inter-rater accuracy level dropped below 90% for any item, the scorers were retrained by expert scorers and retested before being allowed to proceed.

A complete ORCA consisted of two online research activities or 32 possible score points administered on two different days. In the current investigation, item responses for the first LESC completed by each student (16 score points) were analyzed.

A complete report of the context, structure, score points, and the connection of each score point to Common Core Anchor Standards (CCSS, 2012) appears in Leu, et al. (in press). A video of a high-performing student taking the ORCA-Closed assessment, “Are Energy Drinks Heart Healthy?” may be viewed by linking to this URL: http://neag.uconn.edu/orca-video-ira/.

**Administration.** The administration of the assessments took place on MacBook Airs, specifically prepared to rapidly take students to the online assessment location with the format and topic assigned to them. It also initiated the video screen capture. Test administration was led by a trained test administrator and followed a standard protocol. Students had as much time as needed to complete each assessment, though most students finished within about 45 minutes. In between the two assessment sessions, students completed two additional items: 1) an
assessment of traditional, offline reading comprehension; and 2) a student Internet Use Survey. Teachers completed a Teacher Internet Use Survey.

**Prior Domain Knowledge of the Topic**

Prior domain knowledge of the research topic was measured using verbal report and idea unit analysis, an approach with demonstrated reliability (Wolfe & Goldman, 2005). We gathered verbal report data online of prior domain knowledge before students initiated their work on each LESC scenario. Students were prompted to enter all that they knew about the topic in a window on their laptop. After every 15 seconds, they were prompted to enter additional information that they knew about the topic. When 15 seconds went by without an entry, a button appeared that said, “I don’t know anything else.” Prior knowledge entry concluded when students selected this button and began the LESC scenario. Idea unit analysis (the number of propositions provided by a student) was conducted on all entries. Each accurate proposition received one point. Two scorers were trained to 90% accuracy with a sample set and then scored all prior knowledge entries. Disagreements were resolved by discussion.

**Data-Analytic Strategies for Psychometric Analysis**

Several modeling procedures (cf. Kulikowich, 2008) have been used to analyze ORCA scores. To address the research questions of the current study, the investigators employed both classical test score analysis and latent variable modeling strategies to evaluate the reliability and validity of scores (i.e., Research Questions 1 - 3).
Results

Research Question 1: What were the reliability estimates of scores for 24 LESCs that vary by item format (i.e., Multiple-Choice, Closed, and Open) and by science topic?

To evaluate the reliability for each of the 24 LESC scores, we estimated the internal consistency of each 16-point scale using the Kuder-Richardson 20 (KR-20) formula. Tables 3-5 display the KR-20 values for each 16-point scale by research topic (e.g., energy drinks, third-hand smoke) for the separate assessment formats (i.e., Multiple Choice, Closed, Open). As displayed in these tables, KR-20s ranged from .48 to .76 for Multiple Choice, from .48 to .82 for Closed, and from .48 to .76 for Open. Five out of eight LESCs met the general standard of KR-20 estimates of .70 or greater for reliability in each of the Closed and Open formats. Two out of eight LESCs met this level in the Multiple-Choice format. In all, half of the 24 LESCS met this level and two-thirds of all LESC had a KR-20 greater than .65. KR-20 scores were included in our evaluation of which LESCs would remain in the next round of the project and which specific items in each of the remaining LESCs would be targeted for potential revision.

We also tested the degree to which the scales were unidimensional. We used exploratory common factor analysis. Given that this is a pilot study, we anticipated that some items would be problematic and in need of revision. Thus, we used the criterion of factor loadings < .20, to identify problematic items. The number of items targeted for potential revision in each LESC also appears in Tables 2-4. These items were either too easy or too difficult and, for the final LESCs selected for use in the Year Four study, were revised accordingly.
Research Question 2: What were the reliability estimates of scores for the three formats?

To evaluate the reliability for each of the three formats, we also estimated the internal consistency of each format using the Kuder-Richardson 20 (KR-20) formula. Table 6 displays the KR-20 values for each format across each of the eight LESC s. As displayed in this table, KR-20s ranged from .65 (Multiple Choice) to .71 (Closed).

We also tested the degree to which the scales were unidimensional. We used exploratory common factor analysis. Given that this is a pilot study, we anticipated that some items would be problematic and in need of revision. Thus, we used the criterion of factor loadings < .20, to identify problematic items. For the multiple-choice LESC s, three items demonstrated consistently loadings less than .20. These items were: 1) the first communication item, which was the first item on the test; 2) the first location item; and, 3) the last evaluation item. These items were either too easy or too difficult and, for the final LESC s selected for the year four study, were revised accordingly.
For the Closed LESCs, only one item was problematic consistently across research topics: the first communication item. For the Open LESCs, two items had overall factor loadings less than .20. These were for the scores of the second evaluation item and the first synthesis item. Again, these items were either too easy or too difficult and, for the final LESCs used for the Year Four study, were revised accordingly.

**Research Question 3: How much variance is explained for each of the 24 LESC scales?**

To address the amount of variance explained by each LESC or 16-point scale, exploratory factor analysis (EFA) using maximum likelihood estimation was run for each LESC and for each format. MPlus software (Muthén & Muthén, 1998-2012) was used to estimate factor structures and model-data fit. Tables 3-5 present the percentage of variance explained for the scale scores as well as goodness-of-fit indices (GFI) to support an unidimensional interpretation of LESC scales. Generally, estimates of the percentage of variance above 20% are considered satisfactory as are GFIs greater than .90.

As shown in Tables 3-5, results indicated that the percentage of variance accounted for by a single factor ranged from 15.34% - 24.58% for Multiple Choice LESCs, from 13.65% to 29.10% for Closed LESCs, and from 13.65% - 25.30% for Open LESCs. A majority (14 out of 24) LESCs exceeded the 20% level.
Tables 2-4 also indicate that single-model GFIIs ranged from .59 to .91 for Multiple Choice LESC, from .28 to .96 for Closed LESC, and from .65 to .94 for Open LESC. Eight out of 24 LESC had GFIIs of .90 or greater with a unidimensional analysis.

Dimensionality of scores is essential in the interpretation of results. To date, the present analyses aim to establish support for an overall general score for each LESC topic. We are currently examining and comparing unidimensional with multidimensional solutions. The degree to which Location, Evaluation, Synthesis, and Communication scores can be analyzed separately is of particular interest as these online reading comprehension skills may relate differentially to other important variables such as prior knowledge and offline reading comprehension.

**Discussion**

These initial, unidimensional results from our pilot study appear promising. Unidimensional estimates of reliability and validity for many, if not most of the LESC demonstrate reasonable levels, especially given that n’s for individual LESC only averaged about 80 students. We have used these results to select potential LESC for the year four study in these same two states with 1,600 seventh graders (800 from each state), who will again complete two LESC. Since the final number of LESC that have been selected for the Year Four study (eight) is only a third that of the Year Three study (24), we expect substantially greater n’s for each LESC in Year Four, approximately 400.

We have decided to eliminate the Open format in the year four studies, largely because an important aspect of our work is to also develop practical assessments. The Open format requires one to view the video of each student in order to score it and we have concluded that the time required to do this would not be practical for schools and teachers. We initially thought this could be done easily but we were unable to determine a way to capture actually online reading
behavior, outside of our system. In addition, the website locations used in tasks often change or go offline during the data collection period, making it difficult to guarantee that websites will be available to all students who take the ORCA-Open over the duration of testing period.

We have used these results to identify four of the most promising topics and will use these in both Multiple Choice and Closed formats during the work in year four:

1. How do energy drinks affect heart health? (heart)
2. How can snacks be heart healthy? (heart)
3. Do cosmetic contact lenses harm your eyes? (eyes)
4. Do video games harm your eyes? (eyes)

We have also used these results to identify individual items, within each of these LESCs and formats, that would gain from revision and we have completed those revisions. Finally, we are beginning to conclude that collectively, the pilot results indicate that Closed ORCA formats operate best given the psychometric properties of scores. We will evaluate this possibility with the next year’s results.

We are optimistic about the results obtained to date. We expect that continuing work with larger numbers of students, fewer LESCs, measures that have been revised based on initial patterns of the psychometric properties of scores, and multi-dimensional scaling will provide us with robust assessments that are practical for schools to use as they begin to assess students’ ability to conduct online research and comprehend information.

Being able to evaluate student progress in this area will prove useful to policy makers, researchers, and teachers in an age of online information and communication. For policy makers, it will provide student performance data on a new and especially important aspect of learning in our schools, online research and comprehension, that is not fully captured by traditional reading.
assessments. For researchers, it will provide an important instrument to study a new and rapidly expanding area of research. For teachers, it will permit the evaluation of individual students’ online research and comprehension ability in order to inform new elements of classroom instruction.
References


Educational testing Service (2007). iSkills. Princeton: NJ. Downloaded on September 15, 2008 from:http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13be3921509/?vgnextoid=159f0e3c27a85110VgnVCM10000022f95190RCRD&vgnextchannel=e5b2a79898a85110VgnVCM10000022f95190RCRD


Table 1. Number of Participants Who Completed the Online Assessment Activities Each Day by State

<table>
<thead>
<tr>
<th>N of Students Completing Test</th>
<th>Day 1: N of Students</th>
<th>Day 2: N of Students</th>
<th>Days 1 &amp; 2: N of Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>State 1</td>
<td>693</td>
<td>623</td>
<td>591</td>
</tr>
<tr>
<td>State 2</td>
<td>677</td>
<td>567</td>
<td>538</td>
</tr>
<tr>
<td>Total</td>
<td>1370</td>
<td>1190</td>
<td>1129</td>
</tr>
</tbody>
</table>
Table 2. The score point system, showing the four score points in locate (L), evaluate (E), synthesize (S), and communicate (C). The first three in each area are “process score points.” The fourth is a “product score point.”

**Locate**

L 1: Can the student develop an appropriate key word search for a problem, including topic and claim?

L 2: Within a set of search results, can the student identify a link that will take them to information that will answer the question on the first click?

L 3: Within a set of search results, can the student identify a link that will take them to information that will answer the question on the first click?

L 4: Can a student identify and provide two URLs for websites that contain information that will answer a question?

**Evaluate**

E 1: Can the student determine the name of the author for a website?

E 2: Can a student determine if the author is an expert in the area they write about?

E 3: Can the student determine the point of view of the author?

E 4: Can the student determine if a website is reliable or not?

**Synthesize**

S 1: Can the student summarize information, related to the question, from a website?

S 2: Can the student summarize the information from two websites?

S 3: Can the student summarize the information related to the question from two websites?

S 4: Can the student summarize the information related to the question from four websites?

**Communicate**

C 1: Can the student select the correct link to take them to the correct communication area for a task on the first click?

C 2: Can the student determine the correct location in a wiki to edit, based on the research question?/Can the student construct an appropriate subject line for an email message?

C 3: Can the student revise the heading of a wiki appropriately?/Can the student address an email message to the appropriate individual?

C 4: Based on their research, can the student revise a wiki entry appropriately or send an email message, with the correct discourse conventions and argument structure?
Table 3. Reliability estimates (KR-20) for ORCA-MC LESCs, including the percentage of variance explained by each 16-point scale and the number of items targeted for potential revision.

<table>
<thead>
<tr>
<th>Format</th>
<th>Research Topic</th>
<th>Science Topic</th>
<th>KR-20</th>
<th>Mean</th>
<th>SD</th>
<th>% of Variance Explained</th>
<th># Items for Revision</th>
<th>GFI/RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORCA-MC</td>
<td>Energy Drinks</td>
<td>Heart</td>
<td>.70*</td>
<td>12.26</td>
<td>2.73</td>
<td>23.85**</td>
<td>9</td>
<td>.91/.05***</td>
</tr>
<tr>
<td></td>
<td>Healthy Snacks</td>
<td>Heart</td>
<td>.48</td>
<td>9.81</td>
<td>2.36</td>
<td>15.34</td>
<td>6</td>
<td>.59/.09</td>
</tr>
<tr>
<td></td>
<td>Volume Level</td>
<td>Ears</td>
<td>.68</td>
<td>10.95</td>
<td>2.90</td>
<td>21.64**</td>
<td>6</td>
<td>.89/.05</td>
</tr>
<tr>
<td></td>
<td>Ring Tones</td>
<td>Ears</td>
<td>.57</td>
<td>10.31</td>
<td>2.45</td>
<td>19.32</td>
<td>8</td>
<td>.82/.05</td>
</tr>
<tr>
<td></td>
<td>3rd Hand Smoke</td>
<td>Lungs</td>
<td>.67</td>
<td>11.34</td>
<td>2.72</td>
<td>21.02**</td>
<td>5</td>
<td>.86/.06</td>
</tr>
<tr>
<td></td>
<td>Asthma</td>
<td>Lungs</td>
<td>.64</td>
<td>9.37</td>
<td>2.84</td>
<td>17.40</td>
<td>2</td>
<td>.75/.06</td>
</tr>
<tr>
<td></td>
<td>Contact Lenses</td>
<td>Eyes</td>
<td>.59</td>
<td>11.81</td>
<td>2.41</td>
<td>19.12</td>
<td>8</td>
<td>.86/.05</td>
</tr>
<tr>
<td></td>
<td>Video Games</td>
<td>Eyes</td>
<td>.76*</td>
<td>12.07</td>
<td>3.05</td>
<td>24.58**</td>
<td>5</td>
<td>.91/.06***</td>
</tr>
</tbody>
</table>
Table 4. Reliability estimates (KR-20) for ORCA-Closed LESCs, including the percentage of variance explained by each 16-point scale and the number of items targeted for potential revision.

<table>
<thead>
<tr>
<th>Format</th>
<th>Research Topic</th>
<th>Science Topic</th>
<th>KR-20</th>
<th>Mean</th>
<th>SD</th>
<th>%Variance Explained</th>
<th># Items for Revision</th>
<th>GFI/RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORCA-Closed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy Drinks</td>
<td>Heart</td>
<td></td>
<td>.72*</td>
<td>9.15</td>
<td>2.74</td>
<td>20.69**</td>
<td>4</td>
<td>.88/.04</td>
</tr>
<tr>
<td>Healthy Snacks</td>
<td>Heart</td>
<td></td>
<td>.73*</td>
<td>7.50</td>
<td>3.10</td>
<td>22.57**</td>
<td>1</td>
<td>.68/.11</td>
</tr>
<tr>
<td>Volume Level</td>
<td>Ears</td>
<td></td>
<td>.60</td>
<td>7.30</td>
<td>3.55</td>
<td>16.62</td>
<td>4</td>
<td>.74/.07</td>
</tr>
<tr>
<td>Ring Tones</td>
<td>Ears</td>
<td></td>
<td>.48</td>
<td>6.53</td>
<td>1.99</td>
<td>13.65</td>
<td>6</td>
<td>.28/.06</td>
</tr>
<tr>
<td>3rd Hand Smoke</td>
<td>Lungs</td>
<td></td>
<td>.82*</td>
<td>7.32</td>
<td>3.58</td>
<td>28.13**</td>
<td>4</td>
<td>.96/.08***</td>
</tr>
<tr>
<td>Asthma</td>
<td>Lungs</td>
<td></td>
<td>.63</td>
<td>7.61</td>
<td>2.70</td>
<td>17.93</td>
<td>3</td>
<td>.88/.04</td>
</tr>
<tr>
<td>Contact Lenses</td>
<td>Eyes</td>
<td></td>
<td>.77*</td>
<td>7.50</td>
<td>3.17</td>
<td>23.83**</td>
<td>1</td>
<td>.86/.09</td>
</tr>
<tr>
<td>Video Games</td>
<td>Eyes</td>
<td></td>
<td>.82*</td>
<td>8.02</td>
<td>3.80</td>
<td>29.10**</td>
<td>0</td>
<td>.93/.08***</td>
</tr>
</tbody>
</table>
Table 5. Reliability estimates (KR-20) for ORCA-Open LESC, including the percentage of variance explained by each 16-point scale and the number of items targeted for potential revision.

<table>
<thead>
<tr>
<th>Format</th>
<th>Research Topic</th>
<th>Science Topic</th>
<th>KR-20</th>
<th>Mean</th>
<th>SD</th>
<th>%Variance Explained</th>
<th># Items for Revision</th>
<th>GFI/RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORCA-Open</td>
<td>Energy Drinks</td>
<td>Heart</td>
<td>.75*</td>
<td>8.06</td>
<td>3.07</td>
<td>22.27**</td>
<td>3</td>
<td>.94/.07***</td>
</tr>
<tr>
<td></td>
<td>Healthy Snacks</td>
<td>Heart</td>
<td>.69</td>
<td>7.83</td>
<td>2.92</td>
<td>18.91</td>
<td>2</td>
<td>.59/.09</td>
</tr>
<tr>
<td></td>
<td>Volume Level</td>
<td>Ears</td>
<td>.52</td>
<td>6.30</td>
<td>2.11</td>
<td>16.31</td>
<td>5</td>
<td>.65/.09</td>
</tr>
<tr>
<td></td>
<td>Ring Tones</td>
<td>Ears</td>
<td>.48</td>
<td>6.53</td>
<td>1.99</td>
<td>13.65</td>
<td>6</td>
<td>.63/.09</td>
</tr>
<tr>
<td></td>
<td>3rd Hand Smoke</td>
<td>Lungs</td>
<td>.77*</td>
<td>8.56</td>
<td>3.23</td>
<td>25.30**</td>
<td>5</td>
<td>.94/.07***</td>
</tr>
<tr>
<td></td>
<td>Asthma</td>
<td>Lungs</td>
<td>.73*</td>
<td>8.35</td>
<td>3.07</td>
<td>21.28**</td>
<td>2</td>
<td>.93/.04***</td>
</tr>
<tr>
<td></td>
<td>Contact Lenses</td>
<td>Eyes</td>
<td>.76*</td>
<td>8.31</td>
<td>3.19</td>
<td>22.72**</td>
<td>2</td>
<td>.94/.07***</td>
</tr>
<tr>
<td></td>
<td>Video Games</td>
<td>Eyes</td>
<td>.74*</td>
<td>8.43</td>
<td>3.13</td>
<td>21.30**</td>
<td>0</td>
<td>.89/.09</td>
</tr>
</tbody>
</table>
Table 6. Overall Scale Estimates for LESCs by Format

<table>
<thead>
<tr>
<th>Format</th>
<th>Sample Size</th>
<th>KR-20</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple-Choice</td>
<td>497</td>
<td>.65</td>
<td>10.94</td>
<td>2.87</td>
</tr>
<tr>
<td>Closed</td>
<td>474</td>
<td>.71*</td>
<td>7.66</td>
<td>3.15</td>
</tr>
<tr>
<td>Open</td>
<td>343</td>
<td>.69</td>
<td>7.94</td>
<td>2.93</td>
</tr>
</tbody>
</table>
Figure 1. The beginning screen sequences in the assessment, “Are energy drinks heart healthy?” showing the avatar directing the student to an email message containing the research problem and context.
Figure 2. The search engine, “Gloogle,” showing keyword entry and search results from the ORCA-Closed assessment, “Are energy drinks heart healthy?”
Figure 3. The critical evaluation sequence including the identification of the author/creator of the website, the evaluation of the author’s level of expertise, the author’s point of view, and the reliability of the information at the website.
Figure 4. An example of the report a student constructed as she sent an email message to Ms. Marin, the President of the School Board, with an answer to the research question, “Are energy drinks heart healthy?”