THE ONLINE RESEARCH AND COMPREHENSION ASSESSMENT (ORCA) PROJECT

PROJECT REPORT #8

The ORCA Project: Designing Technology-based Assessments for Online Research, Comprehension, And Communication

Donald J. Leu, University of Connecticut
Jonna M. Kulikowich, The Pennsylvania State University
Nell Sedransk, National Institute of Statistical Sciences
Julie Coiro, University of Rhode Island
Chang Liu, The Pennsylvania State University
Weiwei Cui, National Institute of Statistical Sciences
Elena Forzani, Clint Kennedy, and Cheryl Maykel, University of Connecticut


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 authors and do not necessarily represent the position of the U.S. Department of Education.
Abstract

We provide descriptions of technology-based assessments and reliability and validity results for two different formats of online research and comprehension assessments (ORCAs) from the ORCA Project (Leu, Kulikowich, Sedransk, & Coiro, 2012): ORCA-Closed and ORCA-Multiple Choice. ORCAs are performance based measures of students’ ability to conduct online research and write a short report of their results, skills central to the Common Core State Standards Initiative (2012), an online age of learning (Goldman & Scardamalia, 2013; Wiley, Goldman, Graesser, Sanchez, Ash, Hemmerich, 2009), and assessment (Quellmalz, Davenport, Timms, DeBoer, & Jordan, 2013). They are designed in two performance based formats: a simulated Internet environment where students provide largely constructed responses to all questions within an active social network, driven by an avatar student (ORCA-Closed); and a scenario-based, multiple-choice format (ORCA-Multiple Choice) with a more restricted and not fully functional representation of the Internet. Specific skills within the components of location, evaluation, synthesis, and communication are measured, and results demonstrate that each component area contributes to the psychometric properties of scores. The ORCA-Closed appears to have slightly higher reliability levels and lower and more variable performance profiles. Uses of ORCAs for policy initiatives, psychometric developments, research projects, and classroom practices are discussed.
The ORCA Project: Designing Technology-based Assessments for Online Research, Comprehension, And Communication

High-level reading and comprehension skills have become requirements for students in an online age of information (Common Core State Standards Initiative, 2012; Partnership for 21st Century Skills, 2006). As reading shifts from page to screen, and with online information use now such an important and ubiquitous part of our lives for learning and success (International Reading Association, 2009), increasing attention is being given to analyzing the nature of online reading, especially online research and comprehension (Goldman, Braasch, Wiley, Graesser, & Brodowinska, 2012; Hartman, Morsink, & Zheng, 2010; Zhang & Duke, 2008). Knowledge-based societies require citizens to be skilled in the effective use of online information for inquiry and communication (OECD, 2010; Rouet et al., 2009). Assessments are needed to measure the online research and comprehension skills required for students to succeed academically as well as for life-long learning as they consider goals related to the workplace, economic planning, recreation, and health and wellness (OECD, 2010.)

Recently, several online research and comprehension assessments (ORCAs) have been developed in the ORCA Project (Leu, Kulikowich, Sedransk, & Coiro, 2012). The purpose of this investigation is to establish the psychometric properties of scores for two, performance based formats for these assessments (ORCA-Closed and ORCA-Multiple Choice). These formats vary in the extent of the simulation that is used and in the response format. In addition, the study sought to examine the contributions of several components to the explained variance of the scales.
Review of Relevant Literature

Online Research and Comprehension

Considerable research has concluded that reading comprehension on the Internet typically is associated with online research and problem solving and requires additional comprehension skills beyond those required by traditional text reading comprehension (Afflerbach & Cho, 2008; Coiro & Dobler, 2007; Goldman & Scardemalia, 2013). Additional work has focused on identifying the skills that appear to be required during online research and comprehension (Goldman et al., 2012; Leu, Kinzer, Coiro, Castek, & Henry, 2013).

One theory suggests that online research and comprehension is a problem-solving process with at least four components: 1) reading to locate information; 2) reading to evaluate information; 3) reading to synthesize information; and 4) reading and writing to communicate information (Castek, et. al, 2007; Coiro, 2003; Henry, 2006; Leu, Kinzer, Coiro, Castek, & Henry, 2013). This set of four components is integrated into several international and national policy initiatives (Common Core State Standards Initiative [CCSS], 2012; Organization for Economic Cooperation and Development [OECD], 2012).

One example appears in the Common Core State Standards Initiative (2012) of the U.S. A key design principle of the initiative, research and media skills, demonstrates the importance of these four components. It states:

“To be ready for college, workforce training, and life in a technological society, students need the ability to gather, comprehend, evaluate, synthesize, and report on information and ideas, to conduct original research in order to answer questions or solve problems... in media forms old and new. The need to conduct research and to produce and consume media is embedded into every aspect of today’s curriculum.” (CCSS, 2012, p. 4, emphasis added)
We see this design principle operationalized in a number of CCSS anchor standards:

1. Use technology, including the Internet, to produce and publish writing and to interact and collaborate with others.

4. Conduct short as well as more sustained research projects based on focused questions, demonstrating understanding of the subject under investigation.

5. Gather relevant information from multiple print and digital sources, assess the credibility and accuracy of each source, and integrate the information while avoiding plagiarism.

In Australia, the Australia Curriculum, Assessment and Reporting Authority, or ACARA, (n. d.) has developed the Australian Curriculum. The English portion of the Australian Curriculum states:

“Students … develop skills in using information technology when conducting research, a range of digital technologies to create, publish and present their learning, and communication technologies to collaborate and communicate with others both within and beyond the classroom.”

(ACARA, n.d., General Capabilities, Information and Communication Technology Competence section, para. 2)

In another example, this time from Canada, the province of Manitoba has developed an educational framework called “Literacy with ICT Across the Curriculum” (Minister of Manitoba Education, Citizenship, and Youth, 2006). This initiative outlines skills required in the 21st century and in their new curriculum:

“…navigating multiple information networks to locate relevant information; applying critical thinking skills to evaluate information sources and content; synthesizing
information and ideas from multiple sources and networks; representing information and ideas creatively in visual, aural, and textual formats; crediting and referencing sources of information and intellectual property; and communicating new understandings to others, both face to face and over distance…” (Minister of Manitoba Education, Citizenship, and Youth, 2006, p. 18).

It is clear that these four elements are recognized in a number of national and international policy initiatives to prepare students for the challenges and opportunities of online information use. They theoretically and empirically drive the design and development of the measures that we have created to measure online research and comprehension. The four component areas are represented in the 16-item scales we have developed and to which we refer to as LESC (Locate, Evaluate, Synthesize, and Communicate).

**L: Reading to locate online information.** Research findings indicate that online research and comprehension requires the ability to generate effective keyword search strategies (Bilal, 2000; Eagleton, Guinee, & Langlais, 2003; Kuiper & Volman, 2008); to read and infer which link may be most useful within a set of search engine results (Henry, 2006); and to scan efficiently for relevant information within websites (McDonald & Stevenson, 1996; Rouet, 2006; Rouet, Ros, Goumi, Macedo-Rouet, & Dinet, 2011). Thus, reading to locate online information is identified as one critical component of, and possibly a gatekeeper skill of, successful online research and comprehension; if one cannot locate information, one will be unable to solve a given problem (Broch, 2000; Guinee, Eagleton, & Hall, 2002; Eagleton, Guinee, & Langlais, 2003; Educational Testing Service, 2002; Sutherland-Smith, 2002).

**E: Reading to critically evaluate online information.** A second component required by successful online research and comprehension is the ability to critically evaluate information
encountered on the Internet (Center for Media Literacy, 2005; Fabos, 2008). Critical evaluation of online information includes the ability to read and then to evaluate the information’s level of accuracy, reliability, and bias (Graham & Metaxas, 2003; Sanchez, Wiley, & Goldman, 2006; Sundar, 2008). Critical evaluation on the Internet presents additional reading challenges beyond that of traditional print and media sources since the content, format, and sources of online information is even more diverse than that of print (Center for Media Literacy, 2005; Fabos, 2008).

**S: Reading to synthesize online information.** Successful Internet reading during online research and comprehension also requires the ability to read and synthesize information from multiple online sources (Jenkins, 2006; Leu et al., 2004). 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 (Glistser, 2000; Jenkins, 2006; Rouet, 2006). Bulger (2006) highlights it as a key component of online literacy.

**C: Reading to communicate online information.** A fourth component of successful Internet reading during online research and comprehension is the ability to communicate through reading and writing online while interacting with others to seek more information or to share what has been learned (Britt & Gabrys, 2001; Greenhow, Robelia, & Hughes, 2009). The interactive processes of reading and communicating have become so intertwined on the Internet that they often happen simultaneously (Leu, Slomp, Zawilinski, & Corrigan, in press; Leu et al, 2013). Thus, the communication processes involved in using a range of online tools to ask and answer questions on the Internet appear to be inextricably linked to aspects of online reading comprehension.
The Need For Reliable and Valid Tools To Estimate Online Research and Comprehension Performance

Changes to educational frameworks and standards taking place in many locations provide additional impetus to the important need for measures that provide reliable and valid estimates of online research and comprehension ability. The need is especially evident for both research and practice.

Research. The Internet is a powerful, disruptive force, fundamentally altering many aspects of life in this century (Christensen, 1997) including the nature of reading. Reading now includes additional skills and social practices emerging from the new online texts, technologies, and social practices (Afflerbach & Cho, 2008; Coiro & Dobler, 2007; Goldman & Scardemalia, 2013; Kist, 2005; Lankshear & Knobel, 2003). There appears to be a complex layering of old and new that takes place during online research and comprehension (Leu, Kinzer, Coiro, Castek, & Henry, 2013). Since additional skills appear to be required, beyond those required for offline reading comprehension (Coiro, 2011; Coiro & Dobbler, 2007), this presents an important challenge to researchers. We require assessments with reliable and valid scores that include the additional skills and strategies needed during online research and comprehension to guide research efforts in this area.

In particular, we require dependent measures that accurately measure learning and that include the additional skills required during online research and comprehension. These dependent measures are certainly required as outcomes in correlational studies that attempt to unravel the complex interplays among variables such as prior knowledge (e.g., Kintsch, 1988), strategic processing (e.g., Goldman & Saul, 1990), motivation and interest (e.g., Alexander, Kulikowich, & Schulze, 1994), text complexity (e.g., Graesser, McNamara, & Kulikowich,
and offline comprehension (e.g., Cui & Sedransk, 2013) on online research and comprehension.

Just as importantly, we require dependent measures with sound psychometric properties in the test of interventions that promote learning (e.g., Kulikowich, 2008; Kulikowich & Sedransk, 2011). Considerable attention and resources are given to the development and testing of effective interventions (e.g., Institute of Education Sciences [IES], 2013). What is effective is often defined as what results in a sizeable “effect size” when a treatment condition is compared to another form of instruction or to control on one or more outcomes, and as such, has “predictive power” (Wieman, 2014, p. 12). However, if the dependent variables do not have score distributions that are reliable and valid, then any effect size is in question as any mean comparison is arguably meaningless.

Practice. If teachers are to prepare students for the new curriculum standards being developed in nations around the world, then they will need tools to help them understand how their students are doing and where they can assist them instructionally. If used appropriately, assessments of online research and comprehension may lead to more effective instruction in at least two important ways. First, they may be especially useful in helping teachers and parents see the types of higher-level thinking and the types of literacy practices important to online research and comprehension. It is hard to teach something that is unfamiliar. Seeing the specific nature of online research and comprehension assessments will yield greater understanding by teachers and parents about the types of skills, strategies and literacy practices that are required during online research and comprehension.

Second, good instruction depends on knowing what students can do and what they have difficulty doing. Assessments of online research and comprehension provide teachers with a
better understanding of their students’ abilities in important new areas for literacy development. They will provide starting points for instruction in classrooms and can support the development of additional skills, strategies, and practices in classroom lessons and activities. In the very best of worlds, assessments of online research and comprehension will serve to support teachers, parents, and students, showing them what is possible far beyond the specifics of any particular assessment. Good assessments of online research and comprehension demonstrate how skilled online readers may use online research ability to develop a rich and sophisticated understanding of any area of knowledge that interests them and follow any dreams that they have for their future.

**Existing Assessments.** International work around a broadly defined term, *digital literacies*, is taking place under the leadership of the 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 may contain design elements that might limit their ability to represent some of the essential qualities of online research and comprehension.

These limitations restrict performance in at least one of several design elements: the use of a multiple choice, rather than a constructed response, format that restricts the nature of the response space; the use of simulations that are not fully functional and restrict interaction possibilities; and the use of a restricted sequence of discreet and disconnected tasks to measure performance. In the PISA Digital Reading Assessment (OECD, 2011), for example, the large majority of items appeared within a multiple-choice format containing 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 a technology tool. Email message items, for example, provided much of the information to students and only required them to complete the message, not enter an address or subject line or even to construct the message. Thus, students only had to provide the final few words to an email message containing evidence for a 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: the use of restricted simulations that limited interaction possibilities. Problems in this assessment typically were posed within a single website rather than within a more extensive and connected set of sites that defines online information. Typically, they also lacked fully functional online tools such as a search engine, email, text messaging, wiki, blog, and social network.

By restricting the information space in these ways, the PISA Digital Reading Assessment, and sometimes the PIAAC Assessment, may have limited the more complex informational demands of actual online research and comprehension or problem solving. Using a large percentage of multiple choice response items and restricting both websites and interactive tools to solve a problem may have resulted in digital reading and problem solving assessments that shared many commonalities with traditional print reading. In fact, correlations between students’
scores on this assessment and a measure of offline reading in the PISA study suggest this possibility (OECD, 2012).

A third limitation of these assessments is that they represented the nature of the online research or problem solving process as a restricted set of unrelated tasks; rather than evaluating the process in its entirety, different elements were evaluated separately.

While both the PISA Digital Reading Assessment and the Problem Solving in Technology-rich Environments portion of PIAAC must deal with inevitable compromises to ease the scoring demands in large-scale assessments, both are further advanced in nature than any other national assessments. In the U. S., for example, The National Assessment of Educational Progress [NAEP] still does not include online research and comprehension tasks in the assessment of reading (Leu, et al., 2013).

The Study of The Reliability and Validity of ORCA Scores

Most measurement models used to study the reliability and validity of scores assume that not only are scales unidimensional (i.e., one single trait), but also, that each singular dimension or trait is a latent one (Borsboom, 2005; Borsboom, Mellenbergh, & Van Heerden, 2003). As a latent trait, sometimes called a psychological construct, it is thought that the dimension (from the minds of students) cause the responses that examinees select or generate depending on the type of item format used. Graphically, this dimension-response relationship is a familiar depiction to many data analysts who specify and test Confirmatory Factor Analytic (CFA) models (see Figure 1). The oval represents the latent construct or trait. The latent trait directs (indicated by arrows and with weights [i.e., factor loadings]) the responses to items (indicated by squares) on the test.
There are many advantages to the interpretation of test score values when there is evidence of scores as reliable and valid for a singular dimension, and these advantages are known well in the literature (Embretson & Reise, 2000). For example, a simple, stable structure permits scientific study of Differential Item Functioning (DIF), horizontal and vertical equating to study changes in score over time, computer-adaptive testing (CAT) capabilities, and quite obviously, ease of interpretation relative to the well-known normal distribution.

However, online research and comprehension variables may introduce complexity in interpretations of scores that are not singular; either as a latent trait or due to the role of context and the examinee’s interaction with the environment (i.e., Cognition and Technology Group at Vanderbilt [CTGV], 1990, 1992). Scores might represent more than one dimension, and these dimensions may reflect the type of interactivity the student has with the design features of the assessment situation as much as they reflect any latent constructs or traits that characterize response selection (for multiple-choice) or generation (for a simulated Internet environment).

While explorations of the dimensionality of scores and interactivity with assessment media introduce exciting challenges for psychometricians and may lead to promising advances in both theory-building and test design (e.g., Mislevy, Almond, & Lukas, 2003; Quellmalz, Timms, Silberglitt, & Buckley, 2012), the approaches to analysis of the data likely require more extensive and systemic study of scores than are used traditionally to estimate internal consistency and validity. For example, Quellmalz et al. (2013), in a study of more than 1800 middle-school
students’ responses to a dynamic science simulation about ecosystems and food web tasks employed multivariate generalizability analyses (mGENOVA, Brennan, 2001), multitrait-multimethod CFA, and multidimensional Item Response Theory (IRT) in one investigation to address questions of reliability and validity of scores. While our approach to the study of reliability and validity of scores is more exploratory than that used by Quellmalz, et al., some procedures are similar and include, for example, several comparisons of factor-analytic structures.

**Design Features for The Assessment of Online Research and Comprehension**

Given this context, which design features should be incorporated into an assessment of online research and comprehension at the seventh grade level that includes the complex interplay between the components of location, evaluation, synthesis, and communication? Previous work suggests that the following elements are especially important to consider:

1. a performance based assessment design;
2. the use of a simulated Internet context;
3. the use of a social network;
4. the social practices associated with online communication;
5. a disciplinary inquiry task in science;
6. the study of human body systems.

**A performance based assessment design.** A performance based assessment is one in which a complex, authentic task is completed, involving a real-world application of knowledge and skills (Lai, 2011). Often, more extended, constructed responses are included in the evaluation of performance, but performance based assessments can also be designed to incorporate multiple-choice items (e.g., Quellmalz et al., 2013).
We selected a performance based design for multiple reasons. First, it appears to be more engaging to students (Hancock, 2007) and thus more likely to optimize problem-solving effort. Second, performance based assessments are better able to evaluate complex tasks such as writing, critical evaluation, and the completion of a research report that often include complex interrelations between elements not always obtained when separate items are used (Frederiksen, 1984). Third, performance based assessments permit the evaluation of both products and processes (Messick, 1994). Finally, advances in multimedia and technological developments afford test designers opportunity to create assessment environments that allow for authentic, real world practice using fully functional tools like email and search engines to demonstrate skill levels (e.g., Quellmalz et al., 2013) within the context of a completing a complex task.

**The use of a simulated Internet context.** For ecological validity, it would be best to position any assessment of online skills within the actual Internet. During the first three years of development and extensive pilot testing, we explored the use a third format that did this, ORCA-Open. It had students use websites and search engines in the actual Internet but presented problems via an avatar of a student within a social network. Unfortunately, we could not overcome problems associated with the instability of the Internet context as websites and search engines regularly changed. Thus, the assessment space on one day would be different from the assessment space on the next day, creating problems with stability and comparability. This approach also required substantially more time to score, relying on an analysis of a video screen capture for online website and search engine use. As such, it exceeded the parameters we had set for practicality. This format was dropped after pilot testing and not included in the present investigation.

Two of our formats were designed to overcome any potential problems with instability of
the assessment space by simulating the Internet to varying degrees: ORCA-Closed and ORCA-Multiple Choice. Both allowed us to create a more stable assessment within a performance based context. ORCA-Closed contained a fully functioning simulation of the Internet and over 500 websites that were imported. An avatar of a student directed the research task through text(chat) messages within a social network. Students used fully functional tools (a social network, text(chat), email, wikis, a search engine (Gloogle), and a notepad) to conduct their research within a simulation of the Internet. A majority of items required constructed responses.

In addition, we developed a second format, ORCA-Multiple Choice. This was also a performance based assessment within a more restricted and limited simulation of the Internet. It contained the same 16 items but with more restricted functionality of the online resources. ORCA-Multiple Choice presented identical research problems as ORCA-Closed but sequentially staged each decision, separate from other decisions, within a scenario-based format.

**The use of a social network.** We chose to situate the assessments within a social network for several reasons. First, the use of a social network is familiar to most adolescents and thus creates a well-recognized context for the assessment. Eight-one percent of teens report using a social network with the percentage for boys (79%) and girls (84%) being somewhat similar (Madden et. al, 2013). Second, we used a social network since the text messaging elements of these contexts provided us with the opportunity to have an avatar student guide participants through the many elements of a research task, allowing us to evaluate process performance as well as product performance in a natural manner.

**The social practices associated with online communication.** The Internet makes new social practices possible with technologies such as instant messaging, social networks, blogs, wikis, and email, among others (Greenhow, Robelia, & Hughes, 2009). It thus becomes
important to include several channels in an assessment task to adequately represent the use of online communication during online research and comprehension. For the simulated Internet assessments, we included text (chat), email, a notepad, and wikis. Text(chat) was used to both direct students during the research task and to evaluate some of their responses, during replies to the student avatar. Email was used to communicate the research problem (in an email from the school principal) and as one context in which to evaluate students’ ability to communicate a report of their research. A notepad was used as a means to record information during the research task and to evaluate students’ ability to synthesize information from websites. A wiki provided us with a second context in which to evaluate students’ ability to communicate a report of their research.

Disciplinary inquiry in science. We selected the discipline of science for the research tasks because science is an increasingly important subject area (National Research Council, 2011). As several groups have noted (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2011) only four percent of the U.S. workforce is composed of scientists and engineers, but this group disproportionately creates jobs for the other 96 percent. In a competitive global economy, an imperative is the preparation of students in STEM related fields (National Research Council, 2007; President’s Council of Advisors on Science and Technology, 2010).

The study of human body systems. The research problems in these assessments focused on questions related to human body systems. We selected this domain since it is found in the curriculum of all of the states for adolescent youth and because it is an important area of study in science, according to The Framework for K-12 Science Education (National Academy of Sciences, 2011).
The Present Investigation

There were two goals in the present investigation. First, we sought to establish the psychometric properties of scores for two formats, ORCA-Closed and ORCA-Multiple Choice as reliable and valid. Second, we wanted to examine the contributions of Location, Evaluation, Synthesis, and Communication items, respectively, to the explained variance of the scales for each format. Both formats provide performance-based assessments of students’ ability to conduct research and comprehend information in science. One format (ORCA-Closed) had students conduct their research within a fully functioning simulation of the Internet during naturalistic interactions with an avatar, who directed each student during the research task in a social network. Student performance was scored at 16 interaction and decision points as students completed the research activity. The majority were constructed response items.

A second format (ORCA-Multiple Choice) used a scenario-based method to provide context for a similar set of 16 items within a multiple-choice response format. This format used a more restricted and limited simulation of the Internet. The sequence of items in the ORCA-Multiple-Choice was identical to the sequence of items in the ORCA-Closed format. The results reported in this manuscript follow from two years of development with extensive cognitive labs and pre-pilot testing (Leu, et. al, 2012) and a year of pilot investigations with over 1100 students in two states (e.g., Cui et al., 2013; Kulikowich et al., 2013; Leu et al., 2012) that surveyed score properties of initial LESC scale development.

Several item characteristics for each of the formats were determined to be too easy or too difficult. Further, some items had poor discrimination indices. Previously, we reported on these data that were used to revise items before they were administered as part of the present investigation (Leu, et. al, 2012). Additionally, the original set of scales included eight topics
about the human systems. Four of these topics were retained after identifying the scales from the pilot year studies that had the best reliability and validity estimates. As a result, two formats (ORCA-Closed and ORCA-Multiple Choice) and four topics for research in human body systems were selected. The four research topics were:

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)

From the pilot year, Cui et al. (2013) determined that scores of the ORCA-Multiple Choice scales were unidimensional. By comparison, scores of the ORCA-Closed scales were multidimensional, although dimensions were either correlated or the location and communication tasks accounted for unique variance apart from the evaluation and synthesis tasks. Theoretically, as well as by design, these patterns of dimensionality make sense. Location and communication tasks rely significantly on tool use with search engines or composing responses using either email or wiki. In contrast, correct responses for evaluation and synthesis tasks depend heavily on studying carefully and critically evaluating website information within and across sites.

Based on the pattern of results for the study of dimensionality, we hypothesized that ORCA-Multiple Choice scores would again be unidimensional while ORCA-Closed scores would suggest multidimensionality. We also hypothesized that location, evaluation, synthesis, and communicate would contribute in roughly equivalent proportions to score validity. However, the skills set may vary in difficulty. For example, in the pilot year, we determined that the evaluation tasks were most difficult. Surprisingly, and for the multiple-choice scales, synthesis
tasks were easy, and as such, revisions were made to introduce difficulty and more accurately reflect the definition of synthesis as described in our introduction.

**Method**

**Sample**

Response patterns of 1293 students were analyzed. Six hundred twenty-five participants (310 boys, 315 girls) provided responses for the ORCA-Closed assessment. Six hundred sixty-eight 7th grade students (310 boys, 358 girls) completed the ORCA-Multiple Choice assessment. While students completed assessments in both formats on two separate days of testing, the reliability and validity information reported in the present investigation focuses on the first administration of testing to eliminate any carry-over effects such as familiarity with types of LESC tasks included on the scales.

**Instruments**

**ORCA-Closed.** The ORCA-Closed was a performance based assessment, a functioning simulation of the Internet, designed to measure online research and comprehension ability in science using the topic of human body systems. An avatar of another student guided each student through the research task with text messages. This simulation of the Internet included a fully functioning search engine, web pages, email, wiki, text messaging, and a notepad. In the research problem “How do energy drinks affect heart health?” Brianna first asked students to check their email inbox to locate a message from the Principal of a middle school. (See Figure 2.) This email defined the research task. The Principal indicated that the President of the School Board was concerned about having energy drinks at school. She asked students to conduct research on how energy drinks affected heart health, using the Internet, and then send an email to the School Board President with a short report of the findings.
An extensive collection of web pages was imported into the assessment space for use with our simulation of the Google search engine (Gloogle). Fictitious teachers, principals, and students, represented with avatars, prompted each student throughout the research process within the social network interface via text messages. See Figures 3-6.

Two other interactive scenarios, also within a simulation of the Internet, used a wiki for writing the final report of research on two of the research topics: “Do video games harm your eyes?” and “Do cosmetic contact lenses harm your eyes?” Each asked students to take a position on an issue.
All four assessments followed a parallel structure, where students were asked to locate four different websites, synthesize information across them, and critically evaluate one of the sites. Students were then asked to write a short report in an email message or on the class wiki. A video example of a student taking the ORCA-Closed appears at: http://neag.uconn.edu/orca-video-ira.

**ORCA-Multiple Choice.** The ORCA-Multiple Choice format used a scenario context with the identical research problems and the same research questions as the ORCA-Closed. Online tools in multiple-choice scenarios had some, but very limited, functionality. Example items from an ORCA-Multiple Choice appear in Figure 7.

A video more fully describing the ORCA-Closed and ORCA-Multiple choice appears at: http://youtu.be/aXxrR2wBR5Y.

**Scoring**

Each scenario, a full simulation of the Internet in the ORCA-Closed and a limited simulation of the Internet in the ORCA-Multiple-choice, formed a testlet (Wainer, Bradlow, & Wang, 2007) called a LESC to represent the four skill areas assessed (i.e., Locate, Evaluate, Synthesize, Communicate). Each LESC contained 16 total score points (see Table 1), with four points assigned to each skill area. Each of the 16 score points evaluated an online research and comprehension skill that had been identified from previous research, national frameworks for English Language Arts, and through discussions with researchers in this area.
Each of the four skill areas (Locate, Evaluate, Synthesize, Communicate) included three process skills and one product skill. Four experts in online research and comprehension scaled the three process skills by the likely order of difficulty, so that each skill was considered more difficult than the one before. Each of the four product skills was considered to be a culminating task for its given area, and therefore was intended to be the most difficult of the four score points in that area.

The LESC components did not appear in a strictly linear sequence (e.g., the assessment did not begin with Locate tasks, followed by Evaluate tasks, etc.), nor did the four skills that were evaluated within a component. Instead, a more logical and natural sequence of events developed in the scenario. The one exception was the evaluation sequence, which asked students to evaluate one of the web pages with four sequential requests from the student avatar: 1) identify the author (process); 2) evaluate the expertise level of the author (process), 3) evaluate the author’s argument (process), and 4) evaluate the reliability of the website (product).

A back-end, data capture system was developed to record and track students’ online reading decisions for subsequent scoring. Video screen captures were also used for a richer interpretation of student performance and served as a backup for the data capture system.

Two graduate students scored all assessments. They evaluated each assessment following a common rubric for each of the 16 score points. Each score point was evaluated using a binary (i.e., 0 or 1) scoring system. Scorers were initially trained on a common set of 10 assessments. Then, they were each tested for accuracy on another set of 10 assessments, and were required to
reach 90% inter-rater agreement for each one of the 16 score points before being allowed to score the actual student assessments. The scorers compared their scoring at several points throughout to reevaluate their reliability of scoring decisions. Each time this reliability check was conducted, inter-rater reliability met or exceeded 90% for each score point, within each assessment. Any disagreements were resolved through discussion.

**Administration**

The administration of the assessments took place within a wifi context on MacBook Airs, specifically prepared to rapidly take students to the online assessment location with the format and topic assigned to them as well as to initiate the video screen capture. Test administration was led by a trained test administrator and followed a standard protocol with groups of up to 25 students. Students had as much time as needed to complete each assessment, though most students finished within about 45 minutes for the simulated Internet LESCOs, or ORCA-Closed, and within about 25 minutes for the ORCA-Multiple choice. 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.

**Data Analysis Plan**

Research Question 1. To study the validity of scores, we ran exploratory factor analyses (EFAs) separately for each LESC scale and for each format. Specifically, we used the Mplus statistical software package (Muthén & Muthén, 1998-2006) as it permits reduction of tetrachoric correlation matrices and provides model-fit indices to allow for comparing and contrasting factor analytic solutions. An oblique, GEOMIN, rotation was used for all analyses as we hypothesized factors would be correlated.
We decided to use EFAs with model-fit indices rather than CFAs for multiple reasons. First, as Marsh et al. (2013) explained, CFAs test the null hypothesis with factor loadings of 1.00 for items that are theorized to load onto specific dimensions. These unit weights may be significantly larger than those that indicate the actual relation of the unobserved latent trait with the item indicator. Second, and from the pilot study (e.g., Cui et al., 2013), we had evidence that factors would be correlated, and as a result, scores could be collapsed to provide a more parsimonious scale interpretation than for several multiple factors. Finally, and as described previously, several items were revised based on the pilot study. As such, an exploratory approach to the study of structure is acceptable (e.g., Preacher, Zhang, Kim, & Mels, 2013) in an effort to describe a best model to characterize the responses selected or constructed by students.

Several fit statistics were used to report and interpret results. These included the classical $\chi^2$ test. We also studied the comparative fit indices (CFIs) and interpreted values of .90 to .95 as acceptable fit (e.g., Trautwein, et al. 2012). Finally, the root mean square error of approximation (RMSEA) with values less than .06 and the standardized root mean square residual (SRMR) with values less than .09 were used to indicate good fit and as based on research by Marsh, Balla, and Hau (1996) and Marsh, Hau, and Wen (2004).

Based on fit indices as well as the interpretability of solution, we then calculated the KR-20 value for each scale. We also examined how item quality changed from the pilot year to the validation year given the modifications or revisions made to the scales. Specifically, we studied how item difficulty estimates (i.e., too difficult, too easy) and discrimination indices improved or remained the same. Chi-square tests were run to study the degree of improvement based on changes in the construction of items between pilot and validation project years.
Research Question 2. To address how well the location, evaluation, synthesis, and communication items, respectively, contributed to the variance of scales, we studied the sizes of factor loadings and estimated the proportion of variance explained. We also examined the level of difficulty for each set of four items that were designed to represent a specific online research and comprehension skill. Finally, we analyzed the set of four skills relative to one another to determine the average order of difficulty on the assessments. For example, in the pilot year, the evaluation items were the most challenging for students. Other researchers (e.g., Goldman et al., 2012) have documented similarly how difficult evaluation of information can be, yet students must learn how to judge critically sources for their accuracy, consistency, and relevance in 21st century, web-based scientific inquiry and problem solving.

**Results**

Results are reported first for the multiple-choice scales (ORCA-Multiple Choice). We repeated the analyses for the simulated Internet scales (ORCA-Closed).

**The Multiple-Choice Scales**

**Energy Drinks.** The best and most simple solution was a two-factor summary: $\chi^2$ (89) = 96.60, $p = .27$; CFI = .93; RMSEA = .02; SRMR = .11. Only two items, which were communication items (i.e., Can the student include a correct address in an email message?; Can the student identify/select a well-constructed email message?) were correlated with Factor 2. Given the pattern of factor loadings, we determined that a unidimensional scale was the best interpretation of results. KR-20 was .73.

**Video Games.** The model-data fit indices for a single factor model were acceptable for this scale. The $\chi^2$ was not significant, $p > .06$; CFI was equal to .94; RMSEA was .036. While SRMR was .12, only two items had factor loadings less .30. For the remaining 14 items, the
range of factor loadings was .39 to .79. KR-20 was .85.

The two items that did not contribute as much to the variance of scores as the other items were the second synthesis item: “Can the student synthesize important elements from two websites?”; and the second communication item, which was a Wiki task: “Can the student include an appropriate heading for a new Wiki entry?” (see Table 1).

**Heart Healthy Snacks.** Like the Energy Drinks scale, a two-factor summary described item responses best. Goodness-of-fit indices were: $\chi^2 (89) = 101.10$, $p = .18$; $CFI = .94$; $RMSEA = .03$; $SRMR = .09$. As we observed for the Energy Drinks scale, the two items that contributed most to characterizing the second dimension were two of the communication items: 1) Can the student include a correct address in an email message?; and, 2) Can the student identify/select a well-constructed email message? KR-20 for the scale was .77.

**Cosmetic Contact Lenses.** Fit statistics were acceptable for a single-factor summary: $\chi^2 (104) = 121.29$, $p = .12$; $CFI = .95$; and, $RMSEA = .03$. While $SRMR = .10$, only three items had loadings less than .30. These tasks included the first two process-point items for the location skill set as well as for the first communication item (i.e., “Can the students make a Wiki entry in the correct location?”). KR-20 was .80.

**ORCA-Closed Scales**

**Energy Drinks.** Three factors best summarized the scores for this scale: $\chi^2 (75) = 94.74$ was a significant reduction than for the one-factor model, $\chi^2 (104) = 215.86$, $p < .001$. Other goodness-of-fit indices for the three-factor solution were: 1) $CFI = .96$; 2) $RMSEA = .042$; and, 3) $SRMR = .11$. All four evaluation items, three synthesis items, and one communication item were correlated with the first factor. Factor 2 consisted of two location items (i.e., Can students use appropriate key words in a search engine? Can students locate the correct site in a set of
search engine results?) and one synthesis item (i.e., Can the student summarize an important element from one website?). Factor 3 items included two locate tasks (i.e., Can students locate the correct email message in an inbox or the correct section of a wiki? Can students locate and share correct website addresses in two different search tasks?) and three of four communication items. KR-20 for the complete scale was .88.

**Video Games.** As for Energy Drinks, based on goodness-of-fit indices and interpretability of factor structure, a three-factor solution summarized the data best. Reduction of the $\chi^2$ statistical value for the three-factor solution compared to the one-factor solution was significant, $p < .001$. The CFI was .98, and the RMSEA and SRMR values were .05 and .09, respectively. Factor 1 included locate and synthesis items. Factor 2 items were the evaluation tasks, and Factor 3 consisted of the communication items. KR-20 was .90

**Heart Healthy Snacks.** A two-factor solution demonstrated significantly better fit, $\chi^2 (89) = 88.60$, than a one-factor solution, $\chi^2 (104) = 139.10$, $p < .001$. CFI, RMSEA, and SRMR values were 1.00, .00, and .09, respectively, for the two-factor structure. As observed for the Energy Drinks solution, Factor 1 items included two location tasks (i.e., appropriate key word use; identifying correct site in a set of search engine results) as well as one of the communication task: Can the student include an appropriate subject line in an email message? All evaluation and synthesis tasks along with the remaining location and communication items were correlated with Factor 2. KR-20 was .86.

**Cosmetic Lenses.** A three-factor solution best summarized the item scores for this scale: $\chi^2 (75) = 91.061$, CFI = .97; RMSEA = .03; SRMR = .09. Factor 1 included three of the four location items. All evaluation and synthesis items were correlated with Factor 2 along with two communication tasks (i.e., Can the student include an appropriate heading for a new wiki entry?;
Can students compose and post a well-structured, short report of their research, including sources, in a wiki?) Items that loaded onto Factor 3 were the first location task (i.e., Can students locate the correct email message in an inbox or the correct section of a wiki?) and the remaining two communication items (i.e., Can the student make a wiki entry in the correct location? Can the student use descriptive voice in an informational wiki?). KR-20 was .88 for the full scale.

**Summary.** Both formats demonstrated good estimates of reliability with somewhat higher levels of reliability reported for the ORCA-Closed format. KR-20 values ranged from .73 to .85 for the ORCA-Multiple Choice format and from .86 to .90 for the ORCA-Closed format.

Both formats also demonstrated good estimates of validity although factor structures were more complex for the ORCA-Closed than for the ORCA-Multiple Choice. CFI values ranged from .93 to .95 for the multiple-choice scales while RMSEA values were less than .05. These fit statistics match recommendations in the literature for selecting the number of dimensions that best summarize the item response data (e.g., Marsh et al., 1996; Marsh et al., 2004). In addition to the acceptable fit statistics, the solutions have to be interpretable and align either with theoretical expectations or observations made in practice. Across the four multiple-choice solutions, scores reduced as one general factor, which matched expectations. Further, and additionally, if there were a departure from unidimensionality as observed for the Energy Drinks and Heart Healthy Snacks scales, then the communication items were likely to contribute to this deviation from one general scale. Based on our definitions of the four skills, communication appears different than location, evaluation, and synthesis as it requires at minimum a transition from using strategies associated with reading to those related to writing.

For the ORCA-Closed scales, we also reported acceptable fit statistics in the ranges of .96
to 1.00 for CFI and with RMSEA equal to or less than .05. However, solutions were much more difficult to describe for the ORCA-Closed than for the ORCA-Multiple-choice. First, the best interpretations of scores were those for multiple dimensions and not a single one. For 3 of the 4 science topic scales, a three-factor solution was selected as the best one based on fit statistics. Study of the factor loadings revealed that evaluation and synthesis items were likely to load on the same factor while location and communication items would likely characterize additional dimensions. The pattern for communication items is similar to some we observed for the multiple-choice scales. Location items introduce a new pattern in describing departure from unidimensionality. However, and based on the literature, this pattern makes sense. For the ORCA-Closed, location is likely a bottleneck skill as the entire assessment experience is arguably initiated by correct use of keywords or selections of websites to access information required to solve the remaining problems. In fact, these were patterns noted in the study of the factor structures for the Energy Drinks and Heart Healthy Snacks assessments.

Finally, there are two final likely explanations as to why description of score patterns is simpler for multiple-choice than for ORCA-Closed. First, selecting multiple-choice options is arguably easier than constructing or generating responses. Second, the multiple-choice scales required significantly less interactivity with design features and tool use than did the ORCA-Closed scales.

**Improvements In Item Characteristics From Pilot Year To Validation Year.**

Before we report the results to address our second research question, we thought it important to summarize the degree to which item characteristics improved after revision based on our pilot study findings from the previous year. Significant changes were made between the
pilot and validation years to increase difficulty for items that were too easy, reduce difficulty for items that were too difficult, and replace items that had negative item discrimination indices.

Based on the results of the pilot study (Leu, et al., 2012), 36/128 items (28.1%) in ORCA-Multiple Choice and ORCA-Closed fell outside our item difficulty or item discrimination indices and were identified for revision. Revisions were made and 32/34 items (94.17%) improved after revision. Among the 28 multiple-choice items that required revisions, 27 items (96.43%) improved after changes were made. Among the 6 ORCA-Closed items that required revisions, 5 items (83.33%) improved after changes were made.

In the pilot year, 94 (73.4%) out of 128 items were considered to have good item characteristics (i.e., item difficulty or item discrimination indices). After revisions were made, 107 (83.6%) out of 128 items were either improved or maintained as having good quality given the validation year results. Overall, and across both formats, revision of the items significantly improved the proportion of items that fell within our acceptable item difficulty and item discrimination index ranges, $\chi^2(1) = 6.768$, $p=.009$. Revision of the ORCA-Multiple Choice items significantly improved the proportion of good items, $\chi^2(1) = 30.73$, $p<.001$. Revision of the ORCA-Closed items also significantly improved the proportion of good items, $\chi^2(1) = 14.897$, $p<.001$. These results are reported in Table 2.

| TABLE 2 ABOUT HERE |

Compared to the ORCA-Closed scales, the ORCA-Multiple Choice scales had more need for item improvement (43.75%, 28 out of 64 items) based on the findings of the pilot study. Most good items (86.11%, 31 out of 36 items) in the pilot year continued to function well in the validation year. For the ORCA-Closed scales, only 6 tasks were in need of revision. However, 14
of the original 58 good items (24.14%) became more difficult in the validation year. But, this difficulty also contributed to the increased variance of the score distributions, and as such, was related to the increases in the reliability estimates.

Specifically, KR-20 estimates for the multiple-choice Energy Drinks, Video Games, Heart Healthy Snacks, and Cosmetic Lenses were .67, .80, .57, and .59, respectively, in the pilot year. By comparison, validation year KR-20 estimates for the four multiple-choice assessments were .73, .85, .77, and .81, respectively. The KR-20 estimates for ORCA-Multiple Choice increased by an average of .135.

For the ORCA-Closed scales, initial reliability estimates were .72 for Energy Drinks, .82 for Video Games, .69 for Heart Healthy Snacks, and .73 for Cosmetic Lenses. Validation year estimates increased to .87 for Energy Drinks, .90 for Video Games, .86 for Heart Healthy Snacks, and .88 for Cosmetic Lenses. The KR-20 estimates for ORCA-Closed increased by an average of .138.

**Research Question 2**

Results that addressed the second research question focused on the contributions of the Location, Evaluation, Synthesis, and Communication tasks to ORCA scales. Additionally, we wanted to study how these components varied by difficulty.

**Explained variance.** Table 3 and Figure 8 present and display descriptive statistics that illustrate how much variance was explained by each of the four skill sets that were included on both the ORCA-MC and ORCA-Closed assessments.

-------------------------------

TABLE 3 ABOUT HERE

-------------------------------
For the multiple choice tests, on average, communicate items accounted for the greatest percentage of explained variance followed by the synthesis and evaluation items. Location items contributed the least amount of variance explained given the set of four skills. For the ORCA-Closed, on average, synthesis scores contributed the largest amount of variance. The other two categories, evaluation and communication, explained a relatively similar amount of variance.

For the multiple choice scales, the relative amount of variance contributed by the four categories was not consistent across the four science topics, $\chi^2(9) = 27.805$, $p=.001$. For the ORCA-Closed tests, the relative amount of variance contributed by the four categories was also not consistent across the four topics, $\chi^2(9) = 39.601$, $p<.001$. This may suggest differential levels of prior knowledge existed among student in the areas of: energy drinks and heart health, heart health and snack foods, video games and eye health, and cosmetic contact lenses and eye health.

**Skill component difficulty.** Table 4 and Figure 9 present information about the difficulty of each of the four skill components. For the multiple-choice scales, item difficulties are relatively similar across the four skill components; the rates of successful scores ranged from 61.0% for Locate scores to 68.6% for Evaluate scores. Evaluate items were easiest, followed by those that measured synthesis and communicate. Items that assessed skill at locating information were the most difficult.

TABLE 4 ABOUT HERE
For the ORCA-Closed, item difficulty estimates were not similar across the four skill sets. Items that measured synthesis were the easiest with students getting 59.9% of these items correct. This was followed by those that measured locate, with students getting 45.9% of items correct and evaluate, with students getting 35.8% of items correct. Items that measured communicate were most difficult, with students getting only 26.2% of items correct.

The low level of performance on the communicate items may be due to the infrequency with which schools in both states integrated the use of email or other communication tools in classrooms. It may also be due to students bringing texting conventions to other, less familiar, communication tools. In particular, we noticed a number of students in the email task simply clicked the “reply” button to the message from the principal rather than creating a new message to the School Board President. With text messaging, of course, we typically reply to someone's message. In addition, many students left out the subject line to their email messages. Text messages, of course, do not have subject lines. Finally, many students were insensitive to the need for a greeting in a message to an unfamiliar person with somewhat high status, the School Board President. In text messaging, we typically communicate with familiar people without using a greeting in a message. In short, the conventions of a more familiar communication tool (text messaging) may have impeded performance with a less familiar communication tool (email).

Items that measured synthesis had similar difficulty across the formats of the assessments (i.e., multiple choice and ORCA-Closed). However, items measuring locate, evaluate, and
communicate were more difficult on the ORCA-Closed scales than on the multiple choice tests. This may be due to very little online tool use being required to complete a summary, or synthesis, in the notepad, other than keyboard use. Other areas seemed to require more use of online tools such as search engines, links on web pages, email tools, or wiki tools.

**Discussion**

This investigation had two goals. First, we sought to evaluate the psychometric properties of scores with different performance based formats for assessing students’ abilities with online research and comprehension. To this end, reliability and validity were evaluated for ORCA-Closed and ORCA-Multiple Choice, two formats that varied in the nature of item responses and the extent to which they simulated the Internet. The ORCA-Closed was a performance based assessment with interactive scenarios containing a fully functioning simulation of the Internet. A majority of items required constructed responses. The ORCA-Multiple Choice was also a performance based assessment using short scenarios that provided the context for each item with a stem and options that students selected. The online tools in multiple-choice scenarios had some, but very limited, functionality. Multiple-choice scales were designed to be a more restricted and more limited simulation of the Internet than the ORCA-Closed scales.

Second, the study sought to examine the contributions of several components to the explained variance of the scales. Each ORCA contained a similar LESC structure and estimated performance on a set of 16 similar tasks to represent the four skill areas assessed (i.e., Locate, Evaluate, Synthesize, Communicate). Each LESC contained 16 total score points with four points assigned to each skill area, previously identified from several sources: research, national frameworks for English Language Arts, and discussions with researchers in this area.
The Psychometric Properties Of Scores

Reliability. This investigation indicated that scores for the scales in both the ORCA-Multiple Choice and the ORCA-Closed formats are reliable. The ORCA-Closed format demonstrated higher levels of reliability, with KR-20 estimates ranging from .86 to .90. Both formats benefited from the pilot year testing and from the targeted changes that were made to items based on item difficulty and item discrimination information. Gains in KR-20 estimates ranged from .135 (ORCA-Multiple Choice) to .138 (ORCA-Closed). Changes made to items from the pilot year resulted in improvements to their performance in the vast majority of cases.

The reliability estimates in this study suggest that both formats may be useful tools to estimate student performance with online research and comprehension ability. The skills associated with online research and comprehension will be important to students in an online age of information.

Validity. We followed recommendations by Marsh et al. (1996) to compare and contrast exploratory factor analysis (EFA) solutions for each multiple-choice and ORCA-Closed scale. Goodness-of-fit statistics supported that factor structures for the multiple-choice scales were more unidimensional than for the ORCA-Closed scales. Added dimensionality for both types of scale was related to the contributions of communication item scores. Location items also introduced extra dimensions in summarizing score patterns for the ORCA-Closed assessments. Evaluation and synthesis items were correlated for both ORCA formats, and this result conforms with hypothesized relations. Evaluation and synthesis, while they required less interactivity with tool functions than location or communicate, also rely significantly on critical analysis of information on websites that must be judged for accuracy, consistency, and lack of bias.
**Item difficulty by format.** Overall, the ORCA-Multiple Choice format provided an easier task environment for students compared to the ORCA-Closed, which was harder. On average, students generated correct responses about one-third more often for assessments in the ORCA-Multiple Choice format (64.6% of items) compared to the ORCA-Closed format (42.0% of items).

The lower mean scores for ORCA-Closed items may have been due to the additional interactions with various online tools required in a more complete simulation of the Internet, the constructed nature of the responses, a complete online research and comprehension task, or all three of these differences with the ORCA-Multiple Choice format. Each meant that the ORCA-Closed took place in an assessment environment with higher fidelity to the actual Internet.

It is also important to recall that the ORCA-Multiple Choice provided a somewhat extensive scenario description for each separate score point, thus reminding students of the conditions for each of the 16 separate contexts where a response was required. These interpretations and reminders did not appear in the ORCA-Closed. It is possible that the somewhat extensive text scenarios required in ORCA-Multiple Choice may maximize demands for offline reading skills and minimize demands for online reading. This might raise concerns about the ability of scenario strategies to represent the online comprehension experience with high fidelity.

In either case, the results of this study suggest formats that do not present a complete and integrated problem-solving task, do not simulate the Internet context with high fidelity, and do not require constructed responses are likely to estimate actual students’ online research and comprehension performance at a higher level than formats with greater fidelity to the actual Internet. Since both instruction and research benefit from an accurate representation of
performance, this pattern suggests that greater fidelity to Internet contexts may be important to consider when developing assessments of online research and comprehension. We need to continue to investigate how best to create simulations of online research and comprehension that maximize fidelity to online contexts while also providing constructed response opportunities and time-efficient scoring procedures.

**Item difficulty by format and component.** Item difficulty analysis also revealed that the difficulty of LESC components was similar in the ORCA-Multiple Choice format but quite different in the ORCA-Closed format (See Figure 4.) Looking at the item difficulties in ORCA-Closed, synthesis had the highest mean item difficulty score, with students getting nearly 60% of these items correct. On the other hand, students taking the ORCA-Closed generated correct responses only about one-third of the time for evaluate (35.8%) and less than this for communicate (26.2%) items. Noticeably, item difficulty levels in these two areas were approximately twice as high in the format with less fidelity to online contexts using a scenario approach and multiple choice responses (ORCA-Multiple Choice). The low level of performance in these two areas within the format with greater fidelity to online contexts may suggest that these areas are particularly challenging for students, an observation that would not have been made from item difficulty data in the ORCA-Multiple choice format.

The similarity of component difficulty for the format that less successfully simulates the functionality of tools needed for online research and comprehension (ORCA-MC), along with the more substantial variation in component difficulty for the format that more successfully simulates the Internet (ORCA-Closed) is important to note. This greater variation in component scores may suggest that more authentic simulations provide better estimates of student
performance within somewhat distinct skill areas, providing important direction for instruction in online research and comprehension.

**Examining The Contributions Of Several Components To The Explained Variance Of The Scales**

Table 3 presents a summary of information about the contributions of each of the four skills to explanation of score patterns. Based on our report about the EFA structures, we concluded that all skills informed a description about the dimensionality of scores. However, there is additional information about these contributions as shown in Table 3.

For the multiple-choice scales, it is clear that the percentages of variance explained in scores, on average, are related most to the communication items. Across the four scientific topics, the percentages were 20.84 to 28.76%, respectively. However, we believe it important to note that the percentages appeared to change by scientific topic as well as by communication tool featured in the LESC design (i.e., email vs. wiki), and as such, scientific topic and communication medium are likely important moderator variables to study in future research investigations. For example, for the Energy Drinks and Heart Healthy Snacks scales, which both required email correspondence, the contribution of the locate tasks was very low, 18.58% and 9.82%, respectively, compared to the percentages contributed by communication, 30.27% and 39.96%, respectively.

In contrast and by comparison, location items contributed most to the explained variability of the Video Games scores (i.e., 32.59%). This multiple-choice scale is arguably the best of all multiple-choice LESCs given the reliability and validity information reported about it. Video Games scores also depended on the use of a wiki medium, rather than email, given its tasks that required communication of information.
For the ORCA-Closed assessments, one observation given the summary of results presented in Table 3 seems clear. Synthesis items contribute significantly in explaining the pattern of results with percentages that ranged from 34.00 to 38.30%. Only for Video Games did communication play more of a role than for synthesis in helping to describe score patterns. Additionally, and for this scale, evaluation scores did not contribute much to the operational definition of scores (i.e., 8.34%).

We could present other interpretations given the descriptive results displayed in Table 3. The point that we would like to make is that interactions among skill sets (i.e., location through communication) and scientific topics as well as communication media are likely, and we recommend these interactions inform the study for future research investigations.

**ORCA Administration**

ORCAs, in either format, appear to be easy to administer, a factor one might refer to as feasibility or utility validity. One suspects that the auto scoring process of the ORCA-Multiple choice makes this format especially attractive to schools and researchers with limited resources and unable to provide the hand scoring currently required for 9/16 items in the ORCA-Closed. Scoring procedures for ORCA-Closed are easy to follow and lead to consistency in judges’ agreement as to assignment of scores but each one requires about 5 minutes to score and some training time to establish reliability of scoring. These are important considerations. On the other hand, the item difficulty data suggest that the ORCA-Multiple Choice may overestimate performance and may not accurately represent performance in terms of LESC patterns.

**Limitations**

There are several limitations to acknowledge. First, and for a validation study, our sample sizes may be viewed as small to evaluate factor structures for the multiple scales we designed
and for which we studied score patterns. However, we did so in an effort to inform our understanding about item characteristics that would add value to an assessment experience as one more authentic and as that realized in interacting every day with the Internet rather than in the classical, standardized completion of multiple-choice assessments where items are likely perceived as significantly not associated with one another.

In total, we examined summaries for eight scales. There were four scientific topics studied across two formats: ORCA-Multiple Choice and ORCA Closed. Our sample sizes ranged from a low of 142 to a high of 206 across the samples of participants who provided scores for us given study of each scale.

For the 16 items on any one scale, some psychometricians (e.g., Everitt, 1975) recommend a ratio of participants to items as 10 to 1 (i.e., 160 participants per ORCA scale) at minimum in efforts to examine validity and reliability. With the dichotomous-scoring of items, estimation of factor loadings is additionally complex for cognitive and motivational measures as tetrachoric coefficients must be estimated rather than phi coefficients to produce the correlation matrix of item-pair associations to be analyzed statistically via factor analyses (e.g., Bonett & Price, 2005).

Additionally, we acknowledge what can be viewed as only “internal examinations” of score distributions by scientific topic/format scale. Specifically, the results of our present investigation focus on internal consistency estimates and inter-item correlation matrices for each scale considered. As such, we did not include study of any external criteria to add to the interpretation of the psychometric properties as recommended, and for example, in the research of Milligan and Cooper (1988) and as evidenced in the empirical work of Lawless and Kulikowich (1996, 1998) in their examination of how students interacted with hypertext menus.
Other external criteria, in addition to prior knowledge, are as important to consider. While not inclusive, the list would include variables such as offline reading comprehension ability (Coiro, 2011) that may be related to responses selected or constructed for evaluation and synthesis tasks. Additionally, it appears from our results that students must have some perception of the affordances (e.g., CTGV, 1990, 1992) of tool use such as search engines or communication medium to acquire points on the scales designed.

Finally, and more than 20 years ago, literacy researchers like Patricia Alexander (e.g., Alexander et al., 1994), Ruth Garner (e.g., Garner, Gillingham, Alexander, Kulikowich, & Brown, 199x), Suzanne Hidi and her colleagues (e.g., Hidi, Renninger, & Krapp, 1992) and Suzanne Wade and her colleagues, as well as many other researchers (e.g., Schiefele et al.), realized that interest about topics was an important motivation element to learning. Should it be no surprise to us that Video Games, of all topics, has the most simple of factor-analytic solutions to describe as well as the highest estimates of reliability for either ORCA-Multiple-Choice or ORCA-Closed?

Summary

This study described two formats of performance based assessments of online research and comprehension from the ORCA (Online Research and Comprehension Assessment) Project. The formats varied in terms of their fidelity to the Internet context for conducting online research and comprehending information online while reading. Analysis indicates that both formats are reliable and valid. The study also examined the contributions of several components to the explained variance of the scales. Analysis indicates that ORCA-Multiple Choice, the more restricted assessment context with less fidelity to the Internet context, yields performance estimates that are higher, more similar between components, and with unidimensional scales. It
also indicates that the ORCA-Closed format, with a less restricted assessment context and greater fidelity to the Internet context, yields performance estimates that are lower, dissimilar between components, and with multidimensional scales. The results provide possible direction for thinking about developing assessments of online information use in ways that have good fidelity to the Internet while also providing a stable assessment context.

References


Table 1.  

The 16 Score Points In Each Assessment (Both Email And Wiki) And For Both Formats (Orca-Closed And Orca-Multiple Choice) Organized By Locate, Evaluate, Synthesize, And Communicate Components. Items 1-3 Are Process Score Points; Items 4 Are Product Score Points

<table>
<thead>
<tr>
<th>Skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading to Locate Online Information</td>
</tr>
<tr>
<td>1.  Can students locate the correct email message in an inbox or the correct section of a wiki?</td>
</tr>
<tr>
<td>2.  Can students use appropriate key words in a search engine?</td>
</tr>
<tr>
<td>3.  Can students locate the correct site in a set of search engine results?</td>
</tr>
<tr>
<td>4.  Can students locate and share correct website addresses in two different search tasks?</td>
</tr>
<tr>
<td>Reading to Evaluate Online Information</td>
</tr>
<tr>
<td>1.  Can students identify the author of a website?</td>
</tr>
<tr>
<td>2.  Can students evaluate an author’s level of expertise?</td>
</tr>
<tr>
<td>3.  Can students identify an author’s point of view?</td>
</tr>
<tr>
<td>4.  Can students evaluate the reliability of a website?</td>
</tr>
<tr>
<td>Reading to Synthesize Online Information</td>
</tr>
<tr>
<td>1.  Can the student summarize an important element from one website?</td>
</tr>
<tr>
<td>2.  Can the student synthesize important elements from two websites?</td>
</tr>
<tr>
<td>3.  Can the student synthesize important elements from a second set of two websites?</td>
</tr>
<tr>
<td>4.  Can the student synthesize important elements from the websites in the research task to develop an argument?</td>
</tr>
<tr>
<td>Writing to Communicate Online Information</td>
</tr>
<tr>
<td>1.  Email and Wiki: Can the student include the correct address in an email message and make a wiki entry in the correct location?</td>
</tr>
<tr>
<td>2.  Email and Wiki: Can the student include an appropriate subject line in an email message and an appropriate heading for a new wiki entry?</td>
</tr>
<tr>
<td>3.  Email and Wiki: Can the student include an appropriate greeting in an email message to an important, unfamiliar person and use descriptive voice in an informational wiki?</td>
</tr>
<tr>
<td>4.  Email and Wiki: Can the student compose and send a well-structured, short report of their research, including sources, in both email and wiki contexts?</td>
</tr>
</tbody>
</table>
Table 2.

*Improvement in Item Quality from Pilot Year to Validation Year*

<table>
<thead>
<tr>
<th>Pilot Year</th>
<th>Validation Year</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ORCA-Multiple Choice</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Outside</td>
<td>Inside</td>
</tr>
<tr>
<td>Inside</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>Outside</td>
<td>1</td>
<td>27</td>
</tr>
</tbody>
</table>

*Inside* = Number of items that fell within parameters for our item difficulty and item discrimination indices.

*Outside* = Number of items that fell outside parameters for our item difficulty and item discrimination indices.

Note. Each cell displays the number of items counted across the four topics.
Table 3.

Percentage of Contribution by L, E, S, C to Explained Variance (Day 1)

<table>
<thead>
<tr>
<th>Test Format</th>
<th>Topic</th>
<th>LOCATE</th>
<th>Evaluate</th>
<th>Synthesize</th>
<th>Communicate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>% Variance</td>
<td>% Variance</td>
<td>% Variance</td>
<td>% Variance</td>
</tr>
<tr>
<td>ORCA-MC</td>
<td>ED</td>
<td>3.12</td>
<td>18.58</td>
<td>4.62</td>
<td>27.56</td>
</tr>
<tr>
<td></td>
<td>VG</td>
<td>9.04</td>
<td>32.59</td>
<td>7.29</td>
<td>26.28</td>
</tr>
<tr>
<td></td>
<td>HHS</td>
<td>1.86</td>
<td>9.82</td>
<td>3.61</td>
<td>19.07</td>
</tr>
<tr>
<td></td>
<td>CL</td>
<td>5.17</td>
<td>22.36</td>
<td>4.06</td>
<td>17.59</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>4.80</td>
<td>20.84</td>
<td>4.90</td>
<td>22.63</td>
</tr>
<tr>
<td>ORCA-Closed</td>
<td>ED</td>
<td>8.86</td>
<td>26.56</td>
<td>9.52</td>
<td>28.51</td>
</tr>
<tr>
<td></td>
<td>VG</td>
<td>8.19</td>
<td>20.94</td>
<td>3.26</td>
<td>8.34</td>
</tr>
<tr>
<td></td>
<td>HHS</td>
<td>4.50</td>
<td>15.18</td>
<td>5.43</td>
<td>18.29</td>
</tr>
<tr>
<td></td>
<td>CL</td>
<td>5.96</td>
<td>18.10</td>
<td>9.38</td>
<td>28.48</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>6.88</td>
<td>20.20</td>
<td>6.90</td>
<td>20.91</td>
</tr>
</tbody>
</table>

Table 4

*Average Item Difficulty for LESC (Day 1)*

<table>
<thead>
<tr>
<th>Test Format</th>
<th>Topic</th>
<th>Locate</th>
<th>Evaluate</th>
<th>Synthesize</th>
<th>Communicate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORCA-MC</td>
<td>ED</td>
<td>.704</td>
<td>.775</td>
<td>.560</td>
<td>.706</td>
</tr>
<tr>
<td></td>
<td>VG</td>
<td>.680</td>
<td>.622</td>
<td>.605</td>
<td>.622</td>
</tr>
<tr>
<td></td>
<td>HHS</td>
<td>.550</td>
<td>.754</td>
<td>.725</td>
<td>.641</td>
</tr>
<tr>
<td></td>
<td>CL</td>
<td>.507</td>
<td>.591</td>
<td>.702</td>
<td>.586</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>.610</td>
<td>.686</td>
<td>.648</td>
<td>.639</td>
</tr>
<tr>
<td>ORCA-Closed</td>
<td>ED</td>
<td>.509</td>
<td>.370</td>
<td>.786</td>
<td>.161</td>
</tr>
<tr>
<td></td>
<td>VG</td>
<td>.631</td>
<td>.388</td>
<td>.511</td>
<td>.307</td>
</tr>
<tr>
<td></td>
<td>HHS</td>
<td>.433</td>
<td>.345</td>
<td>.554</td>
<td>.249</td>
</tr>
<tr>
<td></td>
<td>CL</td>
<td>.263</td>
<td>.328</td>
<td>.543</td>
<td>.331</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>.459</td>
<td>.358</td>
<td>.599</td>
<td>.262</td>
</tr>
</tbody>
</table>

Figure 1. Measurement model of one latent construct with four indicator variables.
Figure 2. The opening sequence showing the text (chat) interface and the avatar directing the student to an email message containing the research problem and context defined the research task in the ORCA-Closed assessment: “Are Energy Drinks Heart Healthy?”
Figure 3. Locate. The search engine, “Gloogle,” showing keyword entry and search results from the ORCA-Closed assessment, “Are energy drinks heart healthy?”
Figure 4. Evaluate. The sequence of evaluation tasks appeared in a scrolling chat window. (See window A.) Students first had to select the link to “Read about me” in Window B to obtain the information about the author that appears. Score points for Evaluation included: 1) identifying the author; 2) determining the author’s expertise; 3) determining the author’s point of view; and 4) evaluating the reliability of the information at the site.
Figure 5. Synthesis. While locating four sites, the student is asked to synthesize what they learned at each site in their notepad. The final synthesis task appears here – synthesizing information from all four sites. This is reported in the text (chat) conversation. Three other synthesis score points have been scored at earlier locations.
Figure 6. Communicate. At the end, each student is asked to write a short report of the research in either an email message to the President of the School Board or in a classroom wiki. Here is a message sent by a student with the results of her research.
**Question 3**

**Multiple Choice**

You want to find another article. It is called "Energy Drinks Jolt the Heart." The article was published November 6, 2007 by WebMD. You will use a search engine to find it.

**Question 3:** Which keywords would be best to use in the search engine?

- Energy Drinks Jolt the Heart Web MD
- Energy Drinks and the Heart
- Energy Drinks Good For Your Heart November 6, 2007
- Energy Drinks and the Heart Web MD

**Question 4**

**Multiple Choice**

On the right is a set of search engine results. You want to locate the article "Energy Drinks Jolt the Heart." The article was published November 6, 2007 by WebMD.

**Question 4:** Which search engine result would you select from the list below?

- The link marked "A"
- The link marked "B"
- The link marked "C"
- The link marked "D"

---

**Figure 7. ORCA-Multiple Choice Format.** Compare the items for Question 3 and 4 in the ORCA-Multiple Choice format, below, with similar items in the ORCA-Closed format appearing in Figure 2, above. They are similar, but contain a more restricted simulation of the Internet and use a multiple-choice response format.
Figure 8. Proportion of variance contributed by L, E, S, C categories for ORCA-Multiple Choice and ORCA-Closed
Figure 9. Average item difficulties of L, E, S, C categories for ORCA-Multiple Choice and ORCA-Closed