This is the post print version of the article, which has been published in *Communications in Computer and Information Science* 2018, 810, 673-682.
The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-74334-9_69

TamPub This document has been downloaded from TamPub.uta.fi
The Institutional Repository of University of Tampere

A Performance-based Test for Assessing Students' Online Inquiry Competences in Schools

Eero Sormunen[1], Roberto González-Ibáñez[2], Carita Kiili[3], Paavo H.T. Leppänen[4], Mirjamaija Mikkilä-Erdmann[5], Norbert Erdmann[5], María Escobar-Macaya[2]

[1]University of Tampere, Faculty of Communication Sciences, Tampere, Finland
eero.sormunen@uta.fi
[2]Universidad de Santiago de Chile, Departamento de Ingeniería Informática, Santiago, Chile
{roberto.gonzalez.i, maria.escobarm}@usach.cl
[3]University of Oslo, Department of Education, Oslo, Norway
c.p.s.kiili@iped.uio.no
[4]University of Jyväskylä, Department of Psychology, Jyväskylä, Finland
paavo.ht.leppanen@jyu.fi
[5]University of Turku, Department of Teacher Education, Turku, Finland
{mirmik, nwmerd}@utu.fi

Abstract. In this paper, we introduce a performance-based test for measuring adolescents' competences in online inquiry. The test covers four competence dimensions: 1) searching and selecting relevant sources, 2) identifying the main ideas presented in the sources, 3) evaluating the credibility of the sources, and 4) synthesizing information across the sources. We implement a technological solution called NEURONE to carry out this routine. The scoring of the test data is demonstrated by presenting preliminary results of a case study. Finally, we discuss the strengths and limitations of the test.

Keywords: Online inquiry competences, information literacies, performance-based tests, schools, pupils.

1 Introduction

The breakthrough of the Web in the late 1990s gave a boost for research on information literacy (IL). The enormous and ever expanding information resources of the Internet gave a new potential to develop online research as an integrated element of learning [1]. The heterogeneity and complexity of Web information resources directed the interests of various research communities on the issues of efficient searching and critical evaluation of information [2]. Skills related to online research and learning became a genuine multidisciplinary research field.

This paper aims to contribute to IL research by introducing a performance-based online test designed to measure online inquiry competences. Performance assessments are characterized by the use of open-ended tasks and complex problems that require students to display their skills in several types of performance that usually focuses on higher-order skills [3]. Our assessment, targeted for adolescents, measures skills needed when solving a problem with online information. We will introduce the main elements of the assessment and a case study that helps to understand how students' skills are evaluated. The assessment aims at providing timely information of students' skills for improving instructions both in classrooms and in school libraries.

2 Approaches to Assess Online Inquiry Competences

Assessment approaches for IL are divided into three main categories [4]: 1) Fixed-choice tests are good in measuring students' acquisition of declarative knowledge. McCulley [5] calls fixed-choice instruments as knowledge tests. The strength of this approach is that data collection is easy to administer in large volumes and scoring can be automated. 2) Performance assessments (or performance tests) require students to apply their knowledge and skills with tasks simulating the real world tasks. 3) Rubrics are scoring tools for qualitative rating of authentic, often complex student work. Rubrics and performance tests are not exclusive categories since rubrics can also be applied in scoring performance tests.

2.1 Information Literacy Tests

Library institutions in higher education have actively developed fixed-choice tests, such as SAILS [6] applying the IL competency standards. However, there have been only a few attempts to develop instruments to measure younger students' skills. Some fixed-choice tests of IL skills have been developed for primary or secondary students [7, 8]. We did not find any examples of performance tests in the LIS literature.

Fixed-choice tests, such as SAILS, have been systematically validated and are widely used but suffer from three basic limitations: 1) They measure what the student knows rather than what she/he can do; 2) They are strongly contextualized into the library environment neglecting the Web as an information and learning environment; 3) Only limited attention is given to the information use [9].

2.2 Online Research and Comprehension

In educational sciences, research on literacies has increasingly begun to explore online research and comprehension [10, 11]. Online research and comprehension refers to the Web-based activities of locating, critically evaluating, synthesizing, and communicating information when students solve an open-ended problem with the help of online information [10]. Thus, this approach overlaps with IL but research focuses only on skills needed on the Internet environment. This line of research has designed, implemented and applied performance-based assessment for increasing our knowledge on students' online inquiry skills [12].

2.3 Evaluation of Information Interactions

Research on task-based information interaction [see 13] offers a solid theoretical ground to develop tests for IL with authentic or simulated tasks. Further, experimental studies on interactive information retrieval process provide us with models and tools to capture and analyze search logs. Systems, such as Coagmento, collect users' actions during search tasks. The system can collect actions and their outcomes, such as mouse clicks, queries, visited pages, saved pages, and highlighted text extracts [14].

3 A Novel Performance Test for Online Inquiry Competences

Our literature review indicates that the LIS community has neglected the development of assessment methods beyond fixed-choice knowledge tests. This is unfortunate since present instruments do not help us to understand how students apply their conceptual, declarative, and procedural knowledge in authentic online inquiry tasks. The gap in measuring higher-level learning outcomes obviously limits the possibilities to develop IL instruction. Novel performance-based assessment methods are needed also for evaluating the effectiveness of instruction.

Performance-based assessment designs try to build a balance between natural research design, realism, and control. Various dimensions of online inquiry competences require authentic tasks in which testees can demonstrate their mastery of competences. Tasks completed in the actual Internet would be authentic but lack control of the changes in Internet resources and other confounding factors. To control these it is necessary to build closed test environments simulating the Internet [10].

3.1 Tasks

Though simulated online inquiry tasks are complex, they can be divided into stages or phases which appear quite consistently across tasks [13, 15]. Structuring tasks into stages simplifies assessment since each competence dimension can be assessed within one stage where specific scoring methods can be applied. The test instances can be varied by changing the topic, the sources, or the answer required.

In the assessment, students perform a scenario-based task that measures the interrelated online inquiry skills of searching and selecting relevant sources, identifying important information in the selected sources, evaluating credibility of sources, and synthesizing information across multiple sources. After students have been introduced to a simulated task-scenario, students conduct the task in four phases matching the above listed competence dimensions.

3.2 Developing a Task on a Controversial Issue

To demonstrate our approach, we designed a task on a controversial issue. In the beginning of the assessments, students receive an email from a student representing another school. In the e-mail, he gives a tip to students on how to earn some money for a field trip by making a magazine that can be sold to parents and acquaintances. He also suggests that students compose an article titled "Computer-gaming has both advantages and disadvantages" for the magazine. In the article, students should also give a recommendation on how adolescents should use computer-games. Finally, he asks students to search for three Web sources and write an article on the basis of them.

In the first phase (Search and Selection), the student is required to use a search engine to locate three relevant sources. In the closed test environment (see section 3.3 for more details), the document collection contains only three relevant and 17 non-relevant sources. The content and layout of the relevant sources was designed by the research team to simulate authentic Web pages. The non-relevant pages were directly downloaded from the Web but internal links were deactivated. The non-relevant sources also dealt with computer gaming and related themes but did not discuss potential advantages and disadvantages of gaming to players. If the student fails to find all relevant sources in the first phase, the missing ones are given when he/she moves to the second phase. This ensures that all students have equal chances to succeed in the later phases independently of their success in searching.

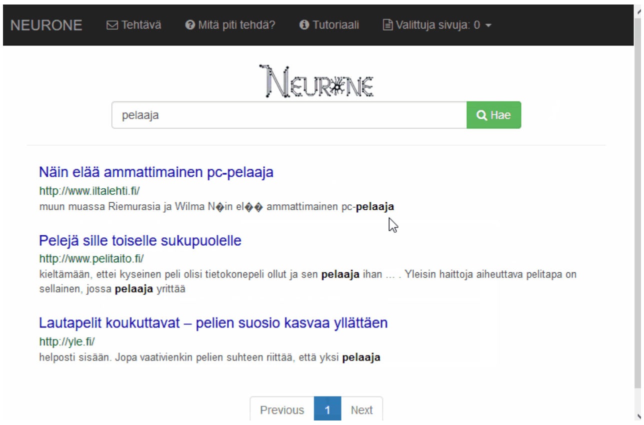
In the second phase (Identifying and Collection), the student is asked to identify two text extracts per page which include relevant information for the task. The aim is that the student reads the relevant sources and tries to make sense of their content. The relevant sources were designed to approach the topic from different perspectives: health, learning, and aggressive behaviour. The sources varied in their argumentation so that one source presents supportive reasons, one counter-reasons, and one both. The student highlights the fragments of text considered as the most important.

In the third phase (Critical Evaluation), the student is asked to evaluate the credibility of each source by rating them with stars (range 1-5) and to justify the ratings. The relevant pages included three different types of sources: a short newspaper article, a press release of a university, and a blog posting of a medical expert. This was supposed to reflect the wide variation of texts that the Internet offers.

In the final phase (Synthesis), the student is required to compose the short article to the school magazine. The fragments of texts identified in the second phase (Identifying and Collection) were available for the student, and she/he could also visit the pages. Students are expected to compose their synthesis by utilizing all sources, and give a balanced review on the advantages and disadvantages of computer gaming.

3.3 Development of the Test System

To implement the above test, we developed an online environment called NEURONE (oNlinE inqUiRy experimentatiON systEm) as part of the Finnish-Chilean iFuCo project (2016-2018). NEURONE comprises four major modules that assess the following online inquiry skills: search, identify main ideas, evaluate information critically, and synthesize information. These modules are materialized through a set of components, resources, and tools deployed in a controlled environment. The first module consists of a search engine and a navigational system that operate on a closed collection of Web pages. Within this component a user can search for information in a system that mimics the appearance of well-known search engines (Figure 1). Moreover, Web pages can be explored and bookmarked for later usage. The second module provides access to the bookmarked Web pages and a tool for collecting fragments of text called snippets from the pages. The third module allows users to revisit Web pages bookmarked so that they can evaluate them. Finally, the fourth module provides access to bookmarked pages, the main snippets collected from them, and a form to compose the text that synthesize information found from the sources.



Näin elää ammattimainen pc-pelaaja
http://www.iltalehti.fi/
muun muassa Riemurasia ja Wilma N�in el�� ammattimainen pc-**pelaaja**

Pelejä sille toiselle sukupuolelle
http://www.pelitaito.fi/
kieltämään, ettei kyseinen peli olisi tietokonepeli ollut ja sen **pelaaja** ihan ... . Yleisin haittoja aiheuttava pelitapa on sellainen, jossa **pelaaja** yrittää

Lautapelit koukuttavat – pelien suosio kasvaa yllättäen
http://yle.fi/
helposti sisään. Jopa vaativienkin pelien suhteen riittää, että yksi **pelaaja**

Figure 1: Snapshot of NEURONE's search module user interface in the Finnish language[1] displaying search results. Other snapshots and additional information of NEURONE can be found at http://www.neurone.info.

To better scaffold students in NEURONE, the modules were connected with contextualized instructions and automated tutorials before each stage. Moreover, to increase the engagement within the system, both the tasks and the tutorials were guided by fictitious characters that were represented as avatars.

NEURONE was developed through an iterative process that let us improve its usability, test structure, and content. Specifically, we conducted three pilots with elementary school students. Overall, 38 students from three schools participated in these pilots completing the test in pairs or individually. The analysis of system logs, on-site researcher observations, and users' comments revealed several software bugs and unclear instructions, tutorials, and task guidelines. All these issues were fixed.

[1] Legend: *Tehtävä* = Task; *Mitä piti tehdä?* = What is the current task?; *Tutoriaali* = Tutorial; *Valittuja sivuja n* = *n* Bookmarked pages; *Hae* = Search.

4. A Case Study

In this paper, we report the results of a case study conducted in two classes of sixth graders from one primary school in Finland. There were 36 students, 19 girls and 17 boys aged from 12 to 13 years. The goal of the case study is to demonstrate the administration of the test and scoring of students' online inquiry competences. Note that the small dataset allows us to make decriptive statistical analysis only.

4.1 Test Arrangements

The test session was arranged at school during a regular 45-minute lesson. Before the students arrived to the class, computers were started and logged into the NEURONE system. A researcher read aloud short written guidelines to students before they started the test. Students were advised to work at their own pace but be aware that the system will alarm when the maximum time in each phase is close to the end. Students were encouraged to ask for help if they faced problems with the system. Spare laptop computers were available to avoid potential problems with the school's equipment.

4.2 Scoring

In scoring students' online inquiry competencies, various approaches need to be applied because the assessed data vary from dimension to dimension. In two first dimensions, we applied automatic procedures in the search log data. In two latter dimensions, scoring was based on a rubric.

Searching and Selecting Relevant Sources. In order to demonstrate the utility of our test approach, we score students' performance at this stage through two specific measures. Given the nature of the task (search three relevant Web pages) a recall-based measure is a natural choice. We applied an adapted version of recall as defined in [16]. Recall(s) corresponds to the ratio between the total number of relevant pages bookmarked by student *s* (relevant coverage) and the total number of relevant documents (here three) in the collection (Eq. 1). However, recall alone turned out to be a vulnerable measure because some students learned to select relevant pages by randomly bookmarking pages and letting NEURONE inform when they got a hit. SearchScore(s) is a process-based measure that takes into account students' selection behaviors. It corresponds to the ratio of the total number of relevant documents bookmarked and the total number of pages bookmarked (Eq. 2). The denominator penalizes for the high number of bookmarked pages. We weigh the ratio by 5 to scale the score between the range 0 – 5.

$$\text{Recall(s)} = \text{RelevantCoverage (s)} / \text{Total no of rel pages} \tag{1}$$

$$\text{SearchScore(s)} = 5 * \text{RelevantCoverage(s)} / \text{ActiveBookmarks(s)} \tag{2}$$

Identifying the Main Ideas from a Source. To evaluate whether students were able to identify two text extracts that contained main ideas from each Web page, an automatic procedure was launched to compare students' snippets and the actual main ideas. The list of main ideas from each document was identified and compared to the text fragments selected by the student. If students' selections match (partial or exact) with the main ideas of the page, he/she receives one point. Additional matches of other snippets with the same idea do not give points. As a result, a student can receive two points from each page and altogether up to six points in this stage.

Justifications for Credibility Evaluations. Evaluation of credibility was scored in terms of the justifications presented by students for their credibility evaluations for each three sources (see e.g., [17]). Students scored 2 points by presenting at least two relevant justifications, 1 point for one relevant justification, and 0 points if students were not able to present any relevant justifications. The maximum score was 6 points.

Synthesizing Information Across Sources. This dimension was scored in terms of quality of source-based argumentation. The scoring rubric took into account the following four aspects in students' articles: 1) how well argumentation relied on multiple sources; 2) whether students' source-based argumentation considered both sides (advantages and disadvantages), 3) quality of source-based arguments, i.e., whether arguments were complete (playing violent computer games increases aggressive behaviour) or incomplete (playing computer games increases aggressive behaviour), and 4) richness of argumentation. The maximum score was 4 points.

4.3 Results

Search and Selection. Results show that 11 students (31%) found all relevant pages (Recall = 1), 16 students (44%) found two (recall = 0.67), and 9 students (25%) found only one of the relevant pages (recall = 0.33) within the given time limit of 8 minutes. Recall does not take into account if the student had bookmarked pages randomly. To address this limitation, we applied our second measure Search Score. The average search score was 2.46 (SD = 1.1). Only three students (8,3%) had bookmarked all three relevant pages straight without miss selections (Search Score=5).

In the group of eleven students, who demonstrated maximum recall (Recall = 1), four students bookmarked 8 or 9 pages. This indicates that they had not really tried or been able to assess the relevance of pages they had retrieved and opened. Their search scores fell below the average (<2.46). In the group of 9 students, who succeed to find only one relevant page, two measures give equal scores (recall = 0,33; Search Score = 1.67) since none of the students practiced with random bookmarkings.

Identifying and Collecting Important Ideas. The students scored on average, 4.36 (SD = 0.96) out of six points in the main idea identification task. All students were able to identify at least two extracts that included main ideas, however only 6% of them (3 students) were able to score the maximum 6 points. The majority of the students identified 4 or 5 relevant extracts. One interesting aspect to note was that the student who achieved the highest score in this stage (6 points) was among the low achievers in the previous phase according to our search score with only 1.66 points. Students with the highest search score obtained 4 or 5 points in the collection stage.

Justification of Critical Evaluations. The results show students' limited abilities to evaluate the credibility of online sources. The total mean score for credibility evaluation was 2.00 (SD = 1.79) out of six points and the mean scores for single sources varied from 0.53 to 0.75 (max. 2 points). One fourth of the students were unable to present any relevant justification to any of the three evaluated sources. Furthermore, only few students were able to present multiple justifications for their evaluations. The proportion of students that presented multiple justifications for their evaluation at least in one of the evaluated sources was 27.8%. Only one student presented multiple justifications in all of his source evaluations. Almost all justifications (n = 75) were confirming the credibility, and only 7% of justifications indicated some skepticism.

Synthesis. Although students performed quite well in the main idea identification task, they had difficulties in source-based argumentation. The mean score on quality of source-based argumentation was 2.06 (SD = 1.07) out of four points. From all students 11% were unable to include any complete source-based arguments into their article. Additionally, 11% of students showed either one-sided argumentation or they relied only on one source in their argumentation. Students may also have faced difficulties in deep level of processing of texts that was shown in quite big proportion of incomplete arguments. Altogether 22% of arguments were incomplete lowering the quality of students' argumentation. From all students, 8% reached four points from their text showing rich and precise argumentation throughout their composed article.

5. Discussion and Conclusions

We argue that the assessment of online inquiry competences is a multidisciplinary research issue, and tackling the challenges of performance-based testing calls for interdisciplinary research efforts. We responded to the call by introducing a test approach that provides new features informed by previous research both in LIS and educational sciences. The notion of inquiry task is bridging researchers in task-based information interaction [13] and online research and comprehension [10]. They provide a solid theoretical ground to the design of test assignments.

Our specific contribution was to apply search log analysis methods developed in interactive information retrieval research [18] to score students' searching performance. To the best of our knowledge, the automatic analysis of search logs has not been used earlier in similar tests. The closed test environment (controlled task and document collection) offers an enormous potential for the search log analysis beyond what can be done in the open Web environment. The analysis may focus both on output variables (e.g. recall and precision) and process variables (e.g. search behaviours such as click-through behavior). For a longer list of measures, see [16, 18].

Expanding the analysis of searching beyond the overall performance scores is important from the view point of developing pedagogical practices. Performance measures tell us the level of searching skills in the tested group of students. However, if we want to develop teaching of online inquiry competences, and especially searching skills, it is more important to know what patterns of searching behaviour make the difference between high and low performers. This is a way to identify unproductive searching practices in particular contexts, design teaching interventions to tackle them, and tools for scaffolding students' learning, see e.g. [19].

The NEURONE-Assessment was able to capture a comprehensive set of complex skills with the task that was reasonable in length for younger learners. In the design, we varied different task types. For example, when students were able to select important ideas with the snippet tool instead of the note-taking tool, they were probably not overwhelmed before the last, the most demanding open-ended task. Varying the task types probably also increases students' engagement in the task.

Our case study demonstrated how the novel test approach, as implemented in NEURONE, can be applied at the level of primary education and how the scoring of students' performance in four dimensions of online inquiry competences can be operationalized. We demonstrated what kinds of scoring can be applied in the obtained dataset. We also demonstrated how a rich search log (and overall system log) can be used in quality and validity control.

The study has several limitations. The paper introduces new ideas based on previous multidisciplinary research and one implementation of those ideas. We demonstrated one set of scoring methods without validating them at this point. In the limited dataset, we could demonstrate how the test works in practice. However, the case study does not allow us to generalize empirical results to a wider population.

In conclusion, we have contributed to the development of performance assessments in online inquiry competences. Based on an interdisciplinary approach, we designed, planned, and implemented a novel test for online inquiry competences. Moreover, we demonstrated that our approach can be applied in the field and that it opens promising avenues to develop performance tests which exploit search logs at full capacity and evaluates performance across the entire online inquiry task.

Our work on the performance test continues. About 340 6th-graders performed the test in ten Finnish schools during the Spring term 2017. In addition, hundreds of Chilean students will take the test during the second half of 2017.

Acknowledgements. The study was part of the iFuCo project funded by the Academy of Finland (grant's no. 294186) and the National Commission for Scientific and Technological Research (CONICYT) in Chile (grant's AKA/EDU-03). The authors thank the schools participating in the study. We are especially grateful to Daniel Gacitúa Vásquez who made an enormous job in building the NEURONE system. We also thank Andras Varga for the layout desings, and Elina Hämäläinen, Nina Kullberg, Mirka Junikka, Teemu Mikkonen and Tuulikki Alamettälä for organizing the tests.

List of References

1. Kingsley, T., Tancock, S.: Internet Inquiry: Fundamental Competencies for Online Comprehension. Read. Teach. 67, 389–399 (2014).
2. Stordy, P.H.: Taxonomy of Literacies Introduction. J. Doc. 71, (2015).
3. Baker, E.L., O'Neil, H.F., Linn, R.L.: Policy and validity prospects for performance-based assessment. Am. Psychol. 48, 1210–1218 (1993).
4. Oakleaf, M.: Using Rubrics to Assess Information Literacy: an Examination of Methodology and Interrater Reliability. J. Am. Soc. Inf. Sci. Technol. 60, 969–983 (2009).
5. McCulley, C.: Mixing and Matching: Assessing Information Literacy Assessing Information Literacy. Commun. Inf. Lit. 3, 171–180 (2009).
6. Lym, B., Grossman, H., Yannotta, L., Talih, M.: Assessing the Assessment: How Institutions Administered, Interpreted, and Used SAILS. Ref. Serv. Rev. 38, 168–186 (2010).
7. Majid, S., Chang, Y.-K., Foo, S.: Auditing Information Literacy Skills of Secondary School Students in Singapore. J. Inf. Lit. 10, 44–66 (2016).
8. Foo, S., Majid, S., Chang, Y.K.: Assessing Information Literacy Skills Among Young Information Age Students in Singapore. Aslib J. Inf. Manag. 69, [pre-print] (2017).
9. Sparks, J.R., Katz, I.R., Beile, P.M.: Assessing Digital Information Literacy in Higher Education: A Review of Existing Frameworks and Assessments With Recommendations for Next-Generation Assessment. (2016).
10. Leu, D.J., Forzani, E., Rhoads, C., Maykel, C., Kennedy, C., Timbrell, N.: The New Literacies of Online Research and Comprehension: Rethinking the Reading Achievement Gap. Read. Res. Q. 50, 37–59 (2015).
11. Cho, B.-Y., Woodward, L., Li, D., Barlow, W.: Examining Adolescents' Strategic Processing During Online Reading With a Question- Generating Task. (2017).
12. Kennedy, C., Rhoads, C., Leu, D.J.: Online Research and Learning in Science: A One-to-one Laptop Comparison in Two States Using Performance Based Assessments. Comput. Educ. Published, (2016).
13. Järvelin, K., Vakkari, P., Arvola, P., Baskaya, F., Järvelin, A., Kekäläinen, J., Keskustalo, H., Kumpulainen, S., Saastamoinen, M., Savolainen, R., Sormunen, E.: Task-Based Information Interaction Evaluation: The Viewpoint of Program Theory. ACM Trans. Inf. Syst. 33, (2015).
14. González-Ibáñez, R., Shah, C.: Coagmento: A System for Supporting Collaborative Information Seeking. Proc. Am. Soc. Inf. Sci. Technol. 48, 1–4 (2011).
15. Kuhlthau, C.C.: Seeking Meaning : a Process Approach to Library and Information Services. Libraries Unlimited, Westport (2004).
16. Shah, C., González-Ibáñez, R.: Evaluating the Synergic Effect of Collaboration in Information Seeking. In: Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '11. pp. 913–922. ACM, New York, NY (2011).
17. Kiili, C., Laurinen, L., Marttunen, M.: Students Evaluating Internet Sources: From Versatile Evaluators to Uncritical Readers. J. Educ. Comput. Res. 39, 75–95 (2008).
18. Shah, C., Hendahewa, C., González-Ibáñez, R.: Rain or Shine? Forecasting Search Process Performance in Exploratory Search Tasks. J. Assoc. Inf. Sci. Technol. [pre-print], (2016).
19. Kiili, C., Coiro, J.L., Hämäläinen, J.: An Online Inquiry Tool to Support the Exploration of Controversial Issues on the Internet. J. Lit. Technol. 17, 31–52 (2016).