This study analyzes the search behavior of Dutch-speaking nursing students with a nonnative knowledge of English who searched for information in MEDLINE/PubMed about a specific theme in nursing. We examine whether and to what extent their search efficiency is affected by their language skills. Our task-oriented approach focuses on three stages of the information retrieval process: need articulation, query formulation, and relevance judgment. The test participants completed a pretest questionnaire, which gave us information about their overall experience with the search system and their self-reported computer and language skills. The students were briefly introduced to the use of PubMed and MeSH (medical subject headings) before they conducted their keyword-driven subject search. We assessed the search results in terms of recall and precision, and also analyzed the search process. After the search task, a satisfaction survey and a language test were completed. We conclude that language skills have an impact on the search results. We hypothesize that language support might improve the efficiency of searches conducted by Dutch-speaking users of PubMed.

Introduction

The growing amount of information makes it paradoxically difficult to stay abreast of current developments in the biomedical domain and to search for information selectively, even with the help of biomedical bibliographic indexes such as MEDLINE and Embase. Many studies have been devoted to the information retrieval (IR) process (Spink, Wolfram, Jansen, & Saracevic, 2001; Sutcliffe, Ennis, & Watkinson, 2000), precision and recall, and ways to make this process more efficient (Bin & Lun, 2001; Muin, Fontelo, Liu, & Ackerman, 2005; Wilson, 1999).

As English has become the lingua franca of science, the “new Latin” (Eisenberg, 1996), it creates continuity in the domain, but may also cause problems in the retrieval of information. Scholars whose mother tongue is not English may experience difficulties when conducting a literature search. General language skills are needed for efficient information retrieval (Lankamp, 1989), as well as domain-specific terminology. In addition, searchers have to be familiar with the language of information and documentation science (Mouillet, 1999) to use the interface of the search system effectively. Most studies focusing on query formulation and on the search process in general have been conducted with native English test groups. The present study,
however, focuses on difficulties caused by the language barrier for Dutch-speaking users of PubMed (http://www.ncbi.nlm.nih.gov/pubmed), a tool designed to search the MEDLINE database and other medical resources through the Internet.

The aim of this study is to describe the efficiency of a PubMed search by Dutch-speaking nursing students (bachelor’s and master’s level), and to explore the impact of Dutch–English translation problems as well as other characteristics (educational background, computer skills, bibliographic skills) on search efficiency. We focus on performance problems in the formulation of efficient queries and on the selection of relevant citations.

**Method**

**Theoretical Framework**


In the problem identification stage, the user is confronted with an uncertainty or problem about which he or she wants to look up information. Need articulation involves parsing of the problem, which is formulated in natural language, into several knowledge structures (Sutcliffe & Ennis, 1998), i.e., into concepts. Dutch-speaking PubMed users with advanced English-language skills will probably do this parsing in English.

The query formulation stage is a crucial step in the IR process, as different types of translation actions take place. For native English users, this step includes a transformation of the concepts that resulted from the need articulation stage into search terms, selecting the correct MeSH terms and combining them with Boolean operators, taking into account the specific query syntax of the search system. In our test case, the language barrier also has to be taken into consideration (see Figure 1): The search question is translated into concepts, which are then translated into English search terms. Based on the search terms, PubMed makes one or more suggestions for MeSH terms, from which the user chooses the most appropriate one(s).

Results evaluation or relevance judgment, i.e., comparing the set of retrieved articles to the initial information need and selecting relevant citations, also involves some translation actions, as the searcher needs to read the retrieved information and base relevance judgments on titles and/or abstracts in a foreign language. A first relevance judgment step takes place when the user skims the results and determines whether the set of articles matches his or her information need. If there are some interesting results, the user will start browsing the citations. If not, a new query will be issued. A second, more thorough relevance judgment takes place when the user runs through the individual citations and decides for each of them whether it is relevant or not. If the searcher is not satisfied with the number of citations that result from this search, he or she will formulate a new query.

**Experimental Design**

We selected a group consisting of about 60 nursing students pursuing their bachelor’s and master’s degrees. The bachelor’s level students were in the third year of their training and the master’s level students in their fifth year. They had to complete a test which consisted of five parts. First, they completed a pretest questionnaire that focused on computer skills, facility, and experiences with
the search system PubMed, and self-assessment of English language skills.

Second, an introduction (10 minutes) was given on the use of MeSH (medical subject headings; http://www.ncbi.nlm.nih.gov/mesh) in PubMed. MeSH is a controlled vocabulary created by the National Library of Medicine for the purpose of indexing journal articles and books in the biomedical sciences. It helps PubMed users to optimize their literature searches. In this introduction, the advantages and usefulness of MeSH were emphasized, and indexed searching was advocated.

Third, the students conducted a literature search for a specific theme in nursing. This bibliographic task was based on a preformulated question in Dutch (translated: “What is the effect of a multifactorial treatment (i.e., a combination of physiotherapy/exercise/medication, etc.) on the risk of falling in elderly living in long-term care facilities, such as nursing homes or homes for the aged?”). We assume that this question was clear to the participants, as it was formulated in their mother tongue. Moreover, we paraphrased the question and explained to the participants what they had to look for orally, and they were free to ask questions at any time during the test. In the posttest questionnaire (see below), we asked whether the search question was formulated in a clear and understandable way.

The participants were advised to use MeSH terms instead of free text and to combine several relevant MeSH terms with Boolean operators to construct a well-formulated query. They had 15 minutes to complete the literature search, which was subsequently assessed in several ways (see Evaluation Methods section).

Fourth, a posttest questionnaire was completed to see how the students experienced the test.

Fifth, the participants completed the vocabulary and reading parts of the DIALANG (http://www.dialang.org) diagnostic language test for English. This test has been internationally validated and was developed by more than 20 major European institutions with the support of the European Commission. It is based on the Common European Framework of Reference for Languages (CEFR; http://www.coe.int/T/DG4/Portfolio/?L=EN&M=/main_pages/levels.html) and is available in 14 European languages, including English. The DIALANG language test allowed us to assess the participants’ English reading and vocabulary skills on a 6-band scale (see Table 1) and to link the results to their performance on the literature search task.

Test Groups

We recruited 31 undergraduate students in the Nursing Department of University College Ghent and 40 master’s level students at the Nursing and Midwifery Department of the University of Antwerp. Both institutions are located in Flanders, the Dutch-speaking part of Belgium. The same test was conducted in both institutions in several sessions from November 2008 to December 2009.

### TABLE 1. Results of the DIALANG test.

<table>
<thead>
<tr>
<th>Reading level</th>
<th>Corresponding skills</th>
<th>Participants (n = 71), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type.</td>
<td>2.8</td>
</tr>
<tr>
<td>A2</td>
<td>Can understand sentences and frequently used expressions related to areas of most immediate relevance.</td>
<td>11.3</td>
</tr>
<tr>
<td>B1</td>
<td>Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.</td>
<td>22.5</td>
</tr>
<tr>
<td>B2</td>
<td>Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization.</td>
<td>45.1</td>
</tr>
<tr>
<td>C1</td>
<td>Can understand a wide range of demanding, longer texts, and recognize implicit meaning.</td>
<td>12.7</td>
</tr>
<tr>
<td>C2</td>
<td>Can understand with ease virtually everything heard or read.</td>
<td>5.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vocabulary level</th>
<th>Corresponding skills</th>
<th>Participants (n = 71), %</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Can introduce him/herself and others and can ask and answer questions about personal details.</td>
<td>0.0</td>
</tr>
<tr>
<td>A2</td>
<td>Can describe in simple terms aspects of his/her background, immediate environment, and matters in areas of immediate need.</td>
<td>7.0</td>
</tr>
<tr>
<td>B1</td>
<td>Can produce simple connected text on topics which are familiar or of personal interest.</td>
<td>9.9</td>
</tr>
<tr>
<td>B2</td>
<td>Can produce clear, detailed text on a wide range of subjects.</td>
<td>62.0</td>
</tr>
<tr>
<td>C1</td>
<td>Can use language flexibly and effectively for social, academic, and professional purposes.</td>
<td>18.3</td>
</tr>
<tr>
<td>C2</td>
<td>Can express him/herself very fluently and precisely, differentiating finer shades of meaning even in more complex situations.</td>
<td>2.8</td>
</tr>
</tbody>
</table>

In the first year of their training, all respondents had taken a compulsory course in which they were briefly initiated into the research domain and learned to search for and understand specialist literature. Additionally, the master’s students had attended a program on scientific research in their master’s degree training, which includes methodological principles of literature searching, among others in PubMed, and systematic review and analysis of literature. As the master’s level students are more experienced searchers, they will be referred to as more experienced compared to the less experienced undergraduate students.

**Development of the Gold Standard**

The gold standard used for the evaluation of the search results was synthesized from the results of three types of searches: the students’ searches, an expert search, and a related-citations search. To qualify for the gold standard, citations had to contain the four main elements of the search question, i.e., falls, elderly, long-term care, and
multifactorial prevention. If one of the components was not present, the citation was not incorporated into the gold standard. The selection of these citations was done by a linguist in consultation with an expert (a medical doctor with professional expertise in bibliographic retrieval and instruction, and with domain expertise about geriatric pharmacology). In accordance with the “union of outputs” principle (Miller, 1971), we filtered the relevant citations from the students’ selections. This resulted in a set of 51 relevant citations. In addition, the search task was executed by the expert, who formulated a gold standard query. This query covered all four concepts of the information need (except for the multifactorial aspect), and it consisted of six terms (“Accidental Falls/prevention and control”[Mesh] AND (“homes for the aged”[Mesh] OR “nursing homes”[Mesh]) AND (“aged”[Mesh] OR “Geriatrics”[Mesh])). The extra relevant articles yielded by this query—11 citations—were added to the students’ selections.

The total set of relevant articles found by our test subjects and by the expert was expanded with citations retrieved with the “related citations” function in PubMed, as Lin and Smucker (2008) showed that tools based on content similarity can increase recall considerably. In our case, only four extra citations were found with this function.

This three-step procedure resulted in a gold standard of 66 articles in total. However, as the test was conducted in several sessions over a time span of 13 months, we had to take the publication date of the articles in our gold standard into consideration. The gold standard was comprised of 62, 64, 65, and 66 records for the test groups of November 2008, February 2009, April 2009, and December 2009, respectively.

The gold standard query had a recall of 71.2% and was used to calculate concept coverage. The precision of the gold standard query was 17.4% (47 citations out of 270 were relevant).

Evaluation

Evaluation of the search process. We used the Morae software (http://www.techsmith.com/morae.asp), a program specifically designed to record and analyze user–computer interaction for the evaluation of the search process. It registers all onscreen actions performed on the computer. In this way, it allows researchers to analyze all operations executed by a user and to log tasks, markers, and marker scores. Tasks take up a period of time, whereas markers are used for events.

We defined several tasks, including “Reading the search question,” “Searching”—a task that usually consists of several individual PubMed searches—and “Final relevance judgment.” One PubMed search includes a querying and a relevance judgment stage. The querying stage is characterized by an alternation of search term formulation and MeSH term selection.

We also logged “hesitations and errors” as a task. It may be questionable to classify hesitations and errors as a task, but this was the only way to mark events that occurred over a period in time. Only those hesitations that were clearly caused by a lack of experience with the search system were logged, i.e., when it was obvious that the participant did not know what to do next, or when he or she made errors (e.g., going back to the PowerPoint presentation about the use of PubMed, or searching for MeSH terms in PubMed instead of in the MeSH section).

Based on these Morae tasks, we evaluated the need articulation, query formulation, and relevance judgment stages (Figure 2). The problem identification stage was not addressed in this study, as the respondents started from an imposed search question. The need articulation stage as such is an implicit process. However, the result of this need articulation is reflected in the search terms used and in the number of concepts covered by queries. Need articulation
was therefore studied in terms of concept coverage, in which
we examined how many of the four main concepts (elderly,
falls, long-term care, and prevention) were used in the
queries. Concept coverage is an indication of how well
the participants analyzed the search question and translated it
into concepts. In this test, a good query was a query that
contained the four main components of the search question,
i.e., falls, elderly, long-term care, and prevention. These
concepts or components can be expressed by several MeSH
terms. Concept coverage is the proportion of those (four)
concepts that were represented in the queries. The query
“("Aged"[Mesh] OR “Frail Elderly"[Mesh]) AND “Acci-
dental Falls"[Mesh]” for instance, has a coverage of 50%
two out of four concepts are covered: elderly and falls.
The time spent on reading the search question is also considered
as an indication of the time spent on need articulation.

The query formulation stage was assessed in terms of the
quality of search and MeSH terms, query identification and
coverage, query complexity and the use of Boolean opera-
tors, hesitations and errors, and zero-result queries.

The final stage of the IR process, relevance judgment,
took place each time a PubMed search was executed. Rele-
ance judgment is therefore seen as a part of the search
task, following query formulation. The time spent on assess-
ing the citations retrieved is considered as an indication of
how thoroughly the relevance judgment process is executed.
The effectiveness of this stage can be measured by precision
(see Search Results section).

Next, we defined 26 different markers for different events
in the search process, the most important of which were
“Search term formulation,” “MeSH term selection,” “Query
submission,” and “Citation selection” (see Figure 3).

Scores were also assigned to the search term formulation
and MeSH term selection events: Each search term
formulated and each MeSH term selected by the partici-
pants was assigned 0 (bad), 1 (medium), or 2 (good).
These scores were the result of consultation between a lin-
guist and our expert in bibliographic instruction. They
were used to assess the quality of the search terms and
MeSH terms (see Query formulation stage subsection in
the Search Process Characteristics section). Bad search
terms included incorrect translations, such as kine,
kinesiotherapy, and kinesics (instead of physiother-
apy; translation of the Dutch word kinesitherapie), movingexercises, or
residetion nursinghome. Also considered as bad search
terms were terms that were not relevant for this informa-
tion search or too general to achieve relevant results (e.g.,
resident or housesettings).

Medium search terms included typographical errors
(e.g., physiotherapy progroms or resiential care). Spelling
is a great source of errors too, even in native English users
of PubMed (Wilbur, Kim, & Xie, 2006). Examples of such
orthographical errors from our data are fsiotherapy or
culfactoriel intervention.

Spelling and language skills in general are not an issue
in the translation into MeSH terms, as the searcher has to
select them from a list of suggestions. Bad MeSH terms
are terms that are not relevant to the search question;
Examples include kinesics and residential treatment.
Medium MeSH terms are terms that can be used in the
context of the search question, but are not specific enough
(e.g., risk factors, hospitals). A list of acceptable MeSH
terms was created by a linguist in consultation with an
expert (the same expert who constructed the gold standard
query; see Development of the Gold Standard section).

Evaluation of the search results. We calculated the effi-
ciency of the search in terms of recall and precision. Cita-
tions that were considered relevant were sent to the
clipboard. The result was a list of citations the students
deemed relevant to the search question, drawn from the
whole search task, which usually consisted of several sepa-
rate searches. These citations had to contain the four main
components of the search question, i.e., elderly, long-term
care facility, falls, and (multifactorial) prevention. All four
components had to be present for the citation to be classified
as relevant. For each participant, the resulting list of cita-
tions was compared to the gold standard, and precision and
recall were deduced (see Figure 4). It may be noted that we
did not intend to measure the performance of the search
engine, but the participants’ ability to find and select relevant
citations in PubMed.

The literature search task came down to a binary classi-
fication task, in which the test participants had to select
relevant articles and discard the irrelevant ones from the list
of citations their query yielded. Precision in our test case
therefore referred to the precision of the selection \(P_r\) of
citations made by the test participants. Citations selected by
the participants that were also in the gold standard were
true-positives \(tp\); false-positives \(fp\) were citations that
were wrongly considered to be relevant. \(P_r\) can be defined as
the proportion of true-positives in the students’ selection:
Analogously, recall in our test case referred to the recall of the final selection ($R_s$) of citations. It represented the proportion of citations in the gold standard that was also retrieved and selected ($tp$) by the test participants. Recall of the students’ selection was defined as follows:

$$R_s = \frac{tp}{\text{GoldStandard}}$$

We used NLM’s E-Utilities (http://www.ncbi.nlm.nih.gov/books/NBK25500/) to simulate the students’ searches to obtain their resulting lists of citations. Taking into account the number of results that were viewed by each participant for each query, we calculated the number of missed citations, i.e., the number of gold standard citations that were returned by a query, but were not selected as being relevant by the participants. This way, we could determine whether false-negatives were the result of a bad query or of bad relevance judgment. The number of false-negatives also allowed us to calculate the potential recall score ($R_{pot}$), i.e., the recall score the participants would have obtained if they had not overlooked any relevant citations:

$$R_{pot} = \frac{tp + fn}{\text{GoldStandard}}$$

The trade-off between recall and precision has been described by many researchers (Alvarez, 2002; Buckland & Gey, 1994; Eysenbach, Tuische, & Diepgen, 2001); it forces users to choose which performance measure to optimize. However, as this task did not focus on either one or the other of the two measures explicitly, we assumed that the participants wanted to keep a balance between precision and recall.

Pre- and posttest questionnaires. The students completed a pretest questionnaire that focused on self-perceived English-language and computer skills, and on facility with PubMed.

The posttest questionnaire was designed to measure the students’ self-perceived test performance. The answers to these questions will be linked to their actual performance on the test to see whether the participants had a realistic view of the quality of their search.

The self-reported skills, attitudes, and opinions were assessed using 7-point Likert scale questions.

Statistical Issues

We analyzed our data with the SPSS PASW 18 package. The Shapiro–Wilk test was used to assess the distribution of the variables. Depending on the types of variables studied, we used the Spearman correlation test, or the Mann–Whitney U (distribution-free) test. The minimum significance level used for these tests was 0.05.

For ranked values, we report the median and interquartile ranges as follows: $Mdn (Q1, Q3; IQR)$—median, first and third quartile, and interquartile range, respectively.

Precision of the user’s selection is a relative notion: A respondent who selected only two citations, one of which was relevant, achieved a precision of 50%. As this may misrepresent the efficiency of the searches, we chose to use Spearman’s rank correlation to assess relationships between precision and recall, and other variables in the test.

Ethical Issues

We asked the Nursing Departments for formal permission to conduct the test. Students were invited to participate in the test by means of an invitation letter, in which we explained the aim and methods of the test. They were also informed that they could leave the classroom at any time if they no longer wanted to participate.

Results

Respondent Characteristics

Seventy-one respondents participated in our test: 31 bachelor’s and 40 master’s level nursing students. The description of the respondent characteristics below is based on the pre- and posttest questionnaires, and on the results of the DIALANG language test.

Language skills. We assume that at least a B2 level is needed to perform this task successfully, as people with this level of language skills can read and produce more technical texts: 63.4% achieved a B2 level or higher for reading, and 83.1% of the participants reached a B2 level or higher for vocabulary.

Self-reported skills. We asked the students to rate their English-language skills and their computer skills on a scale from 1 (very poor) to 7 (excellent). Language skills were
assigned quite a high score, with a median (Mdn) of 5 (4, 5; 1 IQR). With a median of median (Mdn) of 3 (3, 4; 1 IQR), computer skills were assigned lower scores. Although there are very useful biomedical databases, 24% of the students in our test group preferred using Google to look for medical information. More than half of the students indicated that they are used to searching for medical resources in English, as these are also written predominantly in English. However, there is a clear preference for Dutch over English (72%) to read scientific texts.

We asked the participants whether the search question was clearly formulated and understandable. Only one student indicated that the search question was not entirely clear.

**Self-reported test performance.** When asked to assess their performance on the search task, 28% answered that they had made a good selection of citations. Sixty-three percent had difficulties finding the right keywords for their searches, and 62% of students were uncertain about the spelling of the search terms they used. After the literature search, most of the students (73%) were enthusiastic about PubMed and indicated that they would like to learn more about the search system.

**Search Process Characteristics**

**Query formulation stage.** **Quality of search terms and MeSH terms.** On average, half of the search terms entered were good search terms, and 21% of the search terms were scored as bad because they either contained language errors or because they were irrelevant (see Figure 5, chart A). The remaining 29% were medium search terms.

The translation of a search term into a MeSH term is usually an elimination process: One or more suggestions are provided by the search system, and the user selects the most suitable MeSH term for his information need. Consequently, this translation process is less error-prone than the formulation of free-text search terms (see Figure 5, chart B). This mainly results in a larger proportion of good MeSH terms (73%) and a smaller proportion of medium MeSH terms (8%).

About 50% of the search terms were linguistically incorrect or irrelevant and were therefore assessed as bad or medium, depending on the severity of the error. However, as this is only an intermediate step toward finding MeSH terms, many of those incorrect search terms are filtered by the search system. This corrective effect of subject searching with MeSH resulted in an error rate reduction of 25%. This means that the percentage of medium and bad search terms was reduced by half due to the use of MeSH terms. It should be noted, however, that MeSH terms which were not retrieved were not taken into account here.

**Concept identification and coverage.** We assume that the participants understood the search question. Only one student—who achieved a relatively high precision and recall score—indicated in the posttest questionnaire that he or she did not completely understand the search question.

As stated above, a good query has to contain MeSH terms for the four main components of the search, i.e., elderly, long-term care, falls, and (multifactorial) prevention. As there is no MeSH term for the concept “multifactorial,” it could not be translated into a MeSH term. Table 2 shows that the coverage of the concepts “falls” and “elderly” is quite high, and that about half of the participants found a MeSH term for “long-term care.” The word “prevention” was not explicitly in the search question, causing many of the participants to overlook this concept.

To calculate concept coverage, i.e., the number of concepts that were covered by one or more MeSH terms in the participants’ queries, we analyzed the search terms to see which concepts were identified as important (see “Concept Identification” in Table 2). The search term *residention nursinghome*, for instance, which was scored as “bad,”

![Figure 5. Average proportions of good, medium, and bad search and MeSH terms.](image-url)
shows us that the participant did identify “long-term care” as an important component of the search. As no MeSH term suggestions were made for this search term—and the participant failed to formulate a correct search term—the concept was not covered in the participant’s searches. Hence, the absence of a concept does not necessarily indicate that the participant did not identify this concept as important in the search question.

We found three different reasons for noncoverage of concepts. First, sometimes a concept was not identified as important to the search question, and no search terms were formulated for this concept. Consequently, it was not represented in the query. Second, even if a concept was identified as important, the use of an incorrect search term sometimes prevented the participants from finding the correct MeSH term. In other cases, a good search term was formulated, but the participant failed to identify the correct MeSH term.

Figure 6 shows that 56% percent of noncovered concepts were absent in the queries because the participants did not identify these concepts in the search question, and therefore did not search for them. For 16% of the noncovered concepts, the participants did identify the concept, but used a bad search term and consequently did not find an appropriate MeSH term. This category of errors is caused by the lack of active English-language skills. In 28% of the cases, a good search term was formulated, but the participant failed to identify the correct MeSH term.

We can conclude from this data that noncoverage of concepts is caused, in the first place, by the nonidentification of concepts in the search question, and that the number of bad search terms that lead to noncoverage is limited. Selecting the correct MeSH term seems to be a problem, even when a correct search term was entered. This may be due to the lack of experience with the search system, or to the lack of language skills.

Query complexity and the use of Boolean operators. The average query in our test consisted of 3.36 terms. All test participants constructed queries by combining MeSH—or sometimes free-text search—terms with the Boolean operator AND. About 35% of the students used the OR-operator and none of them used the NOT-operator.

The excessive use of the Boolean operator AND (e.g., “Pharmaceutical Preparations”[Mesh] AND “Aged-”[Mesh] AND “Risk Factors”[Mesh] AND “Accidental Falls”[Mesh] AND “Nursing Homes”[Mesh]) often led to zero results, and it was also found to be one of the causes of “unproductive searches” by Walker, McKibbon, Haynes, and Ramsden (1991) and Kingsland, Harbourt, Syed, and Schuyler (1993).

Zero-result queries. A total of 17% of all queries yielded zero results. This is due to either overspecification and the excessive use of AND, or to the incorrect use of MeSH-terms.

Hesitations and errors. We assigned the label “hesitations and errors” when erroneous steps were taken (e.g., searching for a MeSH term in PubMed instead of in the MeSH section), or when the participant clearly hesitated about the next step. Moments of inactivity before formulating a search term were not considered as hesitations. The average total length of hesitations and errors was 2 minutes 4 seconds. The time spent on hesitations and errors can be seen as an indication of search proficiency (see “Associations between respondent and search process characteristics” below).

Relevance judgment stage. Time spent on relevance judgment. During the manual analysis of the screen recordings, we noticed that many participants selected citations too quickly. A combination of the words “elderly” and “falls” in the title was often enough to make them select the citation as relevant. Therefore, we consider the time spent on relevance judgment per search as an indication of how thoroughly this step was executed. The average total time spent on evaluation, i.e., on relevance judgment, during the whole search task was 5 minutes 11 seconds.

Selection of citations. About 1 in 10 participants did not select any citations during the literature search task. On average, the participants selected 6.8 articles with a maximum (max) of 31 and a median (Mdn) of 5 (2, 9; 7 IQR).
TABLE 3. Associations among and between respondent and search process characteristics (n = 71).

<table>
<thead>
<tr>
<th>Associations among respondent characteristics.</th>
<th>Spearman’s rho</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Vocabulary test – self-assessment English-language skills</td>
<td>$r_s = .346$</td>
<td>$p = .003$</td>
</tr>
<tr>
<td>2. Reading test – self-assessment English-language skills</td>
<td>$r_s = .400$</td>
<td>$p = .001$</td>
</tr>
<tr>
<td>3. Self-reported computer skills – self-reported exposure to PubMed</td>
<td>$r_s = .312$</td>
<td>$p = .008$</td>
</tr>
<tr>
<td>4. Self-reported test performance – PubMed = user-friendly</td>
<td>$r_s = .463$</td>
<td>$p = .000$</td>
</tr>
<tr>
<td>5. Vocabulary test – problems finding right keywords</td>
<td>$r_s = .303$</td>
<td>$p = .010$</td>
</tr>
<tr>
<td>6. Vocabulary test – spelling uncertainty</td>
<td>$r_s = .382$</td>
<td>$p = .001$</td>
</tr>
<tr>
<td>7. Reading test – problems finding right keywords</td>
<td>$r_s = .394$</td>
<td>$p = .001$</td>
</tr>
<tr>
<td>8. Reading test – spelling uncertainty</td>
<td>$r_s = .277$</td>
<td>$p = .019$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mann–Whitney U</th>
<th>$z$</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. Education level – self-reported language skills</td>
<td>$U = 381.00$</td>
<td>$z = -2.923$</td>
</tr>
<tr>
<td>10. Education level – self-reported computer skills</td>
<td>$U = 337.50$</td>
<td>$z = -3.646$</td>
</tr>
<tr>
<td>11. Education level – self-reported test performance</td>
<td>$U = 439.50$</td>
<td>$z = -2.141$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Associations among search process characteristics.</th>
<th>Spearman’s rho</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>12. Quality of the first search term – number of bad search terms</td>
<td>$r_s = -.286$</td>
<td>$p = .016$</td>
</tr>
<tr>
<td>13. Hesitations and errors – number of citations selected</td>
<td>$r_s = -.336$</td>
<td>$p = .004$</td>
</tr>
<tr>
<td>14. Time on task: reading – bad MeSH terms in “best” query</td>
<td>$r_s = -.263$</td>
<td>$p = .026$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Associations between respondent and search process characteristics.</th>
<th>Spearman’s rho</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>15. Self-reported exposure to PubMed – query complexity</td>
<td>$r_s = .283$</td>
<td>$p = .017$</td>
</tr>
<tr>
<td>16. Reading test – hesitations and errors</td>
<td>$r_s = -.294$</td>
<td>$p = .013$</td>
</tr>
<tr>
<td>17. Vocabulary test – hesitations and errors</td>
<td>$r_s = -.252$</td>
<td>$p = .034$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mann–Whitney U</th>
<th>$z$</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>18. Education level – query complexity</td>
<td>$U = 428.00$</td>
<td>$z = -2.230$</td>
</tr>
<tr>
<td>19. Education level – total querying time</td>
<td>$U = 406.00$</td>
<td>$z = -2.481$</td>
</tr>
<tr>
<td>20. Education level – language errors in search terms</td>
<td>$U = 432.50$</td>
<td>$z = -2.218$</td>
</tr>
<tr>
<td>21. Education level – hesitations and errors</td>
<td>$U = 444.50$</td>
<td>$z = -2.049$</td>
</tr>
</tbody>
</table>

**Search Results**

**Number of relevant citations in the set of selected citations.** The participants in our test selected 2.2 relevant—max = 13, $Mdn = 1$ (0, 3; 3 IQR)—and 4.6 irrelevant—max = 21, $Mdn = 3$ (1, 7; 6 IQR)—citations. Thirty-seven percent of the test participants did not select any relevant citations, and consequently had a recall score of 0%. In half of those cases, the potential recall score was also zero. This means that these students’ queries did not yield any relevant citations.

In total, 59% of the participants had higher potential than actual recall scores, which indicates that they overlooked relevant citations and hence could have achieved higher recall with the same queries. The average potential recall was 6.8%, almost double the average actual recall score.

**Precision.** On average, only one in three of the citations selected was relevant: The average precision score was 33.30%. Some students achieved 100% precision; however, as mentioned above (see Statistical Issues section), this may misrepresent the performance of these students, as in most cases, this is at the cost of recall.

**Recall.** The average recall score of the selections made by our test participants was 3.7%, and maximum recall was 20%.

**Exploratory Analysis**

**Associations among respondent characteristics.** The students’ self-assessment of their English-language skills was quite accurate: Students with high scores on the reading and vocabulary tests rated their language skills higher in the pretest questionnaire (Table 3; items 1 and 2). Students with better computer skills used PubMed more often to search for medical information (Table 3; item 3), and those who had a positive perception of their retrieval results indicated that PubMed was a user-friendly search system (Table 3; item 4). Students with lower scores on the language test indicated they had problems finding the right keywords for their searches, and that they were uncertain about the spelling of the English words (Table 3; items 5–8).

There were also some differences in respondent characteristics between the bachelor’s and the master’s students. In general, the master’s students seemed to be more confident...
about their skills and performance on the test than the bachelor’s students. The bachelor’s students rated their language skills lower than the master’s students did, bachelor’s students: Mdn = 4 (3.5, 5; 1.5 IQR); master’s students: Mdn = 5 (4.75, 5.25; 0.50 IQR) (see Table 3; item 9). The master’s students were also more confident about their computer skills, bachelor’s students: Mdn = 3 (3, 3; 0 IQR); master’s students: Mdn = 4 (3, 4; 1 IQR) (see Table 3; item 10), and about their performance on the test, bachelor’s students: Mdn = 4 (3, 5; 2 IQR); master’s students: Mdn = 4 (3, 5; 2 IQR) (see Table 3; item 11).

The master’s students used PubMed more often to search for medical information (see Figure 7), whereas most of the bachelor’s nursing students rarely or never used this search engine. In summary, the main differences between the bachelor’s and master’s level students were related to their confidence in their own skills, which is a subjective assessment, and to their experience with PubMed, operationalized as exposure to PubMed and prior training in literature searching. Hence, the division into master’s and bachelor’s level students can be reduced to the division into more and less experienced PubMed users.

Associations among search process characteristics. When the quality of the first search term was low, the rest of the search terms were usually badly formulated as well (Table 3; item 12). This indicates that the effect of human learning (White, Marchionini, & Muresan, 2008) on query formulation was minimal in this test, probably due to the limited time.

As can be expected, hesitations have a negative impact on the number of citations that were selected (Table 3; item 13).

The time the students spent on reading the search task was inversely correlated with the number of bad MeSH terms in their best query (Table 3; item 14), i.e., the query that covered the highest number of gold standard concepts. This indicates that a good understanding, interpretation, and articulation of the information need is crucial for the formulation of a good, comprehensive query.

Associations between respondent characteristics and search process characteristics. The average number of terms used per query, i.e., query complexity, was affected by PubMed experience: Frequent and more experienced users tended to formulate longer queries (Table 3; items 15 and 18). Although the construction of a query involves some translation processes, language skills did not seem to play a role in the coverage of gold standard concepts, nor did it influence the proportions of good, bad, and medium search and MeSH terms.

Also remarkable was the relation between language skills and hesitations and errors. Although hesitations in the query formulation stage were not annotated as hesitations and errors, we see that the lower the scores on the language tests, the more the participants hesitated and made searching errors (Table 3; item 17, 18). This might indicate that there were problems with the language of the interface.

The more experienced searchers in our test group spent less time on the construction of queries than the less experienced searchers (Table 3; item 19), and also produced less language errors in their search terms (Table 3; item 20). The more experienced searchers constructed more complex queries (average number of terms: 3.7 vs. 2.9 in the less experienced group) with a smaller number of bad MeSH terms (16% as opposed to 22% in the less-experienced group) in a shorter querying step, which confirms that they are more experienced in searching PubMed and therefore perform smoother searches. Their level of experience was also reflected in a difference in hesitations and errors (Table 3; item 21).

Associations between respondent characteristics and search results. The main aim of this study was to determine the effect of language skills on the efficiency of literature searches in PubMed. We therefore investigated the relationship between scores on the language tests and performance on the literature search. The test showed a significant relation between language skills—both vocabulary (Table 5; item 1) and reading (Table 5; item 2)—and recall. This means that participants with better English-language skills generally performed better on the literature search task.

Our data did not show a significant correlation between language skills and relevance judgment, which can be measured by precision. Table 4, however, shows a trend: Higher
There were some differences between the less and the more experienced searchers with regard to search results: A maximum of six relevant citations were selected in the less-experienced group versus 13 in the other group.

The more experienced searchers achieved slightly higher recall, mean $M = 4.42$, $Mdn = 2.31$ (0, 7.61; 7.61 IQR), than the less experienced students, $M = 2.69$, $Mdn = 1.59$ (0, 4.73; 4.73 IQR). Although this difference in recall is not significant, we do see that the highest recall scores were achieved by the more experienced searchers. The average precision score in the less experienced group was slightly higher, $M = 37.58$, $Mdn = 27$ (0, 67; 67 IQR), but not significantly, master’s: $M = 29.96$; $Mdn = 27$ (0, 50; 50 IQR). As this group selected a lower number of citations, it was easier to achieve high precision.

**Associations between search process characteristics and search results.** Our test participants were advised to use MeSH terms in their searches. The proportion of good (Table 6; item 3) or bad (Table 6; item 1) search terms did not have an influence on precision and recall. However, the selection of bad MeSH terms (Table 6; item 2) did prove to have a negative effect on performance scores and the selection of good MeSH terms resulted in better retrieval (Table 6; item 4).

Other factors that had an impact on retrieval were the number of corrections (Table 6; item 5), querying times (Table 6; item 6), and total evaluation times (Table 6; item 7). Precision and recall decreased with an increasing number of corrections, which might indicate that these participants had problems finding the right keywords. The total time spent on query formulation is inversely correlated with precision and recall. This means that participants who needed more time to formulate their queries selected a smaller number of relevant

---

**TABLE 4.** Precision and recall level of English-language skills ($n=71$).

<table>
<thead>
<tr>
<th>Reading level</th>
<th>Mean precision</th>
<th>Mean recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>.0882</td>
<td>.0227</td>
</tr>
<tr>
<td>A2</td>
<td>.2791</td>
<td>.0125</td>
</tr>
<tr>
<td>B1</td>
<td>.3386</td>
<td>.0281</td>
</tr>
<tr>
<td>B2</td>
<td>.3226</td>
<td>.0410</td>
</tr>
<tr>
<td>C1</td>
<td>.4161</td>
<td>.0462</td>
</tr>
<tr>
<td>C2</td>
<td>.4357</td>
<td>.0698</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Vocabulary level</th>
<th>Mean precision</th>
<th>Mean recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>.2818</td>
<td>.0322</td>
</tr>
<tr>
<td>A2</td>
<td>.3000</td>
<td>.0111</td>
</tr>
<tr>
<td>B1</td>
<td>.3286</td>
<td>.0337</td>
</tr>
<tr>
<td>B2</td>
<td>.3385</td>
<td>.0544</td>
</tr>
<tr>
<td>C1</td>
<td>.6350</td>
<td>.0873</td>
</tr>
<tr>
<td>C2</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

**TABLE 5.** Associations between respondent characteristics and search results ($n=71$).

<table>
<thead>
<tr>
<th>Spearman correlations</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Vocabulary test</td>
<td>$r_s = .145$</td>
<td>$r_s = .236$</td>
</tr>
<tr>
<td></td>
<td>$p = .229$ (NS)</td>
<td>$p = .048$</td>
</tr>
<tr>
<td>2. Reading test</td>
<td>$r_s = .161$</td>
<td>$r_s = .259$</td>
</tr>
<tr>
<td></td>
<td>$p = .180$ (NS)</td>
<td>$p = .029$</td>
</tr>
<tr>
<td>3. Difficulties finding right keywords</td>
<td>$r_s = .167$</td>
<td>$r_s = .353$</td>
</tr>
<tr>
<td></td>
<td>$p = .163$ (NS)</td>
<td>$p = .003$</td>
</tr>
<tr>
<td>4. Spelling uncertainty</td>
<td>$r_s = .134$</td>
<td>$r_s = .380$</td>
</tr>
<tr>
<td></td>
<td>$p = .266$ (NS)</td>
<td>$p = .001$</td>
</tr>
<tr>
<td>5. Computer skills</td>
<td>$r_s = .154$</td>
<td>$r_s = .092$</td>
</tr>
<tr>
<td></td>
<td>$p = .199$ (NS)</td>
<td>$p = .443$ (NS)</td>
</tr>
<tr>
<td>6. Self-reported exposure to PubMed</td>
<td>$r_s = .060$</td>
<td>$r_s = .118$</td>
</tr>
<tr>
<td></td>
<td>$p = .619$ (NS)</td>
<td>$p = .327$ (NS)</td>
</tr>
<tr>
<td>7. Self-reported performance on search task</td>
<td>$r_s = .540$</td>
<td>$r_s = .551$</td>
</tr>
<tr>
<td></td>
<td>$p = .000$</td>
<td>$p = .000$</td>
</tr>
<tr>
<td>Mann–Whitney $U$ test</td>
<td>$U = 604.00$</td>
<td>$U = 540.00$</td>
</tr>
<tr>
<td></td>
<td>$z = -1.89$</td>
<td>$z = -9.44$</td>
</tr>
<tr>
<td></td>
<td>$p = .850$ (NS)</td>
<td>$p = .345$ (NS)</td>
</tr>
</tbody>
</table>

**TABLE 6.** Associations between search process characteristics and search results ($n=71$).

<table>
<thead>
<tr>
<th>Spearman correlation</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Proportion of bad search terms</td>
<td>$r_s = -.051$</td>
<td>$r_s = -.129$</td>
</tr>
<tr>
<td></td>
<td>$p = .675$ (NS)</td>
<td>$p = .284$ (NS)</td>
</tr>
<tr>
<td>2. Proportion of bad MeSH terms</td>
<td>$r_s = -.252$</td>
<td>$r_s = -.302$</td>
</tr>
<tr>
<td></td>
<td>$p = .034$</td>
<td>$p = .011$</td>
</tr>
<tr>
<td>3. Proportion of good search terms</td>
<td>$r_s = -.040$</td>
<td>$r_s = -.036$</td>
</tr>
<tr>
<td></td>
<td>$p = .738$ (NS)</td>
<td>$p = .776$ (NS)</td>
</tr>
<tr>
<td>4. Proportion of good MeSH terms</td>
<td>$r_s = .307$</td>
<td>$r_s = .333$</td>
</tr>
<tr>
<td></td>
<td>$p = .009$</td>
<td>$p = .005$</td>
</tr>
<tr>
<td>5. Corrections</td>
<td>$r_s = -.333$</td>
<td>$r_s = -.389$</td>
</tr>
<tr>
<td></td>
<td>$p = .005$</td>
<td>$p = .001$</td>
</tr>
<tr>
<td>6. Querying times</td>
<td>$r_s = -.278$</td>
<td>$r_s = -.432$</td>
</tr>
<tr>
<td></td>
<td>$p = .019$</td>
<td>$p = .000$</td>
</tr>
<tr>
<td>7. Total evaluation times</td>
<td>$r_s = .127$</td>
<td>$r_s = .391$</td>
</tr>
<tr>
<td></td>
<td>$p = .290$ (NS)</td>
<td>$p = .001$</td>
</tr>
<tr>
<td>8. Concept coverage</td>
<td>$r_s = .213$</td>
<td>$r_s = .236$</td>
</tr>
<tr>
<td></td>
<td>$p = .074$ (NS)</td>
<td>$p = .048$</td>
</tr>
<tr>
<td>9. Number of citations selected</td>
<td>$r_s = .274$</td>
<td>$r_s = .671$</td>
</tr>
<tr>
<td></td>
<td>$p = .021$</td>
<td>$p = .000$</td>
</tr>
</tbody>
</table>
citations. Long querying times can either indicate that the formulation of the query was done with great consideration, or that the participant hesitated. The second explanation seems more plausible, as precision and recall go down with increasing querying times. This is corroborated by our data, which show positive correlation between hesitations and errors and querying times ($r = .412; p = .000$). The time spent on relevance judgment, on the other hand, was positively correlated with recall. This indicates that a thorough relevance judgment step is crucial for successful retrieval.

Queries covering the four concepts (elderly, falls, long-term care, and prevention) resulted in better recall, but not necessarily in higher precision (Table 6; item 8). This underlines the importance of good relevance judgment: A good query might yield a large number of relevant results, but it is then up to the searcher to make a good selection. The selection of a higher number of citations (Table 6; item 9) resulted in higher recall, which seems logical. However, it also resulted in higher precision, which contradicts the classical trade-off between precision and recall.

Although we did not find a significant correlation between query complexity and search performance, we did see a peak in precision and recall at four to six terms per query (see Figure 8).

The use of more than six search terms in a query caused a steep drop in these scores, and less than four search terms yielded moderately lower scores as well. It seems logical that the optimal query for this search question contains four terms for the four components of the search to be represented. Some concepts can be translated into a combination of terms, which explains the fact that a query containing more than four search terms can also be successful. Overspecification, i.e., more than six terms, may lead to empty result sets. The ideal query for this task would therefore consist of four to six search terms. In accordance with these findings, our gold standard query consisted of six MeSH terms.

**Discussion**

From a methodological point of view, the main strength of this study is that direct observation using the Morae software allowed us to collect both quantitative and qualitative data without interfering in the IR process or affecting the search results. This study distinguishes itself from previous work in the field in that it not only analyzes the query formulation process and the resulting citations, but also two very important human interaction steps: need articulation and relevance judgment.

**Main Findings**

Precision and recall were quite low in the whole test group. The highest recall scores were achieved by master’s students, whose searching skills were also reflected in smoother searches with fewer hesitations. English-language skills were crucial in cross-language literature searching: Recall correlated positively with reading and vocabulary skills, and there was a positive trend in precision scores with increasing language skills.

The English MeSH terms had a corrective effect when compared to free-text searching and can therefore be considered as a very useful search aid also for nonnative speakers of English.

It is self-evident that high concept coverage, i.e., the number of concepts from the information need that are actually translated into MeSH terms and combined into a query, is a prerequisite for a good query. There are several reasons for noncoverage of concepts: The main cause was the nonidentification of concepts in the search question. It is therefore very important that searchers know exactly what they are searching for before they start formulating queries. Other causes were the use of bad search terms, and failure to identify a good MeSH term, even with good search terms.

**Limitations**

A limitation of this study was the relatively short period in which the students had to complete the literature search task. However, as the same amount of time was allowed to all participants, we were able to make a valid comparison. Moreover, finding relevant information in a relatively short period can be important in real-life clinical situations.

According to Wendt (1969) and Jacobson and Fusani (1992), the importance of the information need and the motivation of the users in a test case affect the effort made and the results obtained in the search task. In our study, problem identification was admittedly based on a preformulated question rather than on a spontaneous information need, but as this was true to the same extent for all participants, differences in motivation were unlikely to have a major falsifying influence.

We consider high concept coverage as the result of a well-thought-out articulation of the information need combined with the formulation of linguistically correct search terms, but it can probably also be linked to levels of intelligence. This, however, was not studied in this test.

We acknowledge that, in correlating evaluation times with recall, we did not take into account other sources of
difficulties, such as poorly written abstracts, problems understanding the texts in English, etc. However, as we noticed that many participants decided too quickly that citations were relevant, and as there was a strong correlation between evaluation times and recall, we are convinced that a longer and more thorough evaluation step is crucial to a successful search.

Critical Remarks on Main Findings

The role of search engine experience. Several studies (Aula, 2003; Bernstam et al., 2001; Haynes et al., 1990; Lazonder, Biemans, & Wopereis, 2000) conclude that experienced users obtain better results in literature or information search tasks. Fenichel (1981), on the contrary, found that there are only very small differences in the performance of users with different system experience. The more experienced searchers did not perform significantly better on the literature search task. However, we do see that the top 10 recall scores were achieved by these students. Rather than concluding that search engine experience does not have an impact on the efficiency of PubMed searches, we can say that the distinction between the two test groups does not correspond to the distinction between experts and novices made in the aforementioned literature. In other words, the bachelor’s students may be designated as novice users, as most of them have no experience with PubMed, but the master’s students are not experienced enough to be considered as experts.

Search results. The search results in terms of precision and recall are quite low. This can probably partly be attributed to the limited time in which the participants had to complete the literature search task. An experienced user with a spontaneous, specific information need would try to formulate a query that is as efficient and as comprehensive as possible. In this artificial situation, users who had little experience with the search system had to find very specific information in only 15 minutes. This, together with their limited search skills, resulted in a rather chaotic query formulation stage, mostly based on trial-and-error methods.

Taking into account this time limitation, we considered a search with a yield of five relevant citations or more as a successful search. This list of citations could then be expanded using the related-citations tool. This cutoff was achieved by 3% of the less experienced and by 28% of the more experienced searchers. One in five participants had zero potential recall, which means that they did not submit any queries that yielded relevant results. Almost two out of three students had higher potential than actual recall, which means that they overlooked relevant citations and that they could have achieved higher recall with the same queries.

Mouillet (1999) concluded that the MEDLINE/Ovid users in her test group did not have a realistic view of their search results. They seemed to be quite satisfied with their retrieval, despite the fact that “their MEDLINE/Ovid utili-

zation was often irrelevant.” (p. 454) As some students in our test reported that they were quite pleased with their results, whereas the maximum recall score was 20%, we could conclude that these students, too, have an unrealistic view of their performance. However, we found a positive correlation between user satisfaction and actual performance, expressed in recall and precision, indicating that the better performing students were more enthusiastic about their results than those who had lower scores.

Search process. Our test participants were asked to use MeSH terms to construct their queries. This implies that they first entered free-text search terms and then selected one or more MeSH terms from the list of suggestions made by PubMed. Our data showed that the quality of the free-text search terms does not have an impact on precision and recall. This is not surprising because the actual queries were constructed with MeSH terms and not with free text. Whenever a test participant entered a bad search term (e.g., kinestherapy for physical therapy), a warning message appeared: “The following term was not found in MeSH: kinestherapy. See Details. No items found.” In other cases, the MeSH terms suggested for the search term were not suitable for the search question (e.g., the search term multifactorial yielded the MeSH terms Multifactorial Inheritance, Causality, Nephrogenic Fibrosing Dermopathy, Typhlitis, etc.). In many cases, a new—and usually better—search term was then formulated, and there was no impact on the search results. However, these bad search terms are a cause for noncoverage of concepts, which leads to broader and less precise queries. Other reasons for noncoverage were nonidentification and failure to select the correct MeSH term.

The use of MeSH terms, although only available in English in the PubMed search interface, reduced the number of medium and bad keywords in the queries by half. This indicates that the MeSH terms are a useful search aid, compensating for badly formulated search terms. However, the use of MeSH terms can also be a stumbling block: In more than two out of five cases, participants failed to select a good MeSH term. We assume that the possibility to search in one’s mother tongue might lead to an increase in concept coverage, and consequently also in recall.

Self-reported skills and their effect on search process and results. We investigated the relationship between general computer skills, on the one hand, and query complexity, the quality of search terms, and precision and recall scores on the other. Aula (2003) argues that general computer skills affect the query formulation process. However, we did not find a relation between the self-reported level of computer skills and the quality of the search terms, nor did the subjects’ computer skills affect precision and scores. Aula also observed that more experienced Web and computer users tend to formulate longer, more specific queries. Students in our test case who estimated their computer skills higher, however, did not formulate longer queries.
As opposed to general Web and computer skills, exposure to the search engine PubMed did prove to have an impact on query complexity. This is in accordance with Sutcliffe et al. (2000), who found that searchers with more MEDLINE experience use more complex queries when compared to novices, who keep their queries rather simple.

Facility with the search engine is also reflected in the participants’ pause behavior: Participants who were more familiar with the search system paused less during their literature search. This is in accordance with Huang’s findings (2003).

According to Herskovic, Tanaka, Hersh, and Bernstam (2007) and Lin and Smucker (2008), between 16 and 20% of all queries submitted to PubMed yield zero results. We found similar results in our data. Our data showed that zero results can be due to many factors, including badly formulated terms or the selection of incorrect MeSH terms, inexperience with the search system, or the formulation of queries that are too narrow or complex.

Several studies (Sewell & Teitelbaum, 1986; Sutcliffe et al., 2000; Vakkari, Pennanen, & Serola, 2003) have shown that more experienced searchers tend to use more advanced Boolean operators, as opposed to novices who mostly use the AND-operator. This, however, is not corroborated by our data, probably because the master’s students had not reached this level of expertise yet.

Language skills and search results. Higher scores on the DIALANG language test, and therefore better language skills, resulted in higher precision and recall. We assumed that the language barrier would play a crucial role in the stage where active language skills are needed, i.e., the query formulation stage. However, there was no significant correlation between language skills and query formulation in terms of proportions of good, bad, and medium search and MeSH terms. Although there was no significant correlation between language skills and the quality of the terms used in the queries, the participants with lower scores on the language test did indicate that they had problems finding the right keywords and that they hesitated about the spelling of the English words. However, this was not reflected in the efficiency scores.

So in which stage do these language skills come to play such an important role that they entail higher performance scores? Or, in which stage does the language barrier hamper efficient searching? We already mentioned that nonnative English users of PubMed might have difficulties with the interface. Moreover, participants with better scores on the reading test selected a higher number of relevant citations, which means that language skills play an important role in relevance judgment. The importance of language skills in the relevance judgment stage is also emphasized by Mouillet (1999). She compared the answers of self-trained and librarian-mediated users of MEDLINE/Ovid and Pascal (a French bibliographic database) users to a survey in which she focused on the impact of the language barrier on the understanding of the MEDLINE/Ovid interface. Although her test did not simulate an information need and a resulting information search, she did conclude that the English-language barrier is especially reflected in the erroneous selection of articles.

Conclusions

We conducted an experiment to analyze the search behavior of Dutch-speaking nursing students and the efficiency of their literature searches in PubMed, focusing on query formulation and relevance judgment. We found that searching for information about a given topic within a limited time span is a complex and difficult task, the outcome of which is influenced by many factors.

English-language skills proved to have an impact on the efficiency scores: Students with higher scores on the language test also performed better on the literature search task. Especially the relevance judgment stage benefits from better language skills: Students with better knowledge of the English language were better at detecting highly relevant articles and thus had higher precision and recall scores.

Based on our test data, we cannot conclude that search engine experience has an impact on search efficiency. However, the top recall scores were achieved by the more experienced searchers. Moreover, as they were more familiar with the search system, they hesitated less during the search process and spent less time on querying. Although there was no significant difference in language skills, the more experienced searchers formulated a smaller number of incorrect search terms. In summary, we can state that the students who were more familiar with the search system performed relatively smooth searches, apparently experiencing fewer hitches than less experienced searchers.

An analysis of concept coverage showed us that good need articulation, although implicit in this research, is crucial, as higher concept coverage led to higher efficiency scores. The importance of a good interpretation and articulation of the information need, together with good relevance judgment, is underlined by our findings. The translation of an information need into concepts and from concepts into MeSH terms should therefore be an important part in bibliographic instruction, next to the actual use of search engines.

The medical subject headings proved to be a useful language aid, as they compensated for bad search terms. Conversely, the selection of erroneous MeSH terms resulted in an unproductive query. The medical subject headings can therefore be very helpful, but they can easily become a stumbling block when used incorrectly.

In conclusion, the main factors influencing the efficiency of a biomedical literature search in PubMed across language boundaries are language skills, facility with the search engine, a good parsing of the information need into concepts, a careful selection of MeSH terms, and an in-depth evaluation of the relevance of the articles retrieved.
Future Work

We realize that the current subject matter is quite comprehensive; therefore, not every aspect could be studied. We would like to set up several studies in which we will analyze the query formulation step in more detail. We will have students construct a query in Dutch, which will allow us to study concept identification. Second, we would like to study the Dutch–English translation step by having students translate a good query from Dutch into English. Another interesting task would be to have the students search for good MeSH terms for a given query, formulated in English. To analyze the relevance judgment step, we will give a test group a list of citations from which they have to select the relevant ones.

In addition, a think-aloud protocol study would be interesting to reveal the steps between concept identification and concept coverage.

Acknowledgments

We express our gratitude to Professor Dr. Monique Elseviers and to Liesbeth Van Heck, who helped us organize the test in the Nursing Departments of the University of Antwerp and University College Ghent, respectively. We would also like to thank Marc De Spiegelaere and Annelies Verhaeghe, who offered great help for the statistical analysis of our test data. Thanks are due to Joke Coussement of the Centre for Health Services and Nursing Research, KU Leuven, who gave helpful hints in formulating the search question.

References


