Are Self-Assessments Reliable Indicators of Topic Knowledge?

Michael J. Cole, Xiangmin Zhang, Jinging Liu, Chang Liu, Nicholas J. Belkin, Ralf Bierig, and Jacek Gwizdka

School of Communication and Information

Rutgers, The State University of New Jersey

4 Huntington Street, New Brunswick, NJ 08901, USA

{m.cole, belkin, bierig, jacekg}@rutgers.edu, {jingjing, changl}@eden.rutgers.edu, xiangminz@gmail.com

ABSTRACT

Self-assessment of topic/task knowledge is a human metacognitive capacity that impacts information behavior, for example through selection of learning and search strategies. It is often used as a measure in experiments for evaluation of results and those measurements are taken to be generally reliable. We conducted a user study (n=40) to test this by constructing a concept-based topic knowledge representation for each participant and then comparing it with the participant judgment of their topic knowledge elicited with Likert-scale questions. The tasks were in the genomics domain and knowledge representations were constructed from the MeSH thesaurus terms that indexed relevant documents for five topics. The participants rated their familiarity with the topic, the anticipated task difficulty, the amount of learning gained during the task, and made other knowledge-related judgments associated with the task. Although there is considerable variability over individuals, the results provide evidence that these self-assessed topic knowledge measures are correlated in the expected way with the independently-constructed topic knowledge measure. We argue the results provide evidence for the general validity of topic knowledge self-assessment and discuss ways to further explore knowledge self-assessment and its reliability for prediction of individual knowledge levels.

Keywords

User studied, methodology, knowledge representation

INTRODUCTION

It is not unusual in information science experiments to have an interest in measuring the participants' knowledge of the information task they are asked to perform. User knowledge has been identified to have significant effects on task performance and search behavior (e.g. Hsieh-Yee, 1993; Kelly and Cool, 2002; Toms et al, 2007; Wen et al., 2006; White et al., 2009; Wildemuth, 2004; Zhang et al., 2005). Knowledge is a complex mental state and is hard to measure. One can administer tests or ask participants to take some other actions to demonstrate knowledge, but these methodologies are costly to implement and it is hard to interpret the match between these measurements and the specific tasks the participants will perform. It is common to rely on participants' direct judgments of their knowledge. This paper addresses the question of the reliability and validity of this methodology by looking at a case where an alternative and somewhat objective measure of topical knowledge can be constructed and compared with the self-assessments.

Knowledge self-assessment is also an important metacognitive capacity that is exercised by users in the course of carrying out information behaviors, for example in planning and choosing learning strategies. The recognition of an information need and the struggle to precisely specify that need (Belkin, Oddy & Brooks, 1982) involves such a metacognitive capacity. Other user information behavior depends on metacognitive capacities, for example "I'll know it when I see it" or recognition of affective states.

Taking account of metacognitive capacities in user models is of central importance for some advanced information applications such as automated education systems (Pirrone, Azevedo & Biswas, 2009). The design goals for interactive information retrieval systems include many of the same requirements, for example in decision support capabilities and taking account of affect (Belkin, 2008). In this more general perspective, knowledge self-assessment is a worthwhile object of study for information science even beyond its role in experimental methodology and evaluation of results.

Tasks in information science experiments are often closely associated with a topic. For example, a task is often designed to involve the identification of relevant

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.

documents, which is understood to mean *topically-relevant* documents. These types of common tasks were used in our study and so we roughly conflate task knowledge and topic knowledge in the following. A precise understanding of the distinctions between task and topical knowledge is outside of the scope of this study. The results reported here concern the general action of knowledge self-assessment and do not turn on characteristics that may distinguish task knowledge and topic knowledge. We believe our results can be applied in settings where the research design may be explicitly concerned with task or topic knowledge.

RELATED WORK

A standard technique for eliciting topic or task knowledge is via self-assessment in pre- and post-task questionnaires. It can come in several forms, but commonly involves ratings on a Likert scale to a question along the lines of "How familiar is this topic to you?" or "How difficult do you anticipate this task will be?"

Validation of this methodology has come mostly in the educational psychology field where it has been studied for a long time. Boud and Falchikov (1989) provide a review of the history of work in this area. One reasonably consistent finding has been that accurate self-assessment is associated with experience. Another is that high performing students are more accurate but tend to underestimate their knowledge as compared to others and the opposite is observed for poorly performing students.

Similar results have been obtained in self-assessments associated with evaluation of information retrieval systems. Kelly et al. (2006) conducted an evaluation of alternative information systems for intelligence analysts. One aspect of the evaluation involved self-assessment and cross-assessment of the reports produced by the analysts. They found that the highest scoring analysts, that is, the analysts producing the best reports, were most accurate in ranking the work of others but tended to underestimate their own reports in self-assessment. The opposite was true for analysts who produced the lower ranked reports.

Self-assessment is of interest for intelligent tutoring systems where a design goal is to teach students how to learn. One line of research is to model metacognitive skills, including the ability to accurately self-assess knowledge. Mitrovic and Martin (Mitrovic, 2001; Mitrovic and Martin, 2007) have developed a tutoring system for the SQL database language. User studies showed some students cannot accurately self-assess their topic knowledge and that existing domain knowledge is an important factor for accurate self-assessment.

Ross (2006) reviews work in the area of language learning and concludes self-assessment judgments are generally well-correlated with observable skills such as reading and speaking. Studies related to knowledge-assessment in language learning have reached mixed conclusions. Malabonga et al. (2005) found that nearly all students could choose an appropriate difficulty level for testing in oral language proficiency exams. In contrast, Brantmeier (2006) studied advanced university students learning Spanish and reported that self-assessment of reading ability, measured before and after reading, was not a reliable measure of performance on a computer-based skill test. In a comprehensive follow-up Brantmeier and Vanderplank (2008) studied 359 students using a self-assessment questionnaire before treating them with a series of readings. A variety of performance tests were conducted, both computer- and paper-based, that included written recall, multiple choice, and topic familiarity. After the tests a knowledge self-assessment questionnaire was administered. They concluded self-assessment questionnaires can be reliable predictors of computer-based tests and classroom performance as measured by sentence comprehension and multiple-choice tests, but not for reading comprehension measured by recall tasks. The post task self-assessment was shown to be reliable for all three performance measures.

Vocabulary knowledge can be directly related to knowledge of concepts through an internal representation as concept features or as the mechanism for accessing these concepts. The essential link between meaningfulness of words and concept formation and use is a core aspect of research into the nature of concepts (cf Katz, 1972; Fodor, 1975; Armstrong et al., 1983). Words are also central to several computational representations of concepts (e.g. Landauer, 2002; Landauer et al., 2003; Burgess, 1998).

Despite the difficulty of fixing the precise relationship between psycholinguistics and concept access and use, vocabulary knowledge is well-accepted as an indicator of concept knowledge. This allows vocabulary self-assessment to be related to direct assessment of concept representation. Self-assessment of vocabulary knowledge has received specific attention, again in the area of language learning. Meara and Buxton (1987) showed that binary choices on work knowledge could provide a reliable indication of the overall vocabulary knowledge of the student. Harrington and Carey (2009) found that binary choice vocabulary knowledge self-assessment was approximately as accurate as grammar placement tests.

Self-assessment of knowledge of specific words has been used to initialize user models to personalize systems. Heilman and Eskenazi (2008) had students self-assess word knowledge to personalize an automated tutor for English as a second language. They found that although self-assessment of a collection of target words was not as reliable as cloze questions, where a number of possible word substitutions are embedded in a sentence, it is an attractive method for building user models because of low

demands on resources and time. They concluded that improving the granularity of the self-assessment questions to distinguish between word recognition and knowledge of the use of the word would probably improve personalization accuracy.

In summary, the self-assessment of various forms of knowledge has a long history and there is evidence that the technique is reliable in several situations. Self-assessment appears to be subject to knowledge effects, for example in the tendency for high knowledge individuals to systemically underrate their knowledge while low knowledge individuals tend to overrate their knowledge. There is also evidence that self-assessment of word knowledge can be accurate and is related to knowledge of a language, which is a kind of background or domain knowledge. Word self-assessment can be related to self-assessment of constitutive concepts in a knowledge domain by feature-based theories of concept formation and use. We exploit this idea in the present work

In our study we elicited two types of concept self-assessment. First, participants rated their knowledge of the National Library of Medicine's Medical Subject Headings (MeSH)[1] in several specific areas. They were then asked to self-assess their knowledge of the topic and task before and after the performance of the task. We look at the relationship between self ratings of topic knowledge and a measure indicating participant topic knowledge constructed from their rating of individual MeSH terms. Our main research question is to ask if there is a relationship between participants' direct self-assessment ratings of their topic knowledge and their indirect MeSH term-based self-assessment of topic knowledge.

METHODOLOGY

Experiment system

We implemented a search system using Indri from the Lemur toolkit (http://lemurproject.org) to deliver results to a web search interface presented in Internet Explorer (IE6). The experiment was conducted using the multi-source logging PooDLE system (Bierig et al., 2010).

Study procedure

The participants read and signed a consent form and filled out a questionnaire about their background, computer experience and previous searching experience. They were then asked to rate their knowledge of MeSH terms in three categories related to the search tasks (total – 409 terms). Before each task the participants filled out a questionnaire and then were given up to 15 minutes to conduct the search task. The interaction between the participants and the system was logged by the computer. After completing each task, a post-task questionnaire was administered. Finally, the participants completed an exit questionnaire. The experiment was conducted in a human-computer interaction lab and each participant was tested individually and received $25.

Participants

Forty students from the authors' institution participated in the study. They were recruited from related schools and departments, including biology, pharmacy, animal science, biochemistry, and so on. Undergraduate students, graduate students and post-docs participated in the study. The number of graduate students and undergraduate students was roughly balanced to support elicitation of different levels of knowledge. Post-docs were assigned to the graduates group.

Tasks and data sets

The search tasks and associated gold standard document relevance judgments were taken from the 2004 TREC Genomics track (Roberts et al., 2009). The available tasks were categorized along two dimensions: hard/easy and general/specific where specificity refers to path length from the most general node to the task topic subject in the MeSH tree. Hard topics were those with few relevant documents returned by our search system (precision @ 10) using the task description as the query.

The TREC Genomics track employed expert biologist judges, the majority of whom held a Ph.D. in biological sciences or an M.D. They received training and then provided relevance judgments ("highly relevant", "somewhat relevant", or "not relevant") covering the portion of the documents most likely to be returned by the search system for each task (Roberts et al., 2009). The available TREC judgment data did not include the explicit judgments of "not relevant" and we take all documents not judged as "relevant" to be "not relevant". This includes documents that were never considered by the assessors and is an acknowledged limitation of the TREC methodology.

Table 1 lists the tasks selected for the study and their characteristics. We switched task 49 for task 42 during the study, because task 42 was very easy, so the results for these tasks are based on different participants.

The experiment used a simulated work task approach (Borlund, 2003) to design the tasks as presented to the participants. Participants were asked to find and save all of the documents useful for answering the task questions. The task questions were the TREC genomics track descriptions. The pre- and post-task questions and scale values relating to topic knowledge were:

Pre-task questions

(1-Not at all, 4-Somewhat, 7-Extremely)

1. How difficult do you think the search task will be?

[1] http://www.nlm.nih.gov/mesh/

2. How familiar are you with the topic of this task?
3. How much search expertise do you have with this type of task? (1-None, 4-Some, 7-A great deal)

Post-task questions
(1- Strongly disagree, 4-Neutral, 7-Strongly agree)
1. The topic of this task was easy for me.
2. My previous knowledge on this topic helped me with this task.
3. I learned some new knowledge on this topic during this search task.
4. After doing the search task, the difficulty of the task was: (1-Very easy, 4-Neutral, 7-Very difficult)

The study search collection (n=1.85 million) was the 2000-2004 portion of the TREC Genomics collection, a ten-year, 4.5 million document subset of the MEDLINE bibliographic database (Hersh and Voorhees, 2009).

Topic models

The TREC assessors provide an expertise standard. This was used in two ways. First, the TREC assessments of document topic relevance were used to identify the MeSH terms that stand for concepts included in the topic. Second, their assumed expertise regarding each MeSH concept justifies a common point to compare participant knowledge. Specifically, every participant can be related to a hypothetical expert participant to allow aggregation of the self-assessments.

For this study, there are two critical assumptions about the TREC assessors. We assume every assessor is an expert for all of the task topics and that they would rate their self-knowledge of the topic to be the maximum. The other key assumption is that they would rate their knowledge of every MeSH term to be the maximum.

A topic domain knowledge representation was calculated with the idea of a topic as a single concept represented as a collection of concepts as features. The features are taken to be the MeSH terms that have been used to index the relevant documents associated with each topic. This idea is related to that of building a topic language model based on the relevant documents. In this case, MeSH terms, which stand for concepts in the MeSH taxonomy, are used instead of words that are presumed to point at concepts.

The topic representation was constructed by first gathering the documents deemed relevant by the TREC assessors. The MeSH index terms were then extracted. The aggregated terms (with term frequency preserved) that were also in the MeSH terms rated by participants were taken to be the features of the topic representation for this study (Table 1). Notice that the topic model for task 42 is based on many more documents than the other topics and so has more rated terms in its representation. Tasks 2, 7, and 45 have a similar number of rated terms, but task 49 has only 8 rated terms in its model. One can expect these characteristics will have effects on variances observed in the participant topic knowledge measures.

Participant topic knowledge model

The participant's topic knowledge was calculated using their knowledge ratings of the MeSH terms. At the beginning of the experiment each participant rated their knowledge of selected MeSH terms in the three MeSH trees associated with the topic categories (Table 1). The knowledge level rating choices were: 1 – no knowledge, 2 – vague idea, 3 – some knowledge, 4 – high knowledge, 5 – can explain to others. This calculation is normalized against a hypothetical expert participant, presumed to be the pooled TREC experts who rated the documents used to construct the topic models. This expert participant would rate every MeSH term at the maximum value. So the normalized participant topic knowledge (nPTK) is:

$$P_{taskK} = \frac{\left(\sum_i^m \left(w_{px} * t_i * mp_t\right)\right)}{\left(\sum_i^m \left(5 * t_i * mp_t\right)\right)}$$

where w_{px} are the individual MeSH term ratings elicited from the participants, and mp_t is the model weighting rule to be applied to the term, for example the frequency of the term in the topic representation.

Comparing topic model knowledge and self-assessed knowledge

We have two measures of a participant's topic knowledge. One is a direct self-assessment of their knowledge of the topic. The other is a constructed measure, nPTK, which depends on self-assessment of their knowledge of MeSH thesaurus concepts. Notice that the calculation procedure for nPTK has implicitly assumed an interval scaling of the MeSH term knowledge assessments and that the nPTK itself is a ratio measurement.

The participant questionnaire self-assessment of topic knowledge is ordinal. For each participant we can compare the self-assessed topic knowledge (SATK) with the participant's calculated nPTK. For the purposes of analysis we suppose the participant judgments of their topic knowledge and of the various MeSH concepts are

task	difficulty	specificity	terms	documents	Category
2	hard	2	36	101	Genetic structure
7	easy	1	27	115	Genetic processes
42	very easy	3	87	697	Genetic phenomena
45	hard	2	31	156	Genetic phenomena
49	easy	2	8	73	Genetic structure

Table 1: Task (TREC topic) representation characteristics: terms: number of terms in rated MeSH trees; specificity: median depth of MeSH terms in tree (path length to root)

independent of one another and therefore the scales for SATK and nPTK are orthogonal in a model space. We make the hypothesis that the participant's knowledge of the constitutive concepts in the topic cause, in some manner, their self-assessment judgment of topic knowledge. This means we can make a plot of SATK as a function of nPTK. This function provides a kind of objective measure of the participant's accuracy in their self assessment of topic knowledge. Notice that this argument is valid for a single participant (SATK, nPTK) pair, but fails when the results of many participants are compared if the SATK measurement is only ordinal. For this analysis we make the important assumption that the SATK measurement is a ratio measurement and can be compared across individuals. The discussion will comment further regarding this limitation on validity and suggest ways to address this important question.

path length from the term to the root node.

4. Term frequency and specificity model: The term weight is the product of the term frequency and specificity.

Calculations of these parameterized models were made for each task and participant, including the hypothetical expert. While nPTK values can range from 0.0 to 1.0, notice that they tend to be in the range 0.2 to 1.0, because a term rating of 1 means 'no knowledge'. If a participant rated every term in a topic as 'no knowledge' their nPTK would be 0.2. A participant could score below that only if they failed to rate terms in the topic. Of course a participant's nPTK could be greater than 0.2 and also include unrated terms.

model	min	1st	median	mean	3rd	max	sd
flat	0.175	0.348	0.448	0.460	0.526	0.933	0.159
term frequency (TF)	0.091	0.350	0.460	0.481	0.559	0.975	0.172
term specificity (TS)	0.141	0.298	0.403	0.425	0.496	0.926	0.161
freq*spec (TFTS)	0.118	0.322	0.421	0.448	0.530	0.960	0.171

Table 2: Participant variations in topic knowledge by model parameterization

Normalized participant topic knowledge measures were constructed in two steps. First, we created several versions of the basic topic representation model by weighting the knowledge of the MeSH terms using four rules:

1. Flat topic model: Terms are considered only as categorical dimensions in the representation, so the knowledge rating for the term is unweighted.
2. Term frequency model: The term weight is the frequency of the term in the topic knowledge representation.
3. Term specificity model: The term weight is the specificity of the term in the MeSH tree, i.e. the

RESULTS

Participant discrimination by alternative models

Table 2 shows the differences between the aggregated participant nPTK measures for each of the models. The flat model, which takes no account of the frequency or specificity of the terms in the topic representation, has the least discrimination over users. The term frequency-weighted model shows the greatest differences and standard deviation, while term specificity weighting is rather close to the flat model. The term frequency-weighted model is used in the following analysis, however the difference in discrimination as compared to the flat model is rather small. We make some further comments in the discussion section.

task	difficulty	topic depth	specificity	participants	min	1st	median	mean	3rd	max	sd
2	hard	specific	2	40	0.211	0.331	0.461	0.468	0.522	0.922	0.168
7	easy	general	1	40	0.215	0.369	0.485	0.499	0.569	0.933	0.162
42	very easy	specific	3	19	0.274	0.349	0.409	0.445	0.501	0.722	0.131
45	hard	general	2	40	0.213	0.332	0.423	0.438	0.474	0.787	0.144
49	easy	general	2	21	0.175	0.300	0.400	0.432	0.500	0.875	0.190

Table 3: Participant topic knowledge variation by task for flat model. Specificity: Mean MeSH tree depth of model terms

task	difficulty	topic depth	specificity	participants	min	1st	median	mean	3rd	max	sd
2	hard	specific	2	40	0.209	0.344	0.455	0.465	0.516	0.910	0.171
7	easy	general	1	40	0.091	0.414	0.542	0.563	0.704	0.975	0.193
42	very easy	specific	3	19	0.291	0.368	0.459	0.470	0.544	0.712	0.132
45	hard	general	2	40	0.215	0.371	0.444	0.462	0.546	0.852	0.150
49	easy	general	2	21	0.175	0.300	0.400	0.432	0.500	0.875	0.190

Table 4: Participant topic knowledge variation by task for term frequency model. Specificity: Mean MeSH tree depth of model terms

The nPTK measures show consistent variations between the different topic models (Table 2). For a given model, one can see considerable variations by task for the participants. For example, in the flat (unweighted) nPTK (Table 3) and the term frequency-weighted (TF) nPTK (Table 4), participants generally had less knowledge for the task 45 topic. Remember that tasks 42 and 49 involved distinct participant groups, so it is harder to interpret those variances. The TF model provides somewhat greater discrimination of participant topic knowledge as compared to the flat model but the absolute differences are rather small and due mostly to task 7.

Self-assessed topic knowledge and nPTK

Simple linear regression with Gaussian family smoothing was run on the self-assessments using the term-frequency weighted topic knowledge models. Figure 1 shows the correlation between participant self-assessments of their familiarity with the task topic and their calculated nPTK. The result for the three tasks completed by all participants is shown. The red line for each plot is the fitted residuals for all participants.

As participants had greater calculated topic knowledge (as measured by nPTK), they tended to indicate a higher self-assessment score. Although there is a great deal of variability, the plotted slopes resulting from take-one-out cross validation show every combination of participant (self-assessment, nPTK) pairs has a positive slope, although one case for task 45 is very nearly flat. There is one low nPTK outlier for task 7, which is due to the participant failure to rate a number of the MeSH terms in the topic model.

Plots of other knowledge related questions (not shown) asked on the pre- and post-task questionnaires were consistent and matched intuitions about expected results. For example, participant ratings of "My previous knowledge on this topic helped me with this task." and "I learned some new knowledge on this topic during this search task." were positively correlated with nPTK. In contrast, "After doing the search task, the difficulty of the task was: (1-Very easy, 4-Neutral, 7-Very difficult)" decreased with nPTK.

The other topic knowledge related questions followed the same pattern and matched the expected correlations between the self-assessed rating and the calculated nPTK. The only exception was the anticipated difficulty of task 7. Although it was an easy topic, the anticipated difficulty increased with greater nPTK. The post-task reassessment of difficulty, however, was in line with all of the other tasks: greater topic knowledge was correlated with lower difficulty assessments.

DISCUSSION

Self-assessments are a report of a state in a one dimensional mental space that is at least ordered but of unknown scale, so possible judgments for an individual may not reflect a linear scaling for judgment values. Such

variability can be due to the specific wording of the question, the Likert scale labels, and the intrinsic capacity of the person to make the judgment. Likert scales are generally taken to be interval scales.

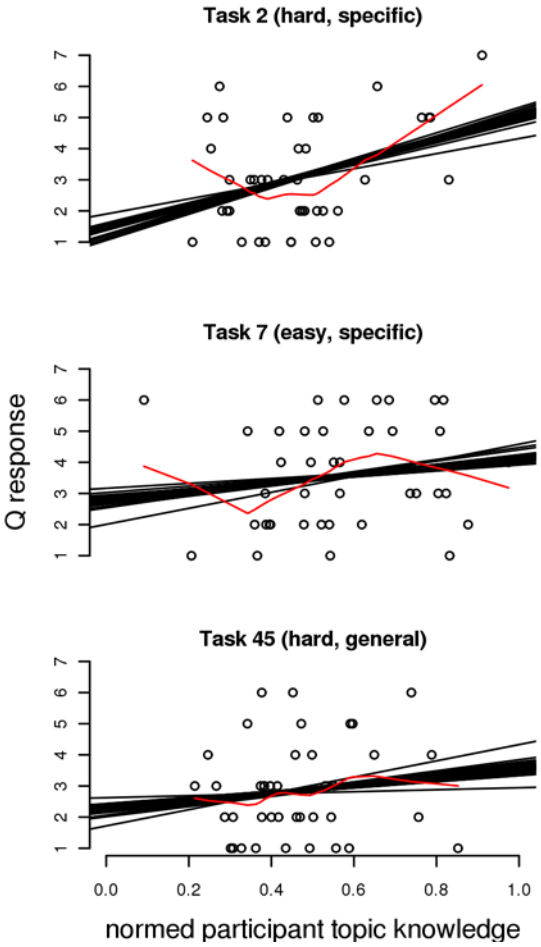

Figure 1 : Self-assessed topic familiarity as a function of calculated topic knowledge (nPTK)

Intuitions about the meaning of levels of knowledge seem to allow for ratio measures. That is, the amount of knowledge held by an individual can be represented as a finely grained interval scale. We have appealed to this intuition in building the nPTK topic knowledge measure on the foundation of the number of concepts known to the participant and the use of the rating in the calculation. However, the self-assessments made in the Likert questions are only valid as interval scales – there can be a big difference between 'high knowledge' and 'can explain to others' as compared to 'vague idea' and 'some knowledge'. This is true both in terms of the value the person associates with their judgment and with their uncertainly about the value selection. To calculate the regression function of SATK and nPTK, we have boldly assumed both measures are valid ratio measures.

Apart from the anticipated difficulty for task 7, the correlations for all of the knowledge-related pre- and post-task questions, as shown by simple regression between self-assessed ratings and the calculated topic knowledge, match intuitions about the interplay between existing knowledge, search experience, and actual and anticipated difficulty. Self-assessment of learning in the task showed stronger correlations with topic knowledge. The fact that the regression lines yield intuitive results is evidence that the Likert-question responses have ratio-like characteristics. Considering the regression results for all of the questions, even the fitted curves do not generally show wild departures from the simple regression lines. In some cases there are interesting kinks for by individuals with particularly low topic knowledge that may indicate some interplay between specific topic properties and self-assessments.

The data shows a great deal of individual variability and the correlations were weak if considered only from a modeling perspective. Our purpose in this work was not, however, to model user self-assessments of knowledge but rather to explore the reliability of self-assessments. The results provide evidence for the validity of the technique but fall far short of showing the reliability of self-assessments where that is understood to be accurate prediction of an individual knowledge level from their questionnaire response. Two types of explanations can be offered for this shortcoming. First, there are a number of significant limitations in the calculated participant topic knowledge (nPTK), for example coverage in the concept space. Second, there are basic issues concerning how knowledge is represented and used that affect the ability to explore the connection between the metacognitive judgment of level of knowledge and the expression of knowledge use in information behaviors.

An important limitation on the validity of this work is the assumption that the SATK is a ratio scale. The functions showing a general relationship between SATK and nPTK over participants depends on this assumption. The claim that SATK is a ratio measure (and that the construction of nPTK is accurate when treating the self-assessments of MeSH concept knowledge) begs important and long-studied questions in psychometrics. We would like to be able to say something about the characteristics of these self-assessed knowledge intervals in order to investigate specific questions about knowledge self-assessments in information science studies, for example, whether relative novices tend to overestimate their knowledge. Ultimately, we would like to know if any two participant ratings can be used to order the participant's actual knowledge with some probability of error. It would be especially useful to establish these relationships against a standard task to assist in allowing some types of comparisons of results across user studies.

Unfortunately, the present work does not address this vital question. We have shown that under the bold assumption of a ratio scale for self-assessed knowledge reasonable relationships are derived when aggregates of participant self-assessments are considered. Ratio scales are generally not thought to exist in psychology. It is possible, though, that certain interval scales, such as those elicited by Likert scales of self-assessed knowledge, may behave close enough to ratio scales with suitable interval scaling to address detailed questions about classes of users. This is a key direction for investigation. A specific step is to explore the knowledge-assessment Likert scales in this study as Rasch models (Bond & Fox, 2007).

Another significant limitation is that in our methodology each term in the topic knowledge representation is taken as an independent dimension in the model. In general, however, there is overlap between concepts. For our study, the constructed participant topic knowledge depends on a collection of judgments about knowledge of a collection of concepts, which are not discrete and disjoint. So there is a collection of judgments about knowledge that in some cases overlap. That is, knowledge of a concept depends on knowledge of another concept, which may or may not be in the rated collection. The specific structure of the topic knowledge representation is unknown and we have ignored the problem of compositionality of concepts. This shortcoming is significant but does not necessarily compromise our approach because we are concerned with relative measures amongst participants rather than absolute measures of topic knowledge. The construction of a hypothetical expert who rates every term at the highest knowledge value underwrites the validity of the comparison and in that sense makes the normalized participant topic knowledge objective.

It is interesting that term-frequency weighting rather than term-specificity weighting provides the greatest discrimination between users. One possible explanation, however, is that when the documents were selected for the construction of the models we took no account of the graded relevance. So the topic models might be dominated (and by varying degrees for each topic) by terms from documents that are only somewhat relevant. A direction for future work is to learn if similar discrimination between participants is achieved when model construction is restricted to the highly relevant documents.

Another observation is the slight difference between the term-specificity and term-frequency weighted models as compared to the unweighted model in discriminating users. There is rather little difference, only 10% or so, between the most and least discriminative model. This might suggest topic knowledge for users tends to be dominated by the number of features (constituent concepts) rather than their specificity. This possibility can be explored using the general methodology for topic representation we developed. One could deliberately select TREC topics for high contrasts in their differences with respect to term frequency and term specificity to see how these variances in topic representations affect the discrimination of participants. One contribution of this work might be that the count of distinct terms is the most important aspect of the topic models for discriminating knowledge levels in users. If so, that implies users are concept coverage-oriented when making judgments about their topic knowledge. That would be a satisfyingly intuitive explanation.

The limited coverage of rated MeSH terms in the entire MeSH space is an important limitation. However, the piecemeal nature of the constructed topic knowledge representation could have easily resulted in poor discriminatory power between participants and across tasks. The fact that a coherent story has emerged from our modeling efforts makes it reasonable to think the observed variances would be reduced with better coverage in the topic model concept space (i.e. having participants rate more MeSH terms). This limitation is therefore also encouraging for the validity of the self-assessment technique and its potential for general reliability. While it is true the participants rated the MeSH trees that might be expected to be most germane to the task topics (Table 1), it is not unreasonable to think that if they had rated more of the MeSH domain terms the model's knowledge discrimination of individual topic knowledge differences would be improved. Testing this hypothesis is one more direction for future work.

Another check on the reliability and validity of self-assessment can come in the substitutability of the measures. In effect, we have constructed another measure of participant topic knowledge and can ask if it provides any information that is not already provided in the direct self-assessment. One way to do this is by looking at activities where topic knowledge is presumed to play a role, e.g. document relevance judgments. If the direct self-assessment measure correlations with such knowledge-based behaviors agree with correlations between the behaviors and the nPTK, one would have additional evidence that direct self-assessment is indeed a good measure of topic knowledge.

Genomics is a specialized domain, so there may be some question about the generalizability of the results. We note that these results are broadly in line with work on self-assessed knowledge in other fields addressing unrelated domains. More generally, self-assessment is a meta-cognitive capacity to make judgments about properties of concepts. In this work, it concerns the ability to judge how much knowledge one possesses for a particular concept. While this capacity may vary across types of

concepts, it is difficult to see how such a meta-cognitive capacity would have domain dependencies.

CONCLUSIONS

This work provides evidence for the validity of self-assessment in eliciting topic knowledge. We compared two types of self-assessments made by participants about their topic/task knowledge. One is a direct judgment of their knowledge of the task/topic concept. The other is an indirect judgment made through self-assessment of constitutive concepts of the task/topic concept.

The regression lines for self-assessed topic knowledge as a function of the calculated topic knowledge for the various pre- and post-task questions were observed to fit well with our intuitions about the relationship between actual knowledge level and the assessment one is likely to give of expected difficulty, amount learned, and so on. Despite the observed individual variability these consistent results and intuitive explanations of the relationships across tasks and knowledge self-assessment questions provides evidence of a real correlation between the constructed topic knowledge model and self-assessed topic knowledge.

There is considerable variability in the correlation for individuals in a task which frustrates the ability to predict the participant response from the calculated topic knowledge. However, the consistency of the relationship over all tasks of varying types is heartening. In view of the rather low coverage of the participant MeSH term self-assessments as compared to the entire MeSH thesaurus, it seems likely that better coverage would lead to less variability in the correlation and further confirm the validity of self-assessments of task and topic knowledge. We have outlined several directions for further exploration of the validity and reliability of the self-assessment technique.

Self-assessment of topic knowledge is an expression of an important metacognitive capacity that affects information behavior. A person's understanding of their own knowledge impacts selection of strategies for information seeking and learning. Further study of knowledge self-assessment in information science is warranted, then, even beyond its methodological use in experiments.

Acknowledgments

This research was supported by IMLS grant LG-06-07-0105-07.

REFERENCES

Armstrong, S., Gleitman, L., and Gleitman, H. (1983). What some concepts might not be. *Cognition*, *13*(3), 263–308.

Belkin, N. J. (2008). Some(what) grand challenges for information retrieval. *SIGIR Forum 42*(1), 47-54.

Belkin, N.J., Oddy, R.N., & Brooks, H.M. (1982). ASK for information retrieval: Part I. Background and theory. *Journal of Documentation 38*(2), 61-71

Bierig, R., Cole, M.J., & Gwizdka, J. (2010). A data analysis and modelling framework for the evaluation of interactive information retrieval. In *Proceedings of ECIR 2010, (Milton Keynes, UK 2009)*, Lecture Notes in Computer Science. Berlin: Springer.

Bond, T.G. & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mawah, NJ : Lawrence Erlbaum.

Borlund, P. (2003). The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Information Research*, *8*(3), Retrieved from http://informationr.net/ir/8-3/paper152.html

Boud, D. & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education*, *18*(5), 529–549.

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System*, *34*(1), 15–35.

Brantmeier, C. & Vanderplank, R. (2008). Descriptive and criterion-referenced self-assessment with L2 readers. *System*, *36*(3), 456–477.

Burgess, C. (1998). From simple associations to the building blocks of language: Modeling meaning in memory with the HAL model. *Behavior Research Methods, Instruments, & Computers*, *30*, 188–198.

Fodor, J. (1975). *The language of thought.* Cambridge, MA: Harvard University Press.

Harrington, M. & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System*, *37*(4), 614–626.

Heilman, M. & Eskenazi, M. (2008). Self-assessment in vocabulary tutoring. *In Proceedings of Intelligent Tutoring Systems 9th International Conference, ITS 2008, Montreal, Canada*, June 23-27, 2008, volume 5091/2008 of LNCS, 656–658, Berlin. Springer.

Hersh, W. & Voorhees, E. (2009). TREC genomics special issue overview. *Information Retrieval*, *12*(1), :1-15.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on online search behavior: Measuring the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science*, *44*(3), 161–174.

Katz, J. (1972). *Semantic theory.* New York: Harper & Row.

Kelly, D. & Cool, C. (2002). The effects of topic familiarity on information search behavior. In *Proceedings of the Second ACM/IEEE Joint*

Conference on Digital Libraries (JCDL '02), Portland, Oregon, 74-75. New York. ACM.

Kelly, D., Kantor, P., Morse, E., Scholtz, J., & Sun, Y. (2006). User-centered evaluation of interactive question answering systems. In *Proceedings of the Interactive Question Answering Workshop at HLT-NAACL 2006 June 9,2006, New York, New York*, 49–56. Stroudsburg, PA. Association for Computational Linguistics.

Landauer, T.K., Laham, D., & Foltz, P. (2003). Automated scoring and annotation of essays with the intelligent essay assessor. In Shermis, M. and Burstein, J., editors, *Automated Essay Scoring: A cross-disciplinary perspective*, pp 87–112. Lawrence Erlbaum, Mawah, NJ.

Landauer, T.K. (2002). On the computational basis of learning and cognition: Arguments from LSA. In Ross, N., editor, *The Psychology of Learning and Motivation*, volume 41, pages 43–84.

Malabonga, V., Kenyon, D., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, *22*(1), 59-82.

Meara, P. & Buxton, B. (1987). An Alternative to Multiple choice tests. *Language Testing*, *4*(2), 142–154.

Mitrovic, A. (2001). Investigating students' self-assessment skills. In *Proceedings of the 8th International Conference for User Modeling, Sonthofen, Germany*, July 13–17, 2001, volume 2109/2010 of LNCS, 247–250, Berlin. Springer.

Mitrovic, A. & Martin, B. (2007). Evaluating the Effect of Open Student Models on Self-Assessment. *International Journal of Artificial Intelligence in Education*, *17*(2), 121–144.

Pirrone, R., Azevedo, R. & Biswas, G. (2009). Why metacognition in modern education systems? In *Proceedings of AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems, Alexandria, Virginia, USA, November 8, 2009*. vii-viii Menlo Park, CA. AAAI Press

Roberts, P.M., Cohen, A.M., & Hersh, W.R. (2009). Tasks, topics and relevance judging for the TREC genomics track: Five years of experience evaluating biomedical text information retrieval systems. *Information Retrieval*, *12*(1), 81–97.

Ross, J.A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment Research & Evaluation*, *11*(10), 1–13.

Toms, E., Mackenzie, T., Jordan, C., O'Brien, H., Freund, L., Toze, S., Dawe, E., & MacNutt, A. (2007). How task affects information search. In *Workshop Pre-Proceedings in Initiative for the Evaluation of XML Retrieval (INEX)*. 337-341.

Wen, L., Ruthven, I., & Borlund, P. (2006). The effects on topic familiarity on online search behaviour and use of relevance criteria. In *Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006 (London, UK)*, 456–459, London. Springer

White, R.W., Dumais, S.T., & Teevan, J. (2009). Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 132–141. ACM New York, NY.

Wildemuth, B.M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, 55(3), 246–258.

Zhang, X., Anghelescu, H., & Yuan, X. (2005). Domain knowledge, search behaviour, and search effectiveness of engineering and science students: An exploratory study. *Information Research*, *10*(2) Retrieved from http://informationr.net/ir/10-2/paper217.html