A Comparison of Techniques for Measuring Sensemaking and Learning within Participant-Generated Summaries

Mathew J. Wilson, Max L. Wilson

Future Interaction Technology Lab,

Swansea University, UK

{csmathew,m.l.wilson}@swansea.ac.uk

While it is easy to identify whether someone has found a piece of information during a search task, it is much harder to measure how much someone has learned during the search process. Searchers who are learning often exhibit exploratory behaviours, and so current research is often focused on improving support for exploratory search. Consequently, we need effective measures of learning in order to demonstrate better support for exploratory search. Some approaches, such as quizzes, measure recall when learning from a fixed source of information. This research, however, focuses on techniques for measuring open-ended learning, which often involve analysing hand written summaries produced by participants after a task. There are two common techniques for analysing such summaries: a) counting facts and statements and b) judging topic coverage. Both of these techniques, however, can be easily confounded by simple variables such as summary length. This article presents a new technique that measures *depth* of learning within written summaries based on Bloom's taxonomy. This technique was generated using Grounded Theory and is designed to be less susceptible to such confounding variables. Together, these three categories of measure were compared by applying them to a large collection of written summaries produced in a task-based study, and our results provided insights into each of their strengths and weaknesses. Both fact-to-statement ratio and our own measure of depth of learning were effective while being less affected by confounding variables. Recommendations and clear areas of future work are provided to help continued research into supporting sensemaking and learning.

1 Introduction

Although there is a continuing research effort towards designing systems that support sensemaking, exploratory search, and learning, it is extremely hard to measure a user's progress in these areas, or to know that one participant has learned 'better' than another (Marchionini & White, 2009). Finding better measures and methods to evaluate exploratory search and learning is often cited as a major open research challenge (Kelly, Dumais, & Pedersen, 2009; Marchionini & White, 2009). Measuring learning, however, is difficult

because: a) the process is primarily internal, b) it can be affected by high variance between individuals, and c) there is divergent opinions over what counts as learning (Anderson, 2000). When trying to measure learning from a fixed source of information, it is common to use a quiz (Chi, De Leeuw, Chiu, & Lavancher, 1994; Kim, Turner, & Pérez-Quiñones, 2009). When trying to support open-ended learning, where the source is not fixed, researchers typically ask participants to demonstrate what they have learned by producing a written[1] summary (Kammerer, Nairn, Pirolli, & Chi, 2009; Nelson et al., 2009). In this article we focus on the latter of these two approaches: techniques for analysing participant written post-task summaries. Summaries can be especially hard to analyse, as participants may write about a broad range of topics, depending on what they have learned. Summaries are further affected by high diversity in prior knowledge (Belkin, 1980), and the often-unconstrained discovery of new information from large digital libraries or resources like the Web.

Two common approaches for measuring learning within written summaries are: a) fact and statement counting, and b) judging breadth and depth of sub-topic coverage. Both measures, however, can be easily confounded by length of written summary; some participants may choose to write many arbitrary facts, while others choose to write only what they believe is pertinent. Measuring topic coverage also usually requires judges to be knowledgeable in the subject, in order to judge elements like breadth or depth of coverage. Further, neither of these approaches measures how well the participants understand what they have written. Consequently, we developed a third measure, described below, that focuses on *depth* of learning based on the levels of learning in Bloom's Taxonomy (Anderson et al., 2000). This third technique was developed using a Grounded Theory approach, based upon a set of written summaries collected in an initial user study described below, and was designed to be both independent of topic and less susceptible to the size of the summary. We then compared all three categories of measure in a larger task-based user study, in order to provide insight into each of their strengths and weaknesses.

[1] We include typed summaries in this general notion of 'written'

Below, we begin by describing previous work related to learning, exploratory forms of searching, and example assessments of learning, before describing our own technique and the Grounded Theory approach used to generate it. After introducing our technique, we describe our main study method, which was used to produce a corpus of participant-generated summaries for the comparative analysis. We then describe the results of all three analytical techniques before discussing the strengths and weaknesses of each approach for different study scenarios. We go on to describe how we intend to improve upon our technique to analysis in the future, before ending with our conclusions.

2 Related work

Learning appears in many domains of research. In much of the Information Science literature, learning is perhaps simplistically related to notions of sensemaking (Dervin, 1992), exploring unfamiliar information spaces (White & Roth, 2009), and decision-making (Klein, Orasanu, Calderwood, & Zsambok, 1993). In psychology, many have studied the process of learning, including seminal papers by Piaget and Vygotsky who each tried to model childhood development (Piaget, 1952; Vygotsky, 1962). Further, the psychology discipline studies how information is stored in short term and long-term memory, and conditions that impede learning (e.g. Cognitive Load Theory (Paas, Renkl, & Sweller, 2003)). Perhaps the domain that is most focused on learning is Education, where people study approaches to teaching, how to encourage active learning, and supporting learners in proactively learning for themselves. It is beyond the scope of this article to cover each domain extensively and so we recommend readers refer to survey papers and much larger texts that focus on these issues (Anderson, 2000; Plass, Moreno, & Brünken, 2010; White & Roth, 2009). The sub-sections below instead try to acknowledge each of these different domains, and provide some key literature from them that relates to the focus of this paper: the information seeking activities performed using the Web as a primary source for learning.

2.1 Human learning

There are many aspects relating to human learning, and this section aims to provide an initial introduction to some of the key sections, including: the structure of memory, the processing capacity of memory, and levels of learning.

2.1.1 The structure of memory

The human memory system is largely considered to have three components: sensory memory, short-term memory, and long-term memory (Atkinson & Shiffrin, 1968). Sensory memory is considered to last less than one second, and primarily works as a buffer of information entering the brain through sensory organs like the eyes. Attention, which can be focused or divided, is the mechanism that causes incoming information to then be processed by short-term memory (Anderson, 2000).

Traditionally, Short-Term Memory (STM), or Working Memory, is generally considered to consist of two main processes and controlled by a third (Baddeley & Hitch, 1974). The visuo-spatial sketchpad is the common term applied to the part of STM that processes visual information, such as shapes and space. The phonological loop is the common term applied to the other part that processes language and speech. The Central Executive controls the attention applied to these two parts and the exchange between them. Researchers believe that STM can handle approximately 7±2 pieces of information at any one time (Miller, 1956), sometimes depending on amount of effort applied. Information, however, can be 'chunked' into fewer individual pieces, such as when remembering phone numbers in sections. Although initial work presumed the Central Executive managed both the processing and storage capacity of STM, research into STM loss indicated that storage capacity was separate (Baddeley & Wilson, 2002). Baddeley (2000) later suggested that a fourth entity, named the Episodic Buffer, must exist in STM, which had the capacity to combine new information from the senses with constructs from Long-Term Memory (LTM). These four elements, however, are abstract representations suggested, with evidence, to explain how STM works. Baddeley (2002) provides a detailed review of work in this area, including

studies that have challenged these four elements, or how they each work independently and together.

LTM is considered to be made up of networks of information, where new information, or the powers of human deduction, can be used to build new connections between information that has already been learned. Early work on memory by Piaget (1952) describes LTM as a series of frames for information, which is a notion that comes up again in literature below. These frames, or instances of them, can be retrieved from LTM and are related to the constructs that Baddeley (2000) suggested were stored in the Episodic Buffer. There is clearly much more research in each of these areas, as many psychologists work to understand the nature of memory, but this section has provided a basic introduction to the concepts.

2.1.2 The processing capacity of memory

Cognitive Load Theory (CLT) (Chandler & Sweller, 1991) describes how the processing capacity of STM (perhaps simply the Central Executive) is limited, and how its load can be reduced such that we have more processing power to 'learn', or to commit information to long-term memory. CLT describes learning as being impeded by the mental workload of the learner, which is made up of 3 parts: Intrinsic load, which is the complexity of what we are trying to learn; Extraneous load, which is the complexity of the device we are learning from; and Germane load which is required to combine new information with existing information (perhaps within the Episodic Buffer) and commit it to LTM. Consequently, users cannot learn new information if they are overloaded by intrinsic and extraneous load. Tools to support learning, therefore, often focus on helping users to break down their problem (reducing intrinsic load) or by making tools simpler to interact with during learning (reducing extraneous load) (Mayer & Moreno, 2003).

2.2 Learning

2.2.1 Levels of learning

There are many approaches to teaching and learning, such as active learning (Bonwell & Eison, 1991), behavioural learning (Skinner, 1974), and discovery learning

(Bruner, 1961). These theories each suggest that there is more to learning than just committing information to LTM to simply recall it later. One popular theory related to our work is Bloom's taxonomy of learning (Bloom & Engelhart, 1956), which describes the different levels or stages of learning, and classifies six levels of complexity in cognitive thinking. The taxonomy is structured in a hierarchical nature so that any person functioning on a higher level will have also mastered the levels below. A revision was later made to the taxonomy by Anderson[2] and his colleagues, to update and add relevance for modern teachers (Anderson, et al., 2000). The revision included minor structural changes and a change in terminology in order to make the levels more distinguishable and less confusing. Below, and for the rest of the document, we refer to the revised form of the taxonomy shown in Figure 1.

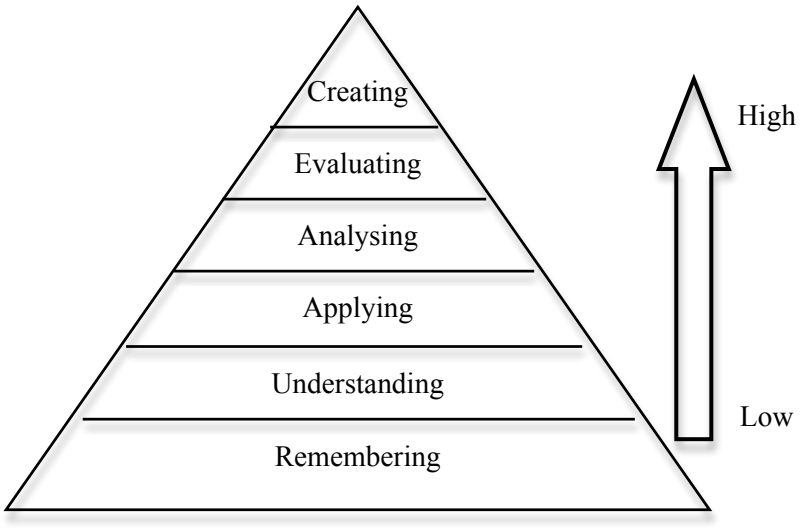

Figure 1: Anderson and Krathwohl's revision of Bloom's Taxonomy of Learning.

While the six levels represent growing complexity, they are split in to two groups, the lower and higher levels of thinking, with three levels in each. The lowest and most basic level of the taxonomy is "remembering", which is simply the recall of relevant knowledge from LTM. "Understanding" is when the learner constructs meaning through several methods of interpretation including summarising, comparing and explaining. The third level, and the highest of the lower levels of thinking, is "applying" and involves the learner carrying out a

[2] Anderson was a former student of Bloom's.

procedure through execution or implementing. The higher levels of learning begin with "analysing", where organising and differentiating allows the learner to break apart information and determine how the parts related to one another as well as the overall structure. "Evaluating" follows this, which includes making judgements based on checking and critiquing. Finally, the highest level of learning is "creating" and this allows the learner to put together and reorganise elements of information to create a coherent whole. In the original taxonomy "evaluating" was considered the highest level, preceded by "creating", but these were reordered in Anderson's revision to create a more logical process where one cannot effectively create without first evaluating the subject.

2.2.2 Information Seeking and Learning

The concept of learning is situated around the acquisition of knowledge, but this can be gathered in a variety of ways. Libraries and Information Science have long looked at ways of storing knowledge, organising it, and helping people to find, use, and learn from it. One definition of Information Seeking is simply to resolve an information need (Marchionini, 1995), although this alone does not necessarily specify learning, as some other theories have. Belkin (1980) described search as being motivated by an Anomolous State of Knowledge (ASK) and explained how this lack of knowledge, depending on its size, may mean that searchers also cannot easily define or describe what they need to know, possibly only recognising it when their ASK is resolved. Shortly after, Dervin (1983) described the Sensemaking problem in terms of users closing a gap (in their knowledge) by trying to build a bridge. Finding or building a bridge involves probing and testing information as possible sources of solutions. Both Belkin and Dervin's work in sensemaking identify the process as containing a need for information (gap) and a method of finding information that attempts to fill this need. Russell et al. (1993) later defined sensemaking as "the process of encoding retrieved information to answer task-specific questions" and went on to define what they call a "learning loop complex". The learning loop complex is a multi-stage process in which the user can gather found information and encode it to a representation that is applicable for the given situation. This notion is similar to that of frames originally described by Piaget.

Several researchers have looked at the process of information seeking, which typically focus on the actions people take (Ellis, 1989), the stages they go through (Marchionini, 1995), and how they feel at different stages (Kuhlthau, 1991). These processes again do not necessarily involve or describe learning, although Kuhlthau studied students preparing for essays; looking at how they made sense of different topics, chose sources of information, and constructed their reports. Belkin, Marchetti and Cool (1993) described 16 different information seeking strategies, using four binary dimensions of searchers' situations. One dimension explicitly highlights searchers who are interested in learning rather than simply finding. Using their other dimensions, more exploratory searchers also do not know if information exists that will solve their problems, and perhaps cannot describe what they need.

More recently, researchers have aimed to define this Exploratory Search to include situations when searchers a) do not know what exactly they are looking for, b) do not know how to describe what they need, or c) do not know much about the systems or domain of information they are in (White, Kules, Drucker, & schraefel, 2006; White & Roth, 2009). In each of these cases, they will need to first learn or make sense of something before they can embark on simply searching for information. While research continues to focus on providing support for sensemaking and learning, research is also trying to find effective ways to evaluate new approaches and measure success in these activities (Kelly, et al., 2009).

2.3 Evaluating Learning

Brookes (1980) detailed what he called his 'fundamental equation' in which learning, or a modified knowledge structure, is described as being comprised of both the learner's initial knowledge structure plus the addition of information. As such, when measuring learning, tasks have to be carefully designed such that they encourage searchers to learn about new things. Approaches typically have to take pre- and post-measures of knowledge, so that the amount learned can be estimated from an individual baseline. Further, White and Iivonen (2001) also looked at whether the form of question given as a task affects a participant's subsequent approach to information seeking. They concentrated on the openness and

predictability of the question to see if this influences a user's selection between using a search engine, directory or direct access to a page. They found that users tended to use source-related reasoning for predictable and search-strategy-related for unpredictable questions. White and Iivonen concluded that failing to consider the characteristics of questions provided to the participant could confound the learning process.

As discussed above, the process of learning is not straight-forward, and so it is often in the interest of the experimenter to control one or more variables. There are often four key variables that can be controlled when evaluating sensemaking and learning: 1) the time given to learning, 2) the size and form of output used to measure learning, 3) the time given to producing this output, and 4) the information source used to learn. Many of the studies described below use experiments that last less than one hour, while some use longer learning tasks such as several hours (Nelson, et al., 2009) or even weeks (Kuhlthau, 1991). Further, (Kalnikaité & Whittaker, 2008) performed a study where memory was tested on the same day, a week later and a month after the initial learning. Some studies below involve quick questions or brief write-ups while others are based on the quality of large essays or presentations. Most studies provide a time limit for producing these outputs such as a short time to answer a questionnaire or a given period to write a summary. Finally, some studies choose to constrain the source of information used for learning so that they can easily identify what has or hasn't been learned with a quiz, while others use essays and written summaries to evaluate individual open-ended learning from unconstrained sources.

2.3.1 Quiz Approaches

One method that looks at information gain is a quiz-based evaluation in which the participants are asked pre-defined questions about a given subject. Kim, Turner and Pérez-Quiñones (2009) looked at note taking systems, comparing paper and electronic forms of note taking. Participants took notes during a presentation, using one of the two methods, and were subsequently given a quiz of six short questions, where they were allowed to consult their notes. In order to know whether quiz answers had been learned or inferred from the fixed

information source, Chi et al. (1994) created a study where each sentence in the passage was coded to what type of information was contained.

Nelson, et al. (2009) studied the effect of social annotations on learning. Participants were asked to research "enterprise 2.0 mashups" for approximately 2 hours to complete a written report about the subject, which had three sub-questions to address. Despite the task involving writing a summary, the learning that occurred was then measured with 20 true-false questions. Notably, the questions were provided by experts and contained an even distribution of easy and hard questions. The questions were rated as easy or hard by 100 random people using Amazon Mechanical Turk. Similarly, Hornbæk and Frøkjær (2003) asked participants to write an essay as part of the learning task and although they rated each essay out of 5, learning was primarily measured using two quizzes. One was focussed on the content of the information provided in the summary and the other involved six incidental learning questions. Inter-rater reliability was used to classify answers as, for example, "an outstanding and well-substantiated answer" or covering "important aspects of questions".

Using a quiz approach runs the risk of becoming more a measurement of recall rather than learning. Numerous articles have detailed how multiple-choice questions should be written to ensure that learning is measured beyond recalling facts or even simply guessing the answers (Bancroft & Woodford, 2004; Haladyna, Downing, & Rodriguez, 2002). Poorly designed questions, therefore, can exaggerate learning, but well-defined questions can potentially be used if the focus of learning has been controlled.

Each of these studies so far have carefully controlled the focus on learning and used quiz based approaches to measure whether specific content had been learned, rather than to see what content had been learned. All of these approaches, therefore, focused on the amount that had been learned from a fixed set of information with an expected output. When measuring learning in situations where the information source is more open and unpredictable, such as the internet, other studies have asked participants to write reports that can be analysed for learning.

2.3.2 Written Summaries

Using written summaries as a means of gathering information from participants allows experimenters to move away from a situation where the output is fixed to predefined criteria. This approach is currently less widely used, as it is harder to measure learning without controlling the source of information.

One approach to gathering written summaries that are created after open-ended learning is to use real world examples, such as students' class notes (Castelló & Monereo, 2005). Using this approach researchers are able to collect data where the participants have created these summaries for their own use rather than to accommodate what a task asks of them. While this allows greater freedom for the participant, it also makes it difficult to measure the knowledge accurately. This method creates a situation, however, where the author of the summary may not actually understand the topic but, instead, simply copied what they have seen from the lecturer.

Sharma (2011) noted in a study involving information gathering from websites that simply collecting information as it is found focuses more on the exploration of online information rather than the knowledge the participant acquires. The study required participants to copy and paste relevant information into a separate document and Sharma found that users with access to a pre-existing representation of information are less likely to adopt their own scheme, making it difficult to differentiate between what is the author's own knowledge and that of an external source. In these situations the written summaries can perhaps be worse than quizzes, since the participant could simply be *copying* information without ever remembering or recalling anything. Consequently many studies ask participants to write summaries after the learning task has taken place, where learning is then measured by counting the amount of information they contain, or by analysing how well they cover a topic.

One approach to measuring learning in a written summary is through simple fact and statement counting. Like quizzes, however, this form of measurement often relies upon the participant remembering information. While perhaps a naïve measure, it is what Bloom considered to be the lowest level of learning (remembering) and should still be present before

any further learning can take place. This allows experimenters to see learning at its most basic level but doesn't test to see whether the participant has understood anything. Wilson et al. (2008), for example, asked participants to recall facts they had found within a faceted browser in order to show that visual highlights could be used to encourage incidental learning with their results found evidence of incidental learning from faceted metadata.

In a study performed by Kammerer et al. (2009) participants were asked to learn about a given topic and their learning was measured in three ways 1) a page collection task, 2) writing a short summary and 3) formulating keywords. This keyword task simply had the participants list keywords that were considered relevant to the topic in a given time limit. These keywords were treated as a form of fact listing and were then simply counted. The summaries, however, were not analysed through fact counting, but through quality and topic coverage, rated by two judges that were already familiar with the topics provided to the participants. In each case, the summary was judged for the number of reasonable topics it covered and each topic was then rated on its quality. Other research, described above (Hornbæk & Frøkjær, 2003), also rated the quality of summaries or written answers.

2.4 Summary

Each of the methods for evaluating learning, described above, have limitations. Quiz based approaches can exaggerate learning because participants can recognise information; Fact recall approaches, especially within a short timeframe, don't necessarily show learning; and written summaries are often measured for the volume or quality of their content, rather than for how they exhibit learning. Further, simply measuring topic depth and breadth approaches can be heavily affected by length of summary. To overcome these limitations, we decided to build an analysis method that applies elements of Bloom's taxonomy of learning to find indicators of higher learning. Such an approach, described in detail below, allows evaluators to study learning independently of topic and size. In the sections below, we describe how the method was developed and validated with inter-rater reliability tests.

3 Analysing Learning with Bloom's Taxonomy

The methods described above typically either count the amount of facts, or use experts to judge the quality of topic coverage, but are not informed by levels of learning. Designed for teachers to devise curriculums and assessment, Bloom's taxonomy describes the levels of learning that students can achieve. In this work, we aimed to develop and evaluate an alternative measure of learning that was influenced by an understanding of learning itself, using Bloom's taxonomy. Other studies of sensemaking and learning have also made use of Bloom's taxonomy. Jansen et al (2007), for example, asked participants to complete 6 tasks that were designed to match each of the levels in the revised version of Bloom's taxonomy. For each task, participants had to answer questions and verify their answers, with each search session being measured by the number of queries, the duration of the session and the number of topics, described as "the information focus of [the query]". Although Jansen's research used Bloom's taxonomy to design study tasks, we have used it to create a measure that can be used to analyse learning within written summaries.

3.1 Method

To create our new measure, we performed a small user study consisting of 12 participants recruited from a University undergraduate course. We acknowledge that this sample was biased toward male students (10) under the age of 23, but the purpose of this study was simply to generate written summaries and, in doing so, pilot a set of Simulated Work Tasks (Borlund & Ingwersen, 1997) that we could use later. Our main study to compare and evaluate different measures, described later, used a larger and broader sample of participants.

3.1.1 Procedure

Participation took approximately 1 hour, including acquiring informed consent and demographic details. Before beginning learning tasks, participants rated their existing knowledge of six topics, described below, on a 7 point scale from low to high. Since this was pre-task, the rating was subjective, with no knowledge test being performed at this point.

Participants then performed three 15-minute sensemaking tasks, which kept participation to less than one hour, to help avoid fatigue. For each task, participants were given either a high or a low prior-knowledge topic, in an alternating fashion, by selecting topics with the highest or lowest rating provided by the participant at the start of the study. Each 15 minute task began with participants spending up to five minutes writing a short summary about the task topic to act as a measure of existing knowledge. Participants then spent a further five minutes researching the topic online and. after searching, participants used the final five minutes to write a post-task summary, without continued access to online resources. After completing three tasks, participants were given a final survey and a short debriefing interview.

3.1.2 Tasks

Six broad sensemaking tasks were selected to cover three main areas, general life (childproofing a home, buying a dog), buying products (E-book readers – shown in Figure 2, home entertainment systems) and technical concepts (anti-virus software, web applications). Tasks were presented to participants using a Simulated Work Task, with a brief scenario and some prompting sub-topics that would serve to initially guide the participant without requiring specific answers. Participants were asked to write their summary as if it would then be passed on to help someone else research the same topic. This scenario encouraged participants to focus on the information that they considered to be the most important for another person to read, increasing ecological validity in the tasks.

Task 4: *Ebook Readers*

Scenario: *You are buying a gift for a close family member. They love reading books but storage is becoming an issue. You have decided to look into buying them an ebook reader.*

Common questions*:*

- *How do ebook readers work?*
- *How are they different from reading text on a monitor?*

Figure 2: Example task from the study, on eBook Readers

3.1.3 Tools

In the study, participants used a custom-made search engine powered by the Yahoo API. The interface itself resembled standard search engine result pages, but creating a dedicated system provided several advantages: 1) interactions could be logged, 2) we could remove distracting elements such as adverts, 3) we could avoid confounding variables such as search engine profiling and personalisation, and 4) we could keep a stable user interface experience for the duration of the study, where services like Google change frequently.

3.2 Our new measure of depth of learning

Before describing our measure of depth of learning, we first describe how we produced this measure from Bloom's taxonomy and the summaries produced from the initial study described above.

3.2.1 Our approach to developing the new measure

Our initial study produced 72 written summaries, made up of 36 pairs of pre- and post-task summaries; 18 pairs were on topics in which the participants had stated they had high prior knowledge, and 18 with low prior knowledge. Using Bloom's taxonomy as an initial model, we developed our new measure using an *inductive* grounded theory technique (Glaser & Strauss, 1967) to highlight the elements of strong and weak summaries. These elements were classified and grouped, and used to develop a scoring scheme to identify the different levels of learning (Tables 1-3).

We began by selecting a stratified sample from the corpus, in order to make sure each topic was covered. We worked with 18 pairs (1 pre- and 1 post-task) of summaries taken from 6 participants each round, in different combinations. The two authors then individually highlighted and coded sentences and facts, agreeing on terminology and rules that could be used to distinguish between them. Separately, the two authors rated each summary with the developing measures and performed Cohen's Kappa calculations to determine inter-rater reliability (Cohen, 1960).

We went through three major iterations of refining our measurements until we reached 'substantial agreement', according to Landis and Koch (1977), between judges. For final validation of our scores, we used Fleiss' Kappa (Fleiss, 1971) to determine the agreement between the two authors and an independent third judge. Our Fleiss Kappa scores are reported inline below as we describe the scales we produced.

3.2.2 The measures produced by our process

Our first measure for depth of learning was 'D-Qual', shown in Table 1, which focused on the quality of recalled facts by their usefulness and was measured on a four-point scale ranging from irrelevant or useless facts (0 points) to facts that showed a level of technical understanding (3 points). The emphasis of usefulness in this measure meant that it was closer to the "understanding" level of Bloom's revised taxonomy, rather than simply "remembering". It was important to differentiate between the two levels as many poor summaries, as determined by the authors during the coding session, simply listed many redundantly obvious facts ("A labrador is a dog") rather than describing them in sentences and summaries. For D-Qual, the judges achieved a Fleiss kappa of 0.64.

Rating	Description
0	Facts are irrelevant to the subject; Facts hold no useful information or advice.
1	Facts are generalised to the overall subject matter; Facts hold little useful information or advice.
2	Facts fulfil the required information need and are useful.
3	A level of technical detail is given via at least one key term associated with the technology of the subject; Statistics are given.

Table 1: Quality of Facts (D-Qual).

Many of the better summaries interpreted facts into more intelligent statements. To identify this, D-Intrp (Table 2) measured summaries in how they synthesised facts and statements to draw conclusions and deductions (Bloom's "analysing") using a 3-point scale. This ranged from simply listing facts with no further interpretation (0 points) to structured combinations in patterns (2 points). The judges achieved a Fleiss kappa of 0.58 for D-Intrp.

Rating	Description
0	Facts contained within one statement with no association.
1	Association of two useful or detailed facts: ‘A -> B’
2	Association of multiple useful or detailed facts: ‘A+B->C’; ‘A->B->C’; ‘A->B∴C’

Table 2: Interpretation of data into statements (D-Intrp).

D-Crit reflected Bloom’s concept of “evaluating” by identifying statements that compared facts, or used facts to raise questions about other statements. The measurement for D-Crit was either true (1 point) or false (0 points), as shown in Table 3. A Fleiss kappa of 0.74 was achieved.

Rating	Description
0	Facts are listed with no further thought or analysis.
1	Both advantages and disadvantages listed; Comparisons drawn between items; Participant deduced his or her own questions.

Table 3: Use of critique (D-Crit).

We did not produce a scale for level three of Anderson’s revised version of Bloom’s taxonomy, “applying”, since the act of writing a summary would not involve the participant to carry out a procedure that has been learned. This level of learning was thus not identifiable in our corpus of summaries. Similarly, the highest level, “creating”, also goes beyond writing about a topic, to more practical elements of learning and so was also left out.

4 Evaluation and Comparison of Measures

Having developed our new measures from our initial sample set of written summaries, we performed a larger user study using a similar protocol. Our new measure was compared with the two other common analytical measures of written summaries: fact counting and topic analysis. We used the same study protocol that was pilot tested in our initial study, refining the Work Task descriptions and procedure slightly. One clear example of the improvements, beyond the wording of tasks, was to change the medium of written

summaries. Summaries in the initial study were hand-written on paper, which led to summaries that were often illegible and hard to analyse. In the main study participants typed their summaries into word processing software, not only making it easier to read but also to clearly determine punctuation for the analysis of statements and spotting listing behaviours. Participants had no access to online resources while writing summaries.

4.1 Method

As before, the study was designed to create open-ended participant-generated summaries, whilst accounting for variations in participants' existing knowledge on different topics. Participants took part for approximately 1 hour, which included performing 3 sensemaking tasks. First, participants were asked to rate their existing knowledge for the six task subjects out of 7. In an alternating pattern, participants were then assigned either a high or low prior knowledge topic (e.g. High-Low-High or Low-High-Low). Participants received a £10 Amazon voucher in appreciation of their time and contribution.

36 participants (17 male and 19 female) were recruited from across a British university, using bulk email to staff and students. Participants ranged in age from 18 to 50. Participants also covered a wide range of job descriptions and across 19 academic departments: 15 were undergraduate students, and the remainder were a mix of administrative support, technicians, and research positions. All but 1 reported their search engine use as daily; 33 reported Google as their primary search engine, where 2 did not answer and one answered with 'Everyclick'.

4.2 Measures

Our evaluation compared three main categories of measures of written summaries: Depth of learning (our new set of scales), Fact and Statement counting (similar to Wilson et al. (2008)), and Topic analysis (similar to Kammerer et al. (2009)). These measures, and their variations, are summarised in Table 4. To measure Depth of learning, we used D-Qual, D-Intrp and D-Crit from our own measure described above. As fact counting measurements, we counted individual facts (assigned 'F-Fact') and individual statements ('F-State') within each summary. Statements typically represented the sentences participants wrote in their summary,

while facts were defined as individual pieces of information either explicitly listed or contained within statements. Finally, using these two sub-measures we also created 'F-Ratio' which represented the ratio of facts per statement.

To measure breadth and depth of topics, we first outlined some common topics that were found in the six tasks of the pilot study (i.e. for buying a dog the topics were history of the breed, health concerns, caring for the dog and personality). Then, to measure breadth ('T-Count'), we counted the number of topics that the participant covered in their summary. To measure depth ('T-Depth'), each topic was measured on a 4-point scale ranging from not covered (0 points) to detailed focused coverage (3 points) and averaged.

As the process of learning is primarily internal it is difficult to measure it objectively. For this reason our measures of learning focused on the difference between pre- and post-task knowledge held by the participant.

Code	Measurement	Scale
D-Qual	Recall of facts	0 – 3 points
D-Intrp	Interpretation of data into statements	0 – 2 points
D-Crit	Critique	0 – 1 point
F-Fact	Number of facts	Count
F-State	Number of statements	Count
F-Ratio	Ratio of facts per statement	Average
T-Count	Number of topics covered (breadth of knowledge)	Count
T-Depth	Level of topic focus (depth of knowledge)	0 – 3 points, averaged

Table 4: Outline of coding scheme used for analysis.

5 Results

Before beginning, the data from two participants were removed from the analysis. A first-pass sanity check over the collected summaries revealed that they had misunderstood the tasks set. One chose to describe their own feelings and history relating to the task topic, rather than trying to answer the task. Another described what they intended to search for in their pre-task summaries, meaning that they could not be compared to other pre-task summaries or measure their information gain. The analyses below relate to the remaining 34 participants. With each participant creating 3 pairs of summaries (pre- and post-task), a total of 204 summaries, or 102 pairs of pre- and post-task summaries, were analysed using all the

measures described above. This set included 51 pairs of high prior knowledge summaries, and 51 pairs of low prior knowledge summaries.

5.1 Searching Behaviour and Task Distribution

In regards to interaction, we found that on average slightly more queries were issued in the high knowledge tasks than the low knowledge tasks, with marginal significance (High: 3.27, Low: 2.67, $t(100) = 1.51$, $p = 0.07$), with participants also submitting significantly more words per query (High: 3.64, Low: 3.14, $t(100) = 1.88$, $p = 0.03$). When looking at how many pages were visited directly from the search engine result page on average (we did not log external navigation) we found that there was no difference between high and low knowledge (3.61 and 3.67, respectively, $t(100) = -0.17$, $p = 0.43$). Participants rarely viewed more than one page of results for any query, from informally observing browsing behaviours, it appeared that much of the sensemaking and learning occurred outside of the search engine while examining specific results. Table 5 shows the distribution of tasks that were performed in each of the high and low knowledge conditions.

Topic	High	Low	Total
Childproofing the home	13	14	27
Buying a dog	7	17	24
Ebook readers	8	9	17
Home entertainment systems	12	4	16
Anti-virus software	8	5	13
Web applications	3	2	5

Table 5: Distribution of topic choice by prior knowledge.

5.2 Comparing the different measures

We analysed the three different categories of measure, and their component parts and variations, using two primary independent variables (IVs). The first IV was pre-task versus post-task summaries, which allowed us to see how well they measured learning. The second IV compared high knowledge summaries with low knowledge summaries, both pre- and post-task, to see if the measures could discern the level of starting knowledge. The measures based on ordinal data (D-Qual, D-Intrp, D-Crit, T-Depth) were tested using Wilcoxon Signed-Rank for the first IV, and Mann-Whitney tests for the second IV. After testing for normality, T-tests

were used on the remaining measures (F-Fact, F-State, F-Ratio, T-Count) in both IV comparisons, paired where appropriate.

5.2.1 How do pre- and post-task summaries compare?

This section compares pre- and post-task summaries using the measurements outlined above. For this IV, we expected to see an increase in every measurement since any amount of research on any topic should hopefully lead to an increase of knowledge.

As expected, we saw significant differences across almost every measure with varying degrees, as shown in the 'All' column of Table 6. The greatest significance was found using our own D-Qual measure that focused on quality of facts recalled ($W(204) = -1172$, $Z = -5.19$, $p < .0001$), F-Fact (number of facts - $t(101) = -7.18$, $p < .0001$) and T-Depth (topic coverage – $W(204) = -2109$, $Z = -4.39$, $p < .0001$). The score that performed least effectively was D-Crit (use of critique – $W(204) = -222$, $Z = -1.74$, $p = 0.04$) but it is possible that given the short amount of time the participants had to both research and write the summaries they simply did not reach this higher level of learning.

	All	High prior knowledge	Low prior Knowledge
D-Qual	W(204) = -1172, p < .0001 *	W(102) = -191, p = 0.002*	W(102) = -423, p < .0001 *
D-Intrp	W(204) = -478, p = 0.003 *	W(102) = -125, p = 0.015 *	W(102) = -115, p = 0.04 *
D-Crit	W(204) = -222, p = 0.04 *	W(102) = -70, p = 0.08	W(102) = -45, p = 0.15
F-Fact	t(101) = -7.18, p < .0001 *	t(50) = -4.23, p < .0001 *	t(50) = -6.18, p < .0001 *
F-State	t(101) = -3.43, p = 0.0004 *	t(50) = -1.81, p = 0.04 *	t(50) = -3.41, p = 0.0006 *
F-Ratio	t(101) = -3.11, p = 0.001 *	t(50) = -2.03, p = 0.02 *	t(50) = -2.35, p = 0.01 *
T-Count	t(101) = -3.03, p = 0.002 *	t(50) = -1.35, p = 0.09	t(50) = -2.98, p = 0.002 *
T-Depth	W(204) = -2109, p < .0001 *	W(102) = -500, p = 0.002 *	W(102) = -564, p = 0.0003 *

Table 6: Comparing pre- and post-task summaries. * Indicates significant results.

We then looked at how the measures varied depending on the participants' prior knowledge. Fewer significant differences were found across the measures when participants had high prior knowledge of the task subject, occurring in six of the eight measures of the 'High' column of

	All	High prior knowledge	Low prior Knowledge
D-Qual	W(204) = -1172, p < .0001 *	W(102) = -191, p = 0.002*	W(102) = -423, p < .0001 *
D-Intrp	W(204) = -478, p = 0.003 *	W(102) = -125, p = 0.015 *	W(102) = -115, p = 0.04 *
D-Crit	W(204) = -222, p = 0.04 *	W(102) = -70, p = 0.08	W(102) = -45, p = 0.15
F-Fact	t(101) = -7.18, p < .0001 *	t(50) = -4.23, p < .0001 *	t(50) = -6.18, p < .0001 *
F-State	t(101) = -3.43, p = 0.0004 *	t(50) = -1.81, p = 0.04 *	t(50) = -3.41, p = 0.0006 *

F-Ratio	t(101) = -3.11, p = 0.001 *	t(50) = -2.03, p = 0.02 *	t(50) = -2.35, p = 0.01 *
T-Count	t(101) = -3.03, p = 0.002 *	t(50) = -1.35, p = 0.09	t(50) = -2.98, p = 0.002 *
T-Depth	W(204) = -2109, p < .0001 *	W(102) = -500, p = 0.002 *	W(102) = -564, p = 0.0003 *

Table 6. This could also have been expected, to some extent, as it is harder for participants to gain knowledge on a topic they are already familiar with from just five minutes of research. Consequently, analysing high knowledge conditions provides insight into the sensitivity of measures. D-Crit (W(102) = -70, Z = -1.4, p = 0.08) and T-Count (t(50) = -1.35, p = 0.09) only showed marginal support for recognising learning. T-Count may have been less sensitive, as High Knowledge pre-task summaries may have already had good topic coverage. The most sensitive measures were F-Fact (t(50) = -4.23, p < .0001), D-Qual (W(102) = -191, Z = -2.9, p = 0.002), T-Depth (W(102) = -500, Z = -2.82, p = 0.002) and D-Intrp (W(102) = -125, Z = -2.16, p = 0.015).

Focusing on participants with low prior knowledge, D-Crit (use of critique) was less sensitive than in High Knowledge conditions (W(102) = -45, Z = -1.05, p = 0.15), again perhaps because participants were not able to reach this level of learning within the 5 minute task. In low knowledge tasks, D-Qual (W(102) = -423, p < .0001), F-Fact (t(50) = -6.18, p < .0001), and T-Depth (W(102) = -564, p = 0.0003) were the clearest indicators of learning.

5.2.2 Can these measures distinguish between high and low prior knowledge?

Next we looked at the independent variable of prior knowledge to see if it is possible to identify whether a participant began with high or low prior knowledge. To examine this, we compared all of the high knowledge summaries against the low knowledge summaries, as well as studying the pre-task and post-task sets separately. We expected that the measures would be able to tell the difference between participants who began with high prior knowledge from those with low prior knowledge of the task.

Taking both the pre- and post-task summaries together, shown in the 'All' column of Table 7, we found that just two of the eight measures showed any significant difference: D-Qual (quality of facts – U(204) = 4290.5, Z = 2.16, p = 0.015) and T-Count (number of topics - t(202) = 1.88, p = 0.03). F-Fact (number of facts - t(202) = 1.41, p = 0.08) was marginally significant. This indicates that overall, higher knowledge participants used higher quality

facts and covered more topics in their summaries, and that simply counting facts can be used[3]. Against our expectations, however, most measures could not differentiate between high and low prior knowledge when all summaries (pre- and post-task) were grouped together. Below, we move on to studying the pre-task and post-task sets separately. We expected that it might be easier to differentiate between high and low prior knowledge before learning had begun.

	All	Pre-task	Post-task
D-Qual	U(204) = 4290.5, p = 0.015 *	U(102) = 888.5, p = 0.003 *	U(102) = 1219.5, p = 0.29
D-Intrp	U(204) = 5184, p = 0.5	U(102) = 1338.5, p = 0.40	U(102) = 1250.5, p = 0.37
D-Crit	U(204) = 5406, p = 0.32	U(102) = 1377, p = 0.31	U(102) = 1326, p = 0.43
F-Fact	t(202) = 1.41, p = 0.08	t(100) = 1.49, p = 0.07	t(100) = 0.71, p = 0.24
F-State	t(202) = 1.08, p = 0.14	t(100) = 1.06, p = 0.15	t(100) = 0.53, p = 0.3
F-Ratio	t(202) = 0.62, p = 0.27	t(100) = 0.52, p = 0.3	t(100) = 0.35, p = 0.36
T-Count	t(202) = 1.88, p = 0.03 *	t(100) = 1.83, p = 0.04 *	t(100) = 0.79, p = 0.22
T-Depth	U(204) = 5143.5, p = 0.44	U(102) = 1248.5, p = 0.37	U(102) = 1316, p = 0.46

Table 7: Comparing high and low prior knowledge. * Indicates significant results.

Against expectations for the pre-task summaries, we again saw that just D-Qual ($U(102) = 888.5$, $Z = 2.75$, $p = 0.003$) and T-Count ($t(100) = 1.83$, $p = 0.04$) showed any significant difference. D-Qual became much more accurate, although T-Count was slightly less able to determine difference. F-Fact became slightly less marginal, focusing on the pre-task summaries only. We saw no significant differences, however, in any of the measures when trying to differentiate between high and low prior knowledge in the post-task summaries. This indicates that all participants reached a similar level of learning regardless of their starting position.

5.3 Effect of summary length in the accuracy of analysis

Previous studies have shown that different measurements of learning can be heavily affected by the size of summary. Jing, et al. (1998) argues that shorter and longer summaries cannot fairly be judged against each other. When comparing summaries of varying lengths (10% and 20% of the length of the article being summarised) they found that the level of agreement between the raters fell in the longer summaries. Jing et al. suggested that people

[3] Although we examine the impact of summary length further below.

are consistent with what is considered to be the most important, but less consistent with the less important aspects.

To investigate length of summary as a factor in the different measurements in our study, we took all pre- and post-task summaries (102 in each) and sorted them by word count. Then, from all the pre-task summaries, the shortest 34 and longest 34 were used to create two sets of short and long summaries respectively. This selection process was repeated for the post-task summaries. Short summaries contained an average of 89 words, while the long summaries had an average of 167. Because these pre-task and post-task sets were no longer correlated by participant, independent statistical measures were used for this analysis. The balance of high and low prior knowledge was also approximately equal, with 67 high and 69 low knowledge summaries used.

5.3.1 Are the measures affected by summary length?

We expected that simple counting measures would be affected by length, but that measures of quality would not necessarily show a difference between long and short summaries. As expected, the fact and statement counting measures, F-Fact and F-State, were highly affected by length of summaries for both pre-task and post-task conditions, as shown in Table 8 and Table 9. Consequently, participants who simply write more will be given higher scores throughout. Perhaps interestingly, the ratio of facts (F-Ratio) was not significantly affected by length in post-task summaries ($t(66) = 1.2$, $p = 0.12$), indicating that longer post-task summaries did not necessarily have more facts per sentence. The reason that this may be interesting is because F-Ratio was good at differentiating between the high and low knowledge and between pre- and post-task summaries.

	All	High prior knowledge	Low prior knowledge
D-Qual	U(68) = 671.5, p = 0.13	U(30) = 147.5, p = 0.08	U(38) = 187.5, p = 0.42
D-Intrp	U(68) = 668.5, p = 0.14	U(30) = 123.5, p = 0.33	U(38) = 218.5, p = 0.14
D-Crit	U(68) = 612, p = 0.34	U(30) = 97.5, p = 0.27	U(38) = 218.5, p = 0.14
F-Fact	t(66) = -3.2, p = 0.001 *	t(28) = -3.39, p = 0.001 *	t(36) = -1.57, p = 0.06
F-State	t(66) = -8.43, p < .0001 *	t(28) = -5.78, p = < .0001 *	t(36) = -5.98, p < .0001 *
F-Ratio	t(66) = 3.85, p = 0.0001 *	t(28) = 2.49, p = 0.009 *	t(36) = 2.88, p = 0.003 *
T-Count	t(66) = -2.08, p = 0.02 *	t(28) = -2.01, p = 0.03 *	t(36) = -0.95, 0.17
T-Depth	U(68) = 706.5, p = 0.06	U(30) = 148.5, p = 0.07	U(38) = 203.5, p = 0.25

Table 8: Comparing short and long summaries, written before the learning tasks were completed.
*** Indicates significant results and, in this instance, is not desirable.**

	All	High prior knowledge	Low prior knowledge
D-Qual	U(68) = 706.5, p = 0.06	U(37) = 222, p = 0.06	U(31) = 132, p = 0.31
D-Intrp	U(68) = 836.5, p = 0.0008 *	U(37) = 275, p = 0.0007 *	U(31) = 144, p = 0.17
D-Crit	U(68) = 629, p = 0.27	U(37) = 178 , p = 0.41	U(31) = 136.5, p = 0.25
F-Fact	t(66) = -5.47, p < .0001 *	t(35) = -4.29, p < .0001 *	t(29) = -3.23, p = 0.002 *
F-State	t(66) = -6.12, p < .0001 *	t(35) = -4.12, p = 0.0001 *	t(29) = -5.05, p < .0001 *
F-Ratio	t(66) = 1.2, p = 0.12	t(35) = 0.76, p = 0.23	t(29) = 0.93, p = 0.18
T-Count	t(66) = -1.49, p = 0.07	t(35) = -3.01, p = 0.002 *	t(29) = 0.67, p = 0.25
T-Depth	U(68) = 829, p = 0.001 *	U(37) = 252, p = 0.007 *	U(31) = 166.5, p = 0.03 *

Table 9: Comparing short and long post-task summaries.
*** Indicates significant results and, in this instance, is not desirable.**

T-Count and T-Depth were also affected by length of summary, especially before the learning task (Table 8), while breadth of topic coverage (T-Count) was only marginally affected by length in post-task summaries (t(66) = -1.49, p = 0.07). D-Qual and D-Intrp, which were effective at differentiating between pre-task and post-task summaries, were here shown not to be affected by length before the learning task. After the task, however, longer summaries contained marginally more useful facts (D-Qual, U(68) = 706.5, Z = -1.57, p = 0.06) and significantly more interpretation of facts (D-Intrp, U(68) = 836.5, Z = -3.16, p = 0.0008). D-Crit was only effective in high knowledge conditions before, and thus it is harder to draw conclusions about how it is affected by length. Because of its nature (identifying if critique is present), however, it is unlikely that D-Crit is affected by length.

5.3.2 Does length make it easier to differentiate pre- and post-task summaries?

We previously compared all pre- and post-task summaries against each other and found significant differences in every measure except for D-Crit. Here, we wanted to see if the length of summaries made it harder or easier for measures to differentiate between pre-task and post-task summaries. Looking at just the shorter summaries (Table 10) we found that fewer measures (only D-Qual, F-Fact, F-State and T-Count) found a significant difference between summaries written before and after learning. This means that if participants only write short summaries, then only the low-level measures would perhaps recognise an increase in knowledge. Of these four measures, only F-Fact (number of facts) could find a significant difference between short pre- and post-task summaries written with high prior knowledge.

	All	High prior knowledge	Low prior knowledge
D-Qual	U(68) = 765.5, p = 0.01 *	U(32) = 159, p = 0.12	U(36) = 224, p = 0.03 *
D-Intrp	U(68) = 588.5, p = 0.45	U(32) = 117.5, p = 0.36	U(36) = 183, p = 0.25
D-Crit	U(68) = 595, p = 0.42	U(32) = 130, p = 0.47	U(36) = 167.5, p = 0.43
F-Fact	t(66) = -2.9, p = 0.003 *	t(30) = -1.9, p = 0.03 *	t(34) = -2.22, p = 0.02 *
F-State	t(66) = -2.61, p = 0.006 *	t(30) = -1.51, p = 0.07	t(34) = -2.42, p = 0.01 *
F-Ratio	t(66) = -0.34, p = 0.37	t(30) = -0.05, p = 0.48	t(34) = -0.44, p = 0.33
T-Count	t(66) = -2.14, p = 0.02 *	t(30) = -0.76, p = 0.23	t(34) = -2.18, p = 0.02 *
T-Depth	U(68) = 708.5, p = 0.06	U(32) = 168.5, p = 0.06	U(36) = 186.5, p = 0.22

Table 10: Comparing *shorter* pre- and post-task summaries. * Indicates significant results.

	All	High prior knowledge	Low prior knowledge
D-Qual	U(68) = 816.5, p = 0.002 *	U(35) = 192.5, p = 0.08	U(33) = 198, p = 0.009 *
D-Intrp	U(68) = 787, p = 0.005 *	U(35) = 230, p = 0.004 *	U(33) = 153, p = 0.24
D-Crit	U(68) = 612, p = 0.34	U(35) = 180, p = 0.16	U(33) = 129.5, p = 0.46
F-Fact	t(66) = -5.44, p < .0001 *	t(33) = -4.46, p < .0001 *	t(31) = -3.11, p = 0.002 *
F-State	t(66) = -1.46, p = 0.07	t(33) = -1.12, p = 0.14	t(31) = -0.71, p = 0.24
F-Ratio	t(66) = -3.95, p < .0001 *	t(33) = -2.79, p = 0.004 *	t(31) = -2.59, p = 0.007 *
T-Count	t(66) = -1.75, p = 0.04 *	t(33) = -1, p = 0.16	t(31) = -0.93, p = 0.18
T-Depth	U(68) = 866.5, p = 0.0002 *	U(35) = 230.5, p = 0.004 *	U(33) = 198, p = 0.009 *

Table 11: Comparing *longer* pre- and post-task summaries. * Indicates significant results.

Looking at the longer length summaries (Table 11), however, we saw significant difference in every measure except for D-Crit; with F-State having only marginal significance. All measures achieved much better significance scores, and thus were better at differentiating between pre- and post-task summaries when they were longer. These findings indicate that encouraging participants to write longer summaries may be required in order to use measures like D-Qual, D-Intrp, and F-Ratio.

5.3.3 Does length make it easier to differentiate high and low prior knowledge?

Earlier we saw that fewer measures were effective at identifying high knowledge participants, with only D-Qual, T-Count, and F-Fact finding marginal or significant differences. We now compare prior knowledge in the short and long summaries to see whether length influences the effect of these measures, where we might expect to see the longer summaries allowing the participant to display their higher level of knowledge, especially in the pre-research summaries.

Like before, none of the measures were particularly effective at differentiating between prior knowledge with shorter summaries, in neither the pre-task nor post-task conditions. Table 12 shows that only topic depth (T-Depth) was able to identify high

knowledge, especially for pre-task summaries, which can possibly be explained that the participants who wrote shorter summaries based on high prior knowledge are more likely to concentrate on a single topic.

	All	Pre-task	Post-task
D-Qual	U(68) = 537.5, p = 0.32	U(34) = 125, p = 0.28	U(34) = 148, p = 0.46
D-Intrp	U(68) = 642, p = 0.21	U(34) = 145, p = 0.47	U(34) = 174, p = 0.16
D-Crit	U(68) = 570, p = 0.47	U(34) = 140, p = 0.47	U(34) = 144.5, p = 0.49
F-Fact	t(66) = -0.4, p = 0.35	t(32) = -0.75, p = 0.23	t(32) = -0.25, p = 0.4
F-State	t(66) = -0.21, p = 0.42	t(32) = -0.4, p = 0.35	t(32) = -0.17, p = 0.43
F-Ratio	t(66) = 0.2, p = 0.42	t(32) = 0.31, p = 0.38	t(32) = -0.04, p = 0.48
T-Count	t(66) = -0.35, p = 0.36	t(32) = 0.43, p = 0.34	t(32) = -1.01, p = 0.16
T-Depth	U(68) = 721, p = 0.04 *	U(34) = 194.5, p = 0.04 *	U(34) = 168, p = 0.21

Table 12: Comparing high and low prior knowledge in shorter summaries. * Indicates significant results.

	All	Pre-task	Post-task
D-Qual	U(68) = 390, p = 0.01 *	U(34) = 89.5, p = 0.03 *	U(34) = 113.5, p = 0.18
D-Intrp	U(68) = 497.5, p = 0.16	U(34) = 158.5, p = 0.29	U(34) = 95, p = 0.06
D-Crit	U(68) = 693.5, p = 0.08	U(34) = 189, p = 0.05 *	U(34) = 154, p = 0.32
F-Fact	t(66) = 1.62, p = 0.06	t(32) = 0.64, p = 0.26	t(32) = 1, p = 0.16
F-State	t(66) = 1, p = 0.16	t(32) = 0.29, p = 0.39	t(32) = 0.79, p = 0.22
F-Ratio	t(66) = 0.86, p = 0.2	t(32) = 0.31, p = 0.38	t(32) = 0.21, p = 0.42
T-Count	t(66) = 3.44, p = 0.0005 *	t(32) = 1.92, p = 0.03 *	t(32) = 2.82, p = 0.004 *
T-Depth	U(68) = 572, p = 0.48	U(34) = 163, p = 0.25	U(34) = 142, p = 0.48

Table 13: Comparing high and low prior knowledge in longer summaries. * Indicates significant results.

Conversely, however, some measures were able to differentiate between high and low prior knowledge, even after the task, when summaries were longer, as shown in Table 13. Looking at the longer pre-task summaries we find that D-Qual shows signs of significant difference along with critique (D-Crit) and the number of topics covered (T-Count). This indicates that use of critique in pre-task summaries is a strong differentiator, but only in longer examples. Like before, however, D-Crit's significance is lost in the post-task summary, perhaps indicating that all post-task summaries include some level of critique. A more sensitive measure of critique (D-Crit) may be required and studied in future work. Unlike in our initial analysis, however, we find that one measure (T-Count) is able to tell the difference between high and low prior knowledge, in both pre- and post-task summaries, if they are longer. Again, this indicates that designing tasks such that participants write longer summaries may make it easier for measures to measure learning.

6 Discussion

Our results section has presented findings in three major areas. First, we showed that the majority of measures were able to differentiate clearly between summaries that were written before and after learning. Conversely, however, very few of the measures were able to tell the difference between summaries written during tasks that participants self selected as having high or low prior knowledge. We also examined whether length of summary had an effect on these two elements, as well as whether length confounded these measures.

6.1 The eight measurements

The simple measures of counting facts and statements (F-Fact and F-State) and their ratio (F-Ratio) were generally successful in identifying learning, but F-Ratio became less successful when focusing only on shorter summaries. When trying to use these measures to determine prior knowledge, only fact counting (F-Fact) showed any sign of significance and this was lost in the post-task summaries. We found that fact and statement counting were both highly affected by length, while the ratio of facts to statements (F-Ratio) was only affected in the pre-task summaries.

Topic breadth (T-Count) and depth (T-Depth) were also successful for identifying learning in most cases. Participants with high prior knowledge, however, often already had good topic coverage (T-Count) before the learning task. This means that topic coverage might actually be a good measure for determining prior knowledge, rather than asking participants to self-identify as having high or low prior knowledge. Both T-Count and T-Depth, however, were affected by summary length, indicating that these measures could be easily confounded by the requirements of the task. T-Count was more successful at identifying high prior knowledge when summaries were longer, while T-Depth was better at determining high prior knowledge in shorter summaries.

Our own measure, based upon the revised version of Bloom's taxonomy of learning, had mixed results. D-Qual (quality of facts) and D-Intrp (interpretation of facts) were able to identify learning after tasks, but when focusing on the shorter summaries only D-Qual held

any success and this was lost when participants had high prior knowledge. Focusing on the longer summaries, D-Qual remained successful while D-Intrp did not find any sign of significance when participants had low prior knowledge. Only D-Qual was successful at identifying prior knowledge, but this was lost in post-task summaries. D-Crit (use of critique) was generally ineffective, only being able to identify prior knowledge in longer pre-task summaries, and identifying learning only when high prior knowledge was involved. When looking for learning in the shorter and longer summaries, D-Crit showed no sign of significance. Generally, we found that our measures were less confounded by summary length, especially in pre-task summaries. D-Intrp, however, was heavily affected by summary length but only in post-task summaries.

6.2 Identifying low and high knowledge

The analysis above would lead us to believe that none of the measurements were particularly sensitive to the prior knowledge that a participant holds when creating the summaries. However, this could be an effect of the short period of time in which the summaries were written in our study. Perhaps this time-frame did not allow a participant enough time to get into the detail they were capable of, but a larger 2 hour task (such as those used by Nelson et al. (2009)) would identify more variation.

One possible limitation of the study is that high and low prior knowledge conditions were based on subjective judgements. Without these judgements, we would not have had a baseline for the study, but this self-assessed method might not accurately reflect the actual held knowledge. Here we examine the similarities between self-selection and the scores produced by the measures. As with producing the sets of long and short summaries, we sorted the 102 pre- and 102 post-task summaries by all eight measures, one at a time, and split them into two groups of 34, representing the extreme high and low scores. We expected to see more low knowledge tasks in the bottom 34 summaries for each measure, and more high knowledge tasks in the top 34 summaries. What we actually saw is an almost even mix of high and low prior knowledge in each extreme.

While some studies might use fact counting as a method to show that one person knew more than another, only one measure showed a clear distribution closer to what would be expected and that was D-Qual (measures of fact quality). This distribution held in both the low and high prior knowledge. Only two other measures had similar distributions but both were dependent on prior knowledge and, to an extent, could be expected. In low knowledge, T-Count showed an expected distribution, implying that those participants struggled to cover a breadth of topics. In high knowledge, F-State showed an expected distribution, which implies that participants with high knowledge wrote a higher number of statements.

These measures, however, only held in the pre-task summaries. When looking at the post-task summaries, none of the measures show any obvious similarities to the subjective ratings provided by participants.

6.3 The effect of length

Many of the measures were directly affected by length, especially those that use counting as an approach, such as fact (F-Fact), statement (F-State) and topic counting (T-Count). Some measures, such as the quality of facts (D-Qual) and the ratio of facts to statements (F-Ratio), were less affected by length, while being highly effective at measuring learning according to our primary independent variables. For longer summaries, all measures performed better, but breadth of topic coverage (T-Count) and the quality of facts (D-Qual) were particularly better at identifying high prior knowledge. With short summaries, we saw that just the quality of topics (T-Depth) was accurate at identifying high knowledge conditions.

Length may not be the only factor. In our study, the only limitation on creating the summaries was a 5 minute time limit, which led to some variation between the quality of the output. During the study, some participants wrote a brief summary but required the full five minutes, some requested to stop before the time was up and others wrote consistently until the five minute window was ended. Consequently, some summaries were much better quality,

despite being shorter, while others were poor quality and much longer. There are situations, therefore, where the length of the summaries may require a more thoughtful consideration.

6.4 Recommendations

To identify learning all measures detailed here were generally effective, but both the length of the summaries and the prior knowledge held by the participant should be taken in to consideration. Table 14 provides an overview of the strengths and weaknesses of each measure and recommendations are made below. While serving as a guide readers should refer back to the full text in our results section for more detail before using them in a study.

	Identifies Learning				Identifies Prior Knowledge				Ignores Length	
	High	Low	Short	Long	Pre	Post	Short	Long	Pre	Post
D-Qual	✓	✓	✓	✓	✓			✓	✓	
D-Intrp	✓	✓		✓					✓	
D-Crit	✓							✓	✓	✓
F-Fact	✓	✓	✓	✓	✓			✓		
F-State	✓	✓	✓	✓						
F-Ratio	✓	✓		✓						✓
T-Count	✓	✓	✓	✓	✓			✓		
T-Depth	✓	✓	✓	✓			✓			

Table 14: Overview of measure suitability.

If participants have written shorter summaries (here averaged to around 90 words) then learning is only really noticeable if those participants began with low prior knowledge, where measures such as the quality of facts (D-Qual), simple fact and statement counting (F-Fact, F-State) and topic coverage (T-Count) can be used to determine an increase of knowledge. If short summaries are written based on high prior knowledge then only simple fact and statement counting (F-Fact, F-State) and the depth of topics (T-Depth) reflected an increase.

If participants have written longer summaries (here averaged to around 180 words) measures such as the quality and number of facts (D-Qual and F-Fact, respectively), ratio of facts to statements (F-Ratio) and topic depth (T-Depth) can be used in both high and low prior knowledge situations. Additionally, when the participant has high prior knowledge the interpretation of facts (F-State) can be used.

When attempting to determine prior knowledge we were only able to use topic depth (T-Depth) effectively when looking at shorter summaries. Using longer summaries allows

low-level measures such as quality of facts (D-Qual) and topic coverage (T-Count) to be used. It is important to note that with the exception of T-Count, these measures were only effective in pre-task summaries and so it is recommended that identifying prior knowledge be done before any learning task takes place as none of the measures were sensitive enough to identify high or low prior knowledge after the learning exercise.

6.5 Future Work

While this work has identified many insights that have allowed us to make initial recommendations for measuring learning and sensemaking in open-ended written summaries, there is still much that can be evaluated in future work.

Our analysis found that many measures were affected by length as a confounding variable. Future work could directly control this as an independent variable in order to make specific assertions as to how length can affect measures. Similarly, it would be interesting to directly examine the relationship between time spent and quality of summary, since some participants carefully wrote short summaries within the five minutes, while others felt compelled to keep writing for the whole time, which ultimately led to less focussed summaries.

We would also like to directly examine the ability for these measures to identify prior knowledge, by correlating them directly, and more formally, with subjective ratings provided by participants.

Despite carefully developing our own measures, we found that D-Crit (use of critique) was the least discerning of the measures, but suggest that this is due to the limitations of the summary condition where the imposed time limit did not allow participants to display that level of learning. By extending the length of the summaries we would like to see if this measure gains any advantage, as well as investigating how critique is used in learning in more depth.

7 Conclusions

In this article, we have evaluated three major approaches to analysing learning in written summaries being created in learning-style work tasks: 1) simple fact and statement counting, 2) breadth and depth of topic coverage, and 3) our own measure of depth of learning based upon the revised version of Bloom's learning taxonomy. To compare these three categories of measurement, we used high and low prior knowledge as an independent variable, as well as analysing pre-task and post-task summaries to check that the measures could detect that learning had occurred. We then looked at how the measurements were affected by length of summary and, in the discussion section, the similarities between the measurements and subjective ratings of prior knowledge provided by participants.

Our results found that a one-size-fits-all approach cannot be used to accurately measure the varying degrees of learning, but that various measures worked in different conditions. Many of the variables, especially those that involve counting facts and statements, were heavily confounded by length, where some long summaries were poorer in quality, while other times participants had carefully constructed a short summary. Generally, however, encouraging longer summaries made it easier to measure learning. Some measurements, especially measures of higher-level learning such as the use of critique, were more applicable to high prior knowledge conditions. The quality of facts, however, was less affected by length than the number of facts, while still being effective at measuring learning.

Overall, this work has presented a novel approach to measuring learning, based upon Bloom's established taxonomy. Further, we presented a detailed evaluation comparing several approaches to measuring learning in written summaries, and identifying strengths and weaknesses for each of them. This research has further identified areas of future work, which can be examined to study these findings in more detail. Our findings, and these clear routes for further work, provide a resource for researchers who are focused on better supporting higher-level work tasks, beyond simple information seeking, such as those involving sensemaking and learning.

8 References

Anderson, J. R. (2000). *Learning and memory*. New York: John Wiley & Sons, Inc.

Anderson, L., Krathwohl, D., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P., . . . Wittrock, M. (2000). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, Abridged Version*: Allyn & Bacon.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: a proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation*: New York: Academic Press.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences, 4*(11), 417-423.

Baddeley, A. (2002). Is Working Memory Still Working? [10.1027//1016-9040.7.2.85]. *European Psychologist, 7*(2), 85-97.

Baddeley, A., & Wilson, B. A. (2002). Prose recall and amnesia: implications for the structure of working memory. *Neuropsychologia, 40*(10), 1737-1743. doi: 10.1016/s0028-3932(01)00146-4

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *The psychology of learning and motivation, 8*, 47-89.

Bancroft, P., & Woodford, K. (2004). *Using Multiple Choice Questions Effectively in Information Technology Education*. Paper presented at the Beyond the Comfort Zone. 21st Annual ASCILITE Conference 2004, Perth, WA. http://eprints.qut.edu.au/24033/

Belkin, N. J. (1980). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*(5), 133-143.

Belkin, N. J., Marchetti, P. G., & Cool, C. (1993). Braque: design of an interface to support user interaction in information retrieval. *Information Processing and Management, 29*(3), 325-344.

Bloom, B. S., & Engelhart, M. D. (1956). *Taxonomy of educational objectives : the classification of educational goals. Handbook I, Cognitive domain*. London: Longmans.

Bonwell, C. C., & Eison, J. A. (1991). *Active learning: creating excitement in the classroom*. Washington, DC :: George Washington University, ERIC Clearinghouse on Higher Education.

Borlund, P., & Ingwersen, P. (1997). The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation, 53*(3), 225-250.

Brookes, B. (1980). The foundations of information science, Part I: Philosophical aspects. *Journal of Information Science, 2*(3-4), 125-133. doi: citeulike-article-id:5730714

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review, 31*, 21-32.

Castelló, M., & Monereo, C. (2005). Students' Note-Taking as a Knowledge-Construction Tool. *L1-Educational Studies in Language and Literature, 5*(3), 265-285.

Chandler, P., & Sweller, J. (1991). Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction, 8*(4), 293-332.

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting Self-Explanations Improves Understanding. *Cognitive Science, 18*(3), 439-477. doi: 10.1207/s15516709cog1803_3

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Dervin, B. (1983). *An overview of sense-making research: Concepts, methods, and results to date*. Paper presented at the International Communication Association annual meeting, Dallas, TX, USA.

Dervin, B. (1992). From the mind's eye of the user: The sense-making qualitative-quantitative methodology *Qualitative research in information management* (pp. 61-84): Libraries Unlimited.

Ellis, D. (1989). A behavioural approach to information retrieval system design. *Journal of Documentation, 45*(3), 171-212.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378-382.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hornbæk, K., & Frøkjær, E. (2003). Reading patterns and usability in visualizations of electronic documents. *ACM Transactions on Human-Computer Interaction, 10*(2), 119-149.

Jansen, B. J., Smith, B., & Booth, D. (2007). *Learning as a Paradigm for Understanding Exploratory Search*. Paper presented at the SIGCHI 2007 Exploratory Search and HCI workshop.

Jing, H., Barzilay, R., Mckeown, K., & Elhadad, M. (1998). *Summarization Evaluation Methods: Experiments and Analysis.* Paper presented at the AAAI Intelligent Text Summarization Workshop.

Kalnikaité, V., & Whittaker, S. (2008). *Cueing digital memory: how and why do digital notes help us remember?* Paper presented at the Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1, Liverpool, United Kingdom.

Kammerer, Y., Nairn, R., Pirolli, P., & Chi, E. H. (2009). *Signpost from the masses: learning effects in an exploratory social tag search browser.* Paper presented at the Proceedings of the 27th international conference on Human factors in computing systems (CHI'09), Boston, MA, USA.

Kelly, D., Dumais, S., & Pedersen, J. O. (2009). Evaluation Challenges and Directions for Information-Seeking Support Systems. *IEEE Computer, 42*(3), 60-66.

Kim, K., Turner, S. A., & Pérez-Quiñones, M. A. (2009). Requirements for electronic note taking systems: A field study of note taking in university classrooms. *Education and Information Technologies, 14*(3), 255-283.

Klein, G., Orasanu, J., Calderwood, R., & Zsambok, C. E. (1993). *Decision Making in Action: Models and Methods*: Ablex Publishing Co.

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5), 361-371.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174.

Marchionini, G. (1995). *Information Seeking in Electronic Environments*: Cambridge University Press.

Marchionini, G., & White, R. W. (2009). Information-Seeking Support Systems. *IEEE Computer, 42*(3), 30-32.

Mayer, R. E., & Moreno, R. (2003). Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist, 38*(1), 43-52.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Nelson, L., Held, C., Pirolli, P., Hong, L., Schiano, D., & Chi, E. H. (2009). *With a little help from my friends: examining the impact of social annotations in sensemaking tasks.* Paper presented at the Proceedings of the 27th international conference on Human factors in computing systems (CHI'09), Boston, MA, USA.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist, 38*(1), 1-4.

Piaget, J. (1952). *The origins of inteliigence in children*. New York: International Universities Press.

Plass, J. L., Moreno, R., & Brünken, R. (2010). *Cognitive load theory*. Cambridge, UK: Cambridge University Press.

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). *The cost structure of sensemaking.* Paper presented at the Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems, Amsterdam, The Netherlands.

Sharma, N. (2011). *Role of available and provided resources in sensemaking*. Paper presented at the Proc. CHI 2011.

Skinner, B. F. (1974). *About behaviorism*. New York: Knopf; [distributed by Random House].

Vygotsky, L. (1962). *Thought and language*: Cambridge, MA, US: MIT Press.

White, M. D., & Iivonen, M. (2001). Questions as a factor in Web search strategy. *Information Processing & Management, 37*(5), 721-740. doi: 10.1016/s0306-4573(00)00043-1

White, R. W., Kules, B., Drucker, S. M., & schraefel, m. c. (2006). Introduction. *Communications of the ACM, 49*(4), 36-39.

White, R. W., & Roth, R. (2009). *Exploratory Search: Beyond the Query-Response Paradigm* (Vol. 1): Morgan & Claypool.

Wilson, M. L., André, P., & schraefel, m. c. (2008). *Backward Highlighting: Enhancing Faceted Search.* Paper presented at the UIST08: Proceedings of the 21st annual ACM symposium on User interface software and technology Monterey, CA, USA.