Building an Educational Recommender System based on Conceptual Change Learning Theory to Improve Students' Understanding of Science Concepts

by

Ifeyinwa Uchechukwu Okoye

B.Sc., California State University Hayward, 2006

M.S., University of Colorado at Boulder, 2010

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science
2015

This thesis entitled:
Building an Educational Recommender System based on Conceptual Change Learning Theory to Improve Students' Understanding of Science Concepts
written by Ifeyinwa Uchechukwu Okoye
has been approved for the Department of Computer Science

Tamara Sumner

James Martin

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Okoye, Ifeyinwa Uchechukwu (Ph.D., Computer Science and Cognitive Science)

Building an Educational Recommender System based on Conceptual Change Learning Theory to Improve Students' Understanding of Science Concepts

Thesis directed by Associate Professor Tamara Sumner

Science misconceptions can be deeply held and difficult to change. Conceptual change learning theory (CCLT) applied in the classroom environment has been effective in helping students remedy their persistent science misconceptions. This work studies CCLT in an online learner-driven environment using five related studies.

The first study creates the online learner-driven environment while the second study evaluates the resulting system. The last three studies create computational models that accomplish a teacher's task per CCLT.

The first study uses participatory design methods to create the online learner-driven environment called CLICK2. This is an effective feedback environment because it can answer the questions: *where am I? where am I going? how am I going?* and *where to next?* in relation to a user's knowledge state. Results show that users were satisfied with the interaction design of CLICK2. The second study is a learning study that investigated how CLICK2 influences learners' processes and outcomes. Results show that using CLICK2 improved users' understanding of seasons and their confidence in understanding this concept.

The last three studies use techniques from machine learning and natural language processing to perform three critical tasks underpinning support for CCLT: prioritizing learners' misconceptions, extracting core concepts, i.e., learning goals and sequencing them. All three studies draw on analyses of human expert processes to inform the design and evaluation of the algorithms. Results show that an alignment of sequenced core concepts to misconceptions in a learner's work is a good feature for prioritizing misconceptions in the learner's work; reducing the extraction rate in a multi-document summarizer produces good core concepts; and dynamically generating useful pedagogical sequences from a list of core concepts is feasible.

This work contributes to the scientific literature by introducing a methodology for automatically prioritizing core concepts and student misconceptions in a pedagogically useful manner. Furthermore, this work shows that conceptual change learning theory can be implemented in an online learner-driven environment.

Dedication

This dissertation is lovingly dedicated to my parents Christian Chukwuemeka Okoye and Nneka Lucy Okoye (nee Onweluzo). Without you, there would be no me.

Acknowledgements

Thank you to all my committee members Tamara Sumner, James Martin, Martha Palmer, Steven Bethard, and David Webb for guiding me through this program. I would like to thank Tammy for her exquisite attention to detail and for her demand for excellence. I would also like to thank Jim for all the support, especially during the last two semesters.

Thank you to Holly Devaul for helping me with all my research studies that involved science experts and for helping me prepare for the defense. Thank you to Heather Leary for all the help with preparing for my defense. Thank you to Lee Becker, Arafat Sultan, Haojie Hang, Keith Maull, Rodney Nielsen, and James Foster for collaborating with me on several research projects and papers.

A huge thank you to Philipp Wetzler for the instant and enduring friendship. To Kemi Jolaoso, Jing Zheng, Jinho Choi, Dmitriy Dligach, Chih How Bong, Ovo Dibie, Jena Hwang, Kirill Kireyev, Daryl Lonnon, Praful Mangalath, Manuel Saldivar, Falguni Shah, Yingdan Huang, Jeff Hoehl, Sudha Verma, Shumin Wu, Philip Ogren, Ashwini Vaidya, Sara MacAlpine, Jane Meyers, Holger Dick, Sivashangari Anandakrishna, Ruth Opara and Chizoba Uzoewulu, thank you for your friendship and encouragement.

To Esther Stowell, Peemin Chen, Robert Sajan, Jenny Yang, Eddie Mulyono and especially Hilary Holz, thank you for believing in me and supporting me to and through CU-Boulder. Thank you to the Denver metro Salsa community for making me one of yours and introducing me to a great way to relieve stress. To Megan Bela, thank you for running to catch the last bus back up to Boulder with me after going salsa dancing at The Avalon.

And thank you for encouraging me through this program, for the continued friendship and for the monthly debriefs at The Cup.

Thank you to my late uncle Hilary who introduced me to computers. Very special thanks to my siblings: Chioma, Chukwuemeka and Chika for their infinite patience, unshakable optimism, and unconditional support throughout this program. To my dogs Charlie and Chloe, thank you for giving me a reason to smile everyday.

The work presented in this thesis was supported in part by the National Science Foundation under Grant Numbers 0835393, 0835381, 0835454 and 0537194. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Contents

Chapter

Appendix

Tables

Table

Figures

Figure

Chapter 1

Introduction

1.1 Motivation

Science misconceptions have proven to be persistent, deeply held and difficult to change (Vosniadou, 2008; Duit, 1999). Conceptual change learning theory applied in the classroom environment has proven effective in helping students remedy their persistent misconceptions about science and in developing a more robust understanding (Vosniadou, 2008). The overarching goal of this research was to develop and study a new approach for creating educational recommender systems (ERS) based on conceptual change learning theory. Recommender systems have been shown to be an effective technique for dealing with the information overload problem that confronts learners, as a result of the abundance of information available online (Adomavicius and Tuzhilin, 2005). An educational recommender system, CLICK2, was created as a result of this research. The online system was designed to improve student understanding of science concepts by recommending resources targeting their misconceptions.

This is a dual dissertation in computer science and cognitive science, hence this work makes contributions to both fields. The computational contribution is creating a framework that extends educational recommender systems from making recommendations based on usage to facilitating science understanding. The cognitive, specifically, learning science contribution is in extending conceptual change learning theory into an online environment. For this research, I drew subjects from students at CU-Boulder for my study population, even though my ideal population is middle and high school students.

In the next section, I give an overview of the research questions and discuss the anticipated contributions of this dissertation in more detail.

1.2 Research Overview and Questions

The following scenario will motivate the need for this research. Mandy is a hardworking and conscientious female high school student in Boulder who is enrolled in an Earth science course. Mandy likes to be prepared for class and so, she usually reads a little bit about the topic for each upcoming class. The topic for her next Earth science class is seasons. Mandy assumes she is ready for this class because she has lived in Boulder all her life. She knows about the four seasons and when they occur.

The next day at school during the discussion on seasons, Mandy is surprised to realize that she was not ready for the class. She did not know that not everywhere on Earth has four distinct seasons like Boulder. Also, she did not know that winter in the southern hemisphere is harsher than winter in the northern hemisphere. At the end of the class session, seeing that the students did not fully understand the lesson, the teacher assigns the class an assignment, to research and write an essay report on seasons, specifically that answers the questions, **why do we have day and night?**, **what causes the seasons?** and **why are climates in the Southern hemisphere slightly milder than those at similar latitudes in the Northern Hemisphere?**.

Mandy goes home and starts trying to piece together all the information she needs to write a good essay. She decides to do more research about the seasons, so she goes online to a search engine and types in her search query **seasons on Earth** and gets back more than 94 million results. She starts browsing through the results, 3 hours later, she is still browsing and still does not have a clear idea of how the seasons come to be.

What Mandy needs is a system that can determine her current knowledge state and recommend resources to help her achieve the scientific understanding of the reason for the seasons. In order to create the automatic system that Mandy uses, I identified five research questions that need to be addressed:

(RQ1) What are design options for creating an educational recommender system (ERS) with research-based support mechanisms for promoting conceptual change?

In the ERS literature, the design of the feedback environment has been routinely ignored. In a review of the research on ERS, 19 out of 20 of the recommenders used a standard e-commerce interface. Only 1 in 20 explored interface design to support education (Manouselis et al., 2011) and some ERS researchers clearly state that it is not an important aspect of the research (Buder and Schwind, 2011). Since the goal of educational recommender systems is to facilitate learning and not just recommend sources of information, the recommendation interface should be designed to facilitate learning.

The research question here concerns how we can design a feedback environment, which has embedded within it, research-based support mechanisms that have proven effective for promoting conceptual change in classroom environments.

(RQ2) How does the educational recommender system with its conceptual change support mechanisms affect users' understanding, interest and perception of science content?

The end goal of the system is to help students improve their understanding of science, by correcting their misconceptions. A research study to examine the effect of the system use on learners' processes and outcomes was conducted. The study also explored if and how the conceptual change support mechanisms were used.

(RQ3) How can we model expert strategies for prioritizing student misconceptions?

According to conceptual change learning theory, it is essential that students address misconceptions associated with basic concepts before moving on to other dependent concepts. It is expected that correcting a basic misconception will induce the student to correct other misconceptions that were produced as a result of the basic misconception.

I chose to model the strategies used by expert Earth science teachers. When expert Earth science teachers pose a question in class and get several answers, they have to structure their explanation to address the various misconceptions that the students have. When they review a student's work output such as an essay, they have to give constructive feedback that will help the student produce a better essay. Skilled teachers do not just list all the problems the student has, they point out the major ones knowing that as those are fixed, the minor ones will be fixed too. Hence the goal of the resulting model is to prioritize the students misconceptions by ordering the misconceptions in a way that lets the learner learn more efficiently within a short period of time.

(RQ4) How well can different computational methods identify the learning goals in a collection of documents?

US schools have been criticized for not emphasizing the big ideas in science. Subsequently, there has been a push to target big ideas in depth rather than a long list of ideas in a superficial manner so as to promote robust understanding and produce citizens that are capable of reasoning correctly about scientific phenomena in the time constraints of the classroom (Vosniadou et al., 2001). Hence in order to guide a learner towards an efficient path of understanding, it is imperative that all learning systems work towards significant learning goals and not just assume that the student needs to see and learn every concept in a resource. So the question here is, given a collection of documents (chapters in a textbook or pages from online resources), can we determine a good method for identifying the main ideas or learning goals in this collection?

(RQ5) How well can machine learning classifiers model the pedagogical sequence of learning goals produced by human experts?

Some concepts serve as building blocks for other concepts, and thus it is essential to learn the basic concepts before moving on to other concepts that depend on them, i.e., the order of acquisition of knowledge can be important in promoting conceptual change (Vosniadou et al., 2001). Research has shown that instruction that follows particular learning progressions is important when trying to understand scientific topics (Plummer and Agan, 2010). For example, students must first understand the concept of the revolution of the Earth around the sun and the concept of the tilt of the Earth before they can understand the concept of seasons in the northern and southern hemisphere. There may exist several different but reasonable pedagogical sequences (also known as learning paths). I focus on generating a single sequence to provide students with a meaningful learning path through the resources. The sequence of learning goals that results from this module will serve as input for a later module that prioritizes students' misconceptions.

1.3 Contributions

This research draws on conceptual change theory to extend the state of art in educational recommender systems. In doing so, this research generates contributions to both computer science and learning science. Below, I discuss the contributions of this dissertation to both fields.

1.3.1 Computer Science Contribution

Using conceptual change theory to extend the state of art in educational recommender systems requires significant advances in algorithms that assess learners' information needs.

Conceptual change learning theory operationalizes learners' information need as prioritized misconceptions that depend on an instructionally sound pedagogical sequence of core learning goals. This work advances the state of art of algorithms that assess learners' information needs, by applying machine learning and natural language processing techniques to create new algorithms that automate the instructional process of identifying core learning goals, sequencing the learning goals and prioritizing misconceptions.

Extending the state of art in educational recommender systems also requires advances in how the interface is designed, in order for educational recommender systems to support conceptual change in learners. This work advances the state of art by designing a new interface that supports conceptual change in users of educational recommender systems.

1.3.2 Cognitive Science Contribution

The cognitive science contribution of this work is studying conceptual change learning theory in the context of an online educational recommender system. This work also extends and refines conceptual change learning theory for use in an informal, online, learner-driven learning context. In addition, this work expands the role of educational recommender systems in the learning process from simply making recommendations to also serving as a formative assessment tool.

1.4 Organization

Chapter 2 examines my research context while chapter 3 explores related research in conceptual change learning theory and educational recommender systems. Chapter 4 presents my conceptual framework; chapter 5 examines my research design and gives an overview of the five studies I carried out as part of my dissertation research. Chapter 6 explores the first research question and corresponding study; chapter 7, the second question and study, chapter 8, the third question and study, chapter 9, the fourth question and study and chapter 10, the fifth and final question and study. Chapter 11 is a discussion of my research results, limitations and significance and offers avenues for future work.

Chapter 2

Research Context and Educational Topic

The research context for this work was an educational recommender system, the Customized Learning Service for Concept Knowledge (CLICK). CLICK has been under development for the past 7 years. In that time, it has become a robust computational infrastructure which can serve as a test bed for studying personalized educational recommendations because the core recommender system algorithms have been created. Building on this existing educational recommender system enabled me to focus on algorithms and designs to study conceptual change.

The educational topic I explored during the course of my research is **understanding the Earth system processes that give rise to the seasons**. This topic has been well studied and research has shown that learners of all ages have persistent misconceptions about seasons (Trumper, 2000, 2001a,b,c; Atwood and Atwood, 1996; Sebastià and Torregrosa, 2005). In the 1989 video, *A private universe* (Schneps et al., 1989), graduating Harvard students were asked to explain why we have seasons; many of their answers are classic misconceptions about seasons. Since this topic is relevant and applicable to everyone, I decided it was a good topic to use in my research.

2.1 CLICK version 1

The overarching goal of CLICK (de la Chica et al., 2008b) was to create a scalable online service that recommends resources to users based on the their conceptual understanding.

Currently, CLICK uses the Digital Library for Earth System Education (DLESE) to support learners' understanding of Earth science content. CLICK automatically constructs a domain knowledge base from digital library resources and evaluates users' conceptual understandings against the domain knowledge through automatic essay analysis. CLICK detects flaws and gaps in users' science knowledge of Earth system concepts and recommends digital library resources to address users' misunderstandings and knowledge gaps. Users are encouraged to visit those resources and upon sufficient review, revise and re-write their essays for re-evaluation.

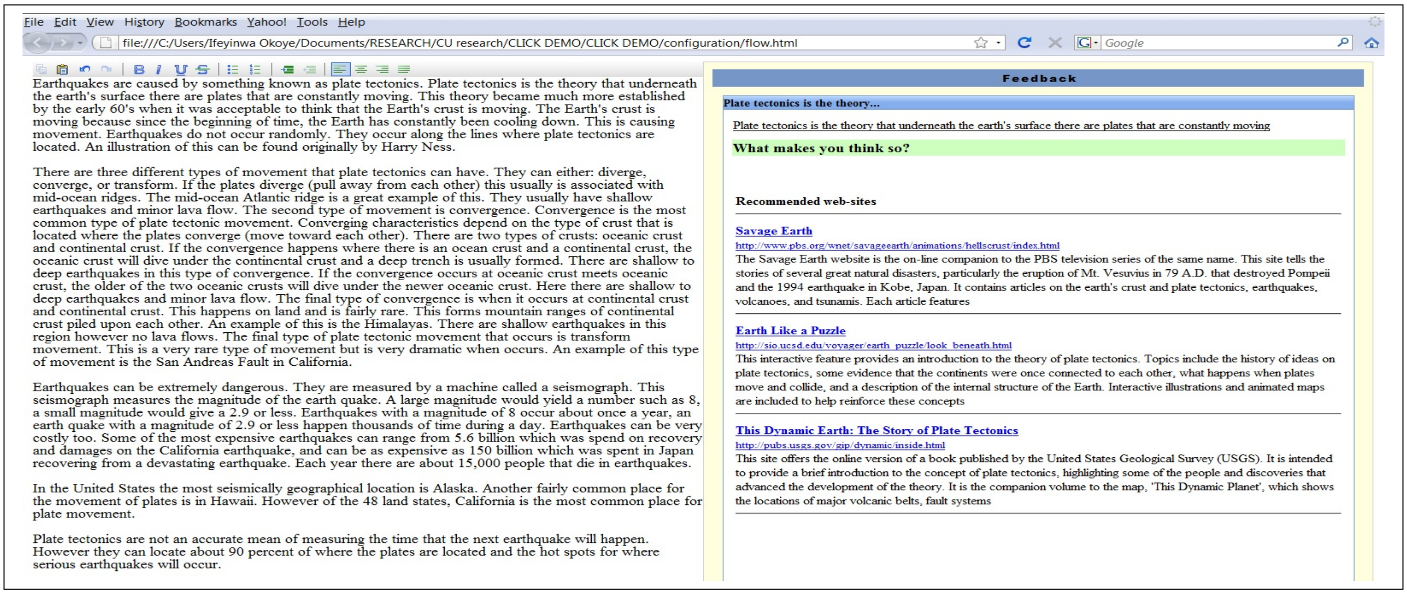
Prior work on CLICK investigated four components of a conceptual educational recommender system: the domain knowledge generator, a misconception identifier, a resource recommender and a preliminary recommender interface.

The domain knowledge generator is COGENT, a multi-document summarizer, optimized for the Earth science domain. Multi-document summarization is a computational technique for analyzing multiple documents and generating a summary of the information contained in the documents. COGENT extends a generalized multi-document summarizer, MEAD (Radev et al., 2004), by adding features such as educational standards, hypertext and content word density to determine which concepts to extract from a collection of resources for use in building a knowledge base (de la Chica, 2009). Although COGENT is capable of identifying all the concepts needed for robust understanding of plate tectonics at the middle and high school levels (de la Chica, 2009), it does not identify the most important learning goals, the concepts that represent big ideas in science.

The initial misconception identifier was produced by graph comparison of a concept map of a student's essay with a domain knowledge concept map. A concept map is a map where the nodes represent concepts and the links show the connections between concepts. The misconception identifier diagnosed three types of misconceptions - incorrect, incomplete (missing) and fragmented (Ahmad, 2009).

The links used in this initial misconception identifier were generated manually i.e., this algorithm relied on links generated by human experts. Since one goal of CLICK is to be a more fully automatic system, this original misconception identifier has been significantly modified over the past 2 years. The current version uses entailment to determine misconceptions in students' work. It does not make use of concept maps. Entailment is a text analytic technique for determining if the information in a text, H, can be inferred from another text, T. In this research, I extend the misconception identifier to automatically prioritize the identified misconceptions. Conceptual change theory highlights the need to focus on the most important concepts and the most important misconceptions first before moving on to other information about a topic.

The resource recommender uses graph analytic techniques to compare the resources and student misconceptions in order to recommend appropriate resources (Gu, 2009). The current recommender system relies solely on knowledge-based recommendations i.e., it only uses information on students' current misconception in its' recommendation.


File
Edit
View
History
Bookmarks
Yahoo!
Tools
Help
file:///C:/Users/Ifeyinwa Okoye/Documents/RESEARCH/CU research/CLICK DEMO/CLICK DEMO/configuration/flow.html
Google
B
I
U
S

Earthquakes are caused by something known as plate tectonics. Plate tectonics is the theory that underneath the earth's surface there are plates that are constantly moving. This theory became much more established by the early 60's when it was acceptable to think that the Earth's crust is moving. The Earth's crust is moving because since the beginning of time, the Earth has constantly been cooling down. This is causing movement. Earthquakes do not occur randomly. They occur along the lines where plate tectonics are located. An illustration of this can be found originally by Harry Ness.

There are three different types of movement that plate tectonics can have. They can either: diverge, converge, or transform. If the plates diverge (pull away from each other) this usually is associated with mid-ocean ridges. The mid-ocean Atlantic ridge is a great example of this. They usually have shallow earthquakes and minor lava flow. The second type of movement is convergence. Convergence is the most common type of plate tectonic movement. Converging characteristics depend on the type of crust that is located where the plates converge (move toward each other). There are two types of crusts: oceanic crust and continental crust. If the convergence happens where there is an ocean crust and a continental crust, the oceanic crust will dive under the continental crust and a deep trench is usually formed. There are shallow to deep earthquakes in this type of convergence. If the convergence occurs at oceanic crust meets oceanic crust, the older of the two oceanic crusts will dive under the newer oceanic crust. Here there are shallow to deep earthquakes and minor lava flow. The final type of convergence is when it occurs at continental crust and continental crust. This happens on land and is fairly rare. This forms mountain ranges of continental crust piled upon each other. An example of this is the Himalayas. There are shallow earthquakes in this region however no lava flows. The final type of plate tectonic movement that occurs is transform movement. This is a very rare type of movement but is very dramatic when occurs. An example of this type of movement is the San Andreas Fault in California.

Earthquakes can be extremely dangerous. They are measured by a machine called a seismograph. This seismograph measures the magnitude of the earth quake. A large magnitude would yield a number such as 8, a small magnitude would give a 2.9 or less. Earthquakes with a magnitude of 8 occur about once a year, an earth quake with a magnitude of 2.9 or less happen thousands of time during a day. Earthquakes can be very costly too. Some of the most expensive earthquakes can range from 5.6 billion which was spend on recovery and damages on the California earthquake, and can be as expensive as 150 billion which was spent in Japan recovering from a devastating earthquake. Each year there are about 15,000 people that die in earthquakes.

In the United States the most seismically geographical location is Alaska. Another fairly common place for the movement of plates is in Hawaii. However of the 48 land states, California is the most common place for plate movement.

Plate tectonics are not an accurate mean of measuring the time that the next earthquake will happen. However they can locate about 90 percent of where the plates are located and the hot spots for where serious earthquakes will occur.

Feedback

Plate tectonics is the theory...

Plate tectonics is the theory that underneath the earth's surface there are plates that are constantly moving

What makes you think so?

Recommended web-sites

Savage Earth
http://www.pbs.org/wnet/savageearth/animations/hellscrust/index.html
The Savage Earth website is the on-line companion to the PBS television series of the same name. This site tells the stories of several great natural disasters, particularly the eruption of Mt. Vesuvius in 79 A.D. that destroyed Pompeii and the 1994 earthquake in Kobe, Japan. It contains articles on the earth's crust and plate tectonics, earthquakes, volcanoes, and tsunamis. Each article features

Earth Like a Puzzle
http://sio.ucsd.edu/voyager/earth_puzzle/look_beneath.html
This interactive feature provides an introduction to the theory of plate tectonics. Topics include the history of ideas on plate tectonics, some evidence that the continents were once connected to each other, what happens when plates move and collide, and a description of the internal structure of the Earth. Interactive illustrations and animated maps are included to help reinforce these concepts

This Dynamic Earth: The Story of Plate Tectonics
http://pubs.usgs.gov/gip/dynamic/inside.html
This site offers the online version of a book published by the United States Geological Survey (USGS). It is intended to provide a brief introduction to the concept of plate tectonics, highlighting some of the people and discoveries that advanced the development of the theory. It is the companion volume to the map, 'This Dynamic Planet', which shows the locations of major volcanic belts, fault systems

Figure 2.1: screen shot of CLICK

The fourth component investigated in CLICK was the preliminary recommender interface shown in Figure 2.1. The text editor where students write their essay, is on the left. The feedback panel, where the recommendations are shown to students, is on the right. For each misconception a student has, the system displays the misconception, a manually constructed cognitive prompt that encourages the student to review the sentence and three recommended resources. For each recommended resource, the title, url and a description of the resource is also displayed. Although this educational recommender interface supports learning, it was not designed to facilitate conceptual change and has no mechanism for users to give feedback on the recommended resources. Although the current components work well on their own, they were not coupled together automatically and have room for improvement. My work will focus on creating this missing link and improving several components.

2.2 Seasons

The educational topic I explored during my research is seasons. I chose this topic because I am interested in understanding why and how the four distinct seasons we experience in Boulder occur. Understanding seasons was not a topic of interest for me in Nigeria because we did not have very different seasons. But after living in Boulder for four years and experiencing the different seasons, I was very interested in exploring this topic.

Core high school Earth science concepts all students should know include a robust understanding of seasons. However, it is difficult for many learners to attain a robust understanding of what causes the seasons, even though it is a phenomena we all experience (Trumper, 2000, 2001a,b,c; Atwood and Atwood, 1996; Sebastià and Torregrosa, 2005).

A survey of different research papers on this topic produced common misconceptions about seasons, some of which are enumerated below:

- Seasons are caused by the distance of the Earth from the sun. So in the summer the Earth is closer to the sun.

- The Earth's orbit around the sun is a highly elongated (skinny) ellipse, making the distance between Earth and sun vary dramatically over the course of a year.

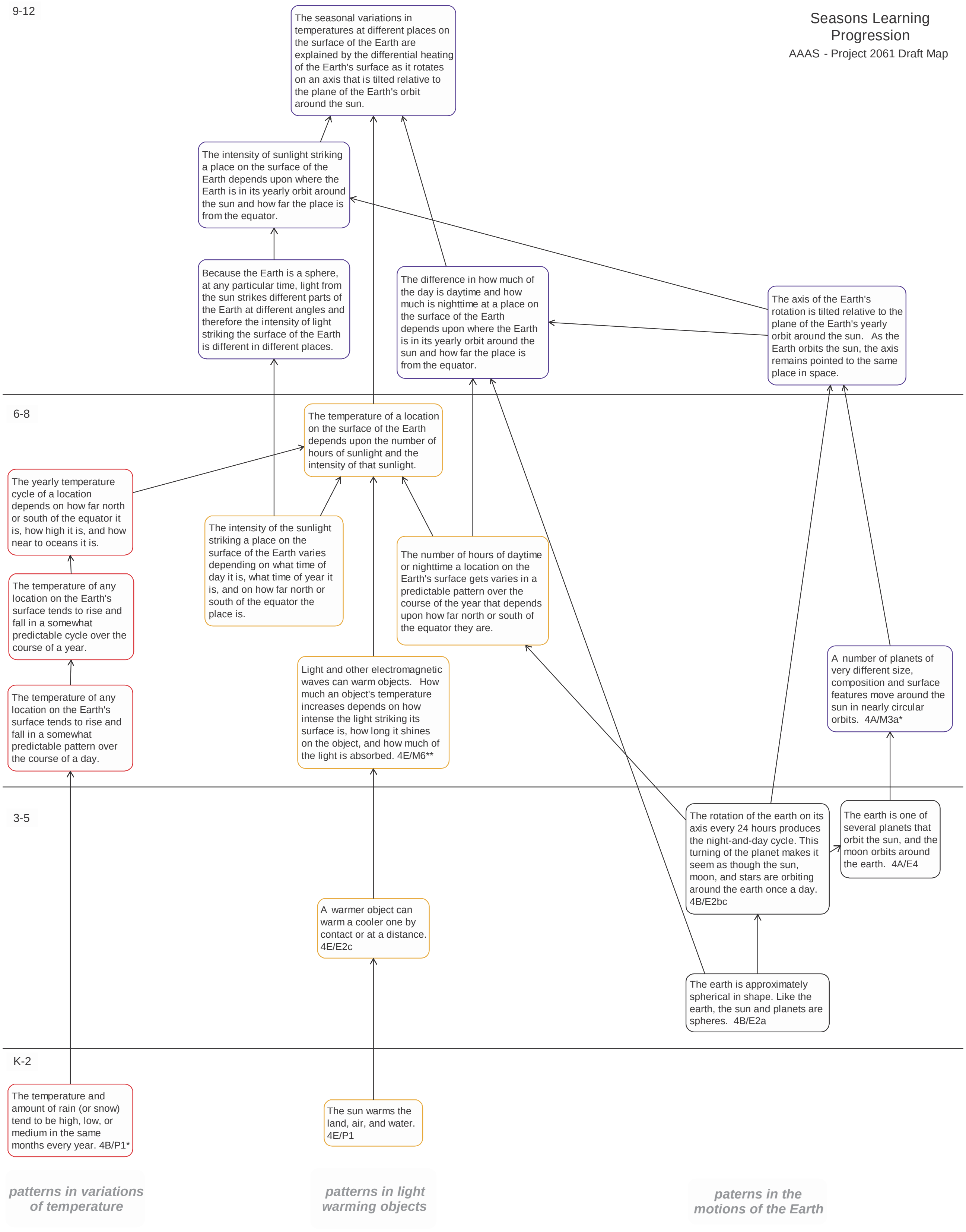

Seasons Learning Progression
AAAS - Project 2061 Draft Map
9-12
The seasonal variations in temperatures at different places on the surface of the Earth are explained by the differential heating of the Earth's surface as it rotates on an axis that is tilted relative to the plane of the Earth's orbit around the sun.
The intensity of sunlight striking a place on the surface of the Earth depends upon where the Earth is in its yearly orbit around the sun and how far the place is from the equator.
Because the Earth is a sphere, at any particular time, light from the sun strikes different parts of the Earth at different angles and therefore the intensity of light striking the surface of the Earth is different in different places.
The difference in how much of the day is daytime and how much is nighttime at a place on the surface of the Earth depends upon where the Earth is in its yearly orbit around the sun and how far the place is from the equator.
The axis of the Earth's rotation is tilted relative to the plane of the Earth's yearly orbit around the sun. As the Earth orbits the sun, the axis remains pointed to the same place in space.
6-8
The temperature of a location on the surface of the Earth depends upon the number of hours of sunlight and the intensity of that sunlight.
The yearly temperature cycle of a location depends on how far north or south of the equator it is, how high it is, and how near to oceans it is.
The intensity of the sunlight striking a place on the surface of the Earth varies depending on what time of day it is, what time of year it is, and on how far north or south of the equator the place is.
The number of hours of daytime or nighttime a location on the Earth's surface gets varies in a predictable pattern over the course of the year that depends upon how far north or south of the equator they are.
The temperature of any location on the Earth's surface tends to rise and fall in a somewhat predictable cycle over the course of a year.
The temperature of any location on the Earth's surface tends to rise and fall in a somewhat predictable pattern over the course of a day.
Light and other electromagnetic waves can warm objects. How much an object's temperature increases depends on how intense the light striking its surface is, how long it shines on the object, and how much of the light is absorbed. 4E/M6**
A number of planets of very different size, composition and surface features move around the sun in nearly circular orbits. 4A/M3a*
3-5
The rotation of the earth on its axis every 24 hours produces the night-and-day cycle. This turning of the planet makes it seem as though the sun, moon, and stars are orbiting around the earth once a day. 4B/E2bc
The earth is one of several planets that orbit the sun, and the moon orbits around the earth. 4A/E4
A warmer object can warm a cooler one by contact or at a distance. 4E/E2c
The earth is approximately spherical in shape. Like the earth, the sun and planets are spheres. 4B/E2a
K-2
The temperature and amount of rain (or snow) tend to be high, low, or medium in the same months every year. 4B/P1*
The sun warms the land, air, and water. 4E/P1
patterns in variations of temperature
patterns in light warming objects
paterns in the motions of the Earth

Figure 2.2: Draft of AAAS idealized learning progression for seasons

- The sun is pretty far off-center within the Earth's orbit, making the distance between Earth and sun vary with time of year even more.
- The distance of different parts of the Earth from the sun, caused by the tilt of the Earth on its axis, causes the seasons
- Seasonal characteristics are the same everywhere on Earth.
- Seasons happen at the same time everywhere on Earth.
- The average temperature in the winter months in the Northern and Southern hemisphere are the same

Figure 2.2 shows the American Association for the Advancement of Science (AAAS) draft of an idealized learning progression for contemporary ideas about seasons (Willard et al., 2007). I hope that users of the CLICK2 system will be able to address and replace their misconceptions with the conventional scientific understandings that are shown in Figure 2.2.

Chapter 3

Background and Related Work

The two main research areas that this work builds on are conceptual change theory and educational recommender systems. Below, I discuss the current state of these two research areas.

3.1 Conceptual Change as Learning Theory

Conceptual change is the process through which people's initial understandings or beliefs are altered and added to, in order to more closely align with scientifically-held understandings through learning and cognitive development (Vosniadou, 2008; Inagaki and Hatano, 2008).

There are competing theories on the nature of these initial understanding or beliefs. But for purposes of this dissertation, I ascribe to the framework theory view. This view of conceptual change sees students naive ideas about science concepts as primitives existing within a naive framework with a distinct ontology that gives rise to predictions and explanations for phenomena that it encounters. This naive framework has usually not been acquired through hypothesis testing but rather has been initialized by life experiences (observing and interacting with physical objects in an environment) and there is usually no metaconceptual awareness within its' owner about its' existence.

Conceptual change is a slow process during which misconceptions can be created. As individuals develop and begin to acquire more sophisticated ideas through experiences and learning, they will try to achieve internal consistency and coherence between their naive mental model and the scientific information they are coming across. Assimilating this new knowledge into an incompatible prior knowledge base creates synthetic conceptions which are referred to as misconceptions (Vosniadou, 2013a).

Learning in conceptual change is seen as a two step process of first, evaluating and correcting the knowledge within the naive framework and second, enriching the new mental model by accommodating new knowledge into a now compatible prior knowledge base (Vosniadou et al., 2008). Conceptual change is inherently a slow and iterative process, therefore, an individual might go through several synthetic frameworks before getting to the scientifically-held understandings.

Conceptual change is a latent variable that cannot be directly observed or measured but is presumed to exert influence on other observable variables such as learning or achievement. Hence, conceptual change has been operationalized as a transformation in learners knowledge, belief and interest (Pintrich et al., 1993; Plummer et al., 2011; Clement and Vosniadou, 2008; Vosniadou et al., 2001; Vosniadou, 2008).

The field of conceptual change research was started by philosophers and historians of science who were trying to explain how scientific theories change (Vosniadou, 2013a). Thomas Kuhn observed that normal science operates within a set of shared beliefs, assumptions and practices that constitute a paradigm. When new discoveries emerge in science that cannot be accommodated within the exisiting paradigm, the scientific community goes into a state of crisis, with several synthetic paradigms being put forward, until the community settles on a new paradigm that explains the existing knowledge and which can accommodate the new observations (Vosniadou, 2013a). An example of a radical paradigm shift is the shift from impetus theory to Newtonian theory in physics.

Conceptual change research was brought to the field of education by Michael Posner and colleagues. They contrasted the paradigm shift in the philosophy of science with Piaget's theory of assimilation and accommodation and came up with the conceptual change learning theory in education aimed at promoting accommodation in students' learning of science concepts (Vosniadou, 2013b)

One of the first theoretical frameworks for conceptual change was also put forth by Posner and colleagues (Posner et al., 1982), where they identified four successive conditions that can induce conceptual change in learners. First, the learners have to be dissatisfied with their current understanding, then the new knowledge has to be intelligible (understandable), plausible(believable) and fruitful (produce correct explanations about related phenomena). This classical approach to conceptual change persisted for a long time but has been criticized for its over-emphasis on logical and rational thinking, its' sole focus on the learner's cognition and not on the learner as a whole and for consequently, ignoring the affective (motivation, values, interests) and other social components of learning (Pintrich et al., 1993; Duit and Treagust, 2003). Newer theoretical frameworks such as the cognitive-affective model of conceptual change (Sinatra, 2005) have incorporated the affective and social components. While the affective and social components of conceptual change are important, I am limiting the scope of this research to the cognitive aspect.

The three basic tenets for how to induce conceptual change in students are (1) students have to be made aware of their misconceptions; (2) students have to be provided with support mechanisms to encourage the disbanding and restructuring of their naive framework into a more correct framework and (3) students have to be supported to enrich the new framework through cognitive accommodation of correct scientific conceptions (Vosniadou, 2008).

The most effective method for eliciting students' alternative conceptions has been by having the students produce an explanatory model (Cartier and Center, 2000; Vosniadou, 2008). An explanatory model is a description of how and why a phenomena is the way it is.

> Such a model is seen as the means by which a theory takes on meaning and, if used flexibly, it gives the theory the power to explain and make predictions for new cases that the subject has not yet seen. Significant changes in an explanatory model are one of the most important types of conceptual change (Clement and Vosniadou, 2008).

Explanatory models can be naive (mainly incorrect), synthetic (a mix of incorrect and correct) and scientific (correct current scientific understanding). An explanatory model of a phenomena can range from simple (containing few entities), such as the model of the circulatory system that a middle school student can produce to very complex, such as the model of the circulatory system that a physician can produce. The importance of explanatory models in promoting understanding is widely recognized; thus explanatory models are being used in several projects aimed at improving students' understanding of science. For example, the Modeling for Understanding in Science Education (MUSE) project (Cartier and Center, 2000; Cartier et al., 2001; Passmore and Stewart, 2002; Stewart et al., 2005) at the University of Wisconsin-Madison, helps students improve their understanding of science by having the students construct an explanatory model for a scientific phenomena and then through different support mechanisms, such as teaching and investigative activities, help the students to gradually produce a more scientific explanatory model.

The most effective method for inducing conceptual change in students has been by using constructive and dissonance strategies to repeatedly criticize students' explanatory models. Students then continually revise their models until they produce a more scientific model, with multiple short cycles needed for complex models (Clement and Vosniadou, 2008; Frede, 2008; Vosniadou, 2008). This method has been effective in a cooperative and facilitative environment where the teacher is very knowledgeable about the topic, co-constructs the knowledge with the students and allows for reflection (Bruning et al., 1999; Scott et al., 1991).

Constructive strategies include analogical, imagistic and simulative modeling. Analogies leverage students' current understanding and makes use of students' prior knowledge in a positive manner. However, sometimes the analog is not well understood and students might transfer all the characteristics including dissimilar ones from the analog to the target. Or, the target might be too far from the analog for the students to make the connection.

Dissonance strategies include discrepant events, contrastive teaching and the use of refutation text. Discrepant events are empirical experiments, data summaries, or demonstrations that provide data that could promote dissonance with students' preconceptions(Vosniadou, 2013b). Refutation texts are texts in which typical misconceptions are refuted directly in juxtaposition to the scientific view (Vosniadou, 2013b). Contrastive teaching involves asking students to explain their understanding about a particular phenomenon and then contrasting these with the scientific view (Vosniadou, 2013b).

Dissonance strategies have been criticized for having the ability to negatively affect students' confidence and self-esteem skills. In addition, they do not encourage new models to be built, they only knock down the old model and students could very well develop another synthetic conception rather than the scientific one. Although constructive and dissonance strategies have their individual failings, when paired together in the right environment, they are the most effective means of producing instructionally induced conceptual change. A meta-analysis has shown that they produce an effect size of up to 1.4999 in the knowledge of users that were taught using these strategies (Murphy and Alexander, 2008).

Conceptual change research has for the most part, been confined to the sciences. It has been targeted at middle school students, high school students, college students and pre-service elementary school teachers. In addition, conceptual change research has mostly been implemented as interventions in traditional classrooms.

3.2 Educational Recommender System as Computational Framework

A Recommender system is any system that produces personalized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options (Burke, 2002). Recommender systems are applicable to a wide variety of domains and tasks.

There are three key recommendation techniques: (a) collaborative filtering: recommending items that match items a user has rated before or recommending items that similar users have rated before (b) content-based filtering: recommending items based on the kinds of items a user has purchased or viewed before, and (c) knowledge-based: recommending items that are based on a user model or profile of interests that may have been captured as a result of explicit feedback or built from behavioral or interaction data. Knowledge-based recommender systems are context-aware recommender systems because they take the user's current knowledge context into consideration rather than simply relying on things the user has done in the past as in collaborative filtering and content-based filtering.

A problem with using collaborative filtering is that it requires the user to use the system for a while before it can start making relevant recommendations to the user. This is known as the new user problem. However most users will want to start getting benefits and recommendations from the system without having to rate a lot of resources first. In addition collaborative filtering can run into the sparsity problem where the ratings for individual resources are sparse. Furthermore, the people doing the rating might not have contemporary science understandings and so the resources being recommended as a result of their ratings might not have good educational value.

Because both collaborative filtering and content-based filtering rely on ratings that have been assigned by users of the system, if a new resource is introduced into the system, it will be difficult for the resource to get recommended since it has no ratings associated with it. This is known as the new item problem. Educational recommender systems will always have new users that want good recommendations right away and it is expected that new items will continually be added to the knowledge repository. Thus, good educational recommendations cannot rely only on collaborative filtering or content. They also have to take the context and knowledge of the user into consideration (Adomavicius and Tuzhilin, 2011). Hybrid Systems built by combining the three recommendation techniques have flourished. They continue to grow in popularity as more profile information is extracted from users' behavior, the longer the users interact and consume items from the same systems.

While recommendation systems have enjoyed a great deal of success in e-commerce systems such as Amazon and Netflix, their benefits are being proven in other areas. Recommendation systems are being developed and deployed in a number of diverse areas - personalized learning and education being one such area. Personalized learning has been recognized as an important advancement for learning in the digital era. It follows that educational recommender systems can form the backbone of personalized learning engines, particularly those that are built on top of the web. These web-based learning engines can provide ubiquitous, instant and continuous access to online learning opportunities that are adapted and customized to each learner's individual needs.

There are several educational recommender systems available right now. Altered Vista uses collaborative filtering to recommend learning resources which have been rated highly by users (Recker and Walker, 2003). QSIA (Rafaeli et al., 2004) is a user-controlled collaborative filtering recommender system. The user can pick the people for example, friends or teaching assistants, whose profile should be used in collaborative filtering to recommend learning resources to him. Or he can decide to let the system decide which group of users to use in the collaborative filtering. Both of these systems rely only on the ratings that users have assigned to a resource. This can lead to the new item problem, where a new item that is relevant does not get recommended because it does not have enough ratings. These systems also do not take the user's profile into consideration. The resource being recommended might be too difficult for the user to comprehend (either maybe because of the wording or because the user needs more background information before tackling the information in the resource).

Shen and Shen (2004) propose a system that uses sequencing rules and an ontology of a domain to guide users through the domain. When non-scientific understandings are identified in the learner's knowledge state, the rules are used to decide which resources to recommend to the learner. This system relies on a hand-crafted knowledge base consisting of rules and ontologies specific to a topic within a domain. Such a system will be very difficult to generalize to other topics and domains. CourseRank (Koutrika et al., 2009) is a hybrid system that recommends classes to take, using collaborative filtering based on a user's profile such as knowledge state (prerequisites taken) and other attributes such as major and area of interest. Huang et al Huang et al. (2009) use a Markov chain model to calculate transition probabilities between learning objects in a sequenced course of study. These two systems target sequencing at a higher level than the system I built. Instead of sequencing the classes or the learning resources within a class, I sequence concepts in order to create a personalized learning path through the content. I target the learner's misconceptions at a much more granular and personal level.

The system I built is similar to ISIS (Hummel et al., 2007), a hybrid ERS which uses ratings of other users and metadata from the learner's profile and learning activity to recommend learning objects. However, I target users' specific information based on their misconceptions instead of simply recommending resources based on users' general learning activity. Another very similar system is a hybrid approach that was implemented in the Virtual University of Tunis (Khribi et al., 2008). It combines collaborative filtering with content-based filtering and also uses the knowledge of the user which it logs and mines from the user's actions. Similar to ISIS, this system approximates the user's specific information need with inadequate features which do not fully characterize those needs. For example, one of its features is to assume a user does not understand the content contained in a resource if the user spends a lot of time on the resource. But it is difficult to determine when a user is actually on the resource, or when the user has left it and just forgot to close the site.

Aside from the already discussed shortcomings of existing systems, a 2011 review of twenty reported educational recommender systems showed that very few ERS are being evaluated on their impact on learners' processes and outcomes (Manouselis et al., 2011). In addition there is no readily available data set to empirically validate the recommendation algorithms. Thus it is difficult to compare and decide which ERS is state of the art.

In summary most of the educational recommender systems have so far tried to use techniques and types of user information found in e-commerce recommender systems to design and create educational recommender systems. It is generally accepted that the goals of educational recommender systems are different from commercial recommender systems. While the objective of e-commerce recommender systems is to provide customers with information to help them decide which products to purchase, the objective of educational recommender systems is to find good items that will address users' knowledge needs. Because the goal of educational recommender systems is very different from e-commerce recommender systems, educational recommender systems should be designed and evaluated in a different way than commercial recommender systems (Buder and Schwind, 2011).

During my research, I did just that. I designed an educational recommender system using conceptual change learning theory in order to facilitate student understanding of science concepts. I evaluated the design of the educational recommender system and its' impact on learners' processes and outcomes.

Chapter 4

Conceptual Framework

My conceptual framework is based on conceptual change learning theory. Conceptual change learning theory advocates targeting student misconceptions in an order such that students deal with basic misconceptions before dependent misconceptions. The basic tenets in conceptual change research are

(1) Awareness of beliefs and presuppositions: make learners aware of their understanding

(2) Dissonance: use dissonance strategies such as refutation text and counter examples to encourage learners to discard their misconceptions

(3) Constructive: Support learners to attain a more scientific understanding by building up correct conceptions using constructive strategies such as experiments, simulations, analogies and imagistic texts.

(4) Knowledge Dependency: Ensure the foundation is solid, i.e., ensure that the building blocks are learned before the dependent blocks of knowledge.

My conceptual framework which is manifested in CLICK2, shown in Figure 4.1 adheres to the preceding principles. CLICK2 takes in resources from a digital library and identifies a knowledge base containing a list of learning goals about a particular topic. Then, students write an essay on that topic. The essay serves a proxy for the students' understanding about the topic. Next, CLICK2 identifies the misconceptions in the essay and confronts the learners

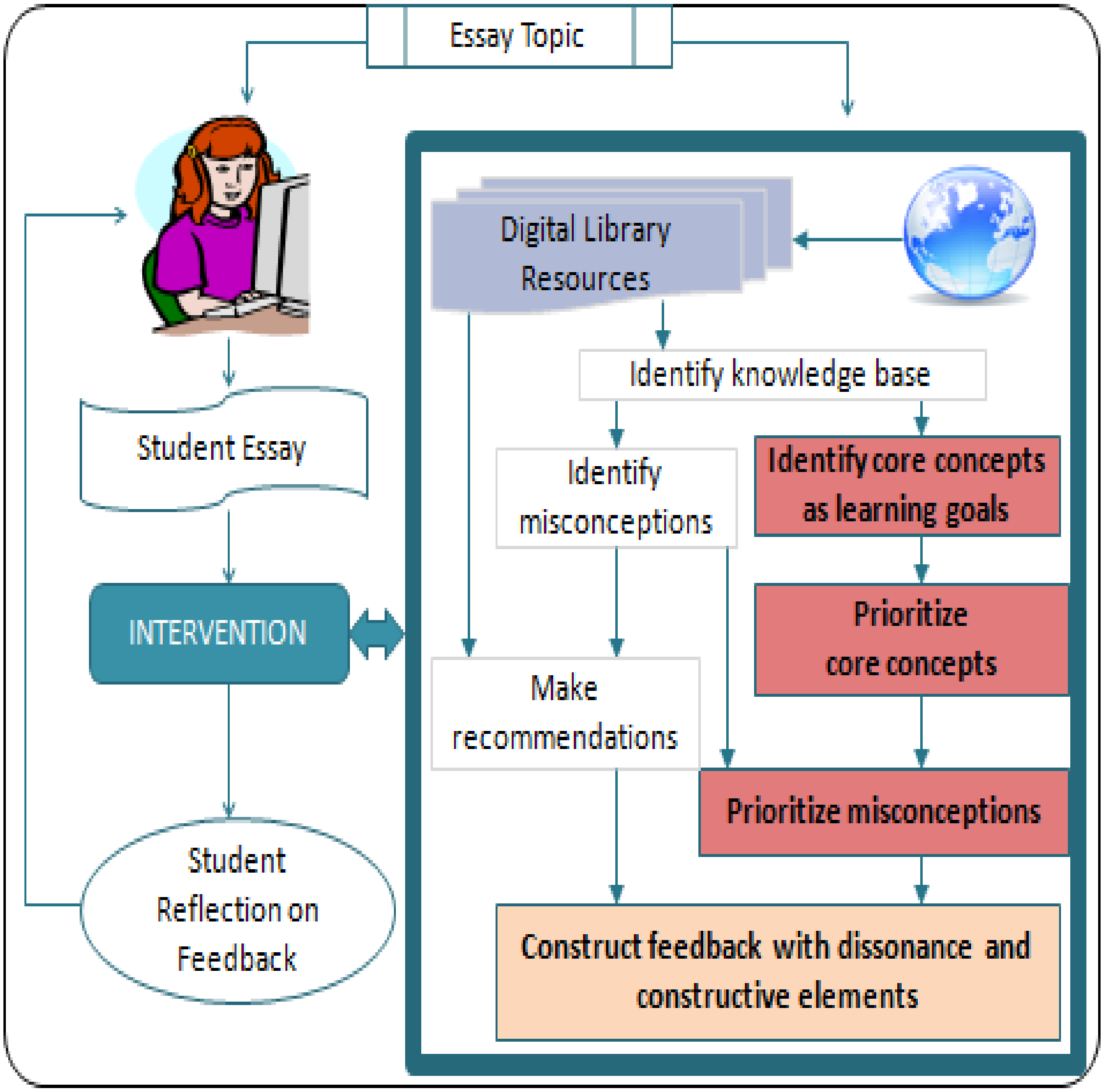

Essay Topic
Digital Library Resources
Identify knowledge base
Student Essay
Identify misconceptions
Identify core concepts as learning goals
INTERVENTION
Prioritize core concepts
Make recommendations
Prioritize misconceptions
Student Reflection on Feedback
Construct feedback with dissonance and constructive elements

Figure 4.1: CLICK2

with it, thereby making learners aware of their understanding. Then CLICK2 uses refutation text as a dissonance strategy and imagistic and simulative representations as constructive strategies to support the learners in addressing their misconceptions. Finally, CLICK2 includes algorithms and computational models to prioritize the knowledge students needs to learn and the misconceptions a particulars student has, thereby satisfying the knowledge dependency requirement of conceptual change learning theory.

My conceptual framework also expands the role of educational recommender systems (ERS) in the learning process from simply making recommendations to also serving as a formative assessment tool. Educational recommender systems serve as a formative assessment tool by providing learners with very targeted feedback on their current work, i.e., their essays. Feedback is a very important aspect of effective formative assessments. Several meta analysis of its' use in classrooms has shown it can induce an average effect size of 0.79 (Hattie and Timperley, 2007).

CLICK2 will be giving feedback to users about their work output (essays).

> Feedback is information provided by an agent regarding aspects of one's performance or understanding (Hattie and Timperley, 2007).

According to the effective feedback model proposed by (Hattie and Timperley, 2007), feedback has to answer three questions; *where am I going?*, *How am I going?* and *Where to next?*, in order for it to be effective in reducing the gap between what the learner understands and what the learner needs to understand. The answers to these questions are built into my conceptual framework.

Referencing the pink boxes in Figure 4.1, **Identify core concepts as learning goals** answers the question *where am I going?*, **Prioritize core concepts** answers *how am I going?* and **Prioritize misconceptions** answers *where to next?*. The feedback interface ensures that the users can answer these questions for themselves as they use CLICK2.

CLICK2, as the name suggests is an enhancement of the CLICK system. The CLICK system is a personalized learning environment that uses graph-based algorithms to perform three main activities: (1) CLICK identifies the smallest subset of sentences that contains the knowledge students in a particular grade should know about Earth Science (de la Chica et al., 2008a); (2) CLICK identifies the misconceptions contained in the students' essays (Ahmad, 2009); and (3) CLICK recommends resources that can help a student address the identified misconceptions in their essays (Gu, 2009).

CLICK2 built on the initial algorithms in CLICK and added some more features that were driven by my analysis of conceptual learning theory. The key differences between the two systems are as follows:

(1) CLICK generates a comprehensive list of all possible learning goals that students in a particular grade level should understand. CLICK2 however investigates how algorithms for identifying learning goals can be optimized to identify core learning goals (Okoye et al., 2013b). It is necessary to identify the core learning goals in a collection of resources because conceptual change learning theory has highlighted the importance of focusing learners on core ideas rather than a plethora of all the ideas about a topic, in order to help learners develop a more robust understanding. In addition CLICK2 includes a pedagogical sequence generator which sequences the core learning goals, thereby generating an ideal order in which the core learning goals should be tackled.

(2) CLICK generates a list of students' misconceptions that are not prioritized while CLICK2 studied and created algorithms that can prioritize students' misconceptions. The misconception prioritization module, which depends on the pedagogical sequence of core learning goals, ensures that learners have the correct conceptions about basic concepts before moving on to tackle their misconceptions about dependent concepts.

(3) CLICK provides resources to help students address their misconceptions, while CLICK2 goes a step further and includes research-based support mechanisms that promote conceptual change. The instructional response that CLICK2 generates incorporates dissonance and constructive strategies from conceptual change theory. The CLICK2 interface is also capable of receiving feedback from learners about the recommended resources. This

feedback can in turn be used to tune the resource recommendation algorithm.

The ultimate goal of CLICK2 is to improve learners' understanding of science content using conceptual change learning theory.

How CLICK2 helps Mandy

The following scenario, which is a continuation of the scenario from chapter 1, shows how CLICK2 can improve learners' understanding of science content.

After unsuccessfully browsing through many websites, Mandy approaches her teacher in class the next day, explaining to her that she can tell her word for word what various online resources say about the reason for the seasons but she does not truly understand how it all works and hence is having difficulty writing an essay in her own words. Her teacher gives her the website of the CLICK2 recommender system and tells her that this will help her improve her understanding of seasons and enable her to write an essay that reflects the contemporary understanding of seasons. Mandy is doubtful about this because she has been to many websites and they have not been able to help her but she is resolved to try this website.

When she gets back home, Mandy goes to the CLICK2 website, where she is prompted to enter her search query, grade level and the number of learning goals about seasons that she would like to see. She types in **why do we have night and day, what causes the seasons, why is winter in the northern hemisphere milder than winter in the southern hemisphere** as the query, selects **high school** as the grade level, and requests to see **six** learning goals. The system searches DLESE using her query and retrieves 100 appropriate resources internally. Then it extracts six learning goals from the 100 resources and creates a learning path through the goals. Next the system prompts Mandy to write down what she knows about seasons in the text editor provided.

Mandy writes **We have night and day because the Earth rotates on its axis every 24 hours and so the places facing the sun will have daylight and the ones**

facing away from the sun will have night. The reason why we have seasons is that the Earth is at different distances from the sun at different times of the year. When it is closer to the sun, we have summer and when it is farthest away from the sun, we have winter. The system processes her answers and annotates her answer for night and day as being correct and the answer for the cause of seasons as a misconception. It also infers that she needs to learn about factors affecting the temperature of any location on the surface of the Earth. So to the system, she has two problematic conceptions (an incorrect and a missing conception). The system indicates these misconceptions in the interface and Mandy is surprised that her understanding of why the seasons occur is wrong.

For her missing conception about winter in the northern and southern hemisphere, CLICK2 recommends three imagistic, simulative and video resources that discuss factors affecting the temperature of any place on Earth and characteristics of the northern and southern hemispheres. For her incorrect conception about what causes the seasons, CLICK2 juxtaposes her incorrect conception with a paragragh containing the correct conception. This paragraph was selected by the system to challenge Mandy's explanation, thereby creating cognitive dissonance within Mandy's mental model, which hopefully will lead her to generate a more scientific explanation. In order to help her address both her misconceptions, the system also displays an interactive simulated model of the Earth as it rotates on its axis and revolves around the sun. The tilt of the Earth is emphasized and the rays of the sun hitting any point on the Earth are illustrated. In addition, the seasons of the northern and southern hemisphere at each point in time is annotated on the model. CLICK2 also recommends some short videos about seasons and a text resource with some images.

Mandy plays with the simulation, reads through the recommended resources and paragraphs and reflects on what she has observed and read. She uses this information to rewrite her essay and goes through several iterations of rewrites and recommendations before the system assures her that her essay now reflects the scientific understanding of seasons. The next week in class, after submitting their essays, the teacher asks what they learned from writing the essay. Mandy is able to say that prior to writing the essay, she thought that seasons were caused by the distance of the Earth from the sun. But from the research she did to write the essay, she learned that the seasons are caused by the tilt of the Earth as it revolves around the sun. The part of the Earth tilted towards the sun receives more direct sunlight, hence it is warmer and experiences summer. The part of the Earth pointing away from the sun receives less,indirect sunlight, thus is colder and experiences winter.

Chapter 5

Research Design

My research design draws on methodologies from human-computer interaction, machine learning and natural language processing. The research design is comprised of five studies. The five studies correspond to the five research questions discussed in chapter 1. Table 5.1 shows the mapping between the research questions and the studies.

Table 5.1: Research questions and studies that will address each question

Research Question	**Study**
(RQ1) What are design options for creating an educational recommender system (ERS) with research-based support mechanisms for promoting conceptual change ?	Study 1 - Design workshop for an educational recommender system that supports conceptual change
(RQ2) How does the educational recommender system, with its conceptual change support mechanisms, affect users?	Study 2 - Qualitative learning study to examine users' processes and outcomes
(RQ3) How can we model expert strategies for prioritizing student misconceptions?	Study 3 - Automatic prioritization of student misconceptions
(RQ4) How well can different computational methods identify the learning goals in a collection of documents?	Study 4 - Automatic Extraction of Core Learning Goals
(RQ5) How well can machine learning classifiers model the pedagogical sequence of learning goals produced by human experts?	Study 5 - Automatic Sequencing of Learning Goals

In the first study, I used participatory design methods to create a recommendation feedback interface. The second study, was a learning study in which I investigated how the CLICK2 system influenced learners' processes and outcomes. The last three studies are concerned with comparing machine learning and natural language processing approaches to performing three critical tasks underpinning support for conceptual change theory: extracting learning goals, sequencing learning goals and prioritizing learners' misconceptions. All three of these studies draw on analyses of human expert processes to inform the design and evaluation of the algorithms.

5.1 Participatory Design

For the first study, I use participatory design to create the recommendation feedback interface. Participatory design is an approach to creating products that endeavors to engage all stakeholders in the design process to ensure the product satisfies their needs and is usable by them (Muller, 2003; Sanders, 2002). Participatory design is not a design technique but instead focuses on the people involved in the design process (Carroll et al., 2000; Sanders, 2002). It started out in Scandinavia as cooperative design and has permeated the design process in many fields including software and hardware design, landscape architecture, city planning and medicine (Bødker and Iversen, 2002).

I decided to use participatory design because the life experience of my target population (middle and high school students) is different from mine. In addition, they have mostly grown up using technology and the internet. I needed to understand how they currently perform several tasks that the ERS can support. Also, I wanted to discern their expectation about an online educational recommender system. Users participated in all stages of the design process from gathering requirements, to creating the design, iteratively evaluating the design and in the final evaluation of the built system.

5.2 Learning Study

The learning study was a pilot learning study with only an experimental group because I was not yet interested in making statistical inferences about the amount and significance of conceptual change in users of CLICK2.

I was more concerned with the qualitative aspects of the study i.e., if and how the conceptual change features I introduced in CLICK2 where used, students' perception of CLICK2, and if there was any improvement in students' understanding after they used CLICK2.

I used instruments such as knowledge perception questionnaires, essay prompts, multiple-choice knowledge questionnaires and usability questionnaires for this study. The instruments were not validated, although they were reviewed by a scientist who has worked in the field of Earth Science education for more than 10 years.

5.3 Study of Human Expert Processes

For the last three studies, I begin by studying human expert processes for achieving the task I want to do computationally. For the tasks of extracting learning goals, sequencing learning goals and prioritizing learners' misconceptions in the domain of Earth science, I chose to study Earth science experts. I define Earth science experts as people who have been involved in Earth science curriculum development or have taught Earth science for more than 10 years.

For an engineering task, we study how experts perform a task. What information sources and other resources they draw from to guide them and their step by step process for performing the task. From this study, we then come up with features and procedures that a computational model would need to extract and go through, in order to perform the task and achieve a similar degree of accuracy as the expert does.

Aside from studying the human expert process to come up with features and procedures, I also use the studies to generate gold standard data that I can analyze. This analysis helps me understand the complexity of the task and calibrate my expections for how accurate the computational model will be. For example, if several experts perform a task and the inter-annotator agreement measured using an appropriate statistic such as Pearson, Spearman's or Kendall's correlation coefficient shows low agreement, then we can assume that the task is complex. In that case, we would not expect a computational model to do well on such a task, since the human experts cannnot agree well on what answer or solution is correct for such a task. The data from studying expert processes is also invaluable for assessing the accuracy of a computational model.

5.4 Building the Machine Learning Models

After studying expert processes, the next step for my last three studies, was creating the computational models using the features and procedures extracted from the expert studies.

Sometimes, I used features that I had not extracted computationally, but which were used by the experts. That is, I use the features *as is* from the expert study. I imagine that down the line, someone will come up with a good way of extracting such features. However, for this research study, my objective was to see how well the features used by the experts would perform when used by a computational model. I used Weka to build the computational models.

> Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset using the interface or called from Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes (Hall et al., 2009).

I was working with very modest data. For the first study, I had about 100 data points, for the second, 30 and for the third, 12. Compared to the thousands of data points that are used to train and test data models, my data set is vastly inadequate. Thus, I could not have separate training and testing data sets. Therefore, for the three tasks, I resorted to using cross validation (Hawkins et al., 2003; Kohavi et al., 1995) to estimate the predictive perfomance of the models. My gold standard data for the studies was the data I got from the study of the human expert process.

Chapter 6

The CLICK2 Environment - Design Study

This design study investigated how personalized learning environments can be enhanced with research-based support mechanisms from conceptual change. Simultaneously tearing down misconceptions and building up correct conceptions have been shown to promote conceptual change in the classroom environment when used together (Murphy and Alexander, 2008). This study investigated how cognitive strategies that encourage the tearing down of pre-existing misconceptions and building up of correct conceptions can be realized in a personalized learning environment that supports undergraduate research writing.

I chose a personalized learning environment that supports undergraduate research writing as the platform where I incorporate the cognitive strategies that support conceptual change as opposed to a speaking or drawing based environment because the state-of-art natural language processing algorithms can only fully support a text-based environment.

This study also investigated how to design the interface and interaction of the personalized learning environment system in a way that makes it seamlessly usable, and capable of supporting varied tasks such as completing an assigned writing task, getting ready for an exam, and learning more about a topic. Since this system was to be a new concept for the target audience (undergraduate students), I also had to make sure the functionality was intuitive. This involves making sure the function triggers (buttons, checkboxes, etc.) can be seen easily, their functions are cued correctly by using appropriate labels and the result of having activated them is easily noticeable.

6.1 Materials

Data for this study were common instruments for design workshops such as scenarios, mock-ups, detailed notes from think-alouds and debrief sessions and a usability questionnaire.

6.2 Methodology

This study had four phases. The first phase was user research, where I researched how potential users currently do research and write a scientific paper. The second phase involved participatory design of the environment. In the third phase, I did multiple iterations and evaluations of the penultimate design and in the fourth phase, I built and ran a usability study on the final prototype.

6.2.1 Phase1 : User Research

My approach to the user research was to observe those writing and research habits which might bear relevance to the interface being designed. I hoped, in that manner, to find data about how a user might use such a system, and how that would relate to their normal habits of writing.

Phase1: Participants I recruited twelve participants that were close in age and experience to our target population i.e., high school students. Seven women and five men were recruited from the SONA pool, a pool of research subjects run by the University of Colorado at Boulder psychology department. The students self-selected into the study and were awarded research points for participating in the study. The research points are a requirement of their psychology class. However a student may choose to work on a research paper in lieu of participating in a study.

Phase1: Method for the User Research As I was seeking to find information about habits of researching and writing, the full process of writing an essay or term paper was simply too lengthy to observe. In lieu of that full process, I sought to get an interviewee to either (a) resume work on an ongoing writing or research project which they were doing, or (b) sit down with a research paper they had recently finished, and walk through its creation. In both cases, there were a number of specific issues I sought to get people to elaborate on, such as:

- How they went about finding the information they had used or were using
- How that information was recorded, tracked and synthesized between its original reading and its use in the paper
- The physical environment around the user, and how they interacted with that environment
- The digital environment - which programs were they using, how they organized desktop space, and how they changed and interacted with that environment
- How, when, and why the user goes about editing their paper
- How, when, and why the user factored trust and relevance into their writing and research process
- What they thought about the sources they were using

The length of interviews ranged from forty five minutes to an hour and fifteen minutes and the questions I asked were:

- Describe how you wrote your last scientific essay OR describe how you write an essay
- How do you go about finding the information you had used or were using?
- How was that information recorded, tracked and synthesized between its original reading and its use in the paper

- What was the physical environment around you, and how did you interact with that environment
- The digital environment - which programs were you using? how do you organize your desktop space, and how do you change it while interacting with it
- If, how, when, and why do you edit the original draft?
- If, how, when, and why do you factor trust and relevance into the writing and research process
- What would have helped you develop a better essay (please consider all issues: time, knowledge of the material, resources you could understand, more personal interest in the topic etc)
- Name a science topic you are comfortable discussing
- How did you get comfortable with it? How did you learn about it? Are you personally interested in it?
- Name a science topic you are currently asked to learn about but are having problems learning about it
- Describe the most memorable feedback you have gotten for learning
- Describe the most memorable essay feedback you have gotten. Why was it memorable?

Phase1: Data Analysis I analyzed this data using the list, lump and label method. To do this, I listed the information that stood out during the interviews on pieces of paper, lumped them according to topic and then labeled that topic. The end result was an affinity diagram shown in Figure 6.1, with six areas that any system designed to support undergraduate research writing needed to support . They are: (1) visualization, (2) tracking information, (3) iteration and revision, (4) pre-existing research habits, (5)trust and relevance and (6) the graphical interface.

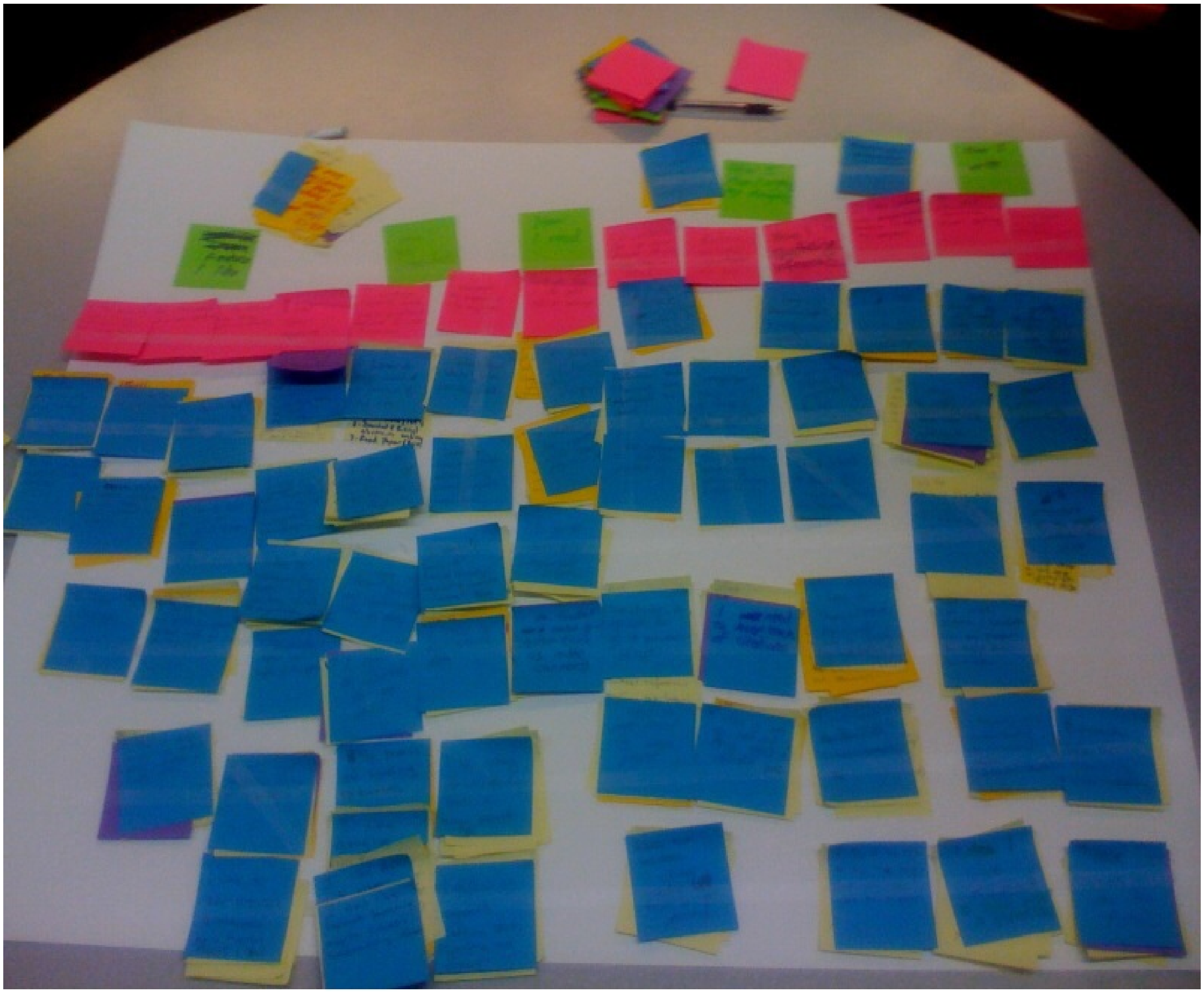


Figure 6.1: Affinity diagram

Phase1: Result

(1) Visualization as Brainstorming and Organization:

Most users made rough outlines, and attempted to organize their data, quotes, etc.

using these outlines - many times using main documents as places of storage and organization for their raw content, unsynthesized quotations, etc. - but showed neither effort nor interest towards any more complicated systems of visualizing data.

I concluded that any attempt to address visualization would, ideally, accommodate or encourage the use of visualization as a planning and brainstorming tool, while allowing for the simpler 'structural outlines' used by the students. I would thus have to address (1) whether to make a manipulable, visual representation of user ideas, and (2) how such a system would interact with writing, annotation, and the sources themselves.

(2) Tracking Information:

While few users had developed systems of tracking information, annotations, citations and ideas, more users often either neglected such tracking entirely (having to therefore retrace their steps when they needed something such as the bibliography information), or kept it in a disorganized manner - such as pasting relevant information into the same document. Such users treated the rough structure of their papers as repositories for unsynthesized information, unintegrated notes, bulleted ideas and direct quotations. In contrast, other users wrote either electronic, in-line comments on relevant sections of papers being read, or scrawled notes in the margins of physical papers. Otherwise, they would make full summaries of works read.

(3) Iteration, and Revision:

Iteration, revision, and the general treatment of papers as an evolving collection of arguments and explanations, was present across the students interviewed, but in highly different forms. While few users tended to produce rough versions of final products, and to iterate upon those versions using revision tracking or commenting systems, more users tended to write one section of a paper at a time, and thus to manage the constant cycle of writing and research without often iterating over any content more than once.

(4) Preexisting Research Habits:

Two major issues involving timing came up in relation to preexisting research habits. Firstly, more interviewees tended toward broad and haphazard surveys of a field of study, in search of a main topic or thesis, aiming only to confirm the viability of their topic or thesis. Such a phase did not necessarily involve deep reading or difficult synthesis of concepts. From there, student habits ranged from reading for comprehension at one extreme, to hunting for concrete arguments or quotes for use within a paper on the other. Secondly, time factors were very important to some students. Many users concluded their initial research, not based upon satisfaction, but purely because of time constraints. Those writers which write while researching tended to do so in a concept-by-concept manner, researching each concept as they needed to explain it.

(5) Trust and Relevance:

While I went in assuming that trust in a resource was a major issue, I found that users did not actively validate the authority of a text nearly as much as they did the relevance of that text to their research. Most users were confident of their own ability to judge the worth of a text, based on the website where the text was published (mistrusting wikipedia and trusting .org websites over .com websites). Many were also more likely to be using sources given to them by their professors (textbooks, recommended readings, etc.) and therefore had little worry of the trustworthiness of their sources. More at issue in the valuation of a source, however, was its perceived relevance. Most users made value judgements of a paper while reading it. These users therefore relied upon abstracts and introductions in order to determine whether a source was worth reading in full.

(6) The Graphical Interface:

Two patterns did emerge concerning the computer environment in which a user worked. Firstly, users tended to reorganize their work, however it was, so that they might see everything that they were dealing with at once; either so that they might look at a document and their notes at the same time, or so that they might see as-yet-unsynthesized information and their main document at the same time. Gratuitous clicking or scrolling was widely panned. Secondly, the predictability of an interface was generally considered paramount, and changes in how an interface worked (such as for Office 2007) were often complained about. Multiple users, instead, expressed desire for an option to revert to ’classic’ or ’legacy’ modes whenever a system did indeed change its interactions.

6.2.2 Phase2 : Design Brainstorm

The design brainstorm was organized so that I could understand what potential users would expect the essay-based recommender system to look and feel like. The end result was a low fidelity system that incorporated the look and feel of what the participants expected, but also paid attention to the six areas that needed to be supported (from phase1 of the design study) in order to support undergraduate research writing.

Phase2: Participants Twelve undergraduate students from University of Colorado, Boulder participated in the design brainstorm. They were recruited from the undegraduate SONA psychology pool at CU-Boulder. These students were different from the ones that participated in Phase 1 of the design study.

Phase2: Method The 12 participants were given the following scenario and asked to design an interface based on it.

Scenario You have written an essay for your Earth science class on a factual topic of why we have seasons. It is factual because there is a scientific explanation of why we have seasons. The systems' task is to give you feedback on your essay: the problems with it, the severity of each problem and how to get to the right solution or understanding by recommending digital (online) resources. No other person sees your essay or the recommendations, only you. The goal of the system is not simply to help you produce a good essay but also to help you understand the phenomena you are writing about. You will be tested on this topic in an in-class quiz later in the semester. Using the sheets of papers and pencils, please indicate how you would like to receive feedback and how you would like to give feedback to the system. You can use text, different text sizes, colors, graphs, pictures, tags, star-rating, lists etc. Feel free to be as creative as you want to be. Imagine I have the ability to implement anything you want. The feedback to you from the system can include:

- Position - an awareness that there is a problem in your essay/understanding
- Problem - what the problem is in your essay/understanding
- Severity of problems - an indication that one problem is more important than another
- How to get there - a suggested pathway for you to achieve the desired understanding
- Progress - an indication of whether your understanding is improving
- Are we there yet? - an awareness of when you have achieved the scientific understanding

The aforementioned feedback that the system could give represented my support of conceptual change in the online personalized learning environment i.e., by ensuring the system tears down pre-existing misconceptions and ensuring the system helps build up correct conceptions.

The participants spent 45 minutes and came up with 12 different designs. However, must of them struggled to come up with something . Many of them ended up with an interface very similar to Microsoft Word. Those people also wanted the conceptual problems to be called upon and displayed the same way Microsoft Word displays grammatical errors.

After the individual design stage, I paired them up into groups of six and had them merge their six individual designs into one, using the blackboard to draw out ideas. This resulted in two designs that were very similar. Then I used the results from phase1 to merge the two designs into one. My biggest change to their designs dealt with tracking information. I included a button to add the information about a source into their bibliography, which could then be imported and included in their research paper.

Phase2: Result The final product of this phase was a low-fidelity design of the recommendation feedback environment.

6.2.3 Phase3: Evaluating and Iterating on Design

This phase involved refining the design using a higher fidelity mockup tool than paper and pencil. I used Balsamiq Mockup(Guilizzoni, 2010) a wireframing and mock up tool to design the mockup for this phase of the design study.

Phase3: Participants I used two female students from the SONA pool (different from the other students that had already been exposed to the system) to evaluate the high fidelity mockup of the design from phase2 during a ThinkAloud session.

Phase3: Method Phase3 consisted of using two inspection methods to validate my design. Inspection methods help to diagnose usability problems in interface and interaction designs. First, I performed heuristic evaluation on my own. Then I went through the think-aloud session with the two students.

Phase3: Result

* Heuristic Evaluation:

Design heuristics are rules of thumb rather than specific guidelines (Nielsen, 1994). They are advice about good and bad design solutions and are based on practical experiences. In heuristic evaluation, you have a set of usability rules or guidelines. For each of the guidelines, you examine your interface and interaction design and try to find the problem described in the guideline. If found, you do further design on the interface and interaction to address the problem. For this study, I used Jakob Nielsen's (Nielsen, 1994) heurisitics which are:

(1) Visibility of system status

(2) Match between system and real world

(3) User control and freedom

(4) Consistency and Standards

(5) Error prevention

(6) Recognition rather than recall

(7) Flexibility and efficiency of use

(8) Aesthetic and minimalist design

(9) Help users recognize, diagnose and recover from errors

(10) Help and Documentation

The most pressing concerns brought up by the heuristic evaluations were those of control and error recovery. My prototype design had no way of ignoring the recommender system if it was activated before being truly useful (before users actually wrote anything). I had always assumed that the users were starting from a pre-existing essay. This inspired me to create the *Missing Information* category in CLICK2. This way, if a user has not written anything yet, the recommender system would recommend resources based on important concepts about the topic that the user was expected to write about.

The remaining heuristic issues were of two kinds. First, issues of clarity in language, such as the main button being labeled *Evaluate*. This made participants believe the system was evaluating for spelling and grammar too, which it was not. So, I changed the button text to read *Evaluate for Conceptual Problems*

Another heuristic issue was that of unpredictable functionality, like what happened if there was no misconceptions left. This issue inspired me to create the color coded *status bar*, that alerted users unobtrusively, to when they had addressed all the misconceptions and missing information within their text, as related to a specific topic for their specific class.

* Think-Aloud Testing: During Think-Aloud testing, the two participants were asked to perform specific tasks and encouraged to elaborate on their actions by vocalizing their thoughts while performing the actions. When they did not, the investigator prompted them with questions like *what are you thinking?*, *what are you trying to achieve?* etc. I asked both participants to use the system to research and write a paper on seasons in the USA. During the Think-Aloud Session, I was simulating the working of the system. So, when the participants made a statement like, ”I want to evaluate my essay, so I press this button”, I would show them the page that would come up.

The think-aloud tests exposed the inadequacy of the interaction design. I had focused a lot on the interface during writing of a text and getting feedback on the text and had not given a lot of thought to the interaction, especially the first use of the system. Using the feedback from the think-aloud testing, I redesigned the system from the perspective of a first user. I included explanations of button functions when they were moused over and added in explantations for manipulatable objects that changed colors, such as *red* for misconceptions that had not been addressed, *yellow* for misconceptions that had been worked on but which the system had not evaluated and *purple* for the misconception that was currently being worked on. In addition, I included a help button, as a last resort, to explain the functionality and use of the system.

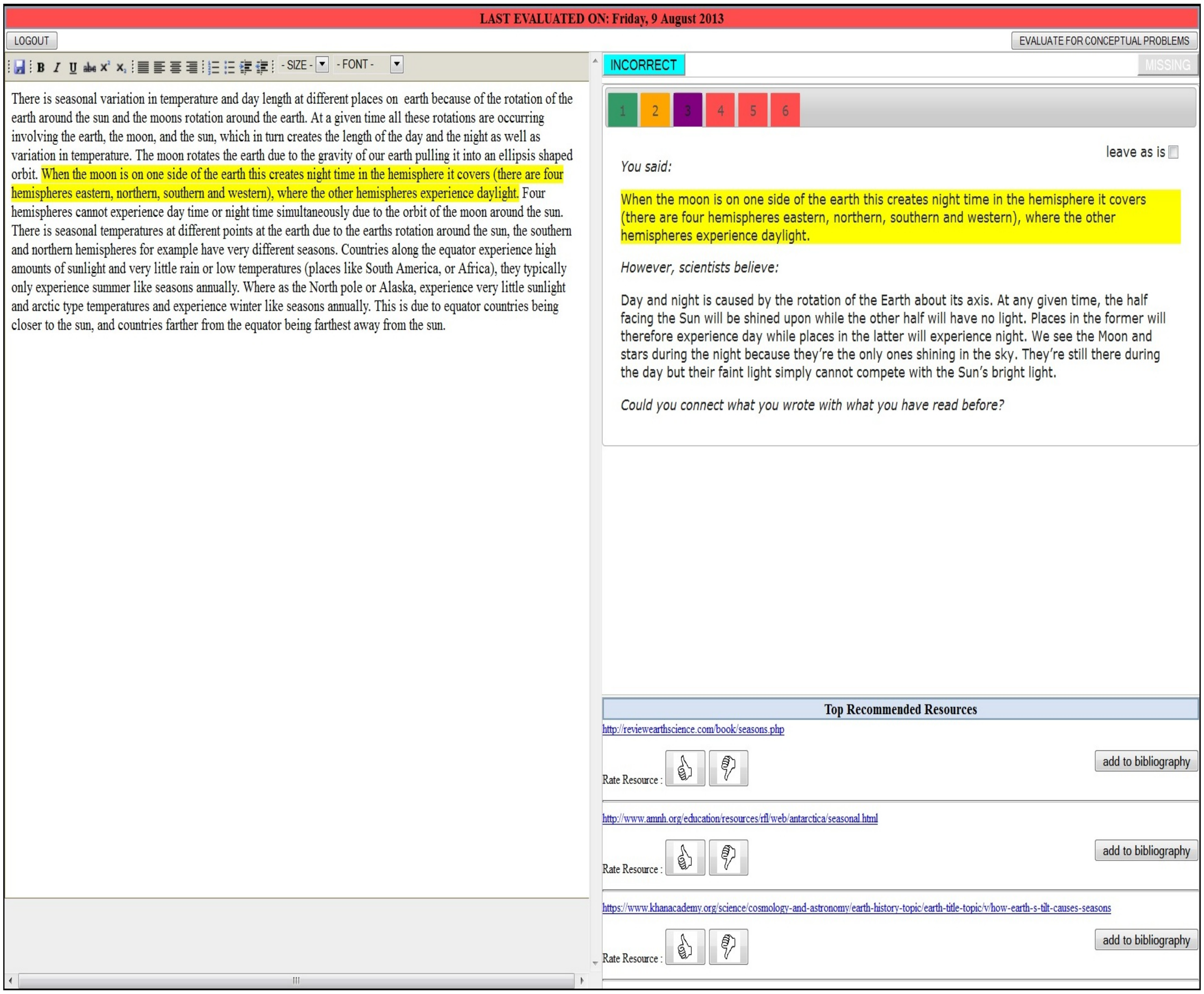
In addition, during think-aloud-testing, one participant disagreed with one of the statements that had been labeled a misconception. The participant wanted a way to indicate that the system had labeled it wrongly. This inspired me to create the *Leave-as-is* check box next to a misconception. When this box is checked, the misconception tab will be colored green, showing that the misconception will not be evaluated by the system and would be assumed to be correct by the system.

6.2.4 Phase4: Usability Study of CLICK2

I built CLICK2 using the Django framework and an sqlite database. CLICK2 was deployed on a Linux server at http://goldfinch.colorado.edu/learning, where it was used to run a learning study. Figures 6.2 and 6.3 show the main features of CLICK2.

The main features of CLICK2 are:

- *The editor*: where users type in their understanding of a topic
- *The status bar*: which displays the status of the text. When red, it means there are still misconceptions in the written text, and when green it means the text is free of misconceptions
- *The evaluate for conceptual change button*: which initiates the algorithm to be run on the text and the feedback displayed
- *The help button*: which explains how to use the system
- *The incorrect feedback pane*: which initiates the display of the feedback for the incorrect sentences
- *The leave-as-is check box*: which marks a misconception as having been misidentified


LAST EVALUATED ON: Friday, 9 August 2013
LOGOUT
EVALUATE FOR CONCEPTUAL PROBLEMS
- SIZE -
- FONT -
INCORRECT
MISSING
1
2
3
4
5
6
leave as is □

There is seasonal variation in temperature and day length at different places on earth because of the rotation of the earth around the sun and the moons rotation around the earth. At a given time all these rotations are occurring involving the earth, the moon, and the sun, which in turn creates the length of the day and the night as well as variation in temperature. The moon rotates the earth due to the gravity of our earth pulling it into an ellipsis shaped orbit. When the moon is on one side of the earth this creates night time in the hemisphere it covers (there are four hemispheres eastern, northern, southern and western), where the other hemispheres experience daylight. Four hemispheres cannot experience day time or night time simultaneously due to the orbit of the moon around the sun. There is seasonal temperatures at different points at the earth due to the earths rotation around the sun, the southern and northern hemispheres for example have very different seasons. Countries along the equator experience high amounts of sunlight and very little rain or low temperatures (places like South America, or Africa), they typically only experience summer like seasons annually. Where as the North pole or Alaska, experience very little sunlight and arctic type temperatures and experience winter like seasons annually. This is due to equator countries being closer to the sun, and countries farther from the equator being farthest away from the sun.

You said:

When the moon is on one side of the earth this creates night time in the hemisphere it covers (there are four hemispheres eastern, northern, southern and western), where the other hemispheres experience daylight.

However, scientists believe:

Day and night is caused by the rotation of the Earth about its axis. At any given time, the half facing the Sun will be shined upon while the other half will have no light. Places in the former will therefore experience day while places in the latter will experience night. We see the Moon and stars during the night because they're the only ones shining in the sky. They're still there during the day but their faint light simply cannot compete with the Sun's bright light.

Could you connect what you wrote with what you have read before?

Top Recommended Resources

http://reviewearthscience.com/book/seasons.php

Rate Resource : 👍 👎 add to bibliography

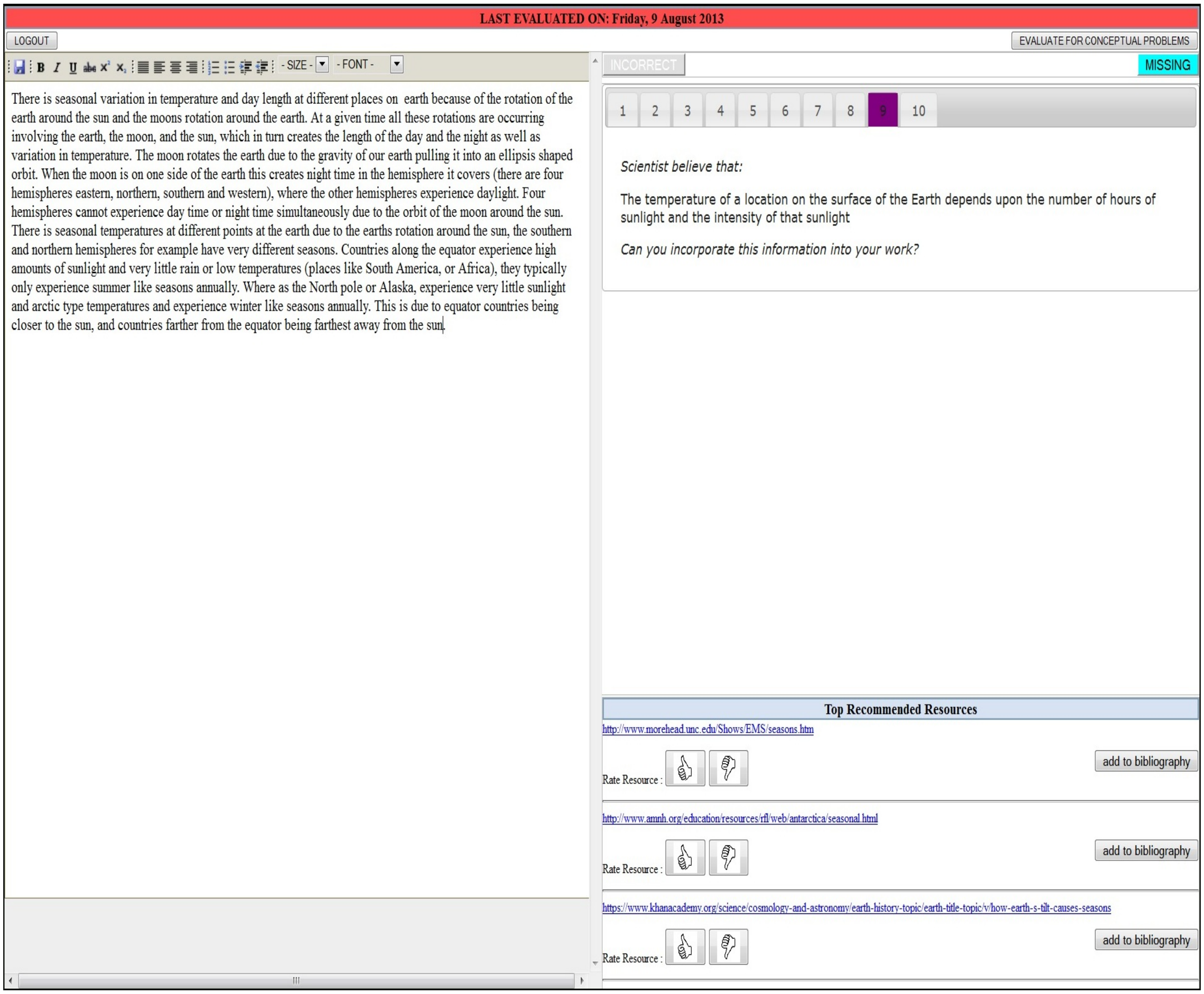
http://www.amnh.org/education/resources/rfl/web/antarctica/seasonal.html

Rate Resource : 👍 👎 add to bibliography

https://www.khanacademy.org/science/cosmology-and-astronomy/earth-history-topic/earth-title-topic/v/how-earth-s-tilt-causes-seasons

Rate Resource : 👍 👎 add to bibliography

Figure 6.2: CLICK2 - with *INCORRECT* information highlighted



LAST EVALUATED ON: Friday, 9 August 2013

There is seasonal variation in temperature and day length at different places on earth because of the rotation of the earth around the sun and the moons rotation around the earth. At a given time all these rotations are occurring involving the earth, the moon, and the sun, which in turn creates the length of the day and the night as well as variation in temperature. The moon rotates the earth due to the gravity of our earth pulling it into an ellipsis shaped orbit. When the moon is on one side of the earth this creates night time in the hemisphere it covers (there are four hemispheres eastern, northern, southern and western), where the other hemispheres experience daylight. Four hemispheres cannot experience day time or night time simultaneously due to the orbit of the moon around the sun. There is seasonal temperatures at different points at the earth due to the earths rotation around the sun, the southern and northern hemispheres for example have very different seasons. Countries along the equator experience high amounts of sunlight and very little rain or low temperatures (places like South America, or Africa), they typically only experience summer like seasons annually. Where as the North pole or Alaska, experience very little sunlight and arctic type temperatures and experience winter like seasons annually. This is due to equator countries being closer to the sun, and countries farther from the equator being farthest away from the sun.

Scientist believe that:

The temperature of a location on the surface of the Earth depends upon the number of hours of sunlight and the intensity of that sunlight

Can you incorporate this information into your work?

Top Recommended Resources

http://www.morehead.unc.edu/Shows/EMS/seasons.htm

Rate Resource :

http://www.amnh.org/education/resources/rfl/web/antarctica/seasonal.html

Rate Resource :

https://www.khanacademy.org/science/cosmology-and-astronomy/earth-history-topic/earth-title-topic/v/how-earth-s-tilt-causes-seasons

Rate Resource :

Figure 6.3: CLICK2 - with *MISSING* information highlighted

as a misconception, by turning that misconception's tab green and ensuring that future evaluations, do not include that text as a misconception.

- *The missing information pane*: which initiates the display of the feedback for the missing concepts
- *The feedback pane*: which displays the feedback. When the incorrect tab is active, it displays the incorrect sentence highlighted in the text and in the feedback pane. It also displays a prompt, a refutation text and recommended resources. When the missing table is active, it displays the missing text with a prompt to include the missing information into the essay.
- *The rate resource pane*: which offers a thumbs up or down button, with which to rate the usefulness of a resource as it relates to a specific incorrect sentence or missing concept.
- *The add to bibliography button*: which discourages students from inadvertently plagiarizing a resource by making it easy to include the bibliography from a resource that was used during the research and writing of a work.

Phase4: Participants I recruited participants using the CU Buff Bulletin, posting fliers on campus and by word of mouth. Although seventeen students signed up for the study, only twelve showed up for the usability study.

Phase4: Method The usability study and the data from the usability questionnaire were part of the learning study described in Chapter 7. I present the results here because they are relevant to this research question. The usability questionnaire was developed by identifying the features whose usage I wanted to learn about, consulting several usability questionnaires and discussing the questions with various people that had conducted usability studies to ensure the questions reflected the information I wanted to learn without biasing the answers I would get. The usability questionnaire which can be seen in Appendix E includes questions about satisfaction, ease of use, utility and questions about the use of some specific elements in the interface.

Phase4: Result

Figure 6.4 show the results for the individual Likert-scale questions that were on the usability questionnaire. The task for this study was the design of an educational recommender system that would provide feedback to users to guide them in their learning. Therefore, I summarize the results of the study based on the features of an effective feedback environment as posited by Hattie (Hattie and Timperley, 2007). However, I break his three questions which an effective feedback environment should answer into four, which are: (1) *Where am I?* : this is the user's current state of understanding, the position. (2) *Where am I going?*: this is the user's target understanding, the target or destination. (3) *Where to next?*: this is the user's learning pathway, the path. (4) *How am I going?*: these are the resources the user has to help him get to the target understanding, the resource recommendations.

In addition, I also summarize the usability result based on satisfaction.

Where am I? : Position

The system clearly indicated when a user had misconceptions i.e., when the user's understanding of a concept was incorrect. The status bar was red indicating the text still had misconceptions. Incorrect sentences were the first type of misconceptions to come up after the *evaluate for conceptual change* button was pressed, so users where able to see the number of misconceptions they had right off the bat. The buttons were all red (except the current open tab), indicating their contained misconceptions had not been addressed. In addition, the preambles that said *You said* and *but scientists believe* indicated that there was a problem with the user's understanding. Furthermore, the text that comes right after *You said..*,was a highlighted sentence, clause, paragraph or word, directly from the user's text. It was highlighted in the feedback pane and in the user's text, so the user could immediately tell which of his or her sentences was incorrect.

100% of participants agreed or strongly agreed it was easy to tell if they had a conceptual problem (Q#1, Figure 6.4). 100% of the participants agree or strongly agree it is easy for them to determine the incorrect sentence in their essay (Q#3, Figure 6.4).

However, participants indicated that it was slightly more difficult to find the missing information. The grayed out button made it seem like the functionality wasn't implemented. It wasn't clear that the button was clickable. 83% of the participants agreed or strongly agreed it was easy to determine the type of misconception while the remaining 17% were unsure, but they didn't disagree (Q#2, Figure 6.4).

In conclusion, users agreed that CLICK2 could easily tell them their position, i.e., the current state of their understanding of the topic they were writing about.

Where am I going? : Destination

Users' destination, in terms of their understanding was evident from the system when one of the *incorrect* misconceptions feedback or the *missing* information feedback was active.

When the *incorrect* misconceptions feedback was active, a preamble was displayed that said *but scientists believe.* The text after this preamble stated the target or destination understanding. When the *missing* information feedback was active, a preamble was displayed that said *Scientists believe that.* After this preamble, the target understanding was presented.

66% of the participants agreed or strongly agreed it was easy for them to tell what the correct answer was to their incorrect one; 17% disagree while the remaining 17% are unsure (Q#4, Figure 6.4). Part of the reason why the positive response is at 66% and not higher is because the feedback was simulated. So, when users got past their first three misconceptions, the rest of their misconceptions didn't have adequate refutation texts.

Where to next?: Path

The sequence or path participants were advised to take when addressing their misconceptions or including missing concepts into their work was presented through the use of a prioritized list, with the priority shown through numbers. The misconception in tab #1 was supposed to be addressed before the misconception in tab #2. When asked, ”how did you determine the order in which to work on your incorrect sentences”, 100% of participants replied that they worked on their incorrect sentences based on the numbered tabs. 100% of participants agreed or strongly agreed that they understand that there is a ordering or suggested sequence in which they should address their misconceptions (Q#6, Figure 6.4).

How am I going?: Resource Recommendations

The refutation text and resource recommendations were the tools through which I showed participants how to get to the correct understanding. I hoped that by interacting with the refutation text and recommended resources, they would come to the accepted understanding.

100% of participants said it was easy for them to find information that supports the correct answer (either through the refutation text or through the recommended resources (Q#5, Figure 6.4)

100% of participants agreed or strongly agreed it was easy for them to give feedback about the recommended resource (Q#9, Figure 6.4). It is important that users are able to give feedback about how useful a resource is in helping the user get to the currently accepted scientific understanding. This, can help improve the systems’ recommendation algorithm.

Satisfaction

In general, the participants were satisfied with the design and interaction in CLICK2. 100% of participants agreed that information in the system interface (messages and button labels) were clear (Q#8, Figure 6.4). 83% of participants felt the that the system will be helpful with their writing (Q#10, Figure 6.4). 83% of participants would use CLICK2 again (Q#11, Figure 6.4). 75% of participants would use CLICK2 frequently (Q#12, Figure 6.4). 100% of participants would recommend CLICK2 to a friend (Q#13, Figure 6.4). 75% of participants said that using CLICK2 improved their understanding of seasons, the topic which they explored while using the system (Q#14, Figure 6.4). And 75% of participants liked the interface (Q#15, Figure 6.4).

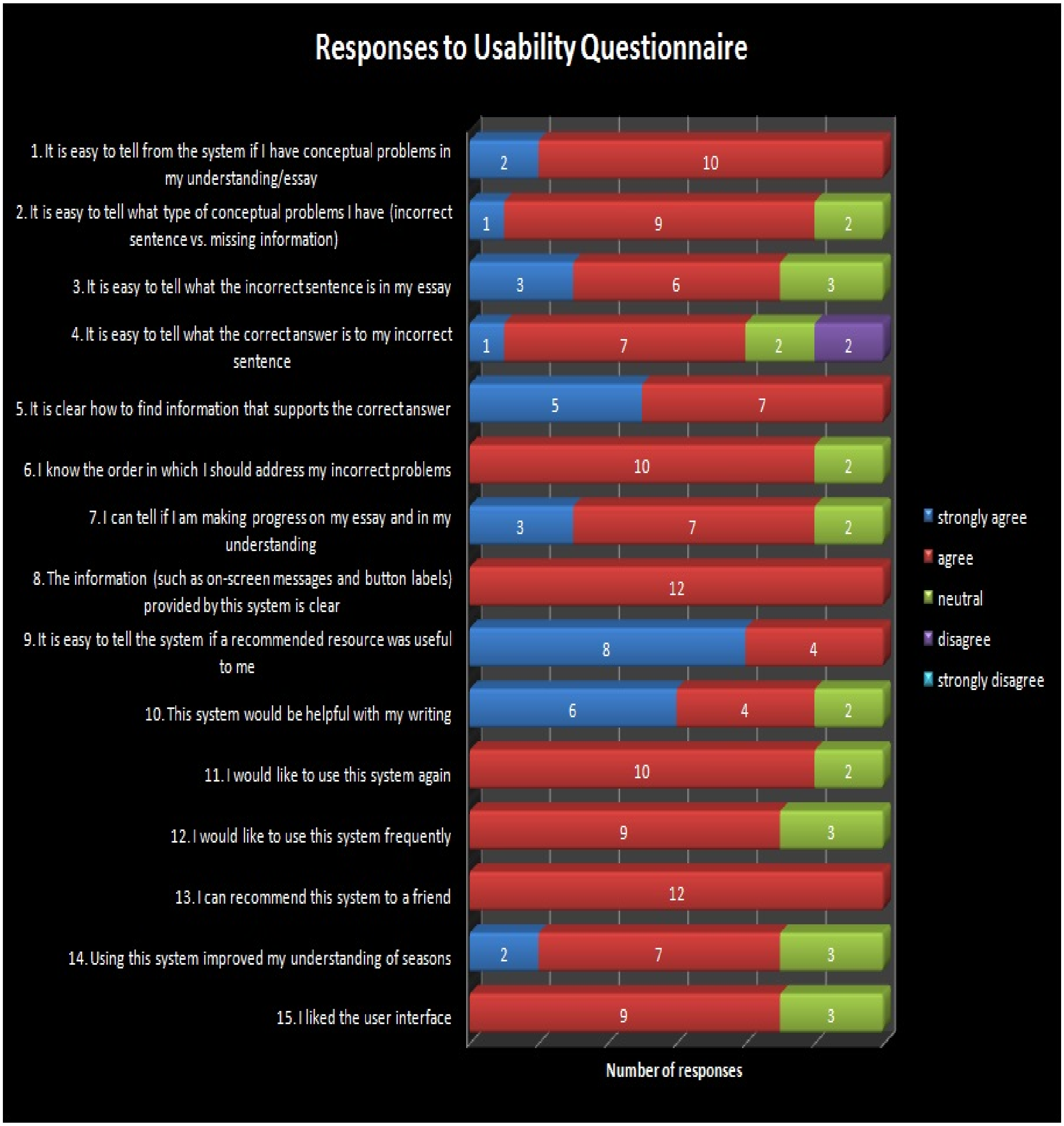


Figure 6.4: Usability Questionnaire Results

6.3 Conclusion

The CLICK2 system was created using the participatory design methodology of interactions design, while also paying heed to the mechanisms that support conceptual change. Overall, the design was good but could use some improvements.

According to users, the top four most helpful features of the final system were:

(1) The highlighted parts of their essay that showed the incorrect sentences

(2) The missing information feedback i.e., being able to see the information missing from their essay

(3) The recommended resources - being able to see what to read, interact with and watch in order to remedy their understanding without having to search for it themselves

(4) The colored tabs that told them at a high level, the status of each misconception. That is, if the misconception had been addressed, left-as-is, not yet addressed or is currently being addressed

The top four problems users had with the system were:

(1) They were not able to draw out ideas or explain their understanding through diagrams since this was a text-only environment.

(2) The missing information feedback tab was not obvious. That is, since the missing information feedback tab was grayed out at the beginning, many users didn't realize that the feature was ”implemented”.

(3) There were only three static recommended resources. Users wanted to be able to say *this resource is not useful to me in addressing this misconception* and have the system recommend another one immediately.

(4) The red status bar was distracting. Many users felt it was not necessary to have the red status bar since the tabs were already red, showing that they had problems in their essay.

Immediate improvements that can be made to the system include: changing the color of the distracting status bar or removing it entirely; changing the color of the misconception-type button [i.e., *missing* or *incorrect*] when they are inactive from grey to a different color that doesn't make people think it is not an implemented feature. Other improvements that can be made in the short term include adding functionality to recommend another resource if a user marks one as not being useful and ensuring that users can call back misconceptions that have been *left-as-is*. Long term improvements, which would involve developing new algorithms include: supporting a drawing environment, and notifying a user, via removal of the corresponding tab, that the missing information in a tab has successfully been included in the essay.

The new CLICK2 system is an effective feedback environment because it can provide answers to users for the questions *where am I?*, *where am I going?*, *where to next?* and *how am I going* in relation to their knowledge state.

In general, users were satisfied with the design and interaction of the system and would recommend it to their friends. In addition, users agreed that using CLICK2 improved their understanding of seasons, the topic they explored during the study.

Chapter 7

Correlations with Learners' Processes and Outcomes - The Learning Study

This purpose of this learning study was to understand how an educational recommender system with conceptual change support mechanisms influences learners' processes and outcomes. The research question guiding this research was: How does the educational recommender system with its conceptual change support mechanisms affect users? understanding, interest and perception of science content?

7.1 Materials

The learning study involved several materials which were developed with assistance from Holly Devaul at the University Center for Atmospheric Research (UCAR), Kirsten Butcher and Lisa Ferrara from the Education department at University of Utah and several professors in the Geology department at University of Colorado Boulder. There were seven main materials we used in this study.

(1) CLICK2

This is the educational recommender system environment that has been imbued with conceptual change support mechanisms. It is the product of the design study and it is where the learning study is run.

(2) Knowledge Perception 1 Questionnaire

This questionnaire included questions about students' perception of their knowledge of seasons in addition to contextual questions such as where the participant grew up and the participant's major in college. This questionnaire contains fifteen questions. The questionnaire is available at Appendix F

(3) Essay Question

This is the essay prompt. It can be seen at Appendix J

(4) Multiple Choice Questionnaire

This questionnaire contains multiple choice questions about seasons. This questionnaire contains fifteen questions. It is available at Appendix H

(5) Application Worksheet

This worksheet contains five short text answer questions about seasons. It is designed so that students have to apply their understanding of seasons when addressing the questions. The worksheet is available at Appendix I

(6) Knowledge Perception 2 Questionnaire

This questionnaire is different from the knowledge perception 1 questionnaire because we do not ask any contextual question here, hence we have only ten questions. We ask only for participants' knowledge perception. The wording of the knowledge perception questions are the same for both knowledge perception 1 and knowledge perception 2. It is available at Appendix G

(7) Usability Questionnaire

This questionnaire contains the twenty usability questions. The questionnaire is available at Appendix E

7.2 Participants

To be eligible for the learning study, participants had to be adult (18+) freshmen or sophomores. We limited the learning study participants to only freshmen and sophomores because we wanted to increase the probability that the subjects that we enroll in the study will still have misconceptions about seasons. The more advanced a student is in college, the greater the possibility that he/she has remedied his/her misconceptions about seasons. If the participants do not have any misconceptions, then we cannot measure the effectiveness of our educational recommender system. Fourteen students participated in the learning study, eight women and six men. They were recruited via the University of Colorado Boulder's Buff bulletin mailing list and via flyers posted at various locations on the University of Colorado Boulder campus. The participants were paid $10/hr and an additional $10 for completing both sessions of the learning study. Two students didn't attend the second session, so I am reporting on only twelve participants.

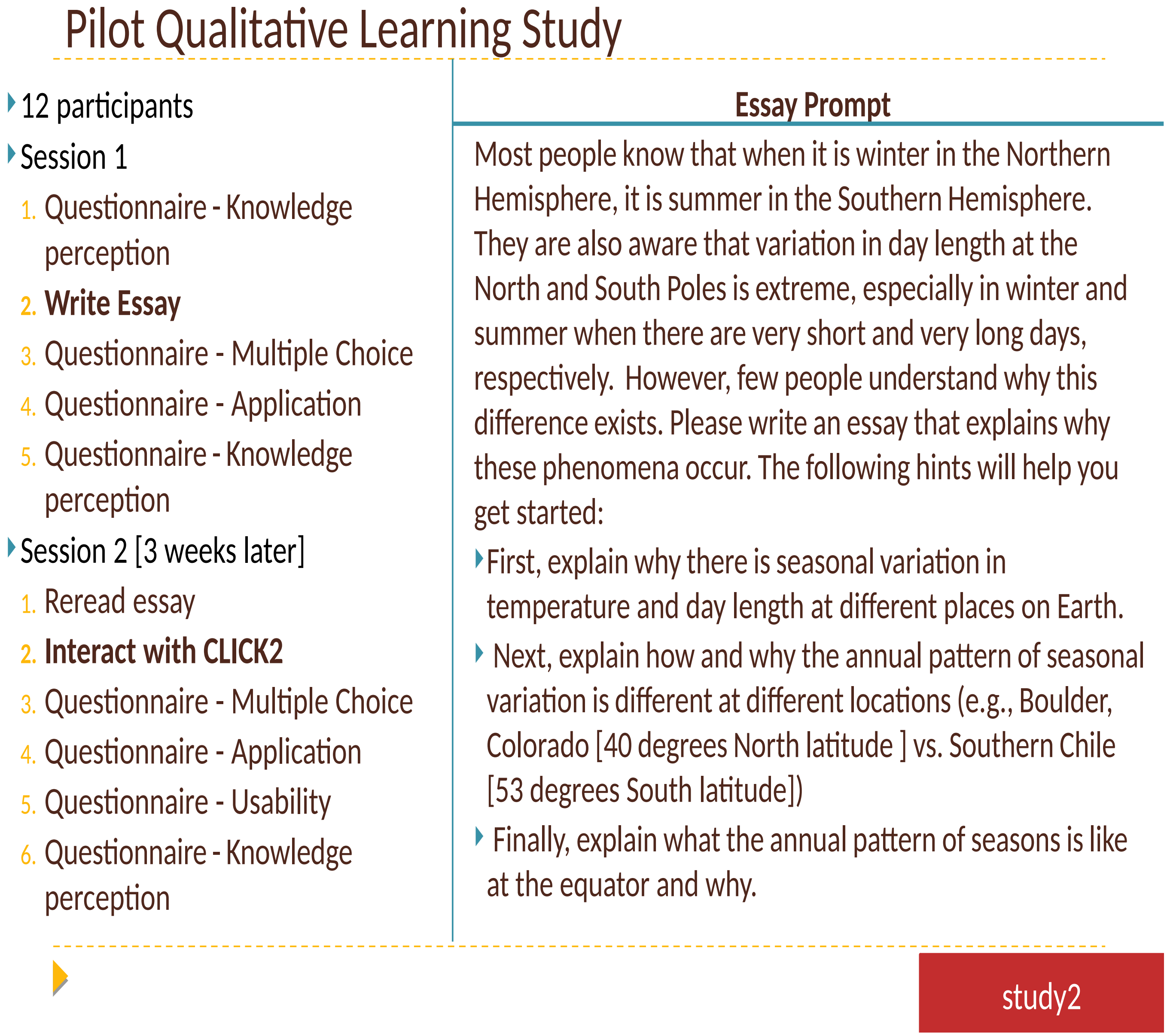
7.3 Methodology

The methodology was a learning study. It was a pilot learning study with two stages and no control, only an experimental group. The study was conducted at the Center for Innovation and Creativity, a University of Colorado at Boulder research facility. The participants were each provided with a laptop, an external mouse and a head phone because some videos were presented as recommended resources. Each stage of the study took a maximum of 90 minutes. The first stage collected information on participants' pre-existing knowledge, perception and attitudes. The second stage was designed to measure the change to these after participants interacted with CLICK2.

In the first stage, participants went through seven steps:

(1) Reading and signing the consent form

(2) Responding to the Knowledge Perception 1 questionnaire



Pilot Qualitative Learning Study

- 12 participants
- Session 1
  1. Questionnaire - Knowledge perception
  2. **Write Essay**
  3. Questionnaire - Multiple Choice
  4. Questionnaire - Application
  5. Questionnaire - Knowledge perception
- Session 2 [3 weeks later]
  1. Reread essay
  2. **Interact with CLICK2**
  3. Questionnaire - Multiple Choice
  4. Questionnaire - Application
  5. Questionnaire - Usability
  6. Questionnaire - Knowledge perception

Essay Prompt

Most people know that when it is winter in the Northern Hemisphere, it is summer in the Southern Hemisphere. They are also aware that variation in day length at the North and South Poles is extreme, especially in winter and summer when there are very short and very long days, respectively. However, few people understand why this difference exists. Please write an essay that explains why these phenomena occur. The following hints will help you get started:

- First, explain why there is seasonal variation in temperature and day length at different places on Earth.
- Next, explain how and why the annual pattern of seasonal variation is different at different locations (e.g., Boulder, Colorado [40 degrees North latitude] vs. Southern Chile [53 degrees South latitude])
- Finally, explain what the annual pattern of seasons is like at the equator and why.

study2

Figure 7.1: Setup of Learning Study

(3) Writing a response to the Essay prompt

(4) Addressing the multiple choice questions

(5) Working on the Application Questions worksheet

(6) Responding to the Knowledge Perception 2 questionnaire

(7) Debriefing and Payment

During the debrief session in stage 1, participants were asked to not read or lookup information on the questions they were asked in stage 1. I emphasized they needed to abide by this in order for us to maintain the integrity of the study results.I also reiterated that the only way I can create an essay-based educational recommender system useful to them is if they abide by this rule.

In the second stage, which occurred three weeks after the first stage, participants went through an additional seven steps, which were:

(1) Rereading the essay written in stage 1

(2) Interacting with CLICK2 to address at least three incorrect and missing information in the essay

(3) Addressing the multiple choice questions

(4) Working on the Application Questions worksheet

(5) Answering the Usability questions

(6) Responding to the knowledge Perception 2 questionnaire

(7) Debrief and Payment

The second stage occurred three weeks after the first stage to give us time to simulate the feedback from the educational recommender system.

7.3.1 Algorithm Simulation

We simulated the algorithms in the educational recommender system in order to provide the best possible responses back to the participants. We wanted to ensure that our evaluation of the system was not limited by the quality of the algorithms. We had two Earth science teachers simulate the algorithms in the educational recommender system. Given a list of important concepts in Earth science and a student's essay, both experts did the following:

(1) Identified the misconceptions in the essay

(2) Commented on each misconception

(3) Prioritized the misconceptions

(4) Created refutation texts for the top three misconceptions.

(5) Identified the important concepts that were missing from the essay

Using pre-identified resources from the Digital Library for Earth System Education (DLESE), for each essay, I identified three resources that contained the knowledge to refute and remedy the top three identified misconceptions and three resources that could provide the missing information in the top three missing concepts, ensuring that the first recommended resource was textual with a diagram within in, the second was a simulation and the third was a short video about the topic.

Each participant had two sets of feedback from the two annotators. However, during presentation of the feedback, initially, I showed the participant the feedback from only one annotator, not both, although the annotator whose feedback was chosen first, varied for the different participants. If the participant then clicked on the button to get more misconceptions, I displayed the other misconceptions that the second annotator had uncovered, providing they still existed in the edited essay. I did this because it was not clear how I could merge the two misconceptions detection and priority list that both annotators had created.

Participants first feedback came from the annotator that found the most misconceptions in the participant's essay.

Only the top three misconceptions from the first annotator had refutation text attached to it. For the rest of the misconceptions, I displayed the comments about the misconception that the annotator made. I cycled among three prompts to prompt the participant to address the misconception. I also cycled among three prompts for the missing conceptions prompt. In addition, the recommended resources for the misconceptions and missing concepts that were not in the top three, were randomly selected from the list of twenty resources, while ensuring that the first recommended resource was textual with a diagram within in, the second was a simulation and the third was a short video about the topic.

7.4 Evaluation

Although I collected a series of data during the learning study, due to time constraints and this being a pilot, I did not do an extensive evaluation on the data. As stated earlier, my goal for this study was to understand how the use of CLICK2 affected students' interest, perception of understanding and actual understanding.

Category: interest in the topic of seasons	**Category: perception of knowledge about a topic**	**Numerical Value**
Not at all interested	Not at all confident	1
A little interested	A little confident	2
Moderately interested	Moderately confident	3
Very interested	Very confident	4
Extremely interested	Extremely confident	5

Table 7.1: Conversion of self-report from categorical to numerical data

To evaluate interest, I analyzed the responses to question 6 in the *Knowledge Perception* questionnaire. Table 7.1 shows how I converted the textual response to a score. As Figure 7.2 shows, participants' general interest in the topic of seasons started out *low* at the start of session 1. Their interest increased to *moderate* by the end of session 1 and stayed the same

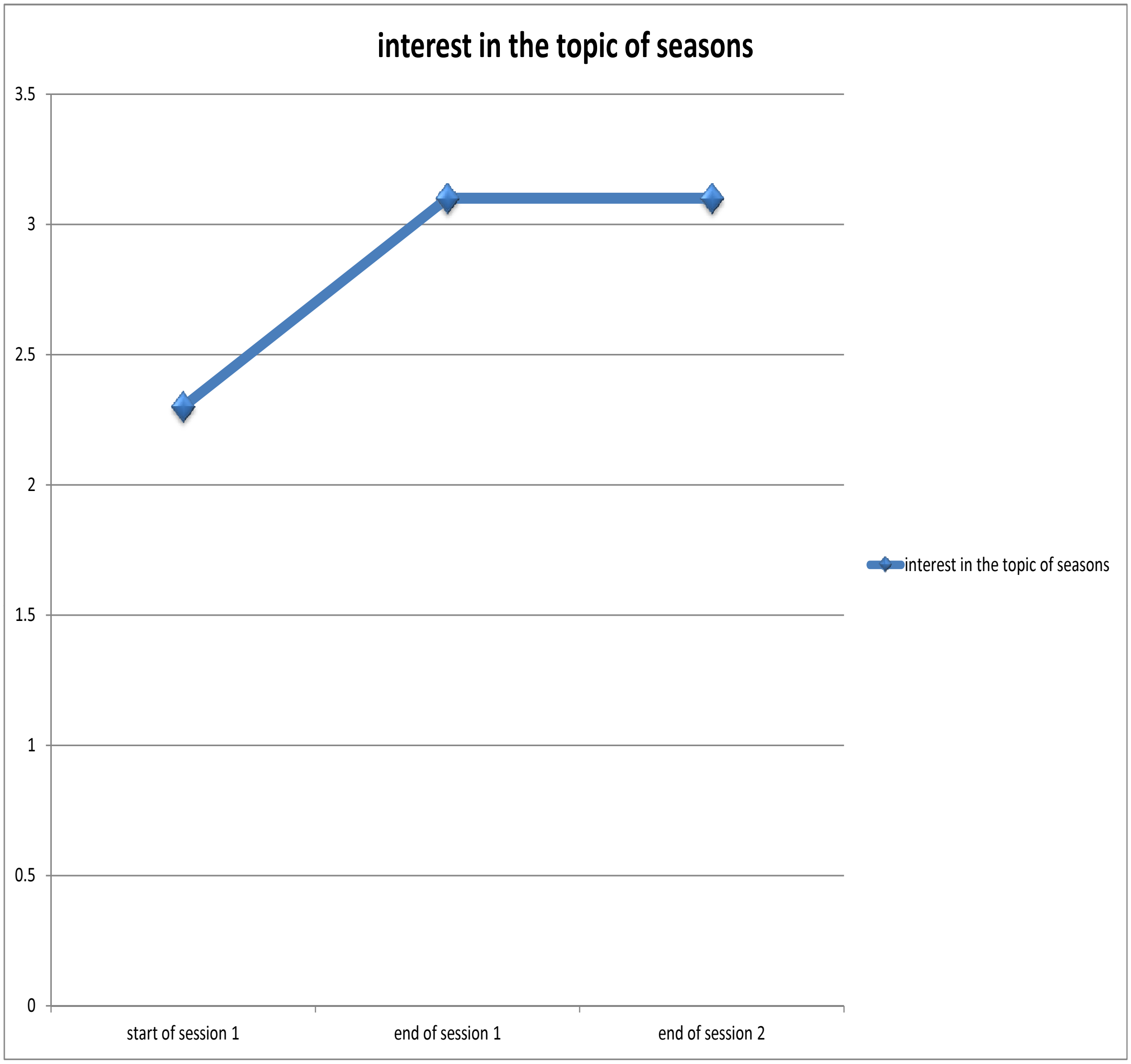


Figure 7.2: Interest trend across sessions

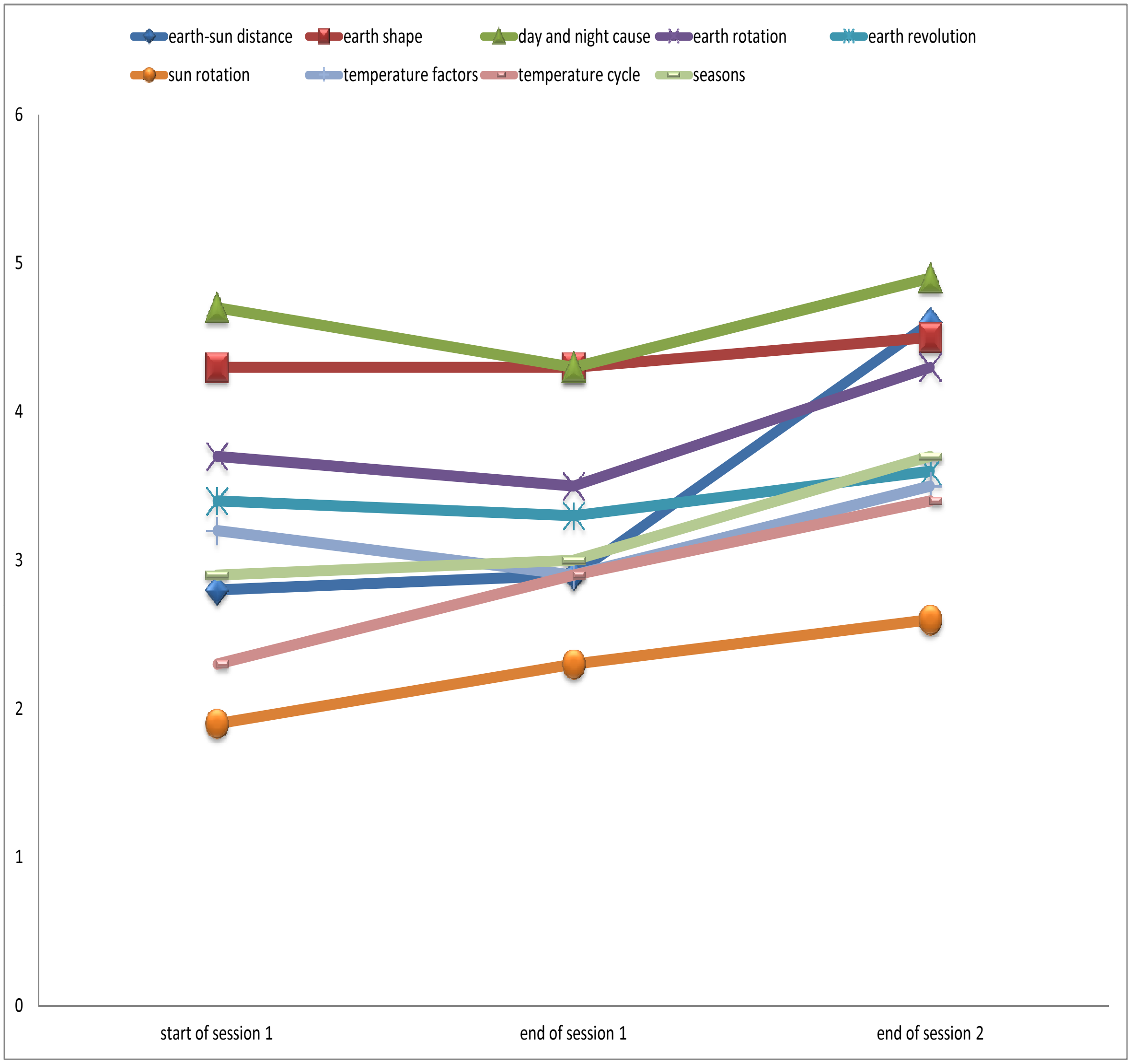


Figure 7.3: Knowledge perception trend across sessions

through the end of session2. Their low interest at the start of session 1 could be explained by the fact that these participants hadn't chosen an Earth science related major in college, so they were most likely people that did not had a strong interest in Earth science, and thereby, people that didn't have a strong interest in the phenomena of seasons. Faced with technical questions about seasons during session 1, I projected that interest would wane, and was pleasantly surprised to see that asking technical questions increased participants' interest in knowing more about the phenomena of seasons by the end of session 1. I speculated that interest would remain the same from end of session1 through the end of session2, and that was the case.

To evaluate perception of understanding, I analyzed the responses to all but question 6 in the *Knowledge Perception* questionnaire. Table 7.1 shows how I converted the textual response to a score. Taking participants initial confidence at the start of session 1 as the baseline, I expected participants' perception of their understanding to decrease by the end of session 1 after being faced with technical questions that they couldn't fully address. I expected their confidence to surpass the baseline at the end of session 2. I expected to see this increase in confidence after participants had used the environment to address their incorrect and missing concepts. I theorized that the feedback from the environment, saying that participants had dealt with their misconceptions would make the participants believe they understood the phenomena of seasons very well, thereby giving their confidence a huge boost.

As Figure 7.3 shows, participants' confidence in their knowledge of seasons started out at different points for different topics in seasons. But in general, participants' confidence in their knowledge declined by the end of session 1. However, confidence in their knowledge of seasons increased and surpassed their baseline confidence by the end of session 2.

To evaluate actual understanding, I analyzed the multiple choice questionnaires and essays. As Figure 7.4 shows, in general participants' knowledge of the phenomena of seasons increased after using CLICK2 as measured by the multiple choice questionnaire. I did a further analysis of the results from the multiple choice questionnaire by grouping the questions into four categories, which represent the main concepts about the Earth and seasons that the questions addressed. They are; (1)path, (2) tilt and pole, (3) daylight and (4) energy. All participants showed improvement in all four categories or stayed the same - in those cases they answered 100% correctly on the pretest.

Results for questions 1, 8, 12, 14 and 15, which were about the earth's path can be seen in Figure 7.5. Results for questions 2, 3, 5 and 11, which were concerned with the earth's tilt and seasons at the poles can be seen in Figure 7.6. Figure 7.7 shows the results for questions 4, 6, 9 and 10 which were concerned with *hours of daylight at different parts of the world in different seasons*. While Figure 7.8 shows the results for questions 7 and 13 which dealt with *energy*.

As Table 7.2 shows, I analyzed three students' pre-CLICK2 and post-CLICK2 essays out of a total of twelve participants. I chose these three because they represent the least, mean and highest number of misconceptions in the set. In addition, I also chose the essays where both annotators agreed on the number of misconceptions, even if they did not agree on the same misconceptions.

Table 7.2: Data information for analyzed essays

user	**number of identified misconceptions**
user_1	2
user_5	4
user_11	6

user_1 had only two misconceptions and they were the same from both annotators. user_1 addressed both misconceptions and also did well on the multiple choice questionnaire. It was clear to me from reading the original and changed misconception sentences that this participant had a very good understanding of the topic of seasons. This participant wanted to know where he or she ranked amongst all the participants and wanted to see the result of the multiple choice questionnaires immediately upon completion. This participant was also

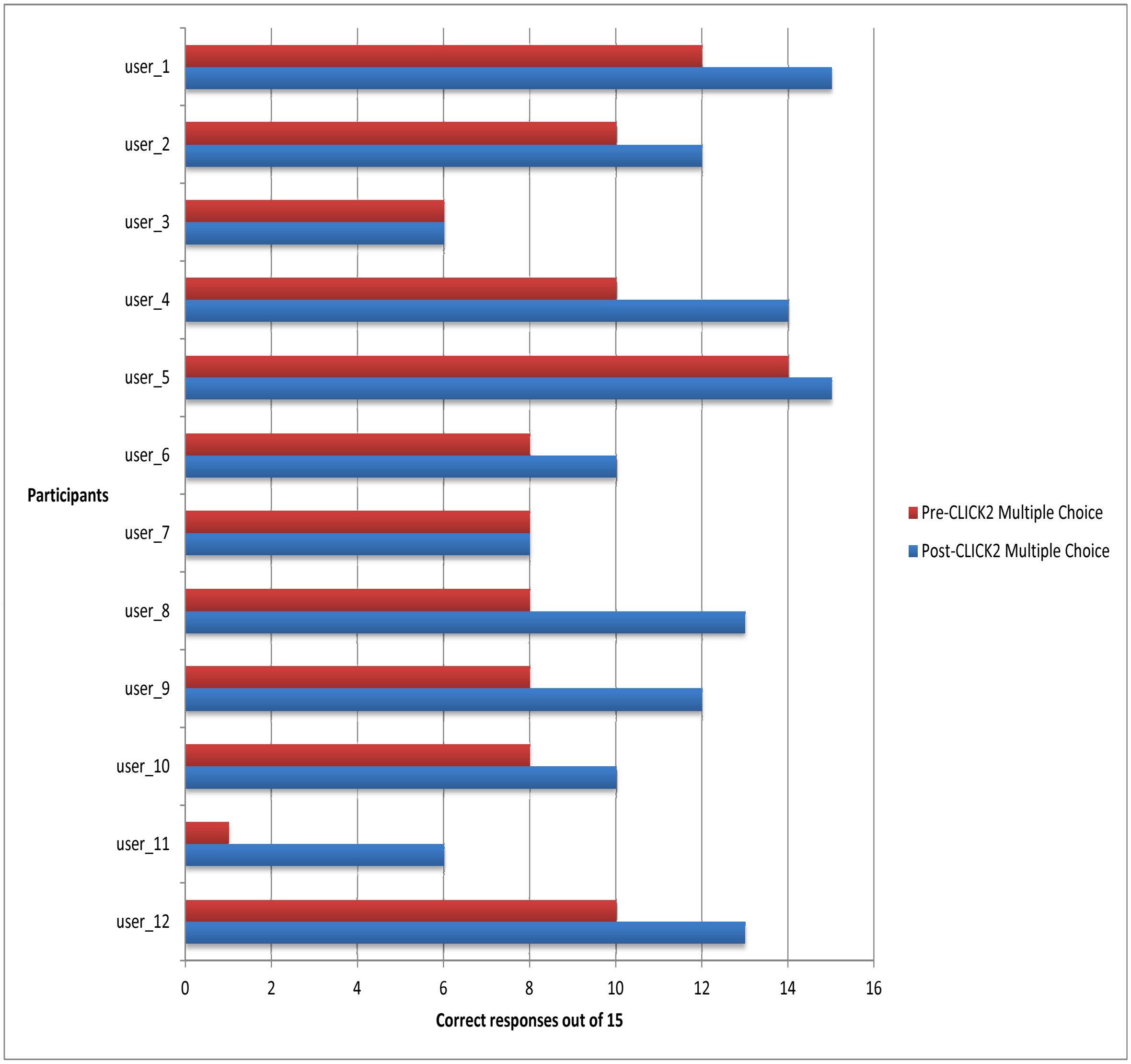


Figure 7.4: Knowledge measured by multiple choice questionnaire

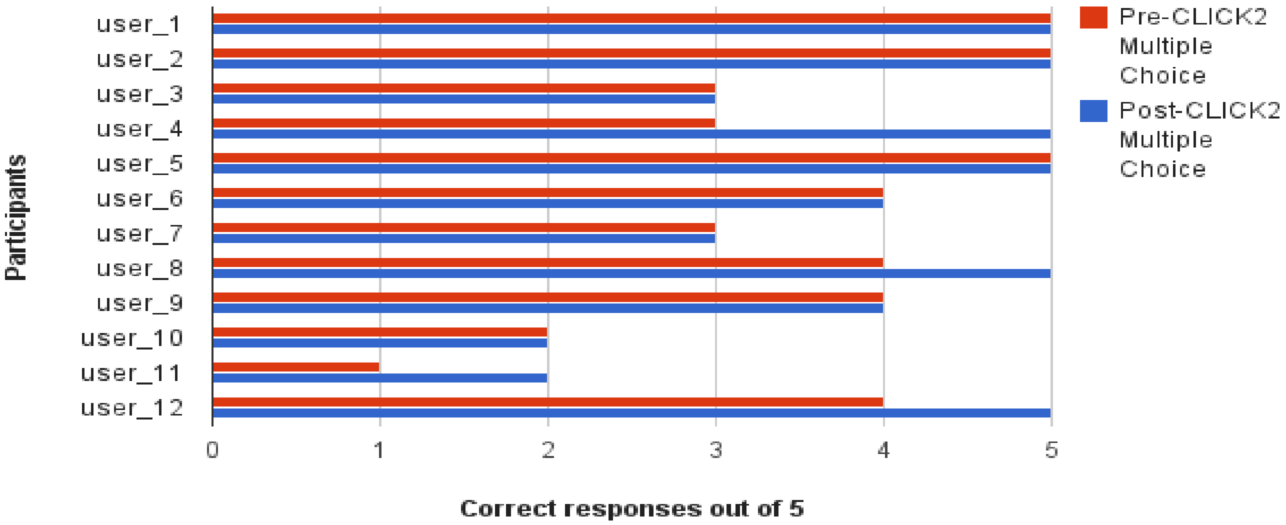


Figure 7.5: Knowledge of "path" measured by multiple choice questionnaire

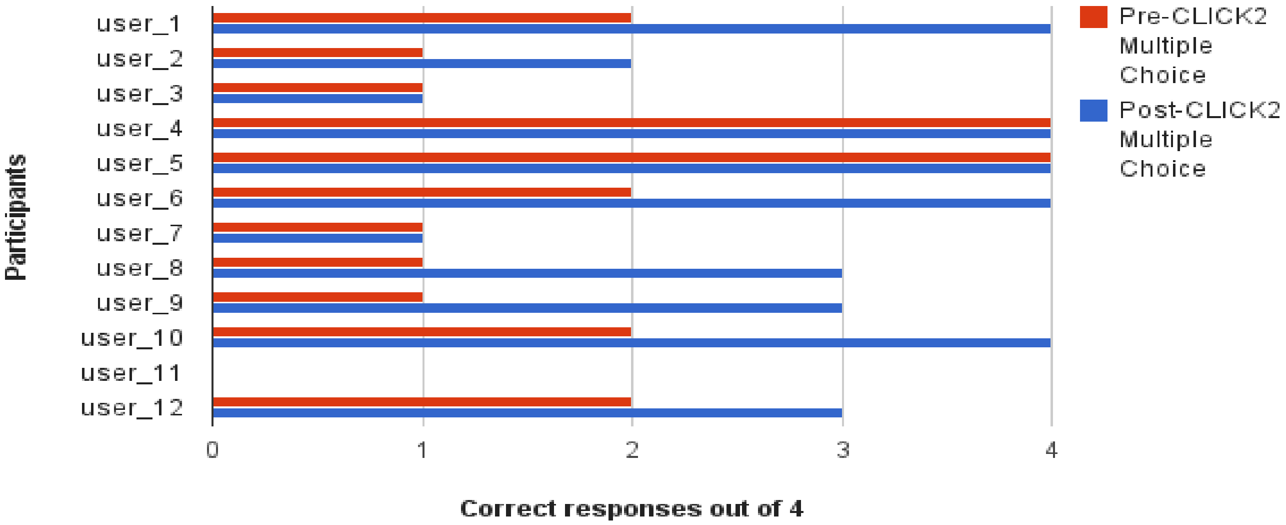


Figure 7.6: Knowledge of "tilt and pole" measured by multiple choice questionnaire

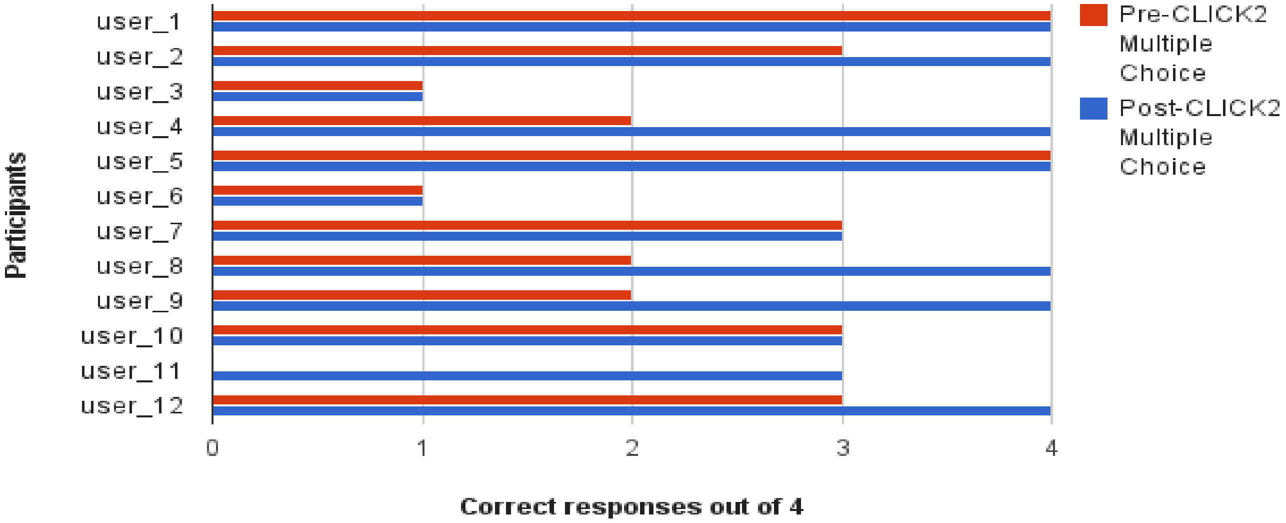


Figure 7.7: Knowledge of ”daylight” measured by multiple choice questionnaire

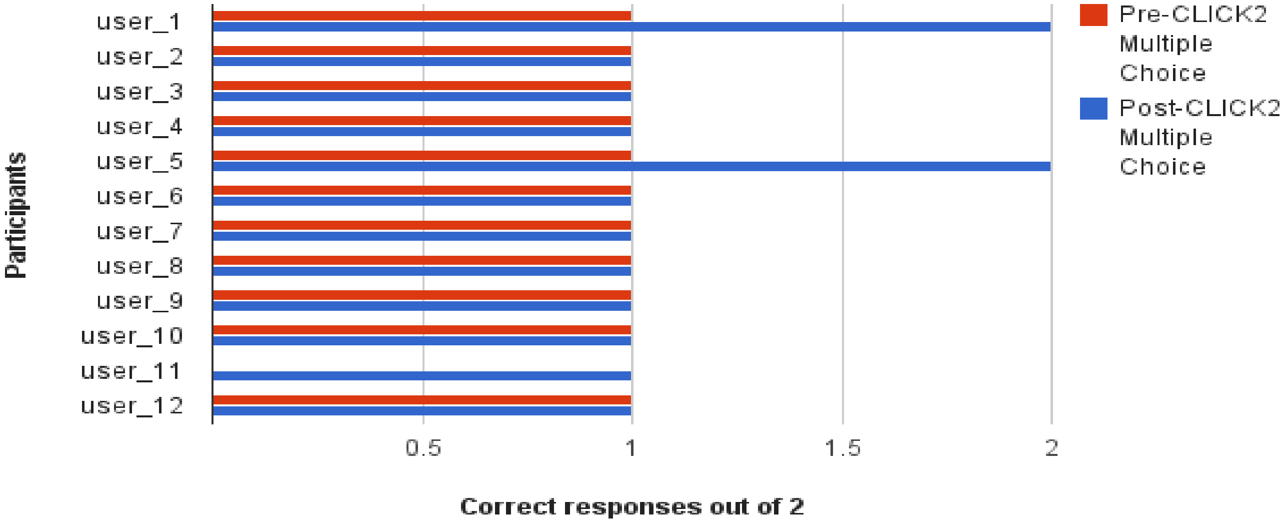


Figure 7.8: Knowledge of ”energy transfer” measured by multiple choice questionnaire

very interested in keeping up with the results of this study and wanted to know how she could pilot the system when the real algorithms were running within.

For user_5, four misconceptions were identified by the first annotator and four by the second annotator. Both annotators had two overlapping misconceptions, so this user really had six misconceptions. user_5 addressed the four misconceptions that were displayed, even though the instructions had them only addressing three. In addition, this participant went on to address the other misconceptions from the second annotator. user_5 was the only participant that addressed any of the missing conceptions.

user_11 did the minimum asked by the learning study instructions. user_11 addressed only the first three misconceptions from the first annotator, even though this participant had six misconceptions from the first annotator. Nevertheless, this participant did show an improvement in understanding between the first and second essays. The multiple choice results in Figure 7.4 also show that the user's understanding did improve, even though this participant still harbored many misconceptions at the end of the intervention. This is not unexpected as this participant expressed *no interest* in the topic of seasons on the self-report on interest within the knowledge perception questionnaires.

For all three users, the changed sentences coherently reflected the correct scientific understanding.

7.5 Conclusion

In this chapter, I sought to understand how the use of CLICK2 affects users' outcomes. I used a pilot learning study with twelve participants and analyzed several questionnaires and essays. The analysis produced strong indications that CLICK2 can improve users' interest, perception of understanding and actual understanding of the topic of seasons. Figure 7.2 shows that interest improved between the start of session1 to the end of session 2. Figure 7.3 shows that perception of understanding or confidence in knowledge did decrease from the start of session 1 to the end of session 1, but it went up past the initial baseline through the end of session 2. Figure 7.4 and the preceding analysis of the essays showed that actual understanding of the topic of seasons also improved.

A deeper analysis of the data and a follow-up study with a larger number of participants over a longer period of time, will undoubtedly yield more answers and questions about the impact of CLICK2 on learners' outcomes. Some remaining questions that future studies might be able to address include:

- Is there a statistically significant change in understanding between participants that use CLICK2 and those that do not?
- Does the interest, perception and understanding of the topic that we saw, remain steady, wane or increase over time?
- How can we discourage paraphrasing the refutation and resource text and instead encourage the creation of a text that reflects participants' understanding?

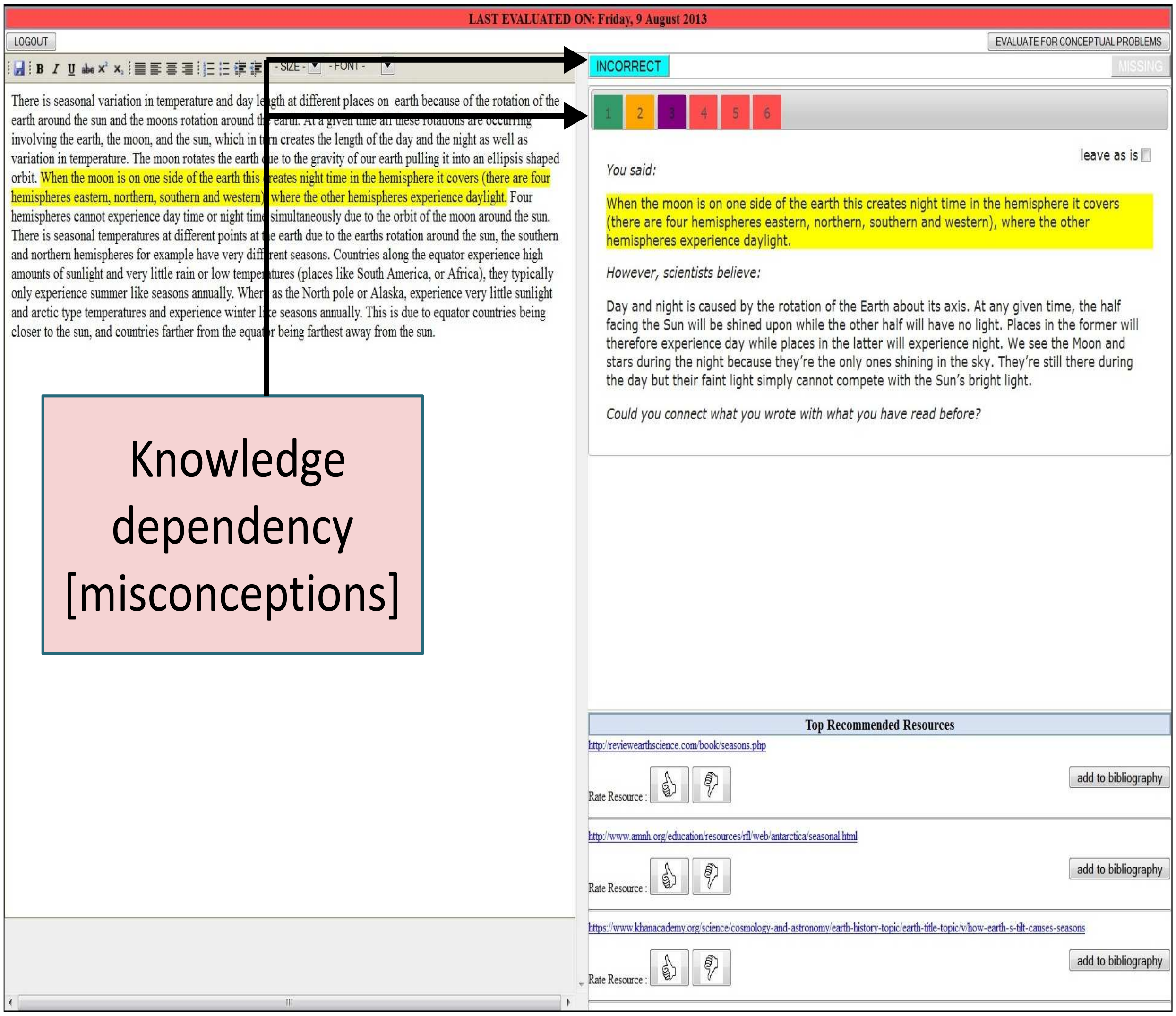
Chapter 8

Automatic Prioritization of Student Misconceptions

The purpose of this study was to automate the discovery of knowledge dependencies between a student's incorrect sentences or misconceptions. Figure 8.1 shows the knowledge dependency between incorrect sentences (misconceptions) in the CLICK2 interface. In Figure 8.1, the **Incorrect** tab is active; it shows the six misconceptions in the essay. The misconceptions are prioritized, which means, misconception 1 (in green) should be remedied before misconception 2 (in orange), which should be remedied before misconception 3 (in purple) and so on. This study is important because conceptual change learning theory advocates constructing a pedagogical sequence of knowledge i.e., acknowledging knowledge dependency between concepts, and by extension, misconceptions. Doing so automatically is crucial for an online learner-directed on-demand learning environment.

I automate the discovery of knowledge dependencies between misconceptions by creating an algorithm that automatically prioritizes student misconceptions in the form of sentences. A misconception is a false belief, flawed mental model, category mistake or missing schema that is usually based on faulty understanding of some knowledge. Examples of misconceptions in Earth science are: **the Earth is flat**, **the reason we have seasons is due to the distance of the Earth from the sun, when the sun is closer to the Earth, we have summer and when it is farthest away, we have winter**.

Misconceptions can be difficult to change if there is no consequence to having them. That is, if an individual can function in a society with the misconception, then there is usually no incentive for the individual to correct the misconception. A lot of science misconceptions also persist because they sound reasonable and it is not easy to reach the correct understanding without resorting to manipulating simulations and viewing pictures and videos.



There is seasonal variation in temperature and day length at different places on earth because of the rotation of the earth around the sun and the moons rotation around the earth. At a given time all these rotations are occurring involving the earth, the moon, and the sun, which in turn creates the length of the day and the night as well as variation in temperature. The moon rotates the earth due to the gravity of our earth pulling it into an ellipsis shaped orbit. When the moon is on one side of the earth this creates night time in the hemisphere it covers (there are four hemispheres eastern, northern, southern and western), where the other hemispheres experience daylight. Four hemispheres cannot experience day time or night time simultaneously due to the orbit of the moon around the sun. There is seasonal temperatures at different points at the earth due to the earths rotation around the sun, the southern and northern hemispheres for example have very different seasons. Countries along the equator experience high amounts of sunlight and very little rain or low temperatures (places like South America, or Africa), they typically only experience summer like seasons annually. Where as the North pole or Alaska, experience very little sunlight and arctic type temperatures and experience winter like seasons annually. This is due to equator countries being closer to the sun, and countries farther from the equator being farthest away from the sun.

You said:

When the moon is on one side of the earth this creates night time in the hemisphere it covers (there are four hemispheres eastern, northern, southern and western), where the other hemispheres experience daylight.

However, scientists believe:

Day and night is caused by the rotation of the Earth about its axis. At any given time, the half facing the Sun will be shined upon while the other half will have no light. Places in the former will therefore experience day while places in the latter will experience night. We see the Moon and stars during the night because they're the only ones shining in the sky. They're still there during the day but their faint light simply cannot compete with the Sun's bright light.

Could you connect what you wrote with what you have read before?

Top Recommended Resources

http://reviewearthscience.com/book/seasons.php

Rate Resource :

http://www.amnh.org/education/resources/rfl/web/antarctica/seasonal.html

Rate Resource :

https://www.khanacademy.org/science/cosmology-and-astronomy/earth-history-topic/earth-title-topic/v/how-earth-s-tilt-causes-seasons

Rate Resource :

Knowledge dependency [misconceptions]

Figure 8.1: CLICK2 environment showing knowledge dependencies for misconceptions

8.1 Materials

The data for this study came from a learning study that I ran at CU-Boulder using CLICK2. CLICK2 is described in more detail in chapter 6 while the learning study is described in more detail in chapter 7. During the learning study, the twelve participants were asked to write an essay on seasons. Then the essays were given to two subject experts to evaluate and annotate. Their annotations became the data set for this study.

Although CLICK2 has a misconception identification module, I chose to have human experts identify the misconceptions because I wanted to ensure the validity of the identified misconceptions so as to build the misconception prioritization model using correct data. In section 8.2.1 below, I discuss how we get this misconception data from the essays.

8.2 Methodology

My approach to automatically prioritizing students' misconceptions was to understand and model Earth science teachers' processes for prioritizing student misconceptions, using supervised machine learning classifiers. To break this down into a tractable task, I cast this as a pair-wise ordering problem i.e., rather than focusing on trying to automatically generate entire pedagogical sequences for the misconceptions in an essay, I focus on developing a model capable of identifying when one misconception precedes (should be remedied before) another.

8.2.1 Producing the human expert evaluation set

Twelve students each wrote an essay on seasons. Then I asked an Earth science teacher and an Earth science curriculum developer to double annotate all 12 essays. For each essay, the annotators:

(1) Identified up to ten misconceptions by sentence

(2) Identified the misconceptions by snippet. A snippet is defined as the smallest level of knowledge needed to identify the misconception

(3) Identified if the snippet needed context (i.e., the entire sentence or paragraph) in order to judge its' validity

(4) Commented on the misconception by explaining why it is a misconception or simply stating a fact that refutes the misconception

(5) Aligned each misconception to one of thirteen middle and high school season concepts which can be seen in Table 8.1. These thirteen concepts came from the AAAS progression map for learning the topic of seasons,

(6) Assigned a priority to the misconceptions without necessarily producing a full ordering i.e., two misconceptions can have the same priority

(7) Commented on the priority by explaining the decision to give a misconception a higher priority over another

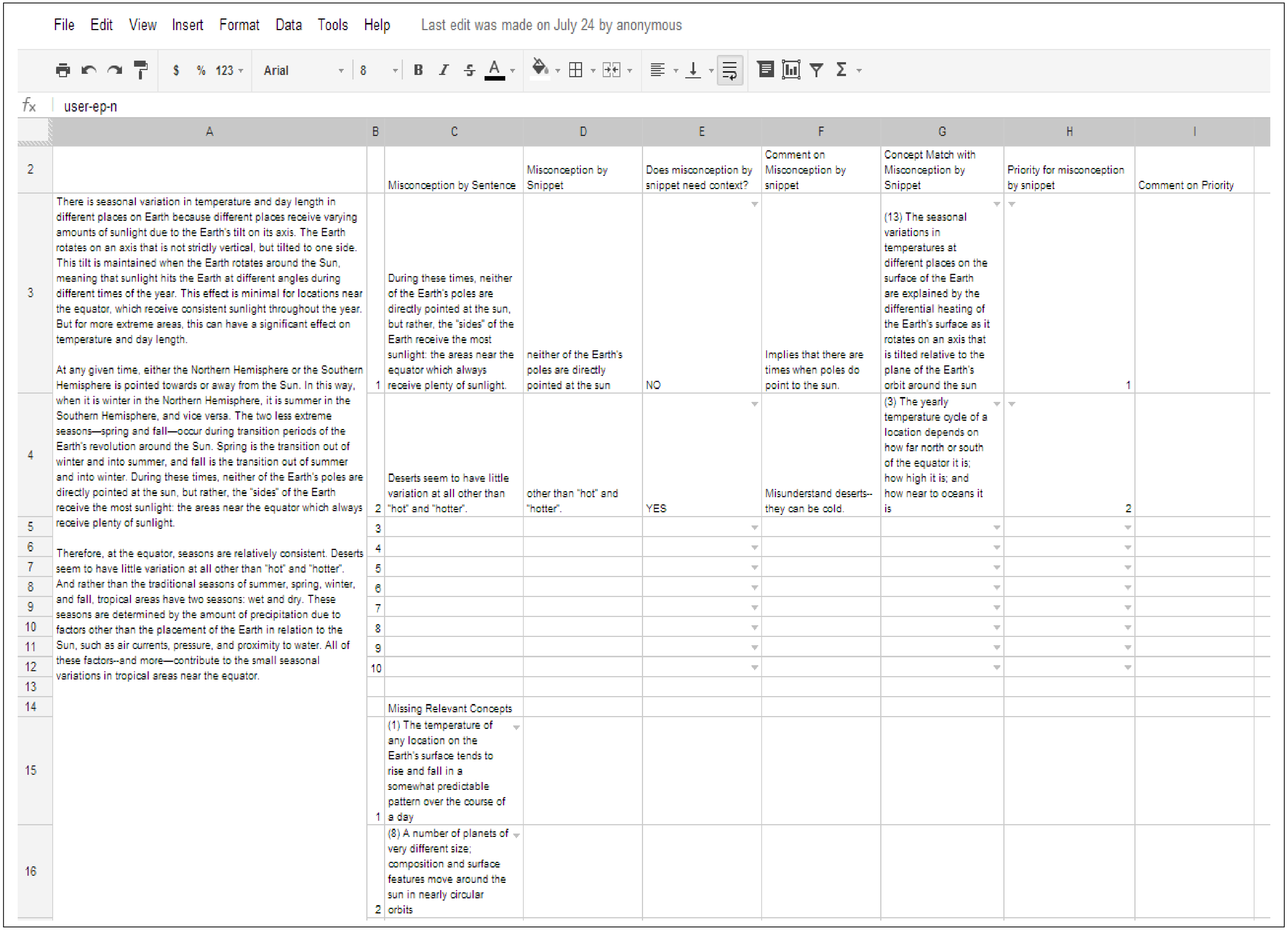
(8) Identified concepts from the thirteen AAAS middle and high school season concepts which were missing from the essay

Figure 8.2 shows an example of an annotation. Because I had only 12 essays, I did not try to create consensus on the identified misconceptions and prioritization. Another reason for not creating consensus in the prioritization is that there are many instructionally sound paths students could take while remedying their misconceptions. There is no reason to think that we have to take the consensus between experts as being the best. I consider each experts' prioritization scheme valid on its' own.

The two data points generated from the annotation in Figure 8.2 are: Label-1 for mis1-mis2, meaning misconception-1 has precedence over misconception-2. And Label-0 for mis2-mis1, which means misconception 2 does not have precedence over misconception 1. I

#	**Core learning goals about seasons**
(1)	The temperature of any location on the Earth?s surface tends to rise and fall in a somewhat predictable pattern over the course of a day
(2)	The temperature of any location on the Earth?s surface tends to rise and fall in a somewhat predictable cycle over the course of a year
(3)	The yearly temperature cycle of a location depends on how far north or south of the equator it is; how high it is; and how near to oceans it is
(4)	Light and electromagnetic waves can warm objects. How much an object?s temperature increases depends on how intense the light striking its surface is; how long it shines on the object; and how much of the light is absorbed,
(5)	The intensity of the sunlight striking a place on the surface of the Earth varies depending on what time of day it is; what time of year it is; and on how far north or south of the equator the place is
(6)	The number of hours of daytime or nighttime a location on the Earth?s surface gets varies in a predictable pattern over the course of the year and that depends upon how far north or south of the equator they are
(7)	The temperature of a location on the surface of the Earth depends upon the number of hours of sunlight and the intensity of that sunlight
(8)	A number of planets of very different size; composition and surface features move around the sun in nearly circular orbits
(9)	The axis of the Earth?s rotation is tilted relative to the plane of the Earth?s yearly orbit around the sun. As the Earth orbits the sun; the axis remains pointed to the same place in space
(10)	Because the Earth is a sphere; at any particular time; light from the sun strikes different parts of the Earth at different angles and therefore the intensity of light striking the surface of the Earth is different in different places
(11)	The difference in how much of the day is daytime and how much is nighttime at a place on the surface of the Earth depends upon where the Earth is in its yearly orbit around the sun and how far the place is from the equator
(12)	The intensity of sunlight striking a place on the surface of the Earth depends upon where the Earth is in its yearly orbit around the sun and how far the place is from the equator
(13)	The seasonal variations in temperatures at different places on the surface of the Earth are explained by the differential heating of the Earth?s surface as it rotates on an axis that is tilted relative to the plane of the Earth?s orbit around the sun

Table 8.1: 13 middle and high school *seasons* core learning goals



File Edit View Insert Format Data Tools Help
Last edit was made on July 24 by anonymous
$ % 123
Arial
8
B I S A
fx

user-ep-n

	A	B	C	D	E	F	G	H	I
2			Misconception by Sentence	Misconception by Snippet	Does misconception by snippet need context?	Comment on Misconception by snippet	Concept Match with Misconception by Snippet	Priority for misconception by snippet	Comment on Priority
3	There is seasonal variation in temperature and day length in different places on Earth because different places receive varying amounts of sunlight due to the Earth's tilt on its axis. The Earth rotates on an axis that is not strictly vertical, but tilted to one side. This tilt is maintained when the Earth rotates around the Sun, meaning that sunlight hits the Earth at different angles during different times of the year. This effect is minimal for locations near the equator, which receive consistent sunlight throughout the year. But for more extreme areas, this can have a significant effect on temperature and day length. At any given time, either the Northern Hemisphere or the Southern Hemisphere is pointed towards or away from the Sun. In this way, when it is winter in the Northern Hemisphere, it is summer in the Southern Hemisphere, and vice versa. The two less extreme seasons—spring and fall—occur during transition periods of the Earth's revolution around the Sun. Spring is the transition out of winter and into summer, and fall is the transition out of summer and into winter. During these times, neither of the Earth's poles are directly pointed at the sun, but rather, the "sides" of the Earth receive the most sunlight: the areas near the equator which always receive plenty of sunlight. Therefore, at the equator, seasons are relatively consistent. Deserts seem to have little variation at all other than "hot" and "hotter". And rather than the traditional seasons of summer, spring, winter, and fall, tropical areas have two seasons: wet and dry. These seasons are determined by the amount of precipitation due to factors other than the placement of the Earth in relation to the Sun, such as air currents, pressure, and proximity to water. All of these factors--and more—contribute to the small seasonal variations in tropical areas near the equator.	1	During these times, neither of the Earth's poles are directly pointed at the sun, but rather, the "sides" of the Earth receive the most sunlight: the areas near the equator which always receive plenty of sunlight.	neither of the Earth's poles are directly pointed at the sun	NO	Implies that there are times when poles do point to the sun.	(13) The seasonal variations in temperatures at different places on the surface of the Earth are explained by the differential heating of the Earth's surface as it rotates on an axis that is tilted relative to the plane of the Earth's orbit around the sun	1	
4		2	Deserts seem to have little variation at all other than "hot" and "hotter".	other than "hot" and "hotter".	YES	Misunderstand deserts-- they can be cold.	(3) The yearly temperature cycle of a location depends on how far north or south of the equator it is; how high it is; and how near to oceans it is	2	
5		3							
6		4							
7		5							
8		6							
9		7							
10		8							
11		9							
12		10							
13									
14			Missing Relevant Concepts						
15		1	(1) The temperature of any location on the Earth's surface tends to rise and fall in a somewhat predictable pattern over the course of a day						
16		2	(8) A number of planets of very different size; composition and surface features move around the sun in nearly circular orbits						

Figure 8.2: Sample Essay Annotation by Subject Experts

generated similar data points from all the annotations that were generated by the experts, resulting in the data set I used for this study.

8.2.2 Features

In machine learning, a feature is an individual measurable property of a phenomenon being observed (Bishop et al., 2006). It is the measurable characteristics or attributes of a phenomenon. For example, the features of a loan applicant might include age, job type, salary, outstanding loans and history of payments on outstanding loans. A machine learning model is created by combining the features in different ways. For example, one model that determines if a loan applicant should be granted a loan might decide to give more negative weight to *history of payments on outstanding loands* than to *outstanding loans*. And more positive weight to *salary* and than to *age*. Another model might not penalize a loan applicant for having preexisting loans if the loan was for education purposes. Part of optimizing machine learning algorithms for specific tasks involves feature selection i.e., selecting a subset of features that can optimally describe the phenomenon.

The four features I used in training the misconception prioritization models are as follows:

(1) Alignment to learning goals that are in a pedagogical sequence [concept-match]

For example if a student has two misconceptions, a and b and they are aligned accordingly to two of the learning goals, L1 and L2 where L2 comes before L1 in the pedagogical sequence, then, b should have a higher priority than a. This was a binary feature, with the value 1 meaning there is a precedence relationship between the misconceptions (i.e., they were aligned to learning goals such as L1 and L2, which have a precedence relationship), while 0 means they were either aligned to the same learning goal or to learning goals that did not have a precedence relationship.

(2) Relative Position where the misconceptions occurred in the students' work [Position]

It might be judicious to prioritize misconceptions that occur earlier in the student's work than later. The reason being that the earlier misconception might have given rise to the later misconception. For coherency sake, it could also be better to prioritize earlier misconceptions so that students can remedy the misconceptions in their essays starting from the top of the essay and going down, rather than jumping around in the essay. This was a binary feature too. So instead of giving the actual sentence number where the misconception occurred as the value of the feature, I normalized that value to 0 or 1. Given a pair of sentences, regardless of when they occurred in the essay, the first sentence to occur got the value of 0, while the second got the value of 1 for this feature.

(3) Sentence Complexity [tf-idf]

tf-idf stands for term frequency-inverse document frequency. It is a measure that is used in computational linguistics to show how important a word is to a particular document (sentence) in a corpus. For this study, term frequency was how many times the word occurred in the misconception sentence. While the document frequency was the number of times the word occurred in the corpus of 20 documents that was used for the extraction study in Chapter 9. In this study, I used tf-idf to approximate sentence complexity. I posit that the more rare a word is, the more difficult the word is and consequently, the more complex the misconception which uses the word. Following this, I theorized that it is better to remedy misconceptions containing simple words first, before misconceptions containing complex words. Hence being able to get the tf-idf of a misconception might help the machine learning classifier.

(4) Word length

I also used word length to approximate sentence complexity. I assume that the shorter the word length, the less complex it is and hence the more easily understandable the sentence that contains it.

8.2.3 Training the models

Given two misconceptions, A and B, in an essay, the classification task is to decide if the first misconception should be remedied before the second misconception. Given the ordered pair AB, if A should be remedied before B, the classifier should put this pair in class1 (A less than B). Otherwise, if B should be remedied before A or if there is no precedence relationship between them, the ordered pair AB should be assigned to class2 (A greater than or equal to B).

I used three machine learning algorithms from Weka (Hall et al., 2009) : SMO, J48 and Naive Bayes to train the classifier models. Because I was dealing with very few features, I used the linear kernel based support vector machine called SMO with the default parameters. I also tried the C4.5 decision tree because most of my data was binary and I wanted to understand how the decision was being made to put a pairing into a particular class. I included Naive Bayes to compare against a basic classification algorithm.

Because I had a small data set of sequenced misconceptions per student and a small number of students, I used leave-one-out cross validation on the data set produced from each expert annotation to train the different machine learning algorithms and evaluate the generated models. So for each expert and on each of the 12 runs, I used 11 students' sequenced misconceptions to train the models and then used 1 student's sequenced misconception to evaluate the generated model. Then I took the average of the 12 runs as the result.

8.3 Evaluation

I evaluated the models using accuracy. Accuracy in this regard is calculated as the percentage of instances that were assigned to the right class. The results of the average accuracy from the 12 runs of the leave-one-out cross validation are shown in Table 8.2. As the result shows, the C4.5 algorithm trained with the concept-match feature only, does the best at classifying an ordered pair of misconceptions. It does better than both baselines, the position feature only and the majority class present in the data set.

The position feature doesn't seem to be as helpful, as the majority classifier does better at prioritizing the misconceptions than it does. Using tf-idf and sentence length as proxy for complexity of a sentence, the result seems to suggest that as I theorized, simple misconceptions should be remedied before more complex ones.

Features	**Dataset**	**NaiveBayes**	**SMO**	**C4.5**	**Majority Class [no precedence]**
Position [baseline1]	expert1	56.2	56.2	56.2	60
	expert2	47.3	37.3	47.3	55
concept-match, word-length, tfidf	expert1	60.8	62.9	64.1	60
	expert2	46.7	57.8	55	55
concept-match	expert1	62.9	62.9	62.9	60
	expert2	52.5	55	55	55
word-length, tfidf	expert1	49.2	53.5	58.2	60
	expert2	43.7	56.3	53	55

Table 8.2: Average accuracy results on pair-wise classification of student misconceptions using leave-one-out cross validation

8.4 Conclusion

The main result from this study is the identification of concept match as a very important feature for prioritizing a set of student misconceptions. Concept match is the alignment of misconceptions to sequenced core concepts.

As Table 8.2 shows, If concept match is available, using it by itself is the best feature for prioritizing misconceptions. However, concept match has proven to be a very difficult feature to acquire automatically as can be seen by the results from the 2014 semantic textual similarity task (Agirrea et al., 2014). If concept match is unavailable for the classifier being built, or has a low degree of accuracy, then using a combination of sentence length and tf-idf would produce the next best prioritization.

The result from this study confirms the need for the next two studies, i.e., being able to extract and sequence core concepts, because we need them for creating the concept-match feature, which is the most useful feature for automatically prioritizing students' misconceptions.

Chapter 9

Automatic Extraction of Core Concepts

The purpose of this study was to compare two different multi-document summarization approaches for identifying **core** learning goals in a collection of documents. The outcomes of this study will be used to improve the core learning goal identifier algorithm in the CLICK2 system. The two methods to be compared are ranking of learning goals and reduction in the number of extracted learning goals. These two methods were selected because they are important variables to be considered when optimizing multi-document summarization techniques for different domains. In this experiment, I report on the scientific topic of plate tectonics. Subsequent studies on weather & climate and biology are reported in Okoye et al. (2010).

9.1 Materials

The first set of materials was twenty resources related to plate tectonics selected from DLESE by subject experts. The second set of materials is a set of extracted domain concepts, which were extracted from the twenty resources using the existing COGENT system with an extraction rate of 5%. 5% was chosen as the extraction rate because de la Chica et al. (2008a) showed that, at an extraction rate of 5%, COGENT is capable of creating a comprehensive knowledge base of Earth science concepts that a high school student should know. When extracting concepts, COGENT is identifying the most promising sentences to represent the learning goal in the system. This extraction yielded a set of 97 concepts which I call the

extracted domain concepts. This study will use these extracted domain concepts as the baseline learning goals from which to examine how well different methods for extracting core learning goals work.

9.2 Methodology

My methodology for this study was to compare and contrast the performance of two algorithms in identifying core learning goals. The measures I used to assess core learning goals were coverage and coreness. My measure of coverage is based on how well the resulting set of identified core learning goals corresponds to the AAAS benchmarks for plate tectonics. My measure for coreness is how well the learning goals identified by the two algorithms correspond to coreness ratings of human subject matter experts. Thus, I used the standard machine learning technique of comparing the output of the algorithms to a gold standard set generated by human experts.

9.2.1 Producing the human expert evaluation set

I worked with two Earth science subject experts to create the evaluation set. The two Earth science subject experts were asked to annotate the 97 concepts in the extracted domain concepts on two dimensions: alignment and coreness to AAAS benchmark learning goals for plate tectonics. Alignment refers to similarity to the 12 AAAS benchmark learning goals for plate tectonics, shown in Appendix A. The experts assigned each concept in the extracted domain concepts to a benchmark learning goal, to which the concept was most related. Coreness in this context is defined as the centrality (degree of alignment) to the AAAS benchmark learning goals for plate tectonics. The experts assigned a coreness rating of 1 to 4 to each concept, with 4 being the most core. This resulted in an alignment and coreness rating for each concept in the extracted domain concepts.

A gold standard set of core learning goals was created by putting all the concepts in the extracted domain concepts which had a rating of 4 into a set. This yielded learning goal data 1 **(LGD1)**, a set of 29 core learning goals which at least one expert had rated as 4.

9.2.2 Algorithm 1 : Reducing COGENT extraction rate

We used COGENT to extract core learning goals from the twenty DLESE resources. We chose to do the extraction at 1% in order to produce a similar number of concepts to those in the gold standard. This resulted in 32 concepts for learning goal data 2 **(LGD2)**, shown in Appendix B. The decision to reduce the extraction rate in COGENT to identify core learning goals came from a study which showed that as the extraction rate for plate tectonics was decreased from 5% to 1% of words, the average coreness of the extracted concepts increased steadily (Foster et al., 2012).

9.2.3 Algorithm 2 : Ranking in COGENT

We use ranking to extract core learning goals from the set of domain concepts. Ranking is a technique used in information retrieval for identifying the most relevant resources. COGENT has a built-in ranker, which it uses during the last stage of the multi-document summarization process, to decide which concepts in a collection of documents to extract to create a summary. We use the COGENT rankings to generate learning goal data 3 **(LGD3)**. The top ranked 29 concepts in the set of domain concepts generated by COGENT became learning goal data 3 **(LGD3)**.

9.3 Evaluation

We evaluate all automatically identified learning goal sets; i.e., **LGD2** and **LGD3** for coverage of and coreness to the AAAS benchmark learning goals for plate tectonics. We used the human subject expert annotations of alignment and coreness to score the output of the extraction and the ranking algorithms. Table 9.1 shows the results for coverage for the automatically identified learning goals sets, the core learning goals produced by reducing the extraction rate from 5% to 1% (LGD2) and the core learning goals produced by ranking

(LGD3). It also shows the coverage of LGD1 for the sake of comparison. The first column in Table 9.1 refers to the 12 AAAS benchmark learning goals for plate tectonics, 6 for middle school and 6 for high school. These benchmarks are available in Appendix A. In Table 9.1, PT-BMK, refers to plate tectonics-benchmark. MS1 refers to middle school 1 and HS1 refers to middle school 1. Thus PT-BMK-HS1, refers to the first high school plate tectonics benchmark learning goal. As Table 9.1 shows, LGD1, the gold standard set, had the best coverage. In the two learning goal sets extracted by the algorithms, LGD2 covered more of the benchmark learning goals than LGD3.

Table 9.1: Coverage results for LGD1, LGD2 and LGD3

#	**AAAS plate tectonics learning goal**	**LGD1 - human annotation**	**LGD2 - extraction algorithm**	**LGD3 - ranking algorithm**
(1)	PT-BMK-MS1	YES	YES	NO
(2)	PT-BMK-MS2	YES	YES	YES
(3)	PT-BMK-MS3	NO	NO	YES
(4)	PT-BMK-MS4	YES	NO	NO
(5)	PT-BMK-MS5	YES	YES	YES
(6)	PT-BMK-MS6	NO	NO	NO
(7)	PT-BMK-HS1	YES	YES	YES
(8)	PT-BMK-HS2	YES	YES	YES
(9)	PT-BMK-HS3	YES	YES	YES
(10)	PT-BMK-HS4	YES	YES	NO
(11)	PT-BMK-HS5	YES	YES	NO
(12)	PT-BMK-HS6	YES	YES	YES
		83% (10/12)	75%(9/12)	58%(7/12)

Because the subject expert annotations for alignment and coreness was done on the 97 domain concept data and not on all the concepts in the 20 resources, we had 4 concepts in LGD2 that were not part of the 97 domain concepts. We asked another domain expert to annotate those 4 concepts using the same rubric. We know that the learning goals in the gold standard set, LGD1 are *core* learning goals because they all have a coreness value of

Coreness Rating	**LGD1 Human Identified**	**LGD2 Extraction Rate algorithm**	**LGD3 Ranking Score algorithm**
1	0	3	9
2	0	1	7
3	0	13	5
4	29	15	8
AVERAGE	4.00	3.25	2.41

Table 9.2: Average Coreness for Learning Goal Data Sets

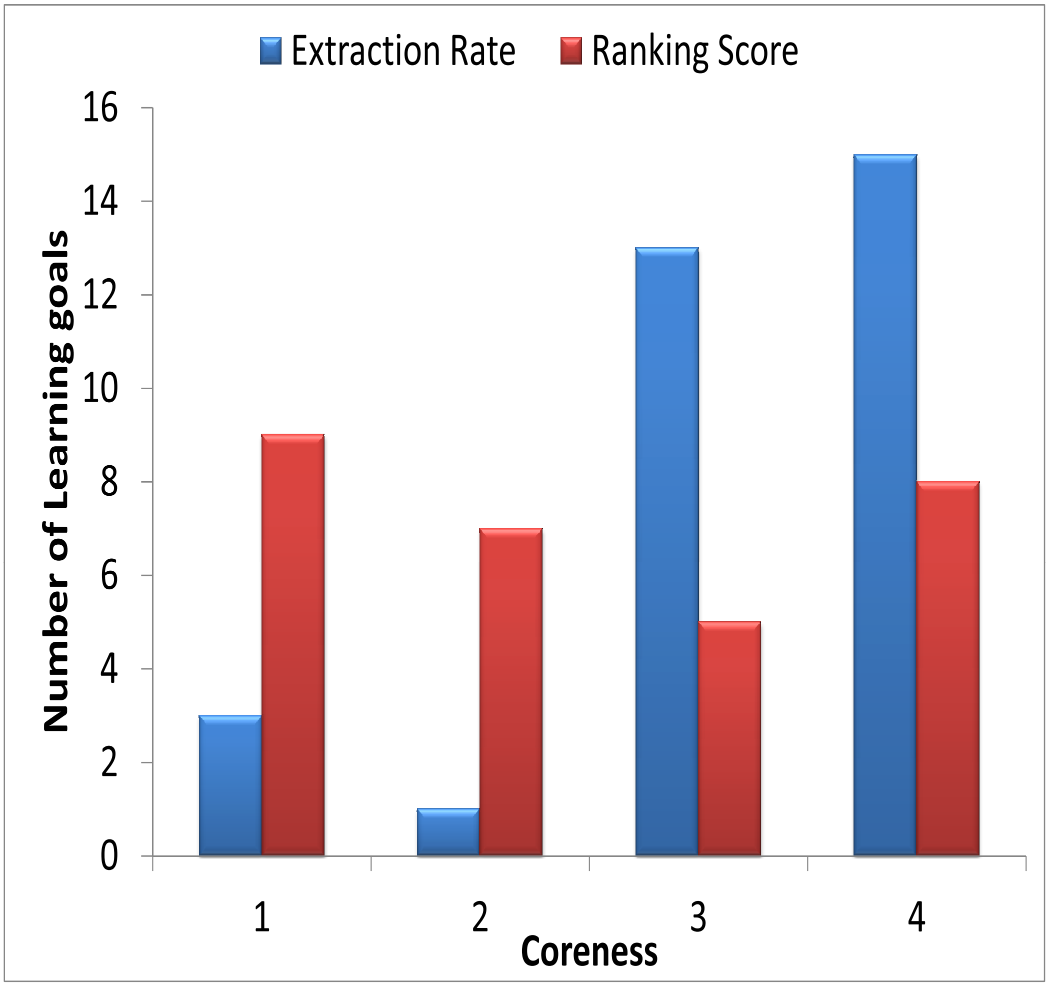


Figure 9.1: Distribution of coreness in Extraction-Rate and Ranking-Score

4 (the highest coreness rating). LGD2 had an average coreness of 3.25 while LGD3 had an average coreness of 2.3. As Figure 9.1 shows, LGD2 identified more core learning goals than LGD3.

9.4 Conclusion

Reducing the extraction rate significantly outperforms the rankings in COGENT. Table 9.1 shows that LGD2 covered more of the AAAS benchmark learning goals for plate tectonics than LGD3. Figure 9.1 shows that LGD2 identified more core learning goals than LGD3. This demonstrates that reducing the extraction rate to generate core learning goals produces better learning goals than using the top K from the default ranking in COGENT to generate core learning goals. By using an algorithm that reduces the extraction rate in COGENT, we can identify concepts with a higher coreness rating and better coverage and thus can identify core learning goals in a collection of resources. Therefore, for this work, I adjust the extraction rate in COGENT to 1% to identify the core learning goals in a collection of resources.

Chapter 10

Automatic Prioritization of Core Concepts

This chapter was published in Okoye et al. (2013b). Many people get their information from online sources such as search engines, portals for dedicated topics and social networks which provide free and ubiquitous access to information. However with free and ubiquitous access to information, comes the potential problem of information overload and learner disorientation (Chen, 2008). One way learning systems can mitigate these cognitive problems is by providing a pedagogical sequence (Yang et al., 2010). A pedagogical sequence is a learning trajectory or route taken by a learner through a range of learning goals in order to achieve understanding of a topic. According to conceptual change learning theory (Vosniadou, 2008) and current research in learning sciences (Margel et al., 2008; Krajcik et al., 2012) a good pedagogical sequence is important because the order in which knowledge is learned is crucial for developing a proper mental model. If the basic building knowledge blocks (or current mental model) is incorrect, it can lead to misconceptions when higher order knowledge blocks are assimilated. For example, if you have unremedied problems with addition and subtraction, you will probably have problems when doing algebra and with unremedied problems in algebra you can expect problems with calculus. Table 10.1 shows learning goals about seasons generated by the Center for Curriculum Materials in Science (Willard et al., 2007). These learning goals state that learning goal 1 should be understood before learning goal 2, learning goal 2 and 3 can be learned in any order and should be mastered, before attempting to understand learning goal 4. So two suggested pedagogical sequences through the learning goals in Table 10.1 are $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$ and $1 \rightarrow 3 \rightarrow 2 \rightarrow 4$.

#	**Core learning goals about seasons**
(1)	Light and other electromagnetic waves can warm objects. How much an object's temperature increases depends on how intense the light striking its surface is, how long it shines on the object, and how much of the light is absorbed.
(2)	The temperature of a location on the surface of the Earth depends upon the number of hours of sunlight and the intensity of that sunlight.
(3)	The axis of the Earth's rotation is tilted relative to the plane of the Earth's yearly orbit around the sun. As the Earth orbits the sun, the axis remains pointed the same place in space.
(4)	The seasonal variations in temperature at different places on the surface of the Earth are explained by the differential heating of the Earth's surface as it rotates on an axis that is tilted relative to the plane of the Earth's orbit around the sun.

Table 10.1: Examples of core learning goals about seasons

The purpose of this study was to establish a computational technique for generating a pedagogical sequence from the core learning goals in a collection of documents. In this study, I compare and contrast different machine learning models for generating a pedagogical sequence. To break this down into a tractable task, I cast this as a pair-wise ordering problem; i.e., rather than focusing on trying to automatically generate entire pedagogical sequences, I focus on developing an algorithm capable of identifying when one learning goal precedes another. The output of the trained models will be evaluated on pair-wise orderings of learning goals. To generate a pedagogical sequence from the resulting pair-wise judgments, first, I construct a precedence table from the pair-wise judgments and then generate a learning path from the precedence table. Table 10.2 is an example of a precedence table that contains four concepts. Three learning paths that can be generated from this table are: (1) $C->A->D->B$, (2) $C->D->A->B$ and (3) $C->A->B->D$.

I use two different pair-wise ordering tasks to create the computational models. One pair-wise ordering task uses the experts' pedagogical sequences of the learning goals. However because that data set is small - only 9 sequences - I needed to create a larger data set to

Table 10.2: Example of a Precedence table

Concept	**Preceded by**
A	C
B	A C
C	
D	C

help train these models. Therefore I created a proxy set using a pair-wise ordering of middle and high school sentences on plate tectonics, with middle school sentences having precedence over high school sentences.

10.1 Materials

This study uses three sets of materials. The first set of materials are the core learning goals identified by using the extraction rate algorithm from Chapter 9. I chose to use LGD2 because it is the best set of core learning goals I can create algorithmically.

The second set of materials are the pedagogical sequences produced by experts using the core learning goals. I collected nine pedagogical sequences from two science experts. As described later, each of these pedagogical sequences was then converted into pair-wise judgments between learning goals. These pair-wise judgments are used for both training and evaluating the machine learning models.

A third set of materials was created to help study pair-wise ordering. This third set of materials is a pair-wise ordering of middle and high school sentences. To construct this data set, I searched the DLESE website for text resources that contained the words **earthquake** or **plate tectonics**. I collected 10 such resources for each of the two grade cohorts: middle school (I allowed anything K-8) and high school (I allowed anything 9+). I downloaded the webpage for each resource, and used COGENT to extract the 20 most important sentences from each. This resulted in 200 sentences for each of the two grade cohorts. I divided the sentences in each grade cohort into three sets, one for training, one for development and the third for testing. To create pairs of grade-ordered sentences, I paired up middle and high school concepts both ways: middle school first (i.e., $\textsc{sequence}(c_m, c_h) = 0$) and high school first (i.e., $\textsc{sequence}(c_h, c_m) = 1$). This resulted in 4356 grade-ordered sentence pairs for each of the three sets (training, development and testing). These pair-wise orderings are only used for training the models.

10.2 Methodology

I used two different approaches to train the machine learning classifier models, in order to compare and contrast their performance on experts' pedagogical sequences. As discussed earlier, because I had a small data set for the experts' pedagogical sequences, I could not create dedicated training and testing sets from them. Thus, I trained the models using a 10-fold cross validation of the experts' pedagogical sequence. In addition, I also explored how the models perform when trained on a larger proxy data set. Each of the resulting models was then evaluated on the test set - expert' pedagogical sequences of the learning goals.

Weka (Hall et al., 2009) is a suite of machine learning algorithms implemented in Java and open sourced under GPL. I used Weka's implementations of the machine learning classifier algorithm: naive Bayes, Logistic Regression, LibSVM and SMO. Using Weka enabled me to concentrate on training the models for the task rather than rewriting the algorithms.

10.2.1 Producing the human expert evaluation set

The human expert evaluation set for this study is the second set of materials, the pair-wise orderings generated from experts' sequencing of LGD2. I asked two subject experts to come up with ideal learning paths, i.e., pedagogical sequences for LGD2. However, I constrained the task. I requested that the first sequence follow an **evidence or research based** learning path while the second sequence follow a **traditional** learning path.

An **evidence or research based** learning path is a pedagogy where students are encouraged to use the scientific method to learn about a phenomena, i.e., they gather information by observing the phenomena, forming a hypothesis, performing experiments, collecting and analyzing data and then interpreting data and drawing conclusions. A teacher that uses this learning path acts as a *guide on the side.* A **traditional** learning path, on the other hand, is the pedagogy where teachers are simply trying to pass on the correct information to students rather than letting the students discover the information themselves. In a classroom environment, a teacher using this learning path would be seen as the classical *sage on the stage.*

Both science education experts agreed that 2 of the 32 learning goals in LGD2, #1 and #3 in Appendix C, were not learning goals, therefore they excluded them when generating pedagogical sequences from LGD2. Table 10.3 shows the pedagogical sequences generated by the two science education experts. The first expert came up with one pedagogical sequence for evidence based (Expert1 Evidence) and two pedagogical sequences for traditional (Expert1 Traditional1 and Expert1 Traditional2) while the second expert came up with one each (Expert2 Evidence and Expert2 Traditional), so we had five pedagogical sequences for 30 of the 32 sentences in LGD2.

The science education experts produced a partial ordering of the learning goals. As shown in Table 10.3, although there are 30 sentences to be sequenced in LGD2, both experts produced only 21 and 26 levels of ordering respectively for the evidence based pedagogical sequences and 20 and 26 levels of ordering respectively for the traditional pedagogical sequences, with more than one learning goal occupying a level. If two learning goals are on the same level, it means they have the same priority, i.e., they do not have any precedence relationship between themselves and can be learned in any order. From these partial orderings, we generated a pair-wise ordering of all the sentences in the learning goal set and assigned each pair to a class ($C1 < C2$) or ($C1 >= C2$). For example as shown in Table 10.3, Expert1 for the evidence based pedagogy (column 5, row 1 and 2), assigned sentence 1 to level 1 and sentences 9 and 27 to level 2. The pair-wise orderings 1-9 and 1-27 will be assigned to class ($C1 < C2$), i.e., sentence 1 should be learned before sentences 9 and 27.

Level	**Expert1 Traditional1**	**Expert1 Traditional2**	**Expert2 Traditional**	**Expert1 Evidence**	**Expert2 Evidence**
(1)	27, 9, 16	27, 9, 16	9,27	1	4
(2)	3	3	5, 14	9, 27	9, 27
(3)	29	29	15	16	5, 14
(4)	5, 15	5, 15	21	5, 15	15, 21, 12
(5)	14	14	12	14	16
(6)	21, 12	21, 12	16	12, 21	22
(7)	28	28	22	28	1
(8)	17, 18, 19	17, 18, 19	1	4	6
(9)	22	22	6	13, 23	17
(10)	6	6	17, 18, 19	8	19
(11)	2, 20	4	4	25	18
(12)	10, 32	13, 23	25	17, 18, 19	25
(13)	30	8	8	24	8
(14)	11, 31	25	31	11, 31	31
(15)	4	2, 20	2	10, 32	2
(16)	13, 23	10, 32	20	30	20
(17)	8	30	3	2, 20	3
(18)	25	11, 31	29	22	29
(19)	24	24	28	6	28
(20)	1	1	10	29	10
(21)			30	3	30
(22)			32		32
(23)			24		24
(24)			13		13
(25)			23		23
(26)			11		11

Table 10.3: Human Experts Pedagogical sequencing of LGD2

The pair-wise ordering 9-27, 9-1 and 27-1 will be assigned to class ($C1 >= C2$) because the partial ordering does not indicate that 9 and 27 should be learned in any specific order, and the partial ordering says they should be learned after sentence 1.

Type	**Instances**	($C1 < C2$)	($C1 >= C2$)
Evidence	589 (65.0%)	49.5%	50.5%
Traditional	613 (70.5%)	49.5%	50.5%

Table 10.4: Distribution of agreement pair-wise pedagogical sequence data

Type	**Number of observed agreements**	**agreements expected by chance**	**KAPPA**	**95% confidence interval**	**Strength of agreement**
Evidence	597 (68.6%)	435 (50%)	0.37	0.31 to 0.43	Fair
Traditional	613 (70.5%)	435 (50%)	0.41	0.35 to 0.47	Moderate

Table 10.5: Inter annotator agreement for pair-wise pedagogical sequence data

Table 10.4 describes the second set of materials, which we will use as test data for our pedagogical sequence models. The **Type** column shows which pedagogical sequence data set we are using, Evidence or Traditional. The sentences are the same for both pedagogies and come from 30 of the 32 LGD2. The **Instances** column is the number of instances that were generated by pairing the core learning goals and then pulling together the instances with which both experts agreed on the class label. For a pair of learning goals **a** and **b**, we pair them both ways to generate two instance **ab** and **ba**. The total number of pair-wise orderings we can generate from 30 core learning goals are 870.

As can be seen from Table 10.4, for the Evidence pedagogy, the experts agreed on the class labels for 65% (589) of the 870 and for Traditional pedagogy, 70.5% (613). For the Traditional pedagogy, there were three sequences, 2 from the first expert and 1 from the second expert. We are using only the pairs for which all three sequences agree, even though 2 sequences were produced by 1 expert. Furthermore, although all core learning goals are paired with all other core learning goals, because the experts produce partial orderings, the number of agreements for each type of ordering may not be the same. Consider sentences 2 and 20 as assigned to levels by Expert1 in Traditional2 (column 3, row 15) and Expert2 in Traditional (column 4, rows 15 and 16) in Table 10.3. For the pair 20-2, they agree on the relation $C1 >= C2$, but disagree on the relation $C1 < C2$ for the pair 2-20. As a result, the class $C1 >= C2$ is slightly larger than the $C1 < C2$ class as can be seen in Table 10.4.

The inter-annotator agreement was measured using Cohen's Kappa (Cohen, 1960). Table 10.5 shows that the agreement was fair for the Evidence data set, while it was moderate for Traditional data set.

10.2.2 Method 1: Training models from proxy task

The proxy task was ordering sentences by grade. In this task, the model is given two sentences s_1 and s_2, one written for middle school, s_1 and another written for high school, s_2, and asked to decide whether $s_1 < s_2$ or $s_2 < s_1$. I expect that a model for ordering sentences by grade should also be a reasonable model for ordering concepts for a pedagogical learning path. And importantly, getting grade ordering data automatically is easy: the Digital Library for Earth System Education (DLESE) contains a variety of Earth science resources with metadata about the grade level they were written for.

Tanaka-Ishii et al. (2010) used local and global word count features only to build a pair-wise classifier for sorting texts by readability that had an accuracy of 90% for sorting English text in which the text vectors were concatenated into a pair. Our task is similar to the Tanaka-ishii task as I am also building a pair-wise classifier for sorting texts so I use the same local and word count features.

So as a first pass on this task, I decided to use word count features only and concatenate the vectors. First, I extracted all the unique words in the training corpus (same topic but different documents), removed stop words (such as **the**, **a** and **and**), and ended up with 1702 words. Then, I indexed the English Gigaword, a background corpus that reflects how often I can expect people to use each word in a normal day to day context and got the word count in Gigaword for each of the corresponding 1702 words. From these I generated two main features;

- **local word count** - the number of times the word appeared in this sentence
- **global word count** - the log of the ratio between the number of times the word occurred in the sentence and the number of times it occurred in the background corpus, Gigaword (Graff, 2002).

I used the development data set to tune the parameters of the classifiers. The best result for the Logistic Regression model was having the ridge set to 0.1. For SMO, the best result was using a polynomial kernel of degree 1, setting the complexity parameter, C, to 0.01 and normalizing the data. For LibSVM, the best result was produced using a polynomial kernel of degree 1, setting the complexity parameter, C, to 0.1 and normalizing the data. Table 10.6 shows the accuracy results for evaluating the four classifiers on the testing data from the third set of materials, the pair-wise ordering of middle and high school sentences.

Table 10.6: Accuracy results for Proxy Task

Model	**Accuracy**
Naive Bayes	84.9%
SMO	80.9 %
LibSVM	83.5%
Logistic Regression	82.1%

10.2.3 Method 2: Training models using 10-fold cross validation

I used 10-fold cross validation on the second set of materials, the agreed pair-wise orderings from the science education experts shown in Table 10.4, to train three models. Cross-validation is a standard technique used in machine learning for training a model and estimating its' predictive performance, especially when there isn't enough data to have dedicated training and evaluation sets. In K-fold cross validation, the data set is split into mutually exclusive subsets of approximately equal size for training and testing. The model is then trained with K-1 parts and evaluated on 1 part K times, using the K possible combinations. The result is the average of the values produced by evaluating the model K times. I used this method because I have a limited supply of subject-expert sequenced core learning goal data.

Given two concepts, a and b, getting the correct class label for one ordering, **ab**, helps with predicting the class label of the second ordering, **ba**. Therefore, when dividing the data set for training and testing, I ensured that both **ab** and **ba** are both contained together in either the training or the testing data set.

I used the same word features that were used in the proxy task, **local** and **global word count**, because the domain is the same. But in addition, I included the COGENT ranking feature, which is 1 if in a pair-wise ordering **ab** of learning goals **a** and **b**, the first learning goal, **(**a) was extracted by COGENT before the second learning goal **(**b) and 0 otherwise. So for each core learning goal (sentence), I calculated 3404 features and a vector in the classifier had 6809 features since I concatenated two sentences into one vector by placing them side by side and then added the cogent ranking feature.

10.3 Evaluation

For evaluating the models, I had two baselines. The first baseline is the majority class. This is the easiest classifier to build. The classifier just looks at the training data and calculates the most frequent class. In this case, it would be $(C1 >= C2)$. Then for all the data in the test set, the classifier says they belong to class $(C1 >= C2)$. The second baseline, cogent rankings, was the order in which COGENT outputs the learning goals. So given a pair of learning goals, 2 and 20, that were the second and twentieth sentences output by COGENT, this classifier assigns the pairing 2-20 to class $(C1 < C2)$ and the pairing 20-2 to class ($C1 >= C2$). The gold standard pair-wise judgments were the pair-wise judgments inferred from the experts sequence of the extraction-rate learning goals (LGD2).

Tables 10.7 and 10.8 show the accuracy results to date. As both Tables show, using the majority baseline classifier resulted in an accuracy of 50.5% for both the Evidence and Traditional pedagogy. While using the cogent rankings baseline classifier resulted in an accuracy of 62% for the Evidence pedagogy and 57% for the Traditional pedagogy.

None of the four proxy task models did consistently better than the baselines. Most times, they were slightly worse than the baselines. However, the 10-fold cross validation models did well. The two support vector classifiers, SMO and LibSVM perform better than both baselines for the two pedagogies using the default SVM parameters. However, Logistic Regression with its default parameters performs consistently better than both baselines and the other two models. So going forward, I will be using Logistic Regression to build pedagogical sequence models.

Table 10.7: Accuracy results on pair-wise classification using Proxy Task

Id	**Majority baseline**	**COGENT rankings**	**NaiveBayes**	**SMO**	**LibSVM**	**Logistic**
Evidence	50.5	62.0	**53.3**	47.4	52.7	47.7
Traditional	50.5	57.0	52.2	56	50.6	57.3

Id	**Majority baseline**	**COGENT rankings**	**NaiveBayes**	**SMO**	**Lib SVM**	**Logistic Regression**
Evidence	50.5	62.0	70.3	76.3	74.9	76.9
Traditional	50.5	57.0	56.8	60.6	59.0	60.6

Table 10.8: Accuracy results on pair-wise classification using cross validation

10.4 Conclusion

In this chapter, I trained eight models from four machine learning classifiers, using two training methods: proxy and cross validation with only local and global word counts as features. The models trained using 10 fold cross validation performed much better than those trained using the proxy task. Cross validation Logistic Regression outperformed the SVM models and the two baselines on the traditional and evidence pair-wise data sets with an accuracy score of 76.9% for the evidence data set and 60.6% for the traditional data set. The dynamically generated pedagogical sequences can provide structure and guidance to digital library users by giving them a pedagogically-meaningful learning sequence through which they can explore related documents they retrieve from a digital library. Furthermore, the dynamically generated pedagogical sequences can support the task of providing personalized feedback to students on what digital library resources they need to explore in order to satisfy a competency requirement or remedy a knowledge deficiency.

Chapter 11

Conclusions and Future Work

The work presented in this thesis took steps towards building a personalized educational recommender system that can support conceptual change in users of the said system. Parts of this thesis have been published in a book chapter (Okoye et al., 2011) and several conference proceedings (Becker et al., 2010; Okoye et al., 2010; Bethard et al., 2012; Okoye et al., 2013a,b).

Section 11.1 summarizes the contributions of this thesis. Section 11.2 revisits the research questions that guided this work and the results from the corresponding studies. Section 11.3 discusses the opportunities for future work while section 11.4 concludes with some final words about the impact of this work.

11.1 Summary

The major contributions of this work are:

(1) Refining and extending conceptual change learning theory from a classroom environment into an online learner-driven environment.

(2) Developing statistical and machine-learning based models to automate the instructional process of identifying core learning goals, sequencing the core learning and goals and prioritizing student misconceptions in a pedagogically useful manner

(3) Creating a framework that extends educational recommender systems from simply making recommendations based on usage to also facilitating science understanding

(4) Demonstrating the utility of such an educational recommender system to actually effect positive change in student outcomes such as understanding and interest.

11.2 Research Questions Revisited

Research Question 1: What are design options for creating an educational recommender system (ERS) with research-based support mechanisms for promoting conceptual change?

CLICK2 was the final design that resulted from the study that was carried out in the bid to address the preceding research question. The incorrect feedback pane in CLICK2 displays: (1) a preamble about science experts believing something different from the identified misconception, (2) a refutation text stating what science experts believe and (3) a list of three recommended resources per identified misconception. All three are research-based support mechanisms for promoting conceptual change. The preamble and refutation text help tear down misconceptions while the recommended resources and refutation text together, help build up the correct conceptions.

CLICK2 is an effective feedback environment as defined by Hattie (Hattie and Timperley, 2007) because it can provide answers to the questions *where am I?*, *where am I going?*, *where to next?* and *how am I going?* in relation to a user's knowledge state. Figure 6.4 shows that users were satisfied with the design and interaction of CLICK2 and would recommend it to their friends. In addition, it shows users agreed that using CLICK2 improved their understanding of seasons, the topic they explored during the study.

Research Question 2: How does the educational recommender system with its conceptual change support mechanisms affect users' understanding, interest and perception of science content?

To address this question, I ran a pilot learning study with twelve participants and analyzed several questionnaires and essays. The analysis produced strong indications that CLICK2 can improve users' interest, perception of understanding and actual understanding of the topic of seasons. Figure 7.2 shows that interest improved between the start of session1 to the end of session 2. Figure 7.3 shows that perception of understanding or confidence in knowledge did decrease from the start of session 1 to the end of session 1, but it went up past the initial baseline through the end of session 2. Figure 7.4 and the analysis of three selected essays show that actual understanding of the topic of seasons also improved.

Research Question 3: How can we model expert strategies for prioritizing student misconceptions?

My approach to modeling expert strategies for prioritizing student misconceptions was to run an annotation study. I used the data from the study to create supervised machine learning classifiers that model how experts prioritize identified student misconceptions. To break this down into a tractable task, I cast this as a pair-wise ordering problem, i.e., rather than focusing on trying to automatically generate entire pedagogical sequences for the misconceptions in an essay, I focused on developing a model capable of identifying when one misconception precedes (should be remedied before) another.

The result in Table 10.8 shows that concept match, i.e., aligning a sequenced core concept to the misconception was the most helpful feature for creating the models. Concept match is a very difficult feature to acquire automatically because it involves three different sequential tasks that are individually difficult. So, in the absence of the concept match feature or if it has a low degree of accuracy, then using a combination of sentence length and tf-idf would produce the next best prioritization sequence. Creating the concept match feature involves three different problems. First, identifying or extracting the core concepts. Second, sequencing the core concepts. And third, aligning the sequenced core concepts to identified misconceptions, i.e., coming up with a textual similarity algorithm. The next two research questions were an attempt to address the first two problems. Research by Sultan et al. (2014) focuses on the third.

Research Question 4: How well can different computational methods identify the learning goals in a collection of documents?

My methodology for this study was to compare and contrast the performance of two algorithms in identifying core learning goals. The measures I used to assess core learning goals were *coverage* and *coreness*. My measure of *coverage* is based on how well the resulting set of identified core learning goals corresponds to the AAAS benchmarks for plate tectonics. My measure for *coreness* is based on how well the learning goals identified by the two algorithms correspond to coreness ratings of human subject matter experts.

The algorithm based on reducing the extraction rate significantly outperforms the rankings algorithm. Table 9.1 shows that the LGD2 data, produced by the extraction-rate algorithm covered more of the AAAS benchmark learning goals for plate tectonics than the LGD3 data produced by the ranking algorithm. And Figure 9.1 also shows that LGD2 identified more core learning goals than LGD3. By using an algorithm that reduces the extraction rate, we can identify concepts with a higher coreness rating and better coverage and thus can identify core learning goals in a collection of resources. This demonstrates an approach to tackling the first problem in producing a concept-match feature automatically, i.e., the problem of core concept extraction.

Research Question 5: How well can machine learning classifiers model the pedagogical sequences of learning goals produced by human experts?

In this study, I trained eight models from four machine learning classifiers, using two training methods: proxy and cross validation with only local and global word counts as features. The models trained using 10 fold cross validation performed much better than those trained using the proxy task. The Logistic Regression model that was trained using ten-fold cross validation outperformed the SVM models and the two baselines on the traditional and evidence pair-wise data sets with an accuracy score of 76.9% for the evidence data set and 60.6% for the traditional data set. This shows that we can model human experts for the task of creating a pedagogical sequence from learning goals, however there is much room for improvement.

This study demonstrates an approach to tackling the second problem in producing a concept-match feature automatically that is, the problem of core concept sequencing.

11.3 Discussion and Future Work

The goal of this dissertation was to build an educational recommender system based on conceptual change learning theory so as to improve students' understanding of science concepts. While it can be argued that this work does build an educational recommender system with conceptual change support mechanisms, which has been shown to positively impact students understanding, there is a lot of room left for improvement. These improvements offer many starting points for future research.

A big issue in applying natural language processing (NLP) algorithms to the domain of education is the lack of decent data both in quality and quantity. The quality of generated machine learning models depend a great deal on the availability of a good amount of quality data for training, development and testing. Therefore, more should be done to collect and disseminate data that can be used when applying NLP to education. In addition, researchers should put their available data to as much use as possible. Data collected for one task should be repurposed if possible for a different task as was done in Becker et al. (2010).

All the algorithms, i.e., core concept extraction, core concept sequencing and misconception sequencing should be improved. Perhaps studying curriculum development experts and teachers grading would reveal more strategies for mining features that can be used to model the experts. The issue of whether to create consensus among different science topic experts for the task of sequencing core concepts or misconceptions is an open question. Personally, I do not think there is value in such a proposition. I believe there are different paths that can be taken when learning a topic and this is what is reflected in the different paths or sequences that different experts generate.

Next, the two sequencing algorithms should be evaluated on at least two more types of data; one from a subject matter within the domain of Earth science such as weather & climate and another from an entirely different domain such as biology. This will address questions as to the generalizability and robustness of the algorithms. I already did this for the core-concept extraction task as reported in Okoye et al. (2010).

This thesis showed that concept-match is a very important feature for modeling experts for the task of sequencing misconceptions and so being able to create a concept match would greatly improve that algorithm. The biggest open question for automatically generating concept match feature is *how can we determine the textual similarity between different texts?*. If this question is addressed adequately, it will also improve other algorithms in CLICK2 such as the misconception identification algorithm, which is discussed in detail in Ahmad (2009).

There were many parts of this work that were simulated, and so would still need to be researched and created for a working system. First is the task of automatically generating the refutation text for a misconception. Second is the task of recommending resources that can help users address their misconceptions. Third is incorporating the feedback from the rated resources into the recommendation algorithm as relates to a specific misconception. Fourth and last, is the task of generating more advanced conceptual change constructive elements for the feedback pane such as analogies.

In addition, a more extensive learning study with an experimental and control group, needs to be carried out in order to assess if there is a statistically significant difference in student learning when using CLICK2 versus the status quo. While this research did yield great preliminary evidence about the utility of my conceptual framework, to limit the scope of my research, I focused on the cognitive aspects of conceptual change theory. Future research should bring in the affective aspects.

11.4 Closing Remarks

This thesis builds on previous work by Ahmad (Ahmad, 2009), De la Chica (de la Chica et al., 2008b) and Gu (Gu, 2009). They developed CLICK which uses a digital library to support learners' understanding of STEM content by using graph-based algorithms to deduce what the learner knows about the domain, infer what the learner should know about the domain, and then present digital library resources that learners can use to address their incorrect, vague and/or missing conceptual knowledge. CLICK2 added the identification of core concepts, sequencing of core concepts to generate a general learning path and sequencing of misconceptions to generate a personalized learning path. In addition, CLICK2 paid attention to pedagogy deeply and in multiple ways.

CLICK2 is designed to be a scalable, personalized learning system that can support many use cases. The dynamically generated pedagogical sequences in CLICK2 can provide structure and guidance to digital library users by giving them a pedagogically-meaningful learning sequence through which they can explore related documents they retrieve from a digital library.

CLICK2 can support teachers by providing personalized support for students at different levels of understanding. Thereby letting teacher maximize the class periods while still giving their students access to personalized support.

Furthermore, CLICK2 can provide tutoring support to students. According to Dzubak (Dzubak and York, 2009), the three main advantages of traditional tutoring are that it is conducted during a personalized, face to face social interaction; it provides immediacy of feedback and it actively engages the tutee in the process of learning. Although CLICK2 cannot provide face-to-face social interaction, it does support personalized interaction, immediacy of feedback and active engagement.

Technology pervades everyday human life and is constantly affecting the way we live for better or for worse. This is an attempt to use technology for the better, by using it to support science literacy.

Bibliography

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE Transactions on, 17(6):734–749.

Adomavicius, G. and Tuzhilin, A. (2011). Context-aware recommender systems. Recommender Systems Handbook, pages 217–253.

Agirrea, E., Baneab, C., Cardiec, C., Cerd, D., Diabe, M., Gonzalez-Agirrea, A., Guof, W., Mihalceab, R., Rigaua, G., and Wiebeg, J. (2014). Semeval-2014 task 10: Multilingual semantic textual similarity. SemEval 2014, page 81.

Ahmad, F. (2009). Generating conceptually personalized interactions for educational digital libraries using concept maps. PhD thesis, UNIVERSITY OF COLORADO AT BOULDER.

Alonzo, A. (2012). Learning progressions: significant promise, significant challenge. Zeitschrift für Erziehungswissenschaft, pages 1–15.

Ashcraft, P. (2006). A comparison of student understanding of seasons using inquiry and didactic teaching methods. In AIP Conference Proceedings, volume 818, page 85.

Atwood, R. and Atwood, V. (1996). Preservice elementary teachers' conceptions of the causes of seasons. Journal of Research in Science Teaching, 33(5):553–563.

Becker, L., Nielsen, R., Okoye, I., Sumner, T., and Ward, W. (2010). Whats next? target concept identification and sequencing. In Proceedings of QG2010: The Third Workshop on Question Generation, page 35.

Bethard, S., Hang, H., Okoye, I., Martin, J. H., Sultan, M. A., and Sumner, T. (2012). Identifying science concepts and student misconceptions in an interactive essay writing tutor. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 12–21. Association for Computational Linguistics.

Bishop, C. M. et al. (2006). Pattern recognition and machine learning, volume 4. springer New York.

Bloom, B. S. et al. (1971). Handbook on formative and summative evaluation of student learning.

Bødker, S. and Iversen, O. S. (2002). Staging a professional participatory design practice: moving pd beyond the initial fascination of user involvement. In Proceedings of the second Nordic conference on Human-computer interaction, pages 11–18. ACM.

Brown, D. and Clement, J. (1989). Overcoming misconceptions via analogical reasoning: Abstract transfer versus explanatory model construction. Instructional Science, 18(4):237–261.

Bruning, R., Schraw, G., and Ronning, R. (1999). Cognitive psychology and instruction. Prentice-Hall, Inc., One Lake Street, Upper Saddle River, NJ 07458. Tel: 800-282-0693 (Toll Free); Web site: http://www. prenhall. com.

Buder, J. and Schwind, C. (2011). Learning with personalized recommender systems: A psychological view. Computers in Human Behavior.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. User modeling and user-adapted interaction, 12(4):331–370.

Butcher, K., Sumner, T., Maull, K., and Okoye, I. (2010). Conceptual Personalization Technology Promoting Effective Self-Directed, Online Learning. In Proceedings of the 10th International Conference on Intelligent Tutoring Systems, page Submitted. Association for Computational Linguistics.

Carroll, J. M., Chin, G., Rosson, M. B., and Neale, D. C. (2000). The development of cooperation: Five years of participatory design in the virtual school. In Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques, pages 239–251. ACM.

Cartier, J. and Center, E. R. I. (2000). Using a modeling approach to explore scientific epistemology with high school biology students. University of Wisconsin-Madison, National Center for Improving Student Learning and Achievement in Mathematics and Science.

Cartier, J., Rudolph, J., Stewart, J., and (US), E. R. I. C. (2001). The nature and structure of scientific models. National Center for Improving Student Learning and Achievement in Mathematics and Science.

Chen, C. (2008). Intelligent web-based learning system with personalized learning path guidance. Computers & Education, 51(2):787–814.

Clement, J. and Vosniadou, S. (2008). The role of explanatory models in teaching for conceptual change. International handbook of research on conceptual change, pages 417–452.

Cohen, J. (1960). A coefficient of agreement for nominal scales. educational and psychological jv! easurement, 20, 3 7-46.

de la Chica, S. (2009). Generating conceptual knowledge representations to support students writing scientific explanations. PhD thesis, UNIVERSITY OF COLORADO AT BOULDER.

de la Chica, S., Ahmad, F., Martin, J. H., and Sumner, T. (2008a). Pedagogically useful extractive summaries for science education. In Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1, COLING '08, pages 177–184, Stroudsburg, PA, USA. Association for Computational Linguistics.

de la Chica, S., Ahmad, F., Sumner, T., Martin, J., and Butcher, K. (2008b). Computational foundations for personalizing instruction with digital libraries. International Journal on Digital Libraries, 9(1):3–18.

Dreyfus, A., Jungwirth, E., and Eliovitch, R. (1990). Applying the cognitive conflict strategy for conceptual changesome implications, difficulties, and problems. Science Education, 74(5):555–569.

Duit, R. (1999). Conceptual change approaches in science education. New perspectives on conceptual change, pages 263–282.

Duit, R. and Treagust, D. (2003). Conceptual change: a powerful framework for improving science teaching and learning. International journal of science education, 25(6):671–688.

Dzubak, C. M. and York, P. (2009). Why tutoring matters: The interaction of a peer tutor and a tutee during scaffolding. Synergy: The Journal of the Association for the Tutoring Profession, 2:1–5.

Foster, J., Sultan, A., Devaul, H., Okoye, I., and Sumner, T. (2012). Identifying core concepts in educational resources. In Proceedings of the12th ACM/IEEE-CS joint conference on Digital libraries, JCDL '12, New York, NY, USA. ACM.

Frede, V. (2008). The seasons explained by refutational modeling activities. Astronomy Education Review, 7:44.

Graff, D. (2002). English Gigaword. Linguistic Data Consortium.

Gu, Q. (2009). Personalized information seeking to support intentional learning. PhD thesis, UNIVERSITY OF COLORADO AT BOULDER.

Gu, Q., Chica, S., Ahmad, F., Khan, H., Sumner, T., Martin, J., and Butcher, K. (2008). Personalizing the Selection of Digital Library Resources to Support Intentional Learning. In Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries, page 255. Springer.

Guilizzoni, G. (2010). balsamiq mockups. Balsamiq, http://www. balsamiq. com.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The weka data mining software: an update. ACM SIGKDD Explorations Newsletter, 11(1):10–18.

Hattie, J. and Timperley, H. (2007). The power of feedback. Review of educational research, 77(1):81–112.

Hawkins, D. M., Basak, S. C., and Mills, D. (2003). Assessing model fit by cross-validation. Journal of chemical information and computer sciences, 43(2):579–586.

Huang, Y., Huang, T., Wang, K., Hwang, W., et al. (2009). A markov-based recommendation model for exploring the transfer of learning on the web. Educational Technology & Society, 12(2):144–162.

Hummel, H., Van Den Berg, B., Berlanga, A., Drachsler, H., Janssen, J., Nadolski, R., and Koper, R. (2007). Combining social-based and information-based approaches for personalised recommendation on sequencing learning activities. International Journal of Learning Technology, 3(2):152–168.

Inagaki, K. and Hatano, G. (2008). Conceptual change in naïve biology. International handbook of research on conceptual change, pages 240–262.

Jurafsky, D., Martin, J. H., Kehler, A., Vander Linden, K., and Ward, N. (2000). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, volume 2. MIT Press.

Khribi, M., Jemni, M., and Nasraoui, O. (2008). Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. In Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference on, pages 241–245. Ieee.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI, volume 14, pages 1137–1145.

Koutrika, G., Bercovitz, B., Kaliszan, F., Liou, H., and Garcia-Molina, H. (2009). Courserank: A closed-community social system through the magnifying glass. In Third International Conference on Weblogs and Social Media (ICWSM).

Krajcik, J. S., Sutherland, L., Drago, K., Merritt, J., Bernholt, S., Neumann, K., and Nentwig, P. (2012). The promise and value of learning progression research. Making it tangible: Learning outcomes in science education, pages 261–284.

Lewis, C. and Rieman, J. (1993). Task-centered user interface design. A Practical Introductio.

Manouselis, N., Drachsler, H., Vuorikari, R., Hummel, H., and Koper, R. (2011). Recommender systems in technology enhanced learning. Recommender Systems Handbook, pages 387–415.

Margel, H., Eylon, B.-S., and Scherz, Z. (2008). A longitudinal study of junior high school students' conceptions of the structure of materials. Journal of Research in Science Teaching, 45(1):132–152.

Marlino, M., Sumner, T., Fulker, D., Manduca, C., and Mogk, D. (2001). The digital library for earth system education: building community, building the library. Communications of the ACM, 44(5):80–81.

Muller, M. J. (2003). Participatory design: the third space in hci. Human-computer interaction: Development process, pages 165–185.

Murphy, P. and Alexander, P. (2008). The role of knowledge, beliefs, and interest in the conceptual change process: A synthesis and meta-analysis of the research. International handbook of research on conceptual change, pages 583–616.

Murray, T., Blessing, S., and Ainsworth, S. (2003). Authoring Tools for Advanced Technology Learning Environments: Toward cost-effective adaptive, inte Kluwer Academic Pub.

Nielsen, J. (1994). Heuristic evaluation. Usability inspection methods, 17:25–62.

NRC, N. R. C. (1996). National Science Education Standards. National Academy Press, Washington DC.

Okoye, I., Bethard, S., and Sumner, T. (2013a). Cu: Computational assessment of short free text answers-a tool for evaluating students understanding. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation, pages 603–607, Atlanta, Georgia, USA. Association for Computational Linguistics.

Okoye, I., Maull, K., Foster, J., and Sumner, T. (2011). Educational recommendation in an informal intentional learning system. Educational Recommender Systems and Technologies: Practices and Challenges: Practices and Challenges, page 1.

Okoye, I., Maull, K., and Sumner, T. (2010). Algorithms for robust knowledge extraction in learning environments. In Intelligent Tutoring Systems, pages 245–247. Springer.

Okoye, I., Sumner, T., and Bethard, S. (2013b). Automatic extraction of core learning goals and generation of pedagogical sequences through a collection of digital library resources. In Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, pages 67–76. ACM.

Passmore, C. and Stewart, J. (2002). A modeling approach to teaching evolutionary biology in high schools*. Journal of Research in Science Teaching, 39(3):185–204.

Pintrich, P., Marx, R., and Boyle, R. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. Review of Educational research, 63(2):167–199.

Plummer, J. and Agan, L. (2010). Reasoning about the seasons: middle school students' use of evidence in explanations. In Proceedings of the 9th International Conference of the Learning Sciences - Volume 2, ICLS '10, pages 464–465. International Society of the Learning Sciences.

Plummer, J., Wasko, K., and Slagle, C. (2011). Children learning to explain daily celestial motion: Understanding astronomy across moving frames of reference. International Journal of Science Education, 33(14):1963–1992.

Posner, G., Strike, K., Hewson, P., and Gertzog, W. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. Science education, 66(2):211–227.

Project2061. (1993). Benchmarks for Science Literacy. Oxford University Press, New York, United States.

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., Otterbacher, J., Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., and Zhang, Z. (2004). MEAD - a platform for multidocument multilingual text summarization. In LREC 2004, Lisbon, Portugal.

Rafaeli, S., Barak, M., Dan-Gur, Y., and Toch, E. (2004). Qsia-a web-based environment for learning, assessing and knowledge sharing in communities. Computers & Education, 43(3):273–289.

Recker, M. and Walker, A. (2003). Supporting word-of-mouthsocial networks through collaborative information filtering. Journal of Interactive Learning Research, 14(1):79.

Sanders, E. B.-N. (2002). From user-centered to participatory design approaches. Design and the social sciences: Making connections, pages 1–8.

Schneps, M., Sadler, P., Woll, S., and Crouse, L. (1989). A private universe. Pyramid Flim & Video.

Scott, P., Asoko, H., and Driver, R. (1991). Teaching for conceptual change: A review of strategies. Connecting Research in Physics Education with Teacher Education, pages 71–78.

Sebastià, B. and Torregrosa, J. (2005). Preservice elementary teachers conceptions of the sun-earth model: A proposal of a teaching-learning sequence. Astronomy Education Review, 4:121.

Shen, L. and Shen, R. (2004). Learning content recommendation service based-on simple sequencing specification. Advances in Web-Based Learning–ICWL 2004, pages 293–323.

Sinatra, G. (2005). The" warming trend" in conceptual change research: The legacy of paul r. pintrich. Educational Psychologist, 40(2):107–115.

Stewart, J., Cartier, J., and Passmore, C. (2005). Developing understanding through model-based inquiry. How students learn, pages 515–565.

Sultan, M. A., Bethard, S., and Sumner, T. (2014). Dls@cu: Sentence similarity from word alignment. In Proceedings of the 8th International Workshop on Semantic Evaluation, pages 241–246, Dublin, Ireland. Association for Computational Linguistics.

Tanaka-Ishii, K., Tezuka, S., and Terada, H. (2010). Sorting texts by readability. Computational Linguistics, 36(2):203–227.

Trumper, R. (2000). University students' conceptions of basic astronomy concepts. Physics Education, 35:9.

Trumper, R. (2001a). A cross-age study of junior high school students' conceptions of basic astronomy concepts. International Journal of Science Education, 23(11):1111–1123.

Trumper, R. (2001b). A cross-age study of senior high school students' conceptions of basic astronomy concepts. Research in Science & Technological Education, 19(1):97–109.

Trumper, R. (2001c). A cross-college age study of science and nonscience students' conceptions of basic astronomy concepts in preservice training for high-school teachers. Journal of Science Education and Technology, 10(2):189–195.

Vosniadou, S. (2008). International handbook of research on conceptual change. Taylor & Francis.

Vosniadou, S. (2013a). The framework theory approach to the problem of conceptual change. International handbook of research on conceptual change, pages 11–30.

Vosniadou, S. (2013b). International handbook of research on conceptual change. Routledge.

Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., and Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. Learning and instruction, 11(4-5):381–419.

Vosniadou, S., Vamvakoussi, X., Skopeliti, I., and Vosniadou, S. (2008). The framework theory approach to the problem of conceptual change. International handbook of research on conceptual change, pages 3–34.

Vygotski, L. (1978). Mind in society: The development of higher psychological processes. Harvard Univ Pr.

Willard, T., Roseman, J., and Plenary, L. (2007). Progression of understanding of the reasons for seasons. Knowledge Sharing Institute of the Center for Curriculum Materials in Science, Washington, DC.

Yang, J., Liu, H., and Huang, Z. (2010). Smap: To generate the personalized learning paths for different learning style learners. In Zhang, X., Zhong, S., Pan, Z., Wong, K., and Yun, R., editors, Entertainment for Education. Digital Techniques and Systems, volume 6249 of Lecture Notes in Computer Science, pages 13–22. Springer Berlin / Heidelberg.

Appendix A

AAAS plate tectonics learning goals

#	**AAAS plate tectonics learning goal**	**Text**
(1)	PT-BMK-MS1	The interior of the Earth is hot. Heat flow and movement of material within the Earth cause Earthquakes and volcanic eruptions and create mountains and ocean basins.
(2)	PT-BMK-MS2	Matching coastlines and similarities in rock types and life forms suggest that today's continents are separated parts of what was long ago a single continent.
(3)	PT-BMK-MS3	The Earth first formed in a molten state and then the surface cooled into solid rock.
(4)	PT-BMK-MS4	There are a variety of different land forms on the Earth's surface (such as coastlines, rivers, mountains, deltas, and canyons).
(5)	PT-BMK-MS5	Some changes in the Earth's surface are abrupt (such as Earthquakes and volcanic eruptions) while other changes happen very slowly (such as uplift and wearing down of mountains).
(6)	PT-BMK-MS6	Vibrations in materials set up wavelike disturbances that spread away from the source. Sound and earthquake waves are examples.
(7)	PT-BMK-HS1	The theory of plate tectonics provides an explanation for a diverse array of seemingly unrelated phenomena, and there was a scientifically sound physical explanation of how such movement could occur.
(8)	PT-BMK-HS2	Earthquakes often occur along the boundaries between colliding plates, and molten rock from below creates pressure that is released by volcanic eruptions, helping to build up mountains. Under the ocean basins, molten rock may well up between separating plates to create new ocean floor. Volcanic activity along the ocean floor may form undersea mountains, which can thrust above the ocean's surface to become islands.
(9)	PT-BMK-HS3	Ocean-floor plates may slide under continental plates, sinking deep into the Earth. The surface layers of these plates may fold, forming mountain ranges.
(10)	PT-BMK-HS4	The Earth's plates ride on a denser, hot, gradually deformable layer of the Earth.
(11)	PT-BMK-HS5	The slow movement of material within the Earth results from heat flowing out from the deep interior and the action of gravitational forces on regions of different density.
(12)	PT-BMK-HS6	The solid crust of the Earth-including both the continents and the ocean basins-consists of separate plates. The crust sections move very slowly, pressing against one another in some places, pulling apart in other places.

Appendix B

Learning goal data 1 (LGD1)

(1) The scraping of one plate on another generates powerful earthquakes; the heating of the plate within the depths of the mantle releases fluids which melt the rock over it, producing blobs of molten rock, or magma, that surface as volcanoes.

(2) In particular, four major scientific developments spurred the formulation of the plate-tectonics theory: (1) demonstration of the ruggedness and youth of the ocean floor; (2) confirmation of repeated reversals of the Earth magnetic field in the geologic past; (3) emergence of the seafloor-spreading hypothesis and associated recycling of oceanic crust; and (4) precise documentation that the world's earthquake and volcanic activity is concentrated along oceanic trenches and submarine mountain ranges.

(3) Though hidden beneath the ocean surface, the global mid-ocean ridge system is the most prominent topographic feature on the surface of our planet.

(4) The gravity-controlled sinking of a cold, denser oceanic slab into the subduction zone (called "slab pull") – dragging the rest of the plate along with it – is now considered to be the driving force of plate tectonics.

(5) Seafloor spreading over the past 100 to 200 million years has caused the Atlantic Ocean to grow from a tiny inlet of water between the continents of Europe, Africa, and the Americas into the vast ocean that exists today.

(6) Some of this newly formed magma rises toward the Earth's surface to erupt, forming a chain of volcanoes above the subduction zone.

(7) The Earth's surface is covered by a series of crustal plates.

(8) The source of heat driving the convection currents is radioactivity deep in the Earths mantle.

(9) lithosphere is the rigid and relatively cool outer layer of the earth, composed of both crust and a portion of the upper mantle.

(10) The chain runs down the middle of the Atlantic Ocean (surfacing at Iceland), around Africa, through the Indian Ocean, between Australia and Antarctica, and north through the Pacific Ocean.

(11) This heat comes mainly from two sources: the radioactive decay of unstable elements in the Earth's mantle and the energy left over from the Earth's formation.

(12) If the magma reaches the surface of the Earth, a volcano forms.

(13) The distinctive rock strata of the Karoo system in South Africa, which consists of layers of sandstone, shale and clay laced with seams of coal, were identical to those of the Santa Catarina system in Brazil.

(14) According to the plate tectonic model, the surface of the Earth consists of a series of relatively thin, but rigid, plates which are in constant motion.

(15) The surface layer of each plate is composed of oceanic crust, continental crust or a combination of both.

(16) Most of the Earth's tectonic, seismic and volcanic activity occurs at the boundaries of neighbouring plates.

(17) The science of the shaping of the Earth's crust goes by the name ”tectonics,” and the process described here is the essence of ”plate tectonics” by the Earth's crust consists of distinct plates which are continually rearranged, sometimes carrying along continents or parts of continents.

(18) Geologists came to the conclusion in the 1960's that the Earth's rigid outer layer (crust and outer, rigid layer of the mantle) was not a single piece, but was broken up into about 12 large pieces called plates.

(19) Some regions in the Earth's mantle are hotter than others; and, like most other substances on Earth, hot mantle rocks are less dense, and thus lighter, than colder mantle rocks.

(20) Plate tectonics is the theory that Earth's outer layer is made up of plates, which have moved throughout Earth's history.

(21) In locations around the world, ocean crust subducts, or slides under, other pieces of Earth's crust.

(22) The Earth's internal heat source provides the energy for our dynamic planet, supplying it with the driving force for plate-tectonic motion, and for on-going catastrophic events such as earthquakes and volcanic eruptions.

(23) Scientists think that the entire crust of the Earth is broken into big pieces called plates.

(24) Plate tectonics tells us that the Earth's rigid outer shell (lithosphere) is broken into a mosaic of oceanic and continental plates which can slide over the plastic aesthenosphere, which is the uppermost layer of the mantle.

(25) These layers include (1) the dense inner core composed largely of solid Fe and subordinate Ni, with radius of about 1200 km, (2) the molten outer core composed largely

of liquid Fe, with subordinate sulfur, with a radius of about 2250 km, (3) the mantle, composed of relatively dense rocky materials, with radius of about 2800 km thick, and (4) the crust which comprises the thin relatively light outer skin of the earth, is divisible into two types: the oceanic crust ($\tilde{7}$ km thick) and the continental crust (about 35 km thick).

(26) Although it feels solid and hard beneath our feet, the outer surface of the Earth is a thin crust of fragile rock, fractured like the cracked shell of an egg.

(27) The pieces of the shell are Earth’s tectonic plates – there are 12 major ones – and they float across a layer of soft rock like rafts in a stream, their motions driven by forces generated deep in the Earth.

(28) These boundaries, the danger lines described in the SAVAGE EARTH program ”Hell’s Crust,” are the most geologically active regions on Earth.

(29) Plate boundaries are found at the edge of the lithospheric plates and are of three types, convergent, divergent and conservative.

Appendix C

Learning goal data 2 (LGD2)

(1) In particular, four major scientific developments spurred the formulation of the plate-tectonics theory: (1) demonstration of the ruggedness and youth of the ocean floor; (2) confirmation of repeated reversals of the Earth magnetic field in the geologic past; (3) emergence of the seafloor-spreading hypothesis and associated recycling of oceanic crust; and (4) precise documentation that the world's earthquake and volcanic activity is concentrated along oceanic trenches and submarine mountain ranges.

(2) In an ocean-continent convergence, the collision of ocean and continental plates causes the accretion of marine sedimentary deposits to the edge of the continent.

(3) The Earth's internal heat source provides the energy for our dynamic planet, supplying it with the driving force for plate-tectonic motion, and for on-going catastrophic events such as earthquakes and volcanic eruptions.

(4) A single seafloor mountain chain circles Earth and contains some of Earth's tallest mountains.

(5) Plate tectonics is the theory that Earth's outer layer is made up of plates, which have moved throughout Earth's history.

(6) Over millions of years, plate tectonics has changed the appearance of the Earth's crust.

(7) **Shaping the Ocean Floor at the Mid-Ocean Ridges**

(8) These ridges, formed as the Earth's plates separate from each other, rise from the deep sea floor as volcanic mountains.

(9) Geologists came to the conclusion in the 1960's that the Earth's rigid outer layer (crust and outer, rigid layer of the mantle) was not a single piece, but was broken up into about 12 large pieces called plates.

(10) This drives the oceanic plates deep into the mantle destroying the oceanic plates.

(11) If the magma reaches the surface of the Earth, a volcano forms.

(12) Plate tectonics tells us that the Earth's rigid outer shell (lithosphere) is broken into a mosaic of oceanic and continental plates which can slide over the plastic aesthenosphere, which is the uppermost layer of the mantle.

(13) Hot volcanic material rises from the Earth's mantle to fill the gap and continuously forms new oceanic crust.

(14) The science of the shaping of the Earth's crust goes by the name "tectonics," and the process described here is the essence of "plate tectonics" by the Earth's crust consists of distinct plates which are continually rearranged, sometimes carrying along continents or parts of continents.

(15) According to the plate tectonic model, the surface of the Earth consists of a series of relatively thin, but rigid, plates which are in constant motion.

(16) The surface layer of each plate is composed of oceanic crust, continental crust or a combination of both.

(17) Most of the Earth's tectonic, seismic and volcanic activity occurs at the boundaries of neighboring plates.

(18) These plates are in constant motion causing earthquakes, mountain building, volcanism, the production of ”new” crust and the destruction of ”old” crust.

(19) The motion of the Earth’s plates help scientists to understand why earthquakes, volcanoes, and mountain building occur.

(20) It moved hundreds of miles in 135 million years at a great speed (4 inches per year!!!) The Indian plate crashed into the Eurasian plate with such speed and force that it created the tallest mountain range on Earth, the Himalayas!

(21) The pieces of the shell are Earth’s tectonic plates – there are 12 major ones – and they float across a layer of soft rock like rafts in a stream, their motions driven by forces generated deep in the Earth.

(22) Some scientists, such as David James of the Carnegie Institution of Washington, believe that the continents are anchored into the mantle by deep keels of rock that extend hundred of miles below the surface, and the continental crust and mantle therefore move in concert).

(23) In the oceans, magma reaches the surface at the boundaries between plates called spreading centers, like the Mid-Atlantic Ridge, and there new oceanic crust forms.

(24) The scraping of one plate on another generates powerful earthquakes; the heating of the plate within the depths of the mantle releases fluids which melt the rock over it, producing blobs of molten rock, or magma, that surface as volcanoes.

(25) As the plates continue to move, and more crust is formed, the ocean basin expands and a ridge system is created.

(26) **Structure of the Earth History of plate tectonics Plates Plate boundaries Forces in the Earth Faults Hypercard Resources**

(27) The Earth's surface is covered by a series of crustal plates.

(28) This heated layer is the source of lava we see in volcanoes, the source of heat that drives hot springs and geysers, and the source of raw material which pushes up the mid-oceanic ridges and forms new ocean floor.

(29) Deep in the Earth's interior, convection of the rocks, caused by temperature variations in the Earth, induces stresses that result in movement of the overlying plates.

(30) As the denser plate of oceanic crust is forced deep into the Earth's interior beneath the continental plate, a process known as subduction, it encounters high temperatures and pressures that partially melt solid rock.

(31) Most of the 600-plus active volcanoes on Earth are associated with the boundaries of the tectonic plates, the seven great plates that carry the oceans and continents.They are especially common in subduction zones, which occur when one plate dips beneath another.

(32) As the plate dives into the mantle – the layer of hot, flexible rock on which the plates glide – it gradually is heated.

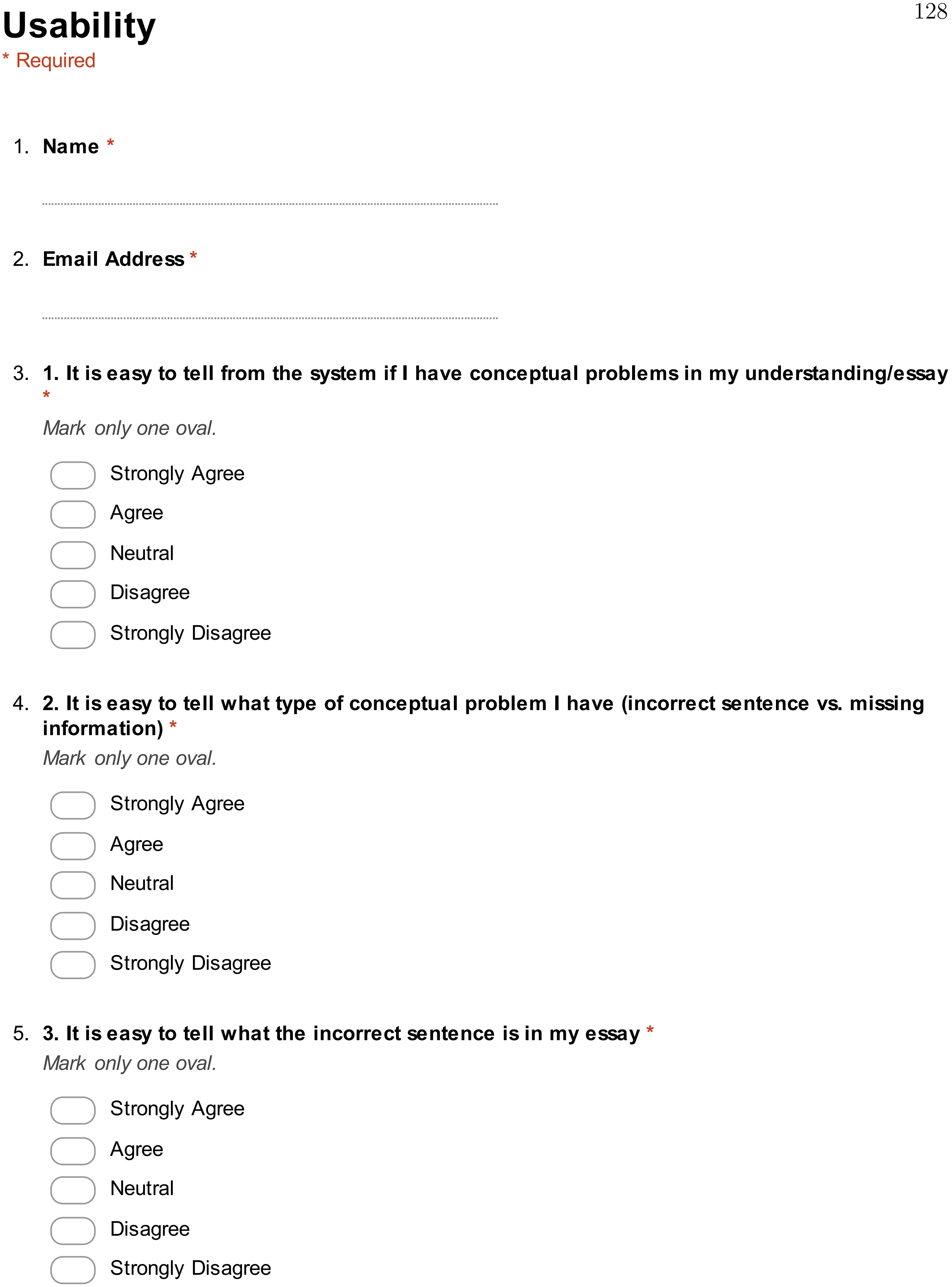
Appendix D

Pedagogical sequencing of LGD2 by two subject experts

Level	Expert1-Traditional1	Expert1-Traditional2	Expert2-Traditional	Expert1-Evidence	Expert2-Evidence
(1)	27, 9, 16	27, 9, 16	9,27	1	4
(2)	3	3	5, 14	9, 27	9, 27
(3)	29	29	15	16	5, 14
(4)	5, 15	5, 15	21	5, 15	15, 21, 12
(5)	14	14	12	14	16
(6)	21, 12	21, 12	16	12, 21	22
(7)	28	28	22	28	1
(8)	17, 18, 19	17, 18, 19	1	4	6
(9)	22	22	6	13, 23	17
(10)	6	6	17, 18, 19	8	19
(11)	2, 20	4	4	25	18
(12)	10, 32	13, 23	25	17, 18, 19	25
(13)	30	8	8	24	8
(14)	11, 31	25	31	11, 31	31
(15)	4	2, 20	2	10, 32	2
(16)	13, 23	10, 32	20	30	20
(17)	8	30	3	2, 20	3
(18)	25	11, 31	29	22	29
(19)	24	24	28	6	28
(20)	1	1	10	29	10
(21)			30	3	30
(22)			32		32
(23)			24		24
(24)			13		13
(25)			23		23
(26)			11		11

Appendix E

Usability Questionnaire



Usability

* Required

1. **Name ***

2. **Email Address ***

3. **1. It is easy to tell from the system if I have conceptual problems in my understanding/essay ***

   Mark only one oval.

   ○ Strongly Agree
   ○ Agree
   ○ Neutral
   ○ Disagree
   ○ Strongly Disagree

4. **2. It is easy to tell what type of conceptual problem I have (incorrect sentence vs. missing information) ***

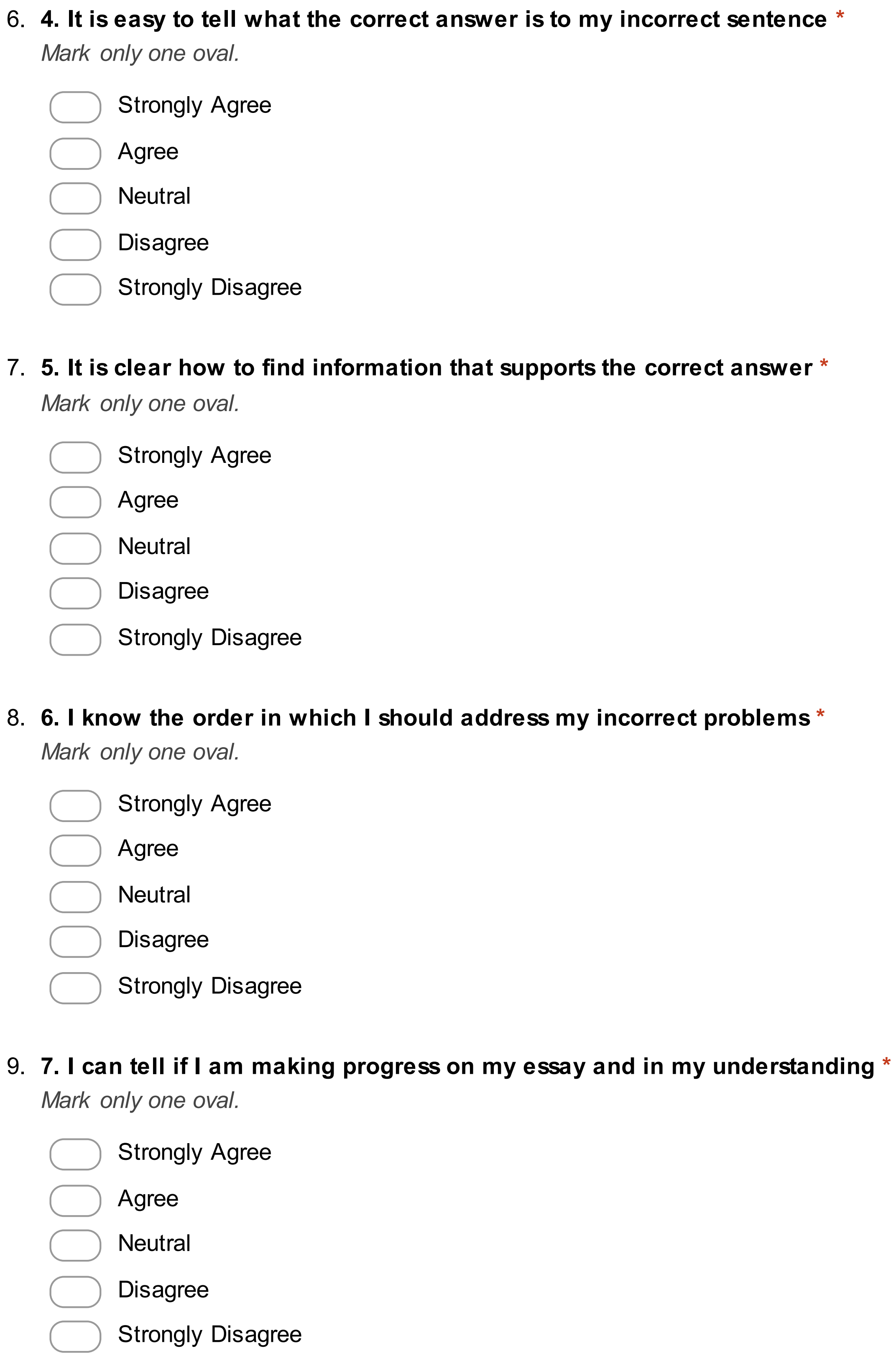
   Mark only one oval.

   ○ Strongly Agree
   ○ Agree
   ○ Neutral
   ○ Disagree
   ○ Strongly Disagree

5. **3. It is easy to tell what the incorrect sentence is in my essay ***

   Mark only one oval.

   ○ Strongly Agree
   ○ Agree
   ○ Neutral
   ○ Disagree
   ○ Strongly Disagree



6. **4. It is easy to tell what the correct answer is to my incorrect sentence** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

7. **5. It is clear how to find information that supports the correct answer** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

8. **6. I know the order in which I should address my incorrect problems** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

9. **7. I can tell if I am making progress on my essay and in my understanding** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

10. **8. The information (such as on-screen messages and button names) provided by this system is clear** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

11. **9. It is easy to tell the system if a recommended resource was useful to me** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

12. **10. This system would be helpful with my writing** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

13. **11. I would like to use this system again** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

14. **12. I would like to use this system frequently** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

15. **13. I can recommend this system to a friend** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

16. **14. Using this system improved my knowledge of seasons** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

17. **15. I liked the system’s interface** *

Mark only one oval.

○ Strongly Agree
○ Agree
○ Neutral
○ Disagree
○ Strongly Disagree

18. **16. I determined I had problems in my essay by:** *

please describe how you determined you had problems in your essay while using the system

....

....

....

....

....

19. **17. How did you determine the order in which to work on your incorrect sentences?** *

please describe how you determined what problem to tackle next in your essay

....

....

....

....

....

20. **18. The things I found most helpful in the system were:** *

....

....

....

....

....

21. **19. The things I found least helpful in the system were :** *

....

....

....

....

....

22. **20. The system can be improved by** *

....

....

....

....

....

Appendix F

Knowledge Perception 1

Knowledge Perception 1

* Required

1. **Name** *

2. **Email** *

3. **1. Where (State, Country) did you grow up?** *
   Feel free to list all the places where you lived for more than one year

4. **2. What is your major in college?** *

5. **3. What is your class level** *
   Mark only one oval.

   ○ College Freshman
   ○ College Sophomore
   ○ College Junior
   ○ College Senior
   ○ Graduated From College
   ○ Graduate Student

6. **4. Have you ever taken a middle, high school or college level class where the topic of seasons was discussed or taught?** *
   Mark only one oval.

   ○ Yes *Skip to question 7.*
   ○ No *Skip to question 8.*

7. **5. When was the last time you took a class where seasons was discussed?** *
   the year will suffice (2013, 2012, 2011, 2010, 2009 etc)

   ....

8. **6. How interested are you in the topic of seasons?** *
   Mark only one oval.

   ○ Extremely interested
   ○ Very interested
   ○ Moderately interested
   ○ A little interested
   ○ Not at all interested

7. Indicate your confidence in your knowledge of the following topics :

9. **(a) Distance of the earth from the sun** *
   Mark only one oval.

   ○ Extremely confident
   ○ Very confident
   ○ Moderately confident
   ○ A little confident
   ○ Not at all confident

10. **(b) The shape of the earth** *
    Mark only one oval.

    ○ Extremely confident
    ○ Very confident
    ○ Moderately confident
    ○ A little confident
    ○ Not at all confident

11. **(c) what causes day and night** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

12. **(d) Rotation of the earth** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

13. **(e) Revolution of the earth around the sun** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

14. **(f) Rotation of the sun** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

15. **(g) Factors affecting the temperature of a place at any given time** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

16. **(h) Factors determining the temperature cycle of a location** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

17. **(i) Seasons** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

Appendix G

Knowledge Perception 2

Knowledge Perception 2

* Required

1. **Name** *

2. **Email Address** *

3. **1. How interested are you in the topic of seasons?** *

   Mark only one oval.

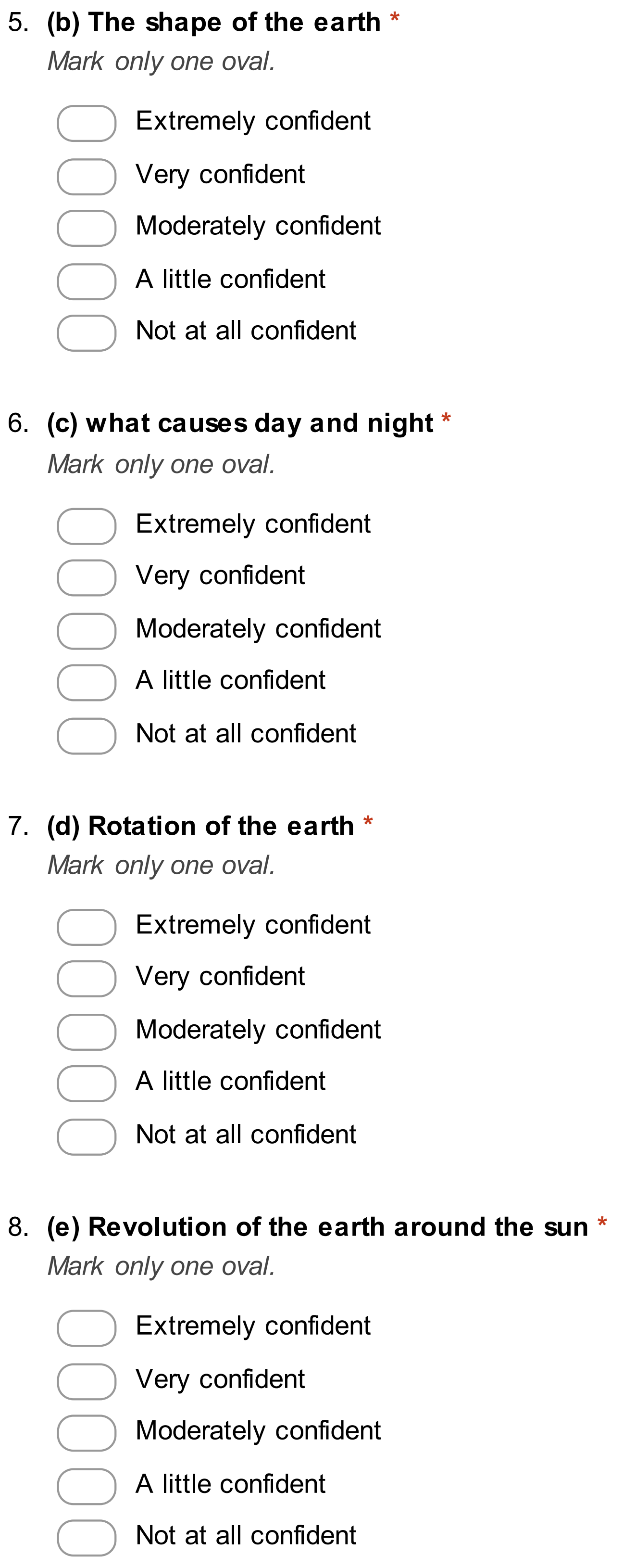
   ○ Extremely interested
   ○ Very interested
   ○ Moderately interested
   ○ A little interested
   ○ Not at all interested

2. Indicate your confidence in your knowledge of the following topics :

4. **(a) Distance of the earth from the sun** *

   Mark only one oval.

   ○ Extremely confident
   ○ Very confident
   ○ Moderately confident
   ○ A little confident
   ○ Not at all confident



5. **(b) The shape of the earth** *

Mark only one oval.

- ○ Extremely confident
- ○ Very confident
- ○ Moderately confident
- ○ A little confident
- ○ Not at all confident

6. **(c) what causes day and night** *

Mark only one oval.

- ○ Extremely confident
- ○ Very confident
- ○ Moderately confident
- ○ A little confident
- ○ Not at all confident

7. **(d) Rotation of the earth** *

Mark only one oval.

- ○ Extremely confident
- ○ Very confident
- ○ Moderately confident
- ○ A little confident
- ○ Not at all confident

8. **(e) Revolution of the earth around the sun** *

Mark only one oval.

- ○ Extremely confident
- ○ Very confident
- ○ Moderately confident
- ○ A little confident
- ○ Not at all confident

9. **(f) Rotation of the sun** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

10. **(g) Factors affecting the temperature of a place at any given time** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

11. **(h) Factors determining the temperature cycle of a location** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

12. **(i) Seasons** *

Mark only one oval.

○ Extremely confident
○ Very confident
○ Moderately confident
○ A little confident
○ Not at all confident

Appendix H

Multiple Choice

Multiple Choice

* Required

1. **Name ***

   ....

2. **1. Diagram 1 below shows the earth's path around the sun as nearly circular and Diagram 2 shows the path as strongly elliptical. What is the actual shape of the earth's path around the sun? ***

   Mark only one oval.

   ○ Nearly circular (slightly elliptical)

   ○ Strongly elliptical

   ○ The shape of the path changes so that some years it is nearly circular and other years it is strongly elliptical

   ○ Neither. The earth does not move around the sun; the sun moves around the earth

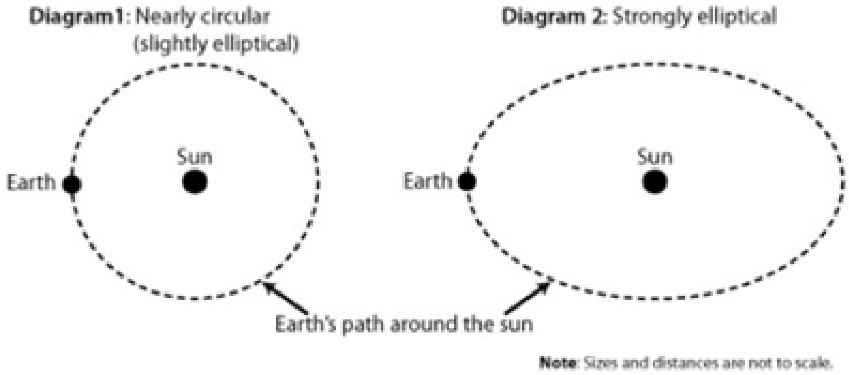


Diagram 1: Nearly circular (slightly elliptical)
Diagram 2: Strongly elliptical
Sun
Earth
Sun
Earth
Earth's path around the sun
Note: Sizes and distances are not to scale.

3. **2. On the first day of spring in the northern hemisphere, which of the following is TRUE ***

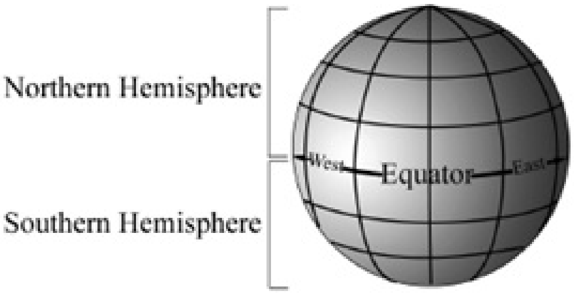
   Mark only one oval.

   ○ The North Pole is tilted 23.5 degrees toward the Sun

   ○ The North Pole is tilted 23.5 degrees away from the Sun

   ○ The North Pole is not tilted toward or away from the Sun

   ○ The North Pole is tilted 12.25 degrees toward the Sun



4. **3. At what time of year is the amount of sunlight reaching the northern hemisphere equal to the amount of sunlight reaching the southern hemisphere** *

Mark only one oval.

○ Every day of the year

○ On a day near the end of March and on a day near the end of September

○ On a day near the end of June and on a day near the end of December

○ The northern and southern hemispheres never receive the same intensity of sunlight

5. **4. Which of the following places has the greatest number of hours of daylight during a day in August?** *

Mark only one oval.

○ The North Pole

○ The equator

○ The South Pole

○ All places on earth have an equal number of hours of daylight in August

6. **5. When does the South Pole have a greater number of hours of daylight than anywhere else on earth?** *

Mark only one oval.

○ When the South Pole is angled away from the sun

○ When the South Pole is angled toward the sun

○ When the South Pole is not angled at all toward or away from the sun

○ Never

7. **6. Which of the following is TRUE about when places on earth receive 12 hours of daylight and 12 hours of darkness?** *

Mark only one oval.

○ There is one day each year when every place on earth receive 12 hours of daylight and 12 hours of darkness

○ There are two days each year when every place on earth receives 12 hours of daylight and 12 hours of darkness

○ There is one day when every place in the northern hemisphere receives 12 hours of daylight and 12 hours of darkness, and there is another day when every place in the southern hemisphere does

○ There is never a day when every place in the northern hemisphere, the southern hemisphere, or every place on earth receives 12 hours of daylight and 12 hours of darkness

8. **7. Thermal energy cannot be transferred between air and water** *

Mark only one oval.

○ TRUE

○ FALSE

9. **8. During the course of the year, the distance between the Earth and Sun currently:** *

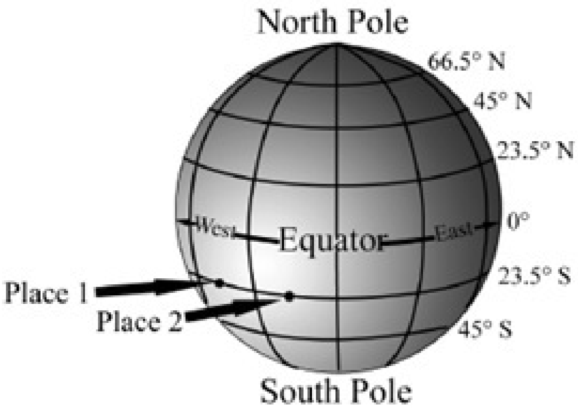
Mark only one oval.

○ changes by about 15%

○ changes by about 4%

○ changes by about 30%

○ is constant



10. **9. What must be TRUE about the number of hours of daylight at Place 1 compared to Place 2?** *

Mark only one oval.

○ There are more hours of daylight at Place 1 than at Place 2 every day of the year

○ There are fewer hours of daylight at Place 1 than at Place 2 every day of the year

○ Both Place 1 and Place 2 have the same number of hours of daylight every day of the year

○ Whether or not there are more daylight hours at Place 1 or Place 2 depends on the time of the year

11. **10. Which of the following statements is TRUE about the number of hours of daylight at the North Pole and at the equator compared to everywhere else on earth?** *

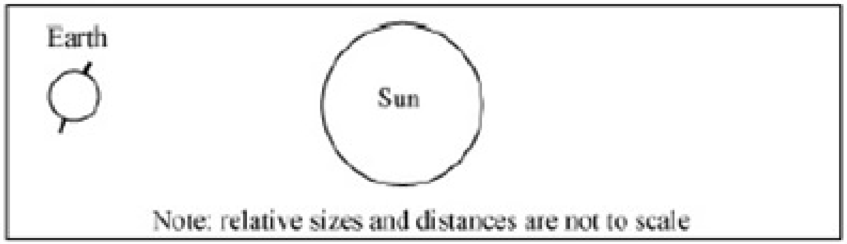
Mark only one oval.

○ The equator always has the most hours of daylight, and the North Pole always has the fewest hours of daylight

○ The North Pole always has the most hours of daylight, and the equator always has the fewest hours of daylight

○ Sometimes the North Pole has the most hours of daylight, and sometimes it has the fewest, but the equator has the same number of hours of daylight everyday

○ The number of hours of daylight on any given days is the same everywhere



Note: relative sizes and distances are not to scale

12. **11. The diagram above shows the earth with its axis of rotation pointed toward the sun. Which of the following diagrams [below] show the earth and sun six months later** *

Mark only one oval.

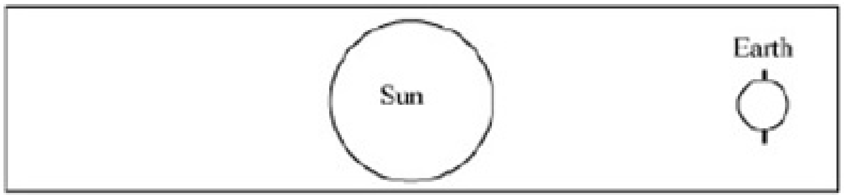
○ A

○ B

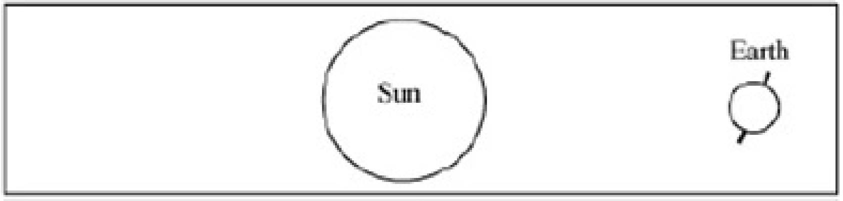
○ C

○ D

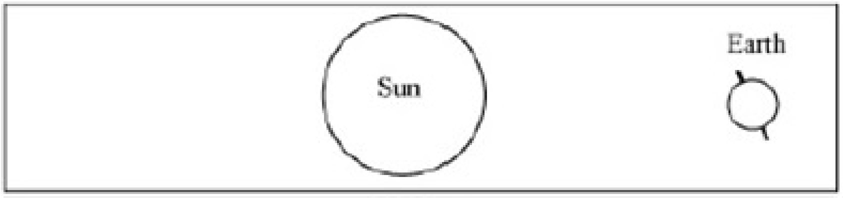
A.



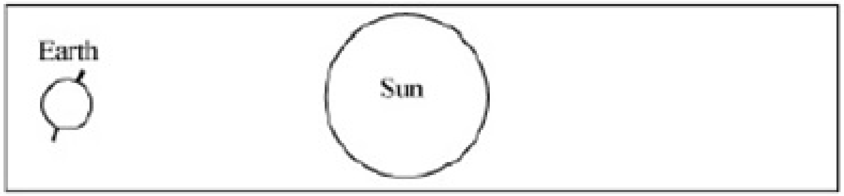
B.



C.



D.



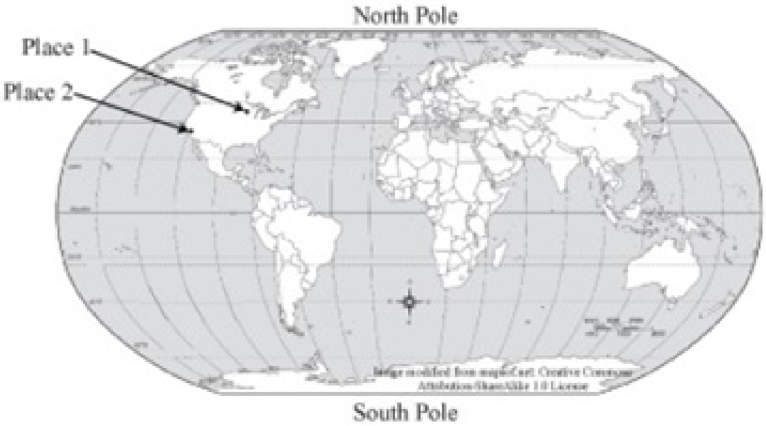


Image modified from mapsOf.net: Creative Commons Attribution-ShareAlike 1.0 License

13. **12. The map above shows two places in the United States, Place 1 and Place 2. Which of the following statements explain why one place has more hours of daylight in July than the other?** *

Mark only one oval.

○ Place 1, because it is farther inland than Place 2

○ Place 1, because it is farther north of the equator than Place 2

○ Place 2, because it is closer to the ocean than Place 1

○ Place 2, because it is closer to the equator than Place 1

14. **13. Does the amount of energy that is transferred from the sun to a given place on earth's surface change during the course of a day** *

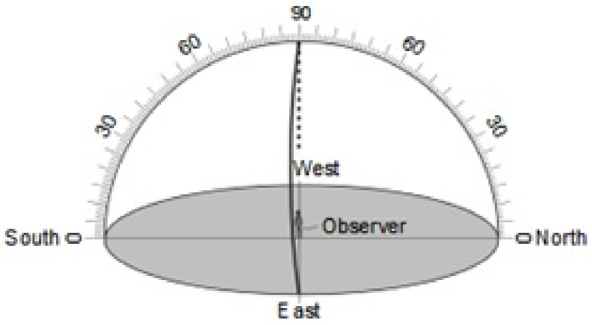
Mark only one oval.

○ The amount of energy transferred to a given place changes depending on where the sun is in the sky

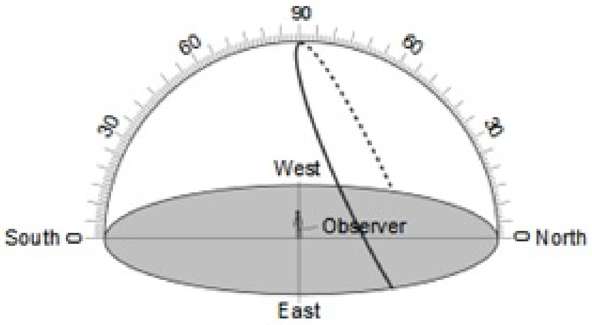
○ The amount of energy transferred to a given place only changes when the sun is blocked by clouds

○ The amount of energy transferred to a given place does not change as long as the sun is above the horizon

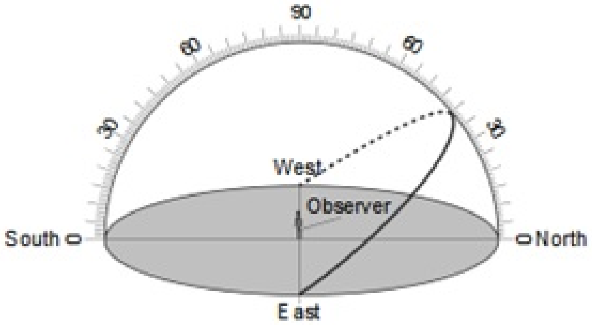
○ Sunlight does not transfer energy to the earth's surface at any time during the course of a day



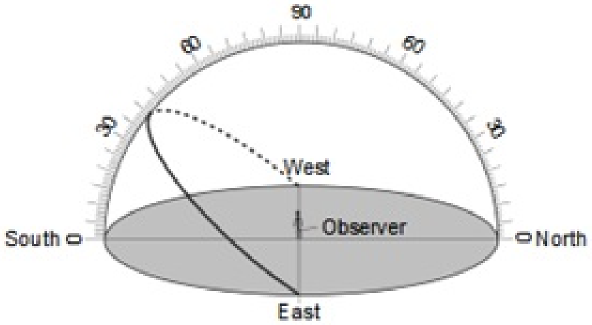
(1)



(3)



(2)



(4)

15. **14. Which diagram represents the apparent path of the Sun on March 21 for an observer at the equator?** *

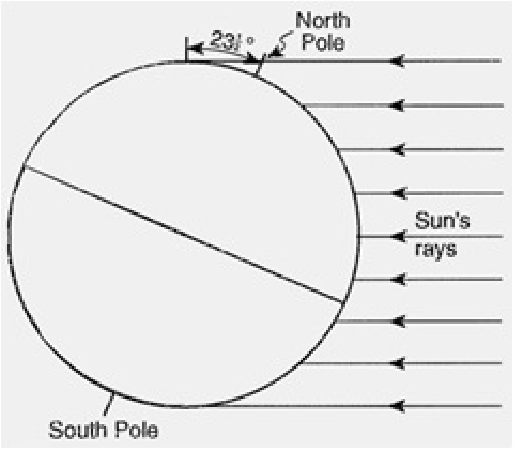
Mark only one oval.

○ 1

○ 2

○ 3

○ 4



16. **15. The diagram above represents the Sun's rays striking Earth at a position in its orbit around the Sun. What month of the year does the diagram represent?** *

Mark only one oval.

○ October

○ December

○ March

○ June

Appendix I

Application Questionnaire

Application

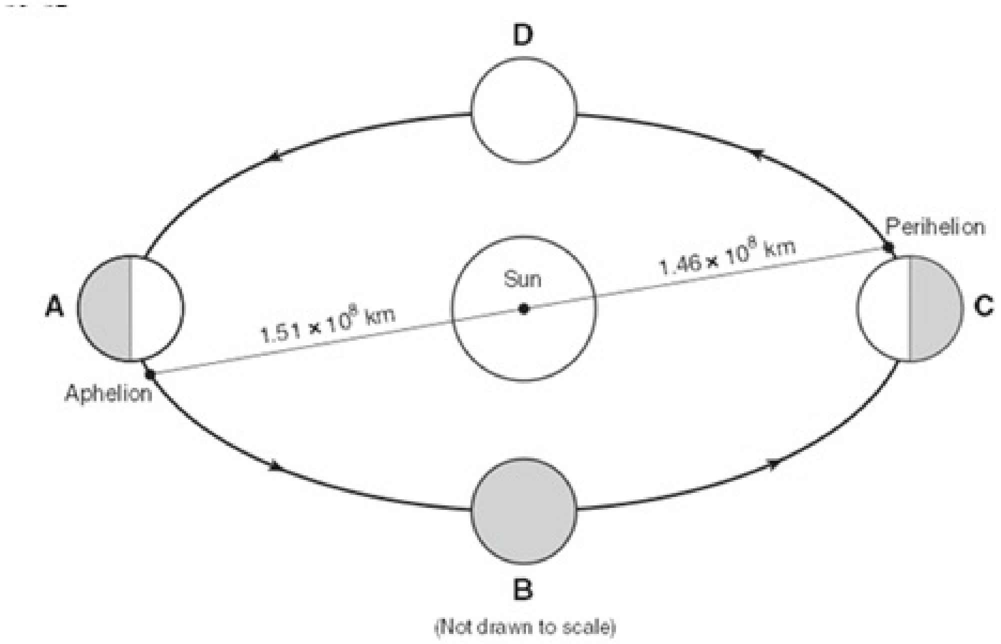
* Required

1. **Name** *

....

2. **Email** *

....



3. **1. The diagram above shows Earth revolving around the Sun. Letters A, B, C and D represent Earth's location in its orbit on the first day of the four seasons. Aphelion (farthest distance from the Sun) and perihelion (closest distance to the Sun) are labeled to show the approximate times when they occur in Earth's orbit. Label each location (A, B,C and D) with the correct season name for the Northern Hemisphere** *

....

....

....

....

....

4. **2a. Draw and label a diagram that illustrates how the path of the sun's rays to the Earth results in difference in sunlight intensity at different locations on the Earth. In your diagram, mark a location where the sun's rays are at the highest intensity and a location where the sun's rays are lower in intensity**

   Draw this on the sheet of paper provided

5. **2b. Imagine that the earth were a cube instead of a sphere. Would this change the ways in which sunlight intensity is experience on the earth? Please explain why or why not.** *

   Answer this in the space provided below

6. **3. Is the number of hours of daylight ever the same at all places on earth over the course of a single day? If so, please explain when and why this occurs. If not, please explain why this is impossible.** *

7. **4. Even though the North Pole and South Pole are "polar opposites," they both get the same amount of sunlight. But the South Pole is a lot colder than the North Pole. For example, In winter, the average temperature in the North pole is -40F(-40C) while in the South pole, it is -76F(-60C). Why?** *

8. **5. What causes day and night on Earth? ***

Appendix J

Essay Question

Most people know that when it is winter in the Northern Hemisphere, it is summer in the Southern Hemisphere. They are also aware that variation in day length at the North and South Poles is extreme, especially in winter and summer when there are very short and very long days, respectively. However, few people understand why this difference exists. Please write an essay that explains why these phenomena occur. Your explanation should be scientifically accurate and as succinct as possible, so that another student (one of your peers) can learn what they need to know to understand these phenomena from your essay. The following hints will help you get started:

- First, explain why there is seasonal variation in temperature and day length at different places on Earth.

- Next, explain how and why the annual pattern of seasonal variation is different at different locations (e.g., Boulder, Colorado [40 degrees North latitude] vs. Southern Chile [53 degrees South latitude])

- Finally, explain what the annual pattern of seasons is like at the equator and why.