The Acquisition of Procedural Skills:

An Analysis of the Worked-Example Effect Using Animated Demonstrations

by

David Lewis

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Secondary Education
College of Education
University of South Florida

Major Professor: Ann Barron, Ed.D.
Michael Coovert, Ph.D.
William Kealy, Ph.D.
Jeffrey Kromrey, Ph.D.

Date of Approval:
November 4, 2008

Keywords: Cognitive Load, Performance Efficiency, Animated Demonstration

DEDICATION

This dissertation is dedicated to my now, late, Grandmother Beuna Lewis who "always wanted to be a teacher." Unfortunately for Grandmother Lewis, World War II intervened, and she could never afford to go to college, but through her hard work, she "made sure her boys could go." This of course, led to my eventual success.

ACKNOWLEDGMENTS

This dissertation has been “under construction” for quite some time, and many people have been both encouraging and helpful during this process. My wife Jennifer in particular, supported me throughout this dissertation. She patiently listened to me whine during the frustrations of variable transformations, was there for the elation of significant differences, and helped me through the defense. She put up with the long hours, and was always there for me when I needed someone. To be honest, I do not think I could have made it without her.

This dissertation could not have been written without the assistance of my committee. Dr. Ann Barron patiently read a series of concept papers, proposals and dissertation drafts. Her suggestions and support were invaluable. The other members of my committee were also very helpful. Drs. Kromrey and Kealy were very kind to provide support during the design and analysis phases. Dr. Coovert was my advocate in the Psychology department, and made it possible for me to take several graduate courses in Psychology, which helped me develop some of the most important aspects of this study. Finally, I want to thank Dr. Lou Carey, who served on my committee until her retirement. She wisely required me to take multivariate statistics, and yes, “it is a multivariate world!” I thank you all for your questions, guidance, and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

THE ACQUISITION OF PROCEDURAL SKILLS: AN ANALYSIS OF THE WORKED-EXAMPLE EFFECT USING ANIMATED DEMONSTRATIONS

David Lewis

ABSTRACT

While many educators suggest active, rather than passive learning, this is not always the best solution, especially when learners are novices. Sweller and Cooper found learners who passively studied worked examples were significantly more efficient than those who actively solved problems (Cooper & Sweller, 1987; Sweller & Cooper, 1985) later described as the "worked-example effect" (Sweller & Chandler, 1991).

The current study tested the claims of Lewis (2005) who suggested animated demonstrations act as worked examples. It compared the performance of groups of pre-service teachers who: studied animated demonstrations (demo); studied animated demonstrations and practiced procedures (demo+practice & demo2+practice), or practiced procedures (practice).

Two MANOVAs were used to compare group performance. During week one, it was hypothesized that the demonstration learners would out-perform those in the practice condition given performance time and accuracy. It was found that there was a significant difference between groups, Wilks' Λ=0.68, F (2, 68) = 6.83, p <0.0001, η^2=0.32. Post hoc comparisons with Scheffé's test (p<0.025) revealed that the demonstration groups (demo+practice and demo2+practice groups) assembled the problem, in significantly less time than the practice group, which is positive evidence for the worked-example effect

(Sweller & Chandler, 1991) given animated demonstrations. During week two, a similar MANOVA revealed no differences between groups.

While this study considered learner performance from a human computer interaction (HCI) perspective, it also considered learners from a cognitive load perspective, by measuring relative condition efficiency (Paas & van Merriënboer, 1993). In addition, it developed a new measure called performance efficiency. During week one, the demonstration conditions were found to be significantly different F (2, 68) = 3.69, p =0.03, given relative condition efficiency. This is positive evidence of the variability effect. However in post hoc comparisons these instructional conditions were not found to differ. Performance efficiency was found to be significantly different, during week one, F (2, 68) = 12.95, p<0.0001, and post hoc comparisons with Scheffé's test (p<0.05) revealed the demonstration learners were significantly more efficient, than the practice learners. During week two, groups were not significantly different, so once learners had practiced procedures, they performed equally well.

CHAPTER ONE - INTRODUCTION

Should we require an unprepared, novice learner to practice a procedure? Some educators would answer this question “No.” Their reasoning is that they should prepare that learner, by first demonstrating the procedure. However, demonstration requires passive learning. So given this reasoning, some educators would answer the question “yes,” because they feel procedure-based learning requires active involvement.

Many well-known educators have questioned passive learning, and instead suggest an active construction of knowledge (Bruner, 1961; Dewey, 1916/1997; Jonassen, 1991; Wittrock, 1974). Even though this philosophy has a rich literature, there is a wealth of empirical evidence to suggest otherwise. A series of empirical studies over the past twenty years have shown that active problem solving during early schema acquisition is a less effective instructional strategy, than allowing learners to learn by studying worked examples (Paas & van Merriënboer, 1993; Sweller, 1988; Sweller, 2006; Tuovinen & Sweller, 1999).

Could discovery learning, a decades-old instructional strategy, be ill-advised? Or is there another explanation? These questions and others are the focus of this study, for the dissertation considers these two instructional strategies, to question the timing of practice during early schema acquisition.

Statement of the Problem

Sweller and Cooper reported that those learners who passively studied worked examples during early schema acquisition, significantly out-performed their peers, who had learned the same procedures through active problem solving (Cooper & Sweller, 1987; Sweller & Cooper, 1985). Sweller and Chandler (1991) described this phenomenon as the “worked-example effect.” This effect has been replicated by many researchers under a variety of circumstances (Carroll, 1994; Paas & van Merriënboer, 1994; Quilici & Mayer, 1996; Zhu & Simon, 1987). Specifically, they described this effect by saying a “decreased solution time was accompanied by a decrease in the number of mathematical errors” (Sweller & Cooper, 1985, p.59), thus this study considers these variables but describes them as performance time and accuracy.

To test the worked-example effect, this dissertation considers the performance of those who study an animated demonstration, a form of animated worked example (Lewis, 2005). Instructional designers may develop animated worked examples by recording computer-based procedures. These animated demonstrations may be designed to make efficient use of both visual and verbal modalities. This allows for multimedia learning (Mayer, 2001).

However, Palmiter (1991) found evidence of a delayed performance decrement given animated demonstrations, later described as Palmiter’s animation deficit (Lipps, Trafton, & Gray, 1998). Tuovinen and Sweller (1999), also proposed retention may be an issue given worked examples, and asked future researchers to consider the durability of learning given worked example based instruction. Although cognitive load researchers have repeatedly found worked examples to be effective, few (if any) have studied animated demonstrations, or the worked-example effect given this presentation form, so this dissertation accepts this responsibility.

Purpose

The purpose of this dissertation is to assess animated demonstrations from both human computer interaction (HCI) and cognitive load perspectives, to (1) consider the worked example and variability effects using animated demonstrations (Paas & van Merriënboer, 1994; Sweller & Chandler, 1991); and (2) determine if demonstration learners exhibit a delayed performance decrement, Palmiter's animation deficit (Lipps et al., 1998; Palmiter, 1991).

To study cognitive load, researchers typically combine measures of performance and perceived mental effort, to assess the relative efficiency of instructional materials (Paas & van Merriënboer, 1993; Paas, Tuovinen, Tabbers, & Van Gerven, 2003). This study considered animated demonstrations given relative condition efficiency (Paas and van Merriënboer, 1993) but also developed a new metric for measuring learner performance, called performance efficiency. In order to compare these results with those of Tuovinen and Sweller (1999), the dissertation considered pre-service teachers.

Rationale

Before the 1990s, Educational researchers were often interested in comparing the effects of media on learning, but this usually led them to find no significant differences, also known as "the no significant difference phenomenon" (Russell, 1999). Eventually a famous set of articles discussed this phenomenon, which later came to be called, the Clark-Kozma debate (Clark, 1983; Clark, 1994; Kozma, 1991; Kozma, 1994).

While the Clark-Kozma debate did not find an immediate resolution, eventually, educational psychologist, Richard Mayer wrote an important article entitled *Multimedia learning: Are We Asking the Right Questions?* (Mayer, 1997). In this article, Mayer concludes that: “Instructional development is too often based on what computers can do, rather than on a research theory of how students learn with technology" (Mayer, 1997 p.17). In doing so, Mayer (1997) repeatedly referred to Sweller’s work and the learners limited working memory load, to develop a cognitive theory of verbal and visual knowledge construction.

Mayer’s theory later came to be called “A cognitive theory of multimedia learning” (Mayer & Moreno, 1998) or simply “Multimedia learning” (Mayer, 1997; Mayer, 2001). Therefore this dissertation considers both cognitive load theory and multimedia learning, to contrast several instructional strategies, rather than the effects of media on learning.

Instructional Strategies and the Research Questions

The literature review (Chapter two) found Tuovinen and Sweller had some reservations about worked examples and retention (Tuovinen & Sweller, 1999). They implied that retention may not be as durable with worked examples. They also asked future researchers to consider retention with worked examples over time. Therefore this dissertation contrasts two main instructional strategies, discovery problem solving versus animated demonstrations (Bruner, 1961).

In addition, Palmiter (1993) proposed a mimicry model of learning with animated demonstrations. However, a review of the psychological literature, found that memories must be reconstructed and are not simply played back (Loftus, 1975; Loftus & Hoffman, 1989).

Palmiter's model for procedural learning from animated demonstrations will be tested, by introducing a group of learners to a mimicry condition in which they encounter an identical animated demonstration to the problem being solved. A week later these same individuals were required to complete a different problem scenario and their performance was contrasted with those that learned from a different animated demonstration, as well as those who learn through discovery problem solving.

Finally learners may also encounter an animated demonstration, but then not practice the learned procedures until sometime later. Therefore a final instructional strategy was considered, one where learners are taught with an animated demonstration, but then were not allowed to practice until one week after initial instruction.

Thus, in order to study the instructional effectiveness of these instructional conditions, the following research questions were analyzed:

Question 1: Is there a significant difference among the instructional strategies, relative to performance time?

Question 2: Is there a significant difference among the instructional strategies, relative to accuracy?

Question 3: Is there a significant difference among the instructional strategies, relative to "relative condition efficiency?"

Question 4: Is there a significant difference among the instructional strategies, relative to "performance efficiency?"

Researchers who study learning via animation, typically only gather data through pencil and paper tests, and therefore usually only assess conceptual or declarative knowledge. Observation is a more practical method to assess procedural learning. In addition, systematic observation offers a higher degree of certainty and replicability than other less-structured methods (Bakeman & Gottman, 1986; Knupfer & McLellan, 1996). Observation is the fundamental basis of science, but is often the most under-used and under-valued means of data collection, given human performance (Pershing, Warren, & Rowe, 2006). Thus, this project studies procedure-based learning with animated demonstrations, with observation as the primary means of data collection. In order to accomplish this goal, the study utilized the screen capture technologies of HCI research, to monitor learner behavior and assess procedure-based learning.

However, observational measures alone are seldom recommended (Gall, Borg, & Gall, 1996; Pershing, Warren, & Rowe, 2006). So in addition, "relative condition efficiency" (Paas & van Merriënboer, 1993), and "performance efficiency" were also documented. Procedures for data collection are outlined in Chapter 3.

Limitations

According to Kirschner, Sweller, and Clark (2006), "Learning, in turn, is defined as a change in long-term memory" (Kirschner, Sweller, & Clark, 2006, p.75). These authors were able to make this statement because of recent advances in the cognitive sciences, specifically in brain imaging technologies, which have mapped the regions of the brain necessary for learning (Anderson, Albert, & Fincham, 2005). Even though this is the case, it is still difficult to directly measure changes in long-term memory. Thus current technologies limit educational researchers to only indirectly measuring learning, by observing behavior. Unfortunately, because learners may stumble across the problem solving operators required to solve a problem, it will not be known if the learner has actually learned how to perform the procedure, only that they have performed the procedure.

The methodology of this study may be described as “computer-supported data collection,” which is the most accurate way to record learner actions (Knupfer & McLellan, 1996). Unfortunately not all user actions may be described using the recording technologies employed in this study. For instance, although mouse clicks (or “mousedown” events) are recorded, a “mouseup” event is not recorded. This is unfortunate, because researchers must decide when the learner ends some procedures. Given this basic limitation of the recording technology, researchers must define some learner actions themselves, allowing for some measurement error.

Delimitations

Delimitations describe the populations to which a study’s results may be generalized (Locke, Spirduso, & Silverman, 2000). The participants in this study were pre-service teachers taking a required, lower level, educational technology course at a large southeastern university. This diverse group of individuals is fairly representative of college-aged adults, although the sample studied contained more females than males.

This study measured learner performance given computer-based instruction. Specifically, it only measured on-screen interaction, a limited form of human-computer interaction. Also, it primarily measured the behavior of novices during learning. Therefore the results of this study may only be generalized to adult learners, specifically novices, engaged in human-computer interaction.

Terminology

This project brings together the research of several fields of study, thus there is a broad array of terms used in this document. If necessary, please consult Table 1 for the definition of common terms.

Table 1

Terminology

Term or acronym	Definition
animated demonstration	A narrated animation depicting procedural tasks
cognitive load	The load placed on working memory. Sweller, Van Merriënboer, and Paas (1998) describe three types of cognitive load – intrinsic, extraneous and germane cognitive load.
cognitive load theory	John Sweller has synthesized several theories (working memory, schema acquisition, and instructional design theory) to derive his own theory of human performance given the information processing requirements of instructional materials (Sweller, 1988). Sweller and others have used this theory to predict and document a number of important learning effects associated with the complexity of instructional materials (e.g. the Worked-example effect, Completion problem effect, Split-attention effect, Modality effect, Redundancy effect, Variability effect, and the Element interactivity effect).
completion problem effect	Paas (1992) found that those learners who study and use partially worked-out examples (completion problems), performed significantly better (took less time with less effort), than their peers who used traditional problem solving strategies.
declarative learning	Declarative learning is concerned with the learning of language-based information (e.g. facts and events) (Squire & Zola, 1996).
discovery learning	A type of learning that became popular in the 1960s. Proponents suggest that when one discovers information for oneself, he or she is more likely to remember it (Bruner, 1961).
element interactivity	According to Sweller, instructional content is

composed of component parts or “elements” (Sweller & Chandler, 1994). Elements may be said to “interact” if there is a relationship between them, thus raising the complexity of the instruction. The total number of elements is not as important as the number of interactions between these elements.

expertise reversal effect

As learners become more competent, the worked-example and other cognitive load effects disappear (Kalyuga, Chandler, & Sweller, 1998); this has been termed the “expertise reversal effect” (Kalyuga, Ayres, Chandler, & Sweller, 2003).

extraneous cognitive load

Extraneous cognitive load is that load not inherent within the activity (Chandler & Sweller, 1991; Chandler & Sweller, 1992), but is load that may be controlled by the instructional designer as they structure and present instructional materials (Pollock, Chandler, & Sweller, 2002).

functional magnetic resonance imaging (fMRI)

This brain imaging technique allows researchers to better understand the cognitive functions of the brain. For instance, learning theorists Anderson, Albert, and Fincham (2005) have used this technique to better understand what areas of the brain are used during problem solving.

germane cognitive load

Germane (or Relevant) cognitive load is load directed toward schema construction (Sweller, Van Merriënboer, & Paas, 1998).

HCI
(human-computer interaction)

Human computer interaction includes the reciprocal events related to the behavior of humans and computers (also known as human-computer interaction, or in instructional settings as learner interaction) (Wagner, 1994; Moore, 1989).

intrinsic cognitive load

Intrinsic cognitive load is the inherent level of difficulty or complexity associated with an instructional activity (Chandler & Sweller, 1991; Chandler & Sweller, 1992).

job aid

This is a text-based list of instructions (Rossett & Gautier-Downes, 1991).

learning efficiency

The combination of perceived mental effort ratings during training, and subsequent test performance scores (Paas, Tuovinen, Tabbers & Van Gerven, 2003).

marker

A small flag placed on the Morae Manger timeline. It represents a researcher designated event or action. For instance, it may represent the end of a user action.

modality effect

This effect suggests learners have superior performance given multimedia (dual modality - visual and verbal) based instructional materials (Moreno & Mayer, 1999; Mousavi, Low, & Sweller, 1995; Mayer, 2001; Penney, 1989).

procedural learning

This is skills-based learning (e.g. learning how to use a computer program) (Squire & Zola, 1996). When one is learning "how to" do something, they are engaged in procedural learning.

problem-solving operator

This is "an action that transforms one state into another state." (e.g. in a maze, the operators are going from one location to another) (Anderson, 1993, p.36)

relative condition efficiency

Relative condition efficiency is "the observed relation between mental effort and performance in a particular condition in relation to a hypothetical baseline condition in which each unit of invested mental effort equals one unit of performance" (Paas & van Merrienboer, 1993, p. 739).

schema

Schema describes "…a structure which allows problem solvers to recognize a problem state as belonging to a particular category of problem states that normally require particular moves." (Sweller, 1988, p. 259).

segment

A section of video within the Morae video file that has been designated by a researcher. It begins with an "in point" and ends with an "end point."

split-attention effect

Chandler and Sweller (1992) found that this learning effect is evident, when learners are required to split their attention between different source of information (e.g., text and diagrams).

variability effect

Paas and van Merriënboer (1994) found that learners who studied high-variability examples performed better than those who learned through problem solving.

worked example

"A worked example is a step-by-step demonstration of how to perform a task or how to solve a problem" (Clark, Nguyen, & Sweller, 2006a, p. 190)

worked-example effect

Sweller and Cooper found learners who studied worked examples performed significantly better than learners who actively solved problems (Cooper & Sweller, 1987; Sweller & Cooper, 1985).

This concludes Chapter one, the introduction. This chapter has outlined the purpose of this study and posed the research questions. This dissertation compares two instructional strategies, given both an HCI and cognitive load perspective. In addition, the study's methodology was briefly described. Finally, this chapter described the limitations of the methodology (behavior analysis) and the tools of data collection. In conclusion, it should be stated that Chapter one was just a brief introduction to the study.

Chapter two is an extensive literature review which describes cognitive load theory, as it relates to the design of instructional materials. Chapter three describes the methodology of the study. Chapter four describes the results, and finally Chapter four concludes the dissertation, by discussing the significance of the results, and relates them to the field of Instructional Technology.

CHAPTER TWO – LITERATURE REVIEW

This dissertation argues that animated demonstrations act as worked examples by reducing extraneous cognitive load, to promote increased learner performance during early schema acquisition. This chapter lays a foundation for the argument. To support this argument the chapter reviews: learning theories (schema and cognitive load theories); instructional design methodologies (including problem solving, worked examples, and discovery learning); and discusses animation as an instructional strategy.

Learning and Memory

Learning and memory are closely related. Cognitive load theory is where they meet instructional design. Each of these topics will be discussed at length in this review. To ensure a comprehensive review of cognitive load theory, the theory is explained in the context of human memory and learning. This section describes the development of the theoretical framework underlying cognitive load theory, what Sweller (2003) terms the "human cognitive architecture."

The Human Cognitive Architecture

Sweller (2003) refers to "the human cognitive architecture" as the theoretical structures within human memory. In particular, he relies heavily on the Atkinson and Shiffrin model (Atkinson & Shiffrin, 1968). This section of the chapter also discusses the Baddeley and Hitch model (Baddeley & Hitch, 1974), and the section closes by discussing these models as they relate to cognitive load theory.

Short-Term Memory

Until the 1950s, memory was considered to be a single unitary system (Baddeley, 2006), but a multistage model of memory was considered as early as 1890, when William James proposed “primary” and “secondary” memory in his classic text *The Principles of Psychology* (James, 1890). However, Psychology steered away from memory theory, to concentrate on behavioral theories until the late 1950s, when George Miller noted that we have a limited ability to process information (Miller, 1956).

Miller found that humans are only able to retain seven plus or minus two “chunks” of information (Miller, 1956). The importance of this observation was that short term memory had a limited capacity. Peterson and Peterson (1959) later found that in addition to limited capacity, short term memory has a limited duration. That is, they found that we can only recall information over brief intervals of time (less than 30 seconds).

Even though our memory is limited, Miller proposed that we have ways around our limitations. He found that we are able to recode, or reorganize information into “chunks” to better recall that information later (Miller, 1956). This idea of chunking will be discussed in great detail, later in this chapter.

The Atkinson and Shiffrin Model

Even though James had proposed a multistage memory model, in the 1960s some researchers were opposed to dividing memory and argued for a unified theory of memory (Melton, 1963; Postman, 1963). It was within this context that Atkinson and Shiffrin (1968) described a three-component memory model, which included a sensory register, a short-term store, and a long-term store (See Figure 1). This model has been generally well received, and is the basis for the memory models taught in many psychology textbooks (e.g. Sternberg, 2002).

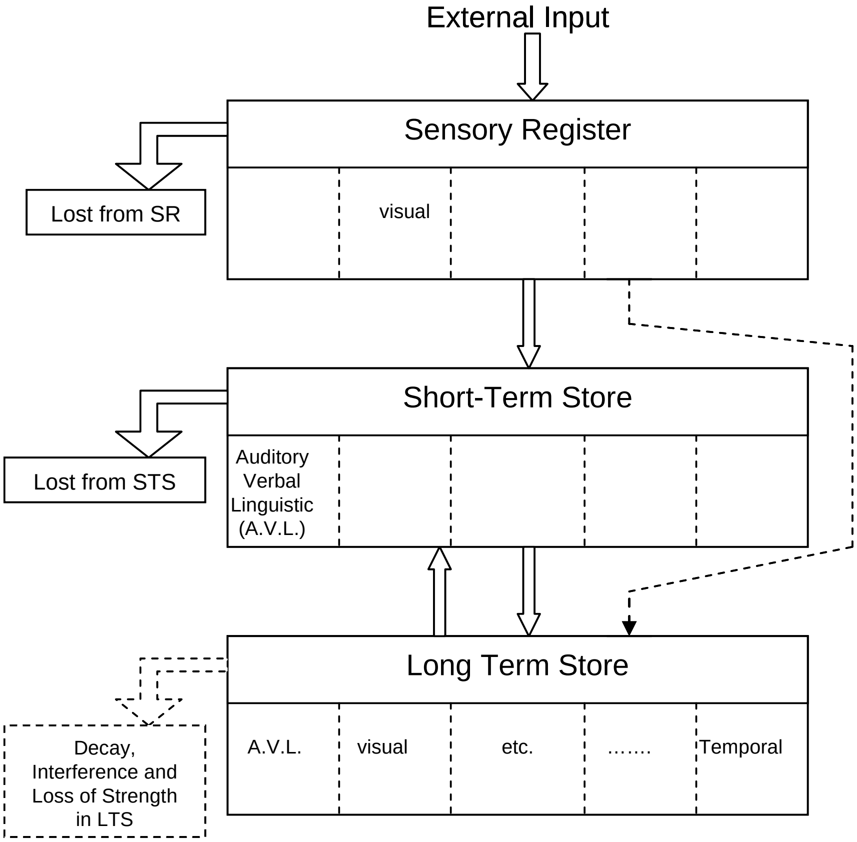


Figure 1. Atkinson and Shiffrin model

Note: Adapted from “Human memory: a proposed system and its control processes,” by Atkinson, R.C. & Shiffrin, R.M. (1968), In *K.W. Spence (Ed.), The psychology of learning and motivation: Advances in research and theory, Vol. 2* (pp. 89–195). New York: Academic Press. p.93

The Atkinson and Shiffrin model describes some components of human memory as being permanent or impermanent. The permanent components are described as “built in,” or innate to the system, whereas impermanent components are learned processes. An example permanent component is the “a-v-l short-term store,” which processes auditory-verbal-linguistic information.

Atkinson and Shiffrin (1968) also proposed impermanent processes that work as “control processes.” They describe these “control processes” as “any schemes, coding techniques or mnemonics that could be used to remember information” (Atkinson & Shiffrin, 1968, p. 106). These are learned strategies for manipulating information and may be unlimited in number. Mnemonics are an example learned strategy (or impermanent component), which allows one to manipulate information within memory.

From a learning perspective, the most important contribution of the Atkinson and Shiffrin model is its description of the processes within, and between each of the stores. Atkinson and Shiffrin (1968) use a computer analogy to describe these processes and how they are associated with each of the information stores. They suggest humans are able to encode information, transform it, and later retrieve that information.

To explain the model’s processes, Atkinson and Shiffrin discussed an example, the complicated processes involved in reading. This example is important because it shows the system at work. During reading, humans “transform” or recode information, the visual information we receive from our eyes, into verbal information (the meaning of the text). This happens in the short-term store. Because of our ability to store information, we are also able to “encode” that short-term verbal information (the meaning of the text) within the “long-term store.” Later, we are able to recall and remember that text, “to retrieve” it from long term memory. Thus, “retrieval processes” allow us to remember what we have read for later use.

Working Memory

In the 1970s, the term “short-term store” was replaced with “working memory,” which was popularized by Baddeley and Hitch (1974). However, Atkinson and Shiffrin had already used this term, “working memory” (Atkinson and Shiffrin, 1968, p.92) to describe their “short-term store.” Baddeley and Hitch (1974) capitalized on this idea to produce their own model, but they concentrated specifically on working memory (See Figure 2). They divide working memory into three subcomponents: the phonological loop (or “articulatory loop”), the visual-spatial sketchpad, and the central executive.

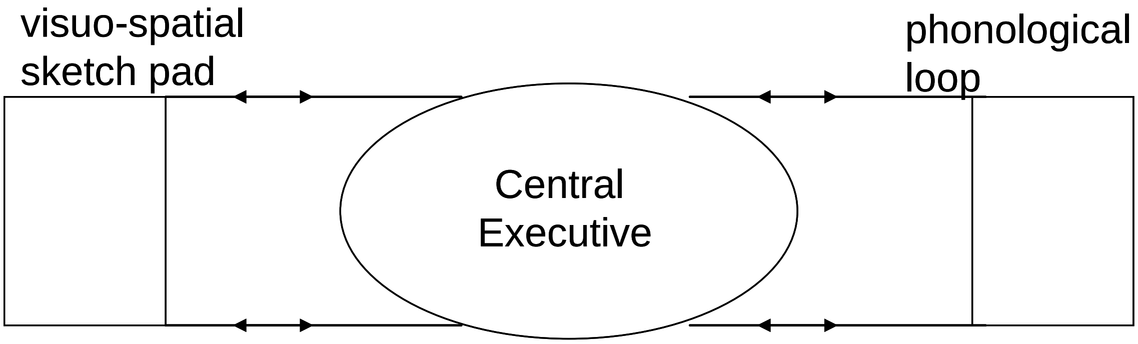


Figure 2. The Baddeley and Hitch working memory model.

Note. Adapted from “The episodic buffer: a new component of working memory?” by A. Baddeley, 2000, *Trends in Cognitive Sciences. 4* p.418

Recent Working Memory Research

The Baddeley and Hitch model is often cited, and during the 1990’s, researchers were even able to find neurological evidence to support this model, using both functional magnetic resonance imaging (fMRI) or positron emission tomography (PET). Researchers were able to use these technologies to find the neural correlates of Baddeley and Hitch’s visual-spatial sketchpad (Jonides, Smith, Koeppe, Awh, Minoshima, & Mintun, 1993), articulatory loop (Paulesu, Frith, & Frackowiak, 1993), and the central executive (D’Esposito, Detre, Alsop, Shin, Atlas, & Grossman, 1995; D’Esposito, Aguirre, Zarahn, Ballard, Shin, & Lease, 1998).

Even though the Baddeley and Hitch working memory model has been well accepted, it is not without its critics. Since its early inception, the central executive has received much attention and criticism. Even though Baddeley and Hitch (1974) gave the central executive high importance in their model, they suggested there was little evidence for this subcomponent of working memory. They included it because Atkinson and Shiffrin (1971) had promoted the idea that a central executive-like entity must coordinate the subroutines of working memory (Baddeley & Hitch, 1974).

Critics of the Baddeley and Hitch model suggest that the central executive is a 'homunculus' — which exists in name only (Parkin, 1998). Parkin (1998) has the most relevant argument, because he suggests there is no specific brain region that plays the role of a central executive. In its place he says, "What emerges instead is a pattern of extensive heterogeneity with different executive tasks associated with different neural substrates" (Parkin, 1999, p.518). As an analogy, one might say there is no specific organ that is responsible for digestion. If you point to the stomach, someone else could easily ask "What about the small intestine?"

Certainly no one is discrediting the idea that the brain or working memory has a central executive function, but Parkin advises us that it is an oversimplification, to conclude there is a specific region in the brain which is responsible for control or consciousness. Sweller (2003) agrees with Atkinson and Shiffrin that there is a central executive function within working memory, but as Atkinson and Shiffrin propose this function is carried out by learned control processes (schemas).

The point of this discussion is that while there is plenty of debate about the exact nature of the processes within working memory, most psychologists are not debating if working memory exists (Miyake & Shah, 1999) and, it is considered to be modal as Baddeley and Hitch proposed (Sweller, 2002). However, exactly how working memory works is still under a great deal of scrutiny (Miyake & Shah, 1999), and this will probably be the case for many decades.

Sweller and most cognitive load theorists refer to the Atkinson and Shiffrin model to describe how memory works (Sweller, 2003). Perhaps this is because this model is open-ended. Even in the 1960s, Atkinson and Shiffrin understood that plenty of research still needed to be conducted before the specifics of memory models could be worked out. Although the neurological work of the past twenty to thirty years has been promising, it seems we still need much work in this area; all the more reason for cognitive load theory.

Long-term Memory and Learning

Long-term memory is a very important component of Sweller's "human cognitive architecture." Cognitive load theory in turn, relies heavily on long term memory and schema theory, as a means of explaining the differences between experts and novices. Therefore this section discusses the literature pertaining to human expertise and schema theory.

Human Expertise

Since short-term or "working" memory had been clarified by the 1960s, cognitive theorists began to focus on long-term memory or how novices become experts. Even though expertise research had begun earlier, it became much more prominent after Chase and Simon published a series of studies on chess expertise.

Chase and Simon (1973a, 1973b) were able to determine that chess masters were not mentally different from novices, but that experts had recorded a vast wealth of experiences in long-term memory. To do so, they replicated a series of studies generated 30 years earlier by De Groot (De Groot, 1965). Through experimentation Chase and

Simon were able to ascertain that an expert's memory of chess piece positions is only limited to game scenarios, not for random piece placement. Thus it is this ability to perceive familiar patterns which sets a master apart from a novice.

Simon and Gilmartin (1973) calculated an expert must have thousands of stored chess patterns in their long-term memories. Chase and Simon (1973a) described these patterns or memory structures as "chunks." Miller (1956) was the first to use this term in this context, but Chase and Simon elaborated on it to further their theories. A chunk describes an amount of information being manipulated in short-term memory. Here is how they describe a chunk in relation to short-term memory:

> Specifically, if a chess master can remember the location of 20 or more pieces on the board, but has space for only about five chunks in short-term memory, then each chunk must be composed of four or five pieces, organized in a single relational structure. (Chase & Simon, 1973a, p.56)

In short, experts have the ability to manipulate more information in a shorter period of time, because they recognize relational structures (patterns) in their domain of expertise. Eventually, the term "chunk" was replaced by another that had already been well established in the literature— schema. Schema theory is perhaps the most important component of cognitive load theory. Its origins and implications are discussed in the next few sections.

Schema Theory

Schema theory is often credited to Sir Frederic Bartlett (1932, 1958). Even though this is the case, Rumelhart (1980) cautions us that Immanuel Kant proposed a schema theory in 1787, and that Kant's theory more closely resembles modern theory than

Bartlett’s schema theory. Regardless of its origins, it should be stated that schema theory is supposed to account for all human knowledge, and because of this ambitious goal, this literature has become a complex theoretical framework.

In the 1970s, authors described schema theory in many different ways. Some described schemas as being similar to theories (Rummelhart & Norman, 1978) or procedures (Rumelhart, 1980), while others suggested they have much in common with conceptual knowledge (Bobrow & Norman, 1975; Rumelhart & Ortony, 1977). Price and Driscoll (1997) present a more modern view, and suggest “Each schema is made up of related concepts, involving both declarative and procedural knowledge” (Price & Driscoll, 1997, p.476). In short, schemas are data structures within long-term memory like Chase and Simon’s chess piece patterns, which are related to concepts or patterns of behavior (Rumelhart, 1980).

Schemas as Problem Categories

In the mid 1970s, Simon continued his work with schema theory by studying learners as they solved algebra problems (Hinsley, Hayes, & Simon, 1976). Simon and his associates reasoned that if learners used schemas to understand and interpret verbal information, they may also use them to categorize problems. Through several experiments, they found that indeed humans tended to categorize problems during problem solving, and more importantly, use their memory of problem categories to solve problems.

Chi, Feltovich, and Glaser (1981) built on Simon’s work to relate schema theory to expertise and problem solving. While studying physicists, they found that physics experts gather information to categorize problems, and once a problem is categorized, the expert uses a set of schema-specific production rules to solve that problem.

Sweller (1988) uses a similar definition to describe problem schemas and describes them in terms of problem representations or problem states. “These cognitive structures will be called schemas where a schema is defined as a structure which allows problem solvers to recognize a problem state as belonging to a particular category of problem states that normally require particular moves” (Sweller, 1988, p. 259). In this article, Sweller was referring to procedures and procedure-based learning. Before discussing cognitive load theory, this review will complete its discussion of expertise, but continue with a focus on procedure-based learning.

Automation and Procedural Learning

Authors from a cognitive perspective make a distinction between many types of learning and memory. Memory maybe considered in relation to how long things can be remembered (short-term and long-term memory) but cognitive psychologists also discuss memory in relation to what is remembered. For instance, Squire (1993) describes two types of learning and memory (procedural and declarative). Declarative learning is concerned with the learning of language-based information (e.g. facts and events), while procedural learning is skills-based learning (e.g. learning how to use a computer program) (Squire & Zola, 1996). This distinction is based upon studies involving the learning capabilities of brain injured patients, primates, and normal humans (Squire, 1986; Scoville & Milner, 1957).

Nearly thirty years ago, scientists studying amnesia patients published the following in the journal *Science*: “Amnesia seems to spare information that is based on rules or procedures, as contrasted with information that is data-based or declarative – ‘knowing how’ rather than ‘knowing that’” (Cohen & Squire, 1980, p.207).

In the 1990s, neuroscientists used brain imaging techniques to find that procedural learning is associated with the striatum (caudate nucleus and putamen) (Poldrack & Gabrieli, 2001; Squire & Zola, 1996), and that declarative learning, relies on the medial temporal lobe (Bear, Connor, & Paradiso, 2001; Grafton, Mazziotta, Presty, Friston, Frackowiak, & Phelpsis, 1992; Squire, 1992; Thompson & Kim, 1996). While brain anatomy and physiology may, or may not, seem relevant to educators, it is important to realize that because these two types of learning occur in different areas of the brain, they must have very different properties. As the next section will show, learning how to use software (the focus of this study) is the acquisition of procedural knowledge.

Procedural Knowledge Acquisition

Anderson’s ACT framework is perhaps the best explanation of procedural skill acquisition (Anderson, 1993, 2005). This framework has changed over the past thirty years, since its original conception in the 1970s (Anderson, 1976, 1983, 1993, 2005). But the underlying basis for this framework, the separation of declarative and production memory, has remained in the model throughout its long history (Anderson, 1976; Anderson, 1983). Even though this is the case, Anderson and Lebiere (1998) were able to back-up this early claim, with the neurological evidence already discussed in this chapter. Figure 3 is useful to make an explanation of the dual nature of human memory.

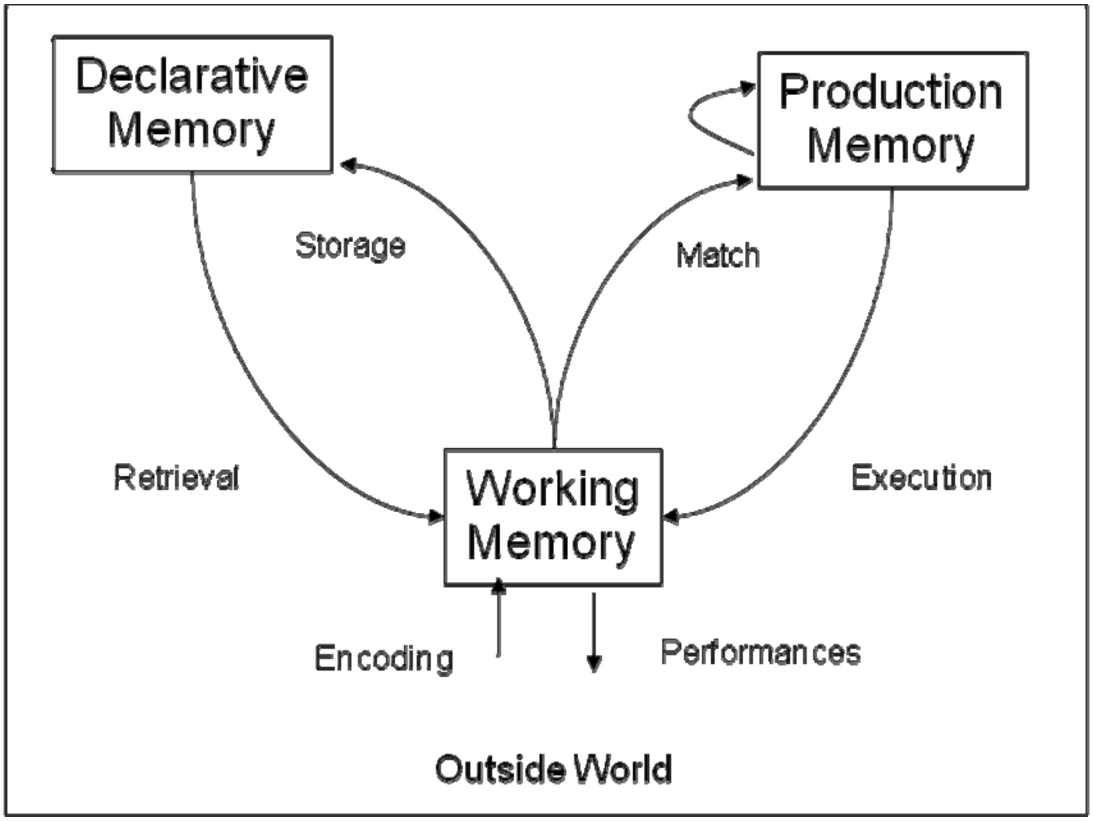


Figure 3. The ACT* model

Note: Adapted from “The architecture of cognition,” by J. R. Anderson, 1983, p.19

The ACT framework involves a series of processes which have multiple implications for human learning. Anderson describes several processes (encoding, retrieval, execution, matching, and storage) (the arrows in Figure 3). The “encoding” and “retrieval processes” of Anderson’s model has some similarities with the Atkinson and Shiffrin model, but Anderson’s model offers an extension of those processes.

Certainly “storage processes” manipulate records in long-term memory, but on the other end of the spectrum, “execution processes” (the performance) must interact with the outside world. More importantly, Anderson (1983) proposes a “matching process,” in which data about productions in working memory must correspond with data in production memory.

According to Anderson's framework the steps of a learned procedure are interpreted in production memory and organized into procedure-specific production rules (Anderson, 1983). So as a learner practices the steps of a procedure, they consider each step in the process, to develop a series of rules for the procedure (the schema). In order to learn schema based production rules, Anderson suggests learners mentally rehearse procedures, and it is even common to observe learners in this phase verbally rehearsing the steps of a procedure. Thus according to Anderson (1983) production rules require declarative if/then statements like the following:

IF the goal is move the cursor,

THEN move the mouse.

According to Anderson a grouping of production rules is called a production (Anderson, 1983). Productions are similar to what a behaviorist would have called a stimulus-response pair (Anderson, 1976), but from a cognitive perspective, because this process requires decision making and memory.

Learning by Example

Anderson's ACT* framework (Anderson, 1983) makes no allowances for example-based learning, and claims that all knowledge is recorded via declarative production rules. However, Pirolli and Anderson (1985) considered example-based processing, and their article became the first study within the ACT framework, to demonstrate the importance of examples in procedure-based learning.

Later, Anderson and Finchman (1994) altered their declarative-only origin of skill acquisition, to include learning by example. They describe this example-based processing as learning by analogy. That is, learners map the steps of an example to the current problem. In this way they can solve that problem, by mapping the procedural steps of the analogous example, to the problem at hand.

Anderson, Finchman, and Douglas (1997) describe examples as the only experience a novice has with the new problem category. Novices tend to draw upon these examples, as they would a reference book. As they practice and experience similar problems novices extract a declarative representation (abstract production rules), and switch from example-based processing to rule-based processing, in order to simplify problem solving (Anderson et al., 1997). So schema acquisition may develop during practice, by forming declarative rule-based statements (production rules) as Anderson (1976) describes, but they may also initially develop from example-based processing (Anderson & Finchman, 1994).

Three years later, Anderson, Finchman and Douglas (1997) complete this transition, to describe a complete version of the framework with four overlapping stages of skill acquisition: (a) an analogy stage, when learners refer to specific examples; (b) later learners begin to describe abstract rules; (c) then production rules and (d) retrieval of examples, that match the target problem.

Eventually, as learners continue to practice, their arduous actions may become automated, and converted into the fluid movements of an expert (Schneider & Shiffrin, 1977; Schunn & Anderson, 2001). So Anderson's stages of skill acquisition are very important, for they provide a descriptive model of how every learner progresses from being a novice, to becoming an expert. But more importantly it includes an explanation of how learners may learn by example.

Relating the Types of Memory to Learning

Before describing cognitive load theory, it is important to state that cognitive load theorists do not believe human intellect and intelligence is the result of one's ability to manipulate material in working memory (Sweller et al., 1998). Quite the contrary, Sweller and his associates believe human intellect comes from one's ability to relate experience (long term memory) to the problems at hand (Sweller et al., 1998). They even describe long term memory as the seat of human intelligence and state "From this view, human intellectual prowess comes from this stored knowledge, not from an ability to engage in long, complex chains of reasoning in working memory" (Sweller et al. 1998, p254).

According to Sweller (1993) all long term memories (1) are processed, and constrained by our limited working memories; (2) originate in working memory, during learning; (3) and finally are consolidated into chunks (or production rules) that are eventually automated. Therefore since "long-term memory is immeasurably large" (Sweller, 1993, p.1), cognitive load theory concentrates on "the weakest link" of the human cognitive architecture – working memory.

A Brief Introduction to Cognitive Load Theory

Cognitive load theory had a theoretical precedence in the educational and psychological literature, well before Sweller's 1988 article (e.g. Beatty, 1977; Marsh, 1978). Even Baddeley and Hitch (1974) questioned "concurrent memory load," but Sweller's cognitive load theory was the first to consider working memory, as it related to learning and the design of instruction.

Sweller's cognitive load theory is in agreement with Anderson's ACT framework, and Sweller and his associates often cite Anderson's work as evidence of the theory (Kirschner, Sweller, & Clark, 2006; Sweller, 1988; Sweller, van Merriënboer, Paas, 1998). However, cognitive load theory extends Anderson's work, to concentrate on initial schema acquisition.

In essence, cognitive load theory proposes that since working memory is limited, learners may be bombarded by information and, if the complexity of their instructional materials is not properly managed, this will result in a cognitive overload. This cognitive overload impairs schema acquisition, later resulting in a lower performance (Sweller, 1988).

When instructional designers develop materials they intentionally choose different means of presenting information. Instructional strategies may vary depending on the content, but they range from organizational strategies, sequencing, cues, feedback, orienting or question techniques, but, may also include different types of media (Fleming & Levie, 1993). These instructional strategies have a variety of effects on learning, depending on the media and strategies being used to present instruction (Mousavi, Low, & Sweller, 1995; Sweller & Chandler, 1991; Sweller & Cooper, 1985). A fundamental claim of cognitive load theory is that these strategies are likely to be random in their effectiveness, unless they consider the underlying cognitive architecture of the learner during instruction (Clark, Nguyen, & Swelller, 2006b).

Schema acquisition is the ultimate goal of cognitive load theory. Recall that Anderson's ACT framework found initial schema acquisition occurs by the development of schema-based production rules, but these production rules may be developed by one of two methods (Anderson et al., 1997), either by developing these rules during practice or by studying examples. As we will see later in this chapter, the second method (studying examples) is the most cognitively efficient method of instruction (Cooper and Sweller, 1987; Paas & van Merriënboer, 1993; Sweller & Cooper, 1985). This realization became one of the central tenets of cognitive load theory.

Later in the process, once learners have acquired a schema, those patterns of behavior (schemas) may be practiced to promote skill automation (Kalyuga, Ayres, Chandler, and Sweller, 2003; Shiffrin & Schneider, 1977). As discussed earlier expertise occurs much later in the process, and is when a learner automates complex cognitive skills, usually via problem solving. Thus it should be clearly stated, this study concentrates on initial schema acquisition, but for this to be a comprehensive discussion of the literature, this section first considers cognitive load theory at all stages of learning, then in later sections, returns to the topic of this dissertation, learning during initial schema acquisition.

Types of Cognitive Load

Now that this discussion has described cognitive load theory, it needs to continue by considering the different types of cognitive load. Cognitive load theorists distinguish between three types of load: intrinsic, extraneous and germane cognitive load. Sweller and his associates clearly defined intrinsic cognitive load this way "Intrinsic load is the mental work imposed by the complexity of the content" (Clark, Nguyen, & Swelller, 2006a, p. 9).

When Sweller (1993) first described intrinsic cognitive load he said "Intrinsic cognitive load is imposed by the basic characteristics of the information rather than by instructional design" (Sweller, 1993, p.6). Later Sweller and his associates described two additional types of load that instructional designers may control, as they structure the manner in which instruction is presented (Sweller, van Merriënboer, & Paas, 1998). These two additional types of load are associated with the presentation of instructional materials, extraneous cognitive load (Chandler & Sweller, 1991; Chandler & Sweller, 1992), and germane cognitive load (Sweller, van Merriënboer, & Paas, 1998).

Sweller and his associates describe "extraneous cognitive load" as that load not inherent within the instruction, but is the load which is imposed by the instructional designer as they structure and present information (Chandler & Sweller, 1991; Chandler & Sweller, 1992). Sweller provides a good example of extraneous cognitive load when he describes how a designer might present a square to a learner (Clark, Nguyen, & Swelller, 2006b). As he describes, an instructional designer may present a square as a visual, and learners would probably understand this graphic representation in a fraction of a second. However the same instructional designer might choose to present a square in verbal form (e.g. one side is vertical to the others at a 90 degree angle, while the next is at a 90 degree angle to the first, etc.). Each of these two forms of instruction present the same material, but the graphic has less extraneous cognitive load associated with it and is much more cognitively efficient. While this is a simple example, other examples are less straightforward, but could have quite different outcomes depending on the learning environment.

Recall that intrinsic cognitive load is due to the complexity of the material, as contrasted with the way an instructional designer presents that material (extraneous cognitive load). Extraneous cognitive load is a concern when intrinsic cognitive load is high (Paas, Renkl, & Sweller, 2003; Paas, Tuovinen, Tabbers, and Van Gerven, 2003). This is because intrinsic and extraneous load are additive (See Figure 4). When intrinsic load (complexity of the material) is low, the learner will probably have less trouble grasping the underlying content (Paas, Renkl, & Sweller, 2003), but instructional designers should always strive to limit extraneous cognitive load.

Finally the third type of cognitive load is germane (or relevant) load. This final type of cognitive load is that remaining free capacity in working memory, which may be redirected from extraneous load toward schema acquisition (Sweller et al., 1998). This will be discussed at length later in this chapter. Next this discussion turns its attention toward the source of intrinsic cognitive load.

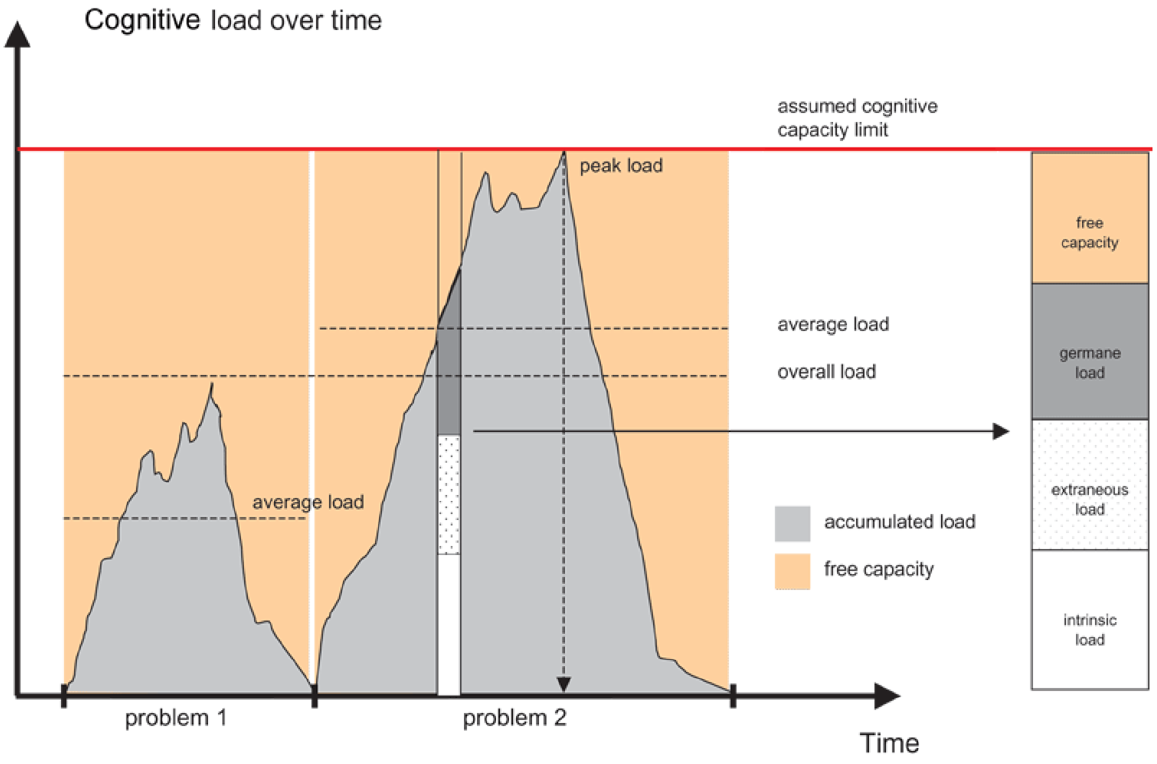


Figure 4. Cognitive load over time

Note: Adapted from “Cognitive Load Measurement as a Means to Advance Cognitive Load Theory,” by F. Paas, J.E. Tuovinen, H. Tabbers, and P. W. M. Van Gerven, 2003, Educational Psychologist, 38, p. 65

Element Interactivity

Certainly the amount of information a learner must process over a period of time is important, but the most important factor given instruction is the complexity of that information (Pollock, Chandler, & Sweller, 2002). According to Sweller and Chandler (1994), instructional content is composed of component parts or “elements;” and these elements may be said to “interact” if there is a relationship between them, raising the complexity of the instruction. Sweller and Chandler (1994) describe this phenomenon as “element interactivity.” Van Merriënboer and Sweller (2005) describe element interactivity concisely, when they mention “Working memory must inevitably be limited in capacity when dealing with novel, unorganized information because as the number of elements that needs to be organized increases linearly, the number of possible combinations increases exponentially” (van Merriënboer & Sweller, 2005, p.149).

Even though, Sweller and Chandler (1994) described the intrinsic structure of information as “unalterable,” Sweller and his associates later argued that even when the cognitive load of instruction is very high, instructional designers may artificially reduce the intrinsic load of instruction, by dividing a lesson into smaller pieces, reducing the intrinsic load of the overall lesson. Sweller describes these smaller pieces as “subschemas” (Clark, Nguyen, & Sweller, 2006b). This method of dividing the presentation of material was first developed by Pollock, Chandler, and Sweller (2002).

However, this method of dividing a lesson into subschemas promotes learning at the expense of understanding, but as Sweller explains, they were never able to understand the full schema anyway (Clark, Nguyen, & Sweller, 2006b). Thus Pollock, Chandler, and Sweller (2002) found that, if learners process the individual elements of instruction serially, rather than simultaneously, that they were able to process that instruction, to recombine these individual subschemas, and eventually understand the whole problem.

It should be noted these researchers were not the first to suggest breaking instructional materials into component parts. Gagné recognized this phenomenon in the 1960s, when he described learning hierarchies (Gagné & Paradise, 1961; Gagné, 1968). However, it is important to realize that Sweller and his associates not only recommended this method of instruction, but were also explained why Gagné's learning hierarchies are an effective means of presenting instruction.

Designing Instruction

Instructional design researchers study the practical uses of learning theory so that these theories may be used to develop instructional strategies that promote efficient, effective learning (Molenda, Reigeluth & Nelson, 2003). Several researchers have devised presentation strategies that help learners to abstract a problem schema. This next section of the literature review describes cognitive load theory as it relates to presentation strategies, problem structure, problem format, and the use of multimedia.

Problem Solving and Cognitive Load

While problem solving skills are highly valued, many problems are complex cognitive tasks that may be difficult for a novice to complete, even when they have the prerequisite skills (Sweller, 1988; van Merriënboer, 1997). Complex cognitive tasks require learners to mentally reorganize what they already know, to restructure a problem, in order to accomplish the overall task (van Merriënboer, 1997). This mental reorganization may impose a high cognitive load on the learner (Sweller, 1988), but this load may be manipulated by an instructional designer, during the design of instruction, to allow the intended learner to grasp the underlying schema (Sweller, 1993).

Means-ends Analysis and the Actions of an Expert

Problem solving has been studied for many decades, Newel and Simon studied problem solving as early as the mid 1950s (Newell, Shaw, & Simon, 1958a). Much of cognitive science and schema theory developed out of their work. Both Newel and Simon, and Sweller suggest that novices usually attempt to solve problems, by using an iterative process called "means-ends analysis."

During means-ends analysis, a learner works backwards from the problem goal, by applying problem solving operators, to achieve a sub-goal of the problem; once this sub-goal has been reached the learner then reassesses the problem, and will continue to apply other problem solving operators until the problem goal is reached (Larkin et al., 1980). Chi et al. (1981) found that experts work somewhat differently, in that they begin by first categorizing a problem, based upon the deep structure of the problem, to work forward toward a problem solution.

Ward and Sweller (1990) describe the use of a means-ends strategy this way: "A heavy cognitive load is imposed because of the need to simultaneously consider and make decisions about the current problem state, the goal state, differences between states, and problem solving operators that can be used to reduce such differences" (Ward & Sweller, 1990, p.3).

Sweller proposes that during the earliest stages of schema acquisition, the actual performance of a procedure may be detrimental to learning, because it adds an additional working memory load to the instruction, in what may be an already complex learning environment (Sweller, 1988). Sweller suggests this is because during early schema acquisition, learners who interact with content are engaging in problem-solving search, a non-schema forming activity (Sweller, 1988).

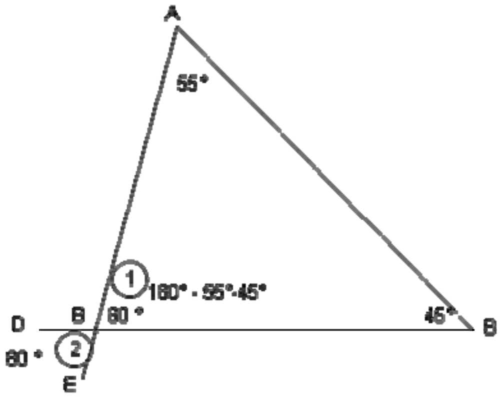
Unfortunately, some learners are required to solve problems when they are not aware of the underlying problem schema. If a learner is required to solve problems, before they understand the problem schema, they may become distracted with irrelevant aspects of a problem, spending their time searching for a problem solution, but still may not be engaged in learning (schema acquisition) (Sweller et al., 1998).

For example, in the case of photo editing (the domain of this dissertation), a novice edits a document using the software interface (using problem solving operators) until they produce the desired product (problem goal). But when novices are learning how to use graphic design tools for the first time, they usually have some difficulty and will make mistakes. The actions of an expert graphic artist are much more rapid and precise, because they work forward with a plan in mind. So even though a novice may have a problem goal in mind, and may know how to use the tools, they may not be fully aware of how to produce that problem goal. Sweller and his associates developed cognitive load theory as a means of explain this behavior. In doing so they discovered several effects that working memory or cognitive load had on memory. Examples of these learning effects are the worked-example effect (Cooper & Sweller, 1987; Sweller & Cooper, 1985), Completion problem effect (van Merriënboer & de Croock, 1992), Split-attention effect (Chandler & Sweller, 1992), the modality effect (Mayer, 2001; Mousavi, Low, & Sweller, 1995; Penney, 1989), and the variability effect (Paas & van Merriënboer, 1994). Each of these learning effects will be discussed in this section beginning with the worked-example effect, but before describing this effect it is important to first define a worked example.

What is a "Worked Example"?

Defining the term "worked example" may seem somewhat difficult, because the underlying root term, example, could be applied to almost anything. However, Atkinson and his colleagues provide a reasonable definition when they describe "worked examples" in terms of problems or procedures. They describe worked examples by saying they "typically include a problem statement and a procedure for solving the problem" (Atkinson et al., 2000, p. 181). Sweller and his associates also provide a definition; Clark, Nguyen, and Sweller, describe a worked example more in terms of a procedure, "A worked example is a step-by-step demonstration of how to perform a task or how to solve a problem" (Clark, Nguyen, & Sweller, 2006a, p. 190). So, to synthesize these two definitions, a worked example is the presentation of a procedural problem and the steps required to solve the problem. Another way to think of the term worked example is to describe it as a "solved problem."

Figure 5 is the epitome of a worked example. The box below the diagram it explains the problem statement and also lists the steps toward solution. In other words it is a solved problem.



Angle DBE = Angle ABC (vertically opposite angels are equal)
= 80 °

Figure 5. The epitome of a worked example

Many people think of their mathematics textbooks when they consider problems and worked examples, but as we all know, not all “problems” are math problems. Therefore researchers have studied the instructional effectiveness of worked examples in a variety of domains [e.g. music, chess, athletics (Atkinson, Derry, Renkl, & Wortham, 2000); physics, mathematics, or programming (Gerjets, Scheiter, and Catrambone, 2004); concept mapping (Hilbert & Renkl, 2007); statistics (Paas, 1992)] and more recently have even begun to considered ill-structured domains like art and design education (Rourke and Sweller, in press).

Types of Worked Examples

Researchers have begun describing several types of worked examples and portraying them in a variety of media. Two groups of researchers in particular (Gerjets, Scheiter, & Catrambone, 2004; van Gog, Paas, & van Merriënboer, 2004) have developed a simple nomenclature to describe worked examples.

Van Gog, Paas, and van Merriënboer (2004) described two types of worked examples, process-oriented or product-oriented worked examples. By process-oriented worked examples they mean those problems that provide a problem solution with some additional information. Specifically they included strategic and principle-based information to several solved problems. They compared the learner performance of those using these process-oriented worked examples to those who received worked examples with no additional information, which they termed "product-oriented worked examples." They found that those using the process-oriented worked examples had increased mental effort during training, but no difference in transfer performance. They later replicated these findings in another study (van Gog, Paas, & van Merriënboer, 2006). Therefore it seems in these studies, that adding strategic or principle-based information only complicated the problems with no performance gains.

Gerjets, Scheiter, and Catrambone (2004) also described two more types of worked examples, which they describe as either "molar" or "modular" worked examples. These terms come to use from the physical sciences and represents the "grain size" of the example. Gerjets et al (2004) define molar worked examples as those which focus on problem categories and their solutions (Gerjets, Scheiter, & Catrambone, 2004, p.33) whereas a modular worked examples "are broken down into smaller meaningful solution elements that can be conveyed separately" (Gerjets, Scheiter, & Catrambone, 2004, p.33). After five experiments they concluded that the processing of modular examples is associated with a lower degree of intrinsic cognitive load. This is line with the element interactivity effect (Pollock, Chandler & Sweller, 2002).

Modular worked examples are useful when the content is so complicated that the intrinsic load imposed is more than the novice can handle. That is the overall schema must be broken down into its subcomponents [described as "subschemas" (Clark, Nguyen, & Sweller, 2006b)]. However schema decomposition comes at a cost, if the material cannot be processed or presented as a whole, learners will not understand how the individual pieces are connected (Clark, Nguyen, & Sweller, 2006a). After this type of instruction, follow-up instruction is needed to allow learners to put the pieces together back together.

Thus researchers have begun to develop a nomenclature for different types of worked examples. This nomenclature is currently based upon the type of information provided within the worked example (either process or product oriented), or the "grain size" of the example (either molar or modular) (Gerjets, Scheiter, & Catrambone, 2004; van Gog, Paas, & van Merriënboer, 2006). Finally, worked examples may be classified based upon the media in which they are presented (discussed in later sections). Before considering media this discussion will first consider the cognitive load learning effects, beginning with the most documented of these effects, the worked-example effect (Sweller, 2006).

The Worked Examples Effect

In the mid 1980s, Sweller and Cooper compared learners who studied worked examples to those learning by traditional problem solving. In a series of five experiments, Sweller and Cooper (1985) measured the performance of high school learners as they learned algebra problems. They found that those students who studied worked examples took less time to process the instructional materials, and subsequently took less time to solve problems. In addition, learners also had a decrease in mathematical errors. This phenomenon has subsequently been described as "the worked-example effect" (Sweller & Chandler, 1991; Sweller et al., 1998).

Later Cooper and Sweller (1987) replicated their earlier findings, but also found evidence that those learners who used worked examples, spent less time solving transfer problems, and made significantly fewer errors on transfer problems.

In 1988, Sweller developed cognitive load theory to explain learner behaviors during early schema acquisition. Specifically, he used cognitive load theory to explain the worked-example effect. He had earlier proposed that learners, who solved problems by means-ends analysis, would have a higher working memory load, as compared with those who were prevented from using a means-ends strategy (Owen and Sweller, 1985). Sweller and Cooper (1985) were perhaps the first to use worked examples to limit means-ends analysis, during schema acquisition. Cooper and Sweller (1987) reported that this instructional strategy was designed to limit problem solving search, and developed to alleviate the cognitive load imposed on a novice. Cooper and Sweller found that by removing problem solving search, learners were more efficient, and made fewer problem solving errors.

Sweller (1988) proposed that solving problems while attempting to learn the underlying problem schema amounted to a dual task problem. He describes it this way:

> *If, as suggested above, problem solving search via means-ends analysis and schema acquisition are independent tasks, then they may be considered as primary and secondary tasks respectively, within a dual task paradigm. Under these circumstances, if a strategy such as means-ends analysis is used to accomplish the primary task (attain the problem goal), then because the strategy imposes a heavy cognitive load, fewer resources may be available for the secondary task (Sweller, 1988, p. 277).*

So, not only did Sweller (1988) provide a theory for describing why problem solving may be detrimental during early schema acquisition, but he also provided a mechanism for why it is less efficient. As he describes it, learners who study worked examples are not burdened with problem solving search. Those that must solve problems during learning must search for sub-goals toward the eventual goal of problem completion, and these two activities constitute a dual task scenario raising the learner's cognitive load. Thus, problem solving search is not necessary for schema acquisition and may even prevent learning (Sweller, 1988).

Finally, it should also be stated that while Sweller and Cooper (1985) initially considered learners who studied, multiple worked examples. Scheiter, Gerjets, and Schuh (2004) later found that multiple worked examples are not necessary, and may even increase the time required for learners to process the underlying schema. So learners may demonstrate the worked-example effect even after viewing a single example.

When are Worked Examples Warranted?

Several authors have found that learners actually prefer to learn from examples, rather than learning from other forms of instruction (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Lefevre & Dixon 1986; Pirolli & Anderson, 1985; van Lehn, 1996), but while worked-examples may be preferred, and have been found to be useful for novices, these learners must eventually practice a procedure to attain expertise (Schneider & Shiffrin, 1977).

As learners gain expertise, some researchers have suggested fading worked examples (Renkl, Atkinson, & Maier, 2000; Renkl, Atkinson, Maier, & Staley, 2002) to replace problems with partially-completed problems (van Merriënboer & de Croock,

1992) and eventually provide practice with whole problems to facilitate skill automation (Kalyuga, Ayres, Chandler, and Sweller, 2003).

Kalyuga et al. (2001) found worked examples actually hindered more advanced learners. Thus it was proposed that once a skill had been acquired, the worked examples became redundant to even overload the working memory of experts (Kalyuga et al., 2001). This was later termed the "expertise reversal effect" (Kalyuga, Ayres, Chandler & Sweller, 2003).

As learners become more competent, the worked-example and other cognitive load effects disappear (Kalyuga, Chandler & Sweller, 1998). Recall that Anderson's ACT framework proposes that, learners learn production rules later in the learning process and no longer need examples (Anderson, Finchman & Douglas, 1997), so a reduction in cognitive load is expected. Cognitive load theorists predict a gradual reduction of the effects of cognitive load, because once an expert automates their skills, the load imposed by a problem dissipates, and worked examples become unnecessary (Kalyuga, Chandler, & Sweller, 1998).

Recall that Chandler and Sweller (1991) suggest that extraneous cognitive load is due to the format of the instruction. In other words, presentation strategies can cause learners to perform poorly. Sweller and his associates have found a series of cognitive load effects due to the presentation techniques employed. The next few sections introduce these learning effects (the problem completion effect, the variability effect, the split attention effect, and the modality effect). What is important to realize while reading this discussion, is that all of these learning effects may be applied to worked examples, but like any instruction, if mismanaged worked examples can even overload a learner (Ward & Sweller, 1990).

The Problem Completion Effect

Van Merriënboer (1990) followed up on Sweller's early worked example studies, to describe another useful instructional strategy. Usually cognitive load is reduced by presenting worked examples as instruction, along with problems for the learner to practice (Sweller, 2003). However, van Merriënboer (1990) considered an intermediate, "problem completion" strategy. He compared two groups of high schools students, learners who generated their own computer code (problem solving), versus those who completed computer programs – problem completion (the use of partially worked examples). Learners using the problem completion strategy studied the worked-out portion of the problem, to abstract that component of the problem schema, and then later complete the problem. Van Merriënboer concluded that learners using the problem completion strategy, had a superior performance, because they had a higher percentage of correctly coded program lines, and also the quality of their programs was higher (van Merriënboer, 1990).

Van Merriënboer and de Croock (1992) later replicated Van Merriënboer's earlier findings, but this time with undergraduates. This study involved learners enrolled in an introductory software programming course. They gave one group of learners a library of partially completed computer programs, and then measured their performance versus another group who had to write their programs. Van Merriënboer and de Croock found that "problem completers" were more successful on both program construction tests, and multiple choice tests. This finding was later replicated by Paas (1992), but he also found that problem completers are better able to transfer their learning to new situations. This was subsequently described as the “problem-completion effect” (Paas, 1992).

The Variability Effect

Paas and van Merriënboer later turned their attention to worked examples, they were among the first to describe another cognitive load learning effect, “the variability effect” (Paas & van Merriënboer, 1994). Paas and van Merriënboer (1994) wanted to see how much of an effect the context of the example would have on learning. Clark, Nguyen, and Sweller (2006a) later described these problems as “varied context examples” (Clark et al., 2006a, p. 222). Cooper and Sweller (1987) had already shown that learners who studied worked examples could subsequently solve similar and transfer problems, more easily than those who learned through problem solving.

Paas and van Merriënboer (1994) found significant results when they compared groups of learners who either studied high or low variability worked examples, versus those who solved high or low variability problems. Specifically they found those learners who studied high variability worked examples invested less time and mental effort during practice, and also had better transfer. This was later termed the “variability effect” (Sweller, van Merriënboer & Paas, 1998). Specifically they describe variability as:

> …different variants of the task over problem situations, or under conditions that increase variability along other task dimensions, such as the manner in which the task is presented, the saliency of defining characteristics, the context in which the task is performed, the familiarity of the task, and so forth (Sweller, van Merrienboer & Paas, 1998, p.287).

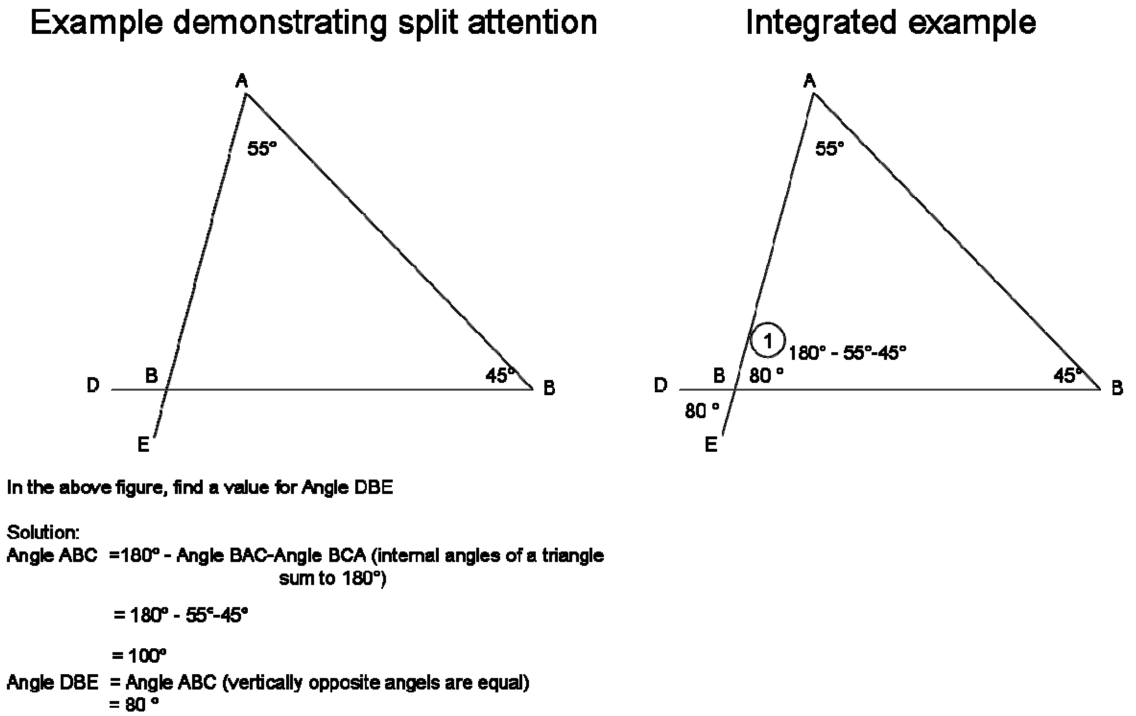
So problem variants, in which the context of the worked example is changed, from example to example, or to different problems solved, are important. This cognitive load learning effect is useful, for it shows learners are able to focus on the underlying structure, to abstract a problem schema, even though the surface features have changed (Van Merriënboer, Schuurman, de Croock, & Paas, 2002).

However, Paas and van Merriënboer (1994) reported that the high variability conditions had higher levels of perceived mental effort, but these learners achieve greater transfer performance. This initially caused some controversy, and was termed the "transfer paradox," for it seemed to contradict cognitive load theory, because typically an increase in cognitive load causes poorer performance (Sweller et al., 1998). Yet, increasing cognitive load is not necessarily a bad idea, because these learners focused their free remaining working memory capacity, toward schema related material, promoting germane cognitive load (Sweller et al., 1998). So, even though the overall load was higher given problem variants, their increased cognitive load was directed toward schema acquisition.

Thus, Sweller et al., (1998) began to refocus the cognitive load literature, from only concentrating on decreasing extraneous cognitive load, to now redirecting the learners' attention from irrelevant material (extraneous load), to germane or relevant materials, promoting germane load.

Split Attention

Tarmizi and Sweller (1988) noticed that the worked-example effect did not work for all worked examples. They compared the learner performance of those using traditional diagrams versus those who used integrated diagrams (like those in Figure 6).



Example demonstrating split attention

Integrated example

In the above figure, find a value for Angle DBE

Solution:
Angle ABC =180° - Angle BAC-Angle BCA (internal angles of a triangle sum to 180°)

= 180° - 55°-45°

= 100°

Angle DBE = Angle ABC (vertically opposite angels are equal)
= 80 °

Figure 6. Split-attention diagrams

Note: Adapted from "Some cognitive processes and their consequences for the organisation and presentation of information," by Sweller, J., 1993, *Australian journal of psychology*. 45(1) p 4-5.

They found learners who used integrated diagrams were better able to process information (Tarmizi & Sweller, 1988, Ward & Sweller, 1990). More specifically, if text (a visual form of instruction) is simultaneously presented to the learner with a diagram (also visual instruction), there is a potential for cognitive overload. This phenomenon was described as the "split-attention effect" (Sweller & Chandler, 1991; Chandler & Sweller, 1992).

Sweller suggests that while structuring materials, instructional designers must be careful how they direct the learner's attention within instruction (Ward & Sweller, 1990). Even worked examples may become ineffective, if they raise a learner's cognitive load to overload levels (Ward & Sweller, 1990).

The split-attention effect is not just limited to geometry. Sweller and Chandler (1991) found that this effect extends to a variety of other disciplines and is simply a limitation of human information processing. Since information may be encoded both as text (visually) and narration (auditorily), split-attention is a potential problem that exists within animated demonstrations, as they would within any type of worked example. Therefore instructional designers should probably remove text from animated demonstrations to limit this split attention effect. The split-attention effect is an important example of how presentation techniques can alter learning. However, this learning effect is limited to single modality instruction.

In the 1990s, several researchers began to study multimodal instruction. This was probably inevitable given the ubiquity of the personal computer and the CDROM. Many of these studies began to consider cognitive load and began to progress from comparing print based visual only conditions, to consider combinations of audio, text and animation. The next section describes this literature as it relates to cognitive load theory.

Cognitive Load and Multimodal Instruction

Recall that Baddeley and Hitch subdivided working memory into two separate visual and auditory subsystems (Baddeley & Hitch, 1974; Baddeley, 1986). This basic plan was also proposed a few years earlier by Paivio (1971) as the dual coding hypothesis, and then later as dual coding theory (Paivio, 1978).

Clark and Paivio (1991) later proposed that the dual nature of working memory is important for those designing instruction. Soon after this proposal, several researchers found empirical evidence that justified this idea. In a series of articles, these researchers found that learners working with multimedia consistently out-performed those learning with single (or mono) media materials (Jeung, Chandler, & Sweller, 1997; Mayer & Anderson, 1991; Mayer & Moreno, 1998; Mousavi, Low, & Sweller, 1995).

Sweller and his associates described it this way “A mixed, audio-visual mode of instruction resulted in superior learning than instructional materials delivered in a purely visual mode” (Jeung, Chandler, Sweller, 1997, p.331). This has since been described as “the modality effect” (Moreno & Mayer, 1999; Penney, 1989) or the Modality principle (Mayer, 2001) and has been one of the most important findings in Instructional design research.

Mousavi, Low and Sweller (1995) were perhaps the first to provide an explanation for the modality effect. This is because they considered dual-modality presentations from a cognitive load perspective. They found that under high load conditions, if an instructional designer moves the instructional message from a visual mode (text) to an auditory mode (narration), learner performance increases. They reasoned that when a lesson is structured so that it uses both modalities, learners are able to use both working memory subsystems simultaneously, to reduce their overall cognitive load by distributing that load to these independent subsystems. Specifically, they propose learner performance is improved because multimodal instruction increases the learners’ “effective working memory capacity” (Mousavi et al., 1995, p319).

Animation

A number of articles have been published that describe the instructional effectiveness of animation as a presentation technique. This section introduces this discussion and then turns to the animated demonstration literature.

Instructional Uses of Animation

Several reviews of the literature have been made concerning instructional animation. Amongst the first was Rieber's (1990) review. At that point, few if any, empirical investigations had considered animation within instructional materials. Later, he updated this review to include the literature of the 1990s (Rieber, 2000).

Rieber's first review concluded that before 1970, it was often thought that graphics did not aid learning, and could even distract learners. However, following this date, evidence began to mount in favor of the use of visuals to support learners. The studies he reviewed had mixed results, but he was able to draw some conclusions and provide several useful guidelines concerning the use of animation in instructional materials (Rieber, 1990).

Rieber proposed two important guidelines concerning the use of animation. First, like all graphics, if an instructional designer intends to include animated instruction, there must be "a need for 'external visualization'" (Rieber, 2000, p. 162). This guideline came from static graphics research, but animation and static graphics do have their differences. Specifically, Rieber (2000) suggested the use of animation when the learning requires changes in object motion or trajectory, or both. He proposed learning will be greater if both motion and trajectory changes are a part of the instructional materials – be it for procedure, concept, or principle-based learning.

Hegarty, Kriz, and Cate (2003) advise readers that narrated animations have the ability to convey more information which is not easily conveyed with static graphics. Hegarty et al (2003) mention that critics may argue that this additional information is a confounding factor, when comparing animation to static graphics, or that this contribution to learning is not a result of the animation. This is probably the case as much of the instructional message is in the verbal channel, and the verbal message directs the learner's attention to specific areas within the scene during the animation (Hegarty et al., 2003).

Animated Instruction that Matches the Task

Rieber (1990) proposed yet another important guideline, that animation should only be included, when the attributes of the animations match the task. Rieber and Parmley (1992, 1995) developed animated instructional materials to teach learners the basic principles of Newtonian physics (specifically the laws of motion). They used interactive animation (or simulation) to teach learners how to control a simulated space shuttle, given structured and unstructured lessons. Their structured lessons allowed learners increasing levels of control; that is, new "subskills" were taught during successive levels of instruction. They compared this "structured" tutorial to unstructured activity (discovery-based learning) with full control from the beginning, and found no significant differences in learner performance given their instructional conditions.

Even though Rieber and Parmley's tutorial was in part procedure-based, it also had a conceptual or principle-based component. Sometimes, in complex domains like physics, it is difficult to separate the two. In this case, however, Rieber and Parmley (1992) measured outcome measures that were primarily principle-based, while their instruction was primarily procedure-based. It is understandable that Rieber and Parmley wanted to teach learners physics principles given simulation, but perhaps they should have chosen another medium or another type of animation. Like many researchers, they compared groups of learners, given their performance on a multiple choice test (in a pretest/posttest design). This use of animation was innovative, but procedure-based learning requires observational data collection.

Animated demonstration is a form of animated instruction that has a main purpose, to teach learners how to perform procedures. This form of instruction may be used to teach learners other types of learning (e.g. concepts), but this could be a misuse of the medium. This study will endeavor to use this form of instruction properly and to use observational data to measure learner performance.

This discussion implies an important distinction for this dissertation project. Certainly animation may primarily be used to teach learners procedures (animated demonstrations). However, animation can also be used to teach learners conceptual material (animated explanation).

Animated Explanation and Animated Demonstration

It is thought that animated demonstrations act as animated worked-examples (Lewis, 2005). Sweller and his colleagues even define worked examples as a form of demonstration when they describe them this way: “A worked example is a step-by-step demonstration of how to perform a task or how to solve a problem” (Clark, Nguyen, Sweller, 2006a, p. 190). While this is the case, little to no cognitive load research has been conducted using animated demonstration.

Some would quickly dismiss this statement to begin describing the work of Richard Mayer, for Mayer and his colleagues have been quite prolific over the last decade. Mayer and his colleagues have indeed used animation extensively in their instructional conditions and are well known for their contributions to the modality effect.

However, the instructional materials used in this literature are more aptly described as animated explanation rather than animated demonstration.

Mayer describes his work well when he explains that his experiments ask questions about scientific explanation: “By ‘explanation’ we mean a description of a causal system containing parts that interact in a coherent way, such as a description of how a pump works or how the human respiratory system works” (Mayer & Sims, 1994, p.389). Clark and Mayer (2003) even describe these as “two different e-learning goals” that teach learners to “inform and perform “(p.17). The argument then, in their terms becomes: learning “how to do” something (perform), is taught via animated demonstration, as opposed to teaching a learner about something (inform), which is taught via animated explanation.

Given Anderson’s well respected ACT framework, it is likely to expect that these two forms of learning are dramatically different. Thus this study argues that animated explanations and demonstrations are different because of the type of learning that occurs, declarative versus procedural learning (Squire, 1992).

Interestingly enough, Mayer’s studies with animated explanations only found significant differences between instructional conditions given transfer. Other researchers (e.g. Cooper & Sweller, 1987) have found significant differences in other important outcome variables, such as completion times and the number of errors.

Animated demonstration represents a more cognitively demanding form of learning from animation. This is because by its very nature, demonstration assumes the eventual action of the learner. Rather than just encoding information which describes a system, the learner is encoding rules based upon a sequence of actions which they will have to perform. Performance is when the true cognitive load of the situation is highest, for it is then that the learner will have to recall the actions taught during the animation (if, then rules), but it is then that they will also apply those rules, in sequence, to produce the goal of the instruction.

Given this heightened cognitive load, learners need well-constructed instructional materials that reduce the extraneous cognitive load imposed by the learning environment. This distinction leads this discussion to the animated demonstration literature, which is of primary importance to this literature review.

Animated Demonstration

The animated demonstration literature extends back to the early 1990s (e.g. Palmiter & Elkerton, 1991a). So this literature was written before much of the modality or cognitive load literature existed. Thus this section will first review some of this literature, but also concentrate on placing this form of instruction in context with more recent instructional design literature. Specifically, it considers animated demonstrations given the modality and split-attention effects.

The animated demonstration literature is at times complex and contradictory (please note Appendix A). This is because an animated demonstration may be produced with or without audio, and it may or may not include text annotations. There are studies comparing animated demonstrations using each of these types of media and combinations of these media.

Animated Demonstrations in the Early 1990s

During the early 1990s, the “World Wide Web” was in its infancy, and most animated instruction was presented with personal computers, via CDROM. In addition, many investigators used a Macintosh hypertext environment called HyperCard. It was in this context that Waterson and O'Malley (1993) conducted a study comparing various forms of animated instruction.

This is a good example of an early 1990s animated demonstration study, because at that time, researchers were very interested in comparing "media effects." Recall this was the height of the famous Clark-Kozma debates (Clark, 1983; Clark, 1994; Kozma, 1991; Kozma, 1994; Jonassen, Campbell, & Davidson, 1994). While this was the case in the early 1990s, in the past decade, researchers have begun to consider learning from a different perspective, "learner-centered rather than media-centered" (Jonassen, Campbell, & Davidson, 1994, p. 31). Cognitive load theory considers learning from a learner-centered perspective, as it focuses on the learners' concurrent memory load. Today, Waterson and O'Malley's results may be reinterpreted, given this learner-centered perspective, to consider the split-attention and modality effects.

Waterson and O'Malley (1993) evaluated the effectiveness of several forms of animated demonstrations (given a set of six discrete tasks). Their instructional conditions included animation with text, animation only (no text or narration), and a combination group (animated demonstration with text and narration). Participants were taught a Macintosh graphing application called Cricket Graph (via HyperCard). They measured performance time given three instructional conditions, with three types of tasks (identical, similar, or different tasks from those that learners had initially learned).

The data revealed a significant main effect with respect to group. The combination text-narration group outperformed the text-only, and no-narration groups (See Figure 7, a graph of the performance times of their participants).

Waterson and O'Malley's data also revealed a significant interaction between type of instruction, and task group, as the combination group completed tasks sooner than the other groups. They also found a trend in the data which suggested learners using the text-only instruction, were slower than either of the other groups (animation only and animation with text and narration).

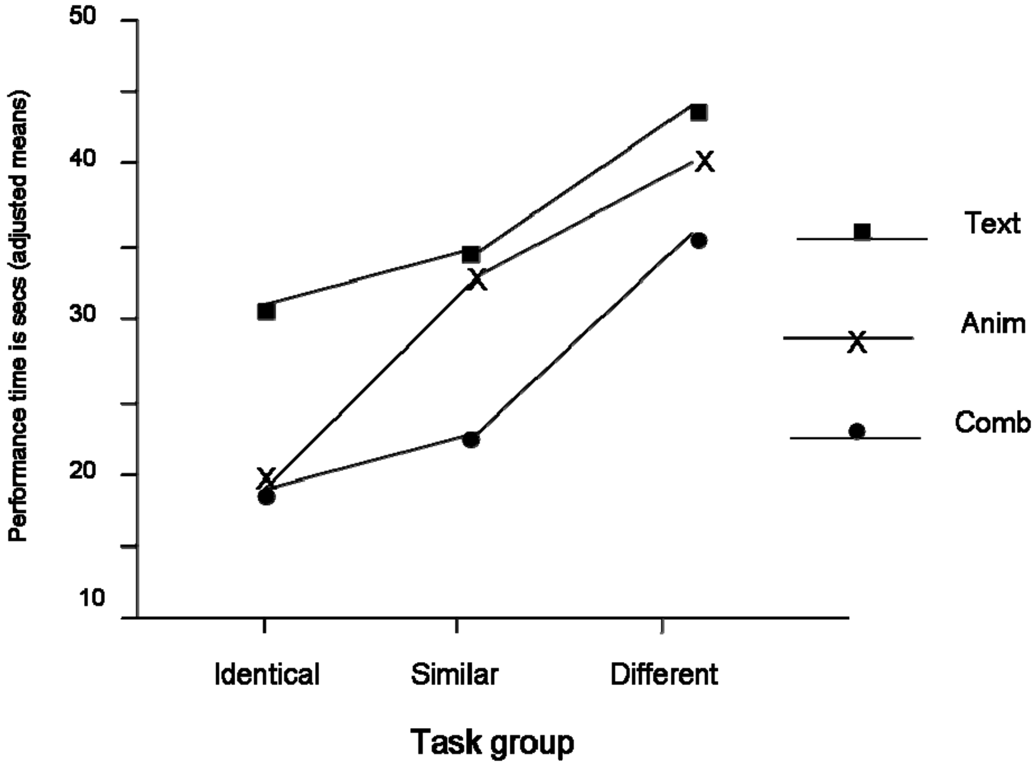


Figure 7. Performance time by task group

Note. Adapted from "Using animated demonstrations in multimedia applications: Some suggestions based upon experimental evidence," by P. Waterson & C. E. O'Malley, 1993, In *the Proceedings of the Fifth International Conference on Human-Computer Interaction, 2* p. 546

The data revealed a significant main effect with respect to group, since the combination text-narration group outperformed the text-only, and no-narration groups (See Figure 7).

Waterson and O’Malley’s data also revealed a significant interaction between type of instruction, and task group, as the combination group completed tasks sooner than the other groups. They also found a trend in the data which suggested learners using the text-only instruction, were slower than either of the other groups (animation only and animation with text and narration).

The instructional conditions in the Waterson and O’Malley studied show some interesting cognitive load effects. The fact that the slowest performance times were from the text-based animated demonstration group, is evidence that the split-attention effect can negatively affect learner performance, given animated demonstration. It is also interesting that the combination group had decreased performance times. This is evidence for the modality effect. The fact that the combination group had a text redundancy for the narrated message may be little reason for concern, since this group out-performed the other groups. However, it is possible that learner performance may still be increased by removing this redundancy. Interestingly enough, these learners may have ignored this redundancy to have benefited from the modality effect.

Even though Waterson and O’Malley conducted a series of repeated measures ANOVAs, demonstrating that animated demonstrations show some promise, the study unfortunately only had 30 participants (10 per condition). This perhaps is sufficient for a pilot study to test the instruments, but the results of this study are somewhat suspect due to a lack of power. Although the Waterson and O’Malley study is interesting from a cognitive load perspective, another study should be considered.

Palmiter's Animation Deficit

Palmiter's dissertation project is perhaps the most cited animated demonstration study (Palmiter, 1991). In a series of experiments, Palmiter compares the learner performance of those who study animated demonstrations with those who use text-based instruction. She used a repeated measures design to study learners as they performed a set of discrete HyperCard tasks. Her study measured four dependent variables – performance time, accuracy, retention and transfer during a training session, an initial test, and a delayed test (Palmiter, 1991).

Palmiter (1991) found several significant session x media interactions. These results are quite interesting, for she found that during the training session learners who studied animated demonstrations sped skill acquisition, and performed tasks in less time and more accurately than their peers using text-based instruction (Palmiter, Elkerton, & Baggett, 1991). These results are similar to those by Sweller and Cooper (1985) who demonstrated that studying worked examples requires "considerably less time to process than conventional problems, but that subsequent problems similar to the initial ones also were solved more rapidly"(p.59).

However, Palmiter (1991) noted that, one week later, learners using animated demonstrations took longer to perform tasks, as compared with learners using text-based instruction. She reported that their skill acquisition may have been quicker during the initial training session, but their retention was lacking one week later. This phenomenon, later described as an "animation deficit" (Lipps, et al., 1998) could not be replicated by either Waterson and O'Malley (1993) or Lipps et al. (1998).

Palmiter et al. (1991) also measures transfer. She measured performance time while performing tasks similar to those trained (e.g. Copy Button vs. Copy Field). She found that demonstration groups took significantly less time than the text only groups during the “immediate test session.” This again is very similar to the results by Cooper and Sweller (1987) who wrote: “The results indicated that subjects whose training included a heavy emphasis on worked examples or an extended acquisition period were better able to solve both similar and transfer problems than were those subjects trained with conventional problems” (Cooper & Sweller, 1987, p 347). However, again, Palmiter (1991) notes a significant increase in performance time for the demonstration groups, between the training and delay sessions.

Palmiter (1993) developed a “Model for Procedural Acquisition for Animated Demonstration: the Mimicry Model” (Palmiter, 1993, p.77). She used the following to describe this model “During initial training, demonstration users seemed to make a recording of what they saw and they play back a 'tape' of the recorded procedure during initial training.” (Palmiter, 1993, p.77). This is in direct contrast with how Sweller and his associates describe how learners abstract information from instructional materials. “Information is not remembered in the way a tape recorder might be considered to 'remember' material, in a form identical to its presentation form. Because we must restructure or construct a representation of material presented to us...” (Sweller & Chandler, 1991, p.357). So from a cognitive perspective, animated demonstrations are not expected to be played back in memory. Since the 1970s, it has generally been accepted in the psychological literature, that memories must be reconstructed and are not simply played back (e.g. Loftus, 1975; Loftus & Hoffman, 1989).

Since Palmiter's early study, many researchers have evaluated animated demonstrations as a presentation form (Cornett, 1993; Harrison, 1995; Reimann & Neubert, 2000). However, these researchers do not describe animated demonstrations as animated worked examples, nor do they describe a phenomenon similar to Palmiter's animation deficit. Quite the opposite, they found learners using narrated demonstrations were faster and more accurate than those using text-based instruction (Cornett, 1993; Harrison, 1995).

Worked Examples and the Design of Animated Demonstrations

From a novice perspective, many cognitive tasks are complicated. Experts may have a sense of a problem solution, and design animated demonstrations as a series of discrete tasks as in Palmiter's study (Palmiter, 1991), but this only teaches the step in the process, not when to use those steps. For a learner to solve authentic problems they must know when and how to use the steps toward solution.

Teaching learners how to solve complex problems via discrete steps is useful, if it is in the context of an authentic problem. This can be accomplished via a narrated demonstration of a problem solution. In this way the instructor is able to communicate the problem steps (and when to use them), limiting the cognitive load of the learner because the learner simply has to watch the demonstration. Later, once a learner understands a problem schema, they can be allowed to perform or practice the problem steps. This is the sequence of instruction when teaching learners with worked examples, and is an instructional sequence which elicits the worked-example effect (Cooper & Sweller, 1987; Sweller & Cooper, 1985), but the literature has yet to show this effect for animated demonstrations.

Atkinson et al. (2000) proposed worked examples “typically include a problem statement and a procedure for solving the problem” (Atkinson et al., 2000, p. 181). Lewis (2005) proposed similarly designed animated demonstrations should act as worked examples “Animated demonstrations clearly fall under this category as they describe a problem and its solution in a series of steps. Because animated demonstrations are goal directed and procedure-based they act as animated worked examples” (Lewis, 2005, p.371).

Several important guidelines should be considered at this point. First animated demonstrations should (1) include a problem statement and the procedure for solving the problem; (2) provide a learner with a verbal commentary directing the learner’s attention during the animation; and (3) describe the procedural steps in the context of a realistic problem. So given these guidelines are followed one may expect to find the worked-example effect given animated demonstrations.

Conclusions about Animated Instruction

Rieber (1990) suggests the use of animation in instruction is relatively new, and has only been made available given computer-based instruction. Rieber (2000) suggested it has often been used in a gratuitous manner and is only useful under a certain range of conditions. In his terms, animated instruction must “pass the test for a need for ‘external visualization’” (Rieber, 2000, p. 162). In addition, Rieber (1990) also suggests animation should be included only when the attributes of the animations match the task.

Early animated demonstration researchers were interested in “media effects”. For instance, Waterson and O’Malley (1993) studied how learners learned given text, animation and narration, and even provided evidence of the modality effect (Mayer, 2001; Penney, 1989). Since these early animated demonstration studies were first published, researchers have begun to follow Mayer’s advice to be concerned with the cognitive processing within the learner, rather than the effect of the media (Mayer, 1997).

The most notable finding from the animated demonstrations literature is the potential for an “animation deficit” (Lipps et al., 1998; Palmiter, 1991). Lipps et al (1998) describes Palmiter’s animation deficit as a short term performance gain by learners using animated demonstrations (during early skill acquisition), but a significant loss in long term performance.

Palmiter found that learners using animated demonstrations would acquire skills in significantly less time during the acquisition phase. But one week later, these same learners had difficulty retaining those skills, and took significantly longer to reproduce the same performance (Palmiter, 1993). However, there is reason to question this “animation deficit,” as other researchers could not replicate these findings (Lipps et al., 1998; Waterson & O’Malley, 1993). In addition, Palmiter (1993) described a mimicry model which is in direct contrast to the writings of Sweller and Chandler (1991). This mimicry model will also be considered and questioned as part of the instructional conditions.

Finally, none of these studies compared learner performance with discovery learning. Only Rieber and Parmley (1992, 1995) compared the learner performance of those using an animated condition to those using a discovery learning condition, but their study uses simulations (dynamic animation) rather than animated demonstrations. This instructional strategy, “discovery learning,” became very popular in the 1960s, and sparked the discovery learning movement. The next section considers this instructional strategy in two ways: as an alternative to studying animated worked examples, but also from a learning perspective.

Discovery Learning

Jerome Bruner initiated the discovery learning movement with his 1961 article "The Act of Discovery." Bruner wrote extensively about discovery, perhaps as an alternative to what was then, the latest advancement in educational technology, direct instruction (Skinner, 1958). Bruner's article proposed that learners should be allowed to discover new rules and underlying principles, rather than be required to memorize material. He contrasted discovery-based instruction his "hypothetical mode" (Bruner, 1961, p.23), with teacher-led instruction which he described as the "expository mode" (Bruner, 1961, p.23). Bruner's "expository mode" was not just a general category for describing Skinner's teaching machines or behavior modification, for it is a much broader classification of instructional strategies. Bruner described the expository mode when he says "the decisions concerning the mode and pace and style of exposition are principally determined by the teacher" (Bruner, 1961, p.23). In contrast Bruner (1961) clearly describes discovery learning as:

> It is, if you will, a necessary condition for learning the variety of techniques of problem solving, of transforming information for better use, indeed for learning how to go about the very task of learning. Practice in discovering for oneself teaches one to acquire information in a way that makes that information more readily viable in problem solving (Bruner, 1961, p.26).

Bruner's article may have described discovery learning, and set off the discovery learning movement, but this was not a new philosophy. This movement had much earlier origins, and originated in the works of early 20th century educators like John Dewey, William Heard Kilpatrick, and Maria Montessori. Although the ideas of these earlier educators are reiterated here by Bruner, unfortunately like many proponents of this philosophy, he did not define "discovery learning."

What is Discovery Learning?

Klahar and Nigam (2004) report that, for almost 100 years authors of this literature have had a consistent problem defining *discovery learning*. In the 1960s, several well known authors convened in an important volume edited by Shulman and Keiserler, to discuss discovery learning from a critical perspective. In this volume several well known authors provided conflicting definitions, like the following: "…learning by discovery is defined usually as teaching an association, a concept, or rule which involves 'discovery' of the association, concept, or rule" (Glaser, 1966, p.14). This somewhat circular definition gives some indication that "learning by discovery" or "discovery learning" is an instructional strategy. While this may be the case, many others suggest discovery learning is accomplished by autonomous learners (e.g. Gagné, 1966).

Glaser (1966) continues by contrasting discovery learning with traditional instruction to suggest that one of the most important characteristics of discovery learning is that it makes use of induction during the process of learning. However, Wittrock (1966) explains induction is not a prerequisite for discovery learning. He proposes that it is equally possible for a learner to (1) begin with a higher order generalization to discover specific conclusions, as it is (2) to discover generalizations or rules from specific examples; or, in his words, "Induction has no exclusive identity with discovery learning" (Wittrock, 1966, p43).

It is important to define discovery learning for this review to remain unambiguous. Thus, this review makes a somewhat complicated definition of discovery learning. Discovery learning is defined as an instructional technique in which the instructor provides an environment for learning, to be accomplished by autonomous or semi-autonomous learners. On some level then “discovery learning” is a bit of a misnomer, since this form of “learning” is really an instructional technique, and described as such by several very prominent researchers (Gagne, 1966; Glaser, 1966; Taba, 1963).

While it is difficult to define discovery learning, the discovery learning movement became the philosophy of many American educators, and has often been heralded as a means of teaching inquiry-based learning. According to Mayer (2004), since the 1960s the arguments for and against discovery learning have waxed and waned, and are still with us today, as constructivism. This review provides the arguments both for and against discovery or expository instruction, and as it turns out there is no simple answer, for neither is the perfect solution, under all conditions.

Proponents of Discovery Learning

One of the earliest proponents of discovery learning was well-known curriculum theorist Hilda Taba. Taba built on Bruner’s foundation to provide a rationale for discovery learning by stating:

> Several proponents of this method argue that a premature verbalization of the generalization or the rule deprives the individual of the essential learning, namely, the reorganization of his own cognitive structure, and puts the student in the position of absorbing the generalization without necessarily understanding what it stands for or how to work it (Taba, 1963, p.312).

Taba's criticism is valid. Understanding is a very important part of learning. This is the central argument of the proponents of discovery learning. Unfortunately understanding is not easily measured, thus this argument has led to the controversy that exists between those for and against discovery learning.

Even today, mathematics educators are battling in the so called "Math Wars" (Schoenfeld, 2004). Mathematics educators, particularly in the 1990s, struggled with standards-based curriculum reform, to find ways of teaching learners how to use mathematics effectively. Schoenfeld (2004) explains that in 1989 the National Council of Teachers of Mathematics (NCTM) published a series of standards with the intent of providing mathematics education reform. These reformers described their underlying philosophy by saying: "This constructive, active view of the learning process must be reflected in the way much of mathematics is taught" (NCTM, 1989, p. 10). These educators took a constructivist perspective and felt very strongly that it was not simply enough to teach learners how to perform a mathematical procedure, but that they must understand the procedures and principles underlying the problems they were solving.

Critics of Discovery Learning

Ausubel (1963) was perhaps one of the most outspoken critics of Bruner's instructional technique. He devoted a sizable portion of his early cognitive textbooks to this instructional strategy (Ausubel, 1963, 1968). Ausubel proposed there is a time and place for discovery learning, but that it is a highly inefficient means of conveying large amounts of information. He contends that learners must learn vast amounts of information in their lifetimes, more than they could ever discover on their own, and after a great deal of discussion concludes that discovery learning “is as unfeasible as it is unnecessary” (Ausubel, 1963, p.151).

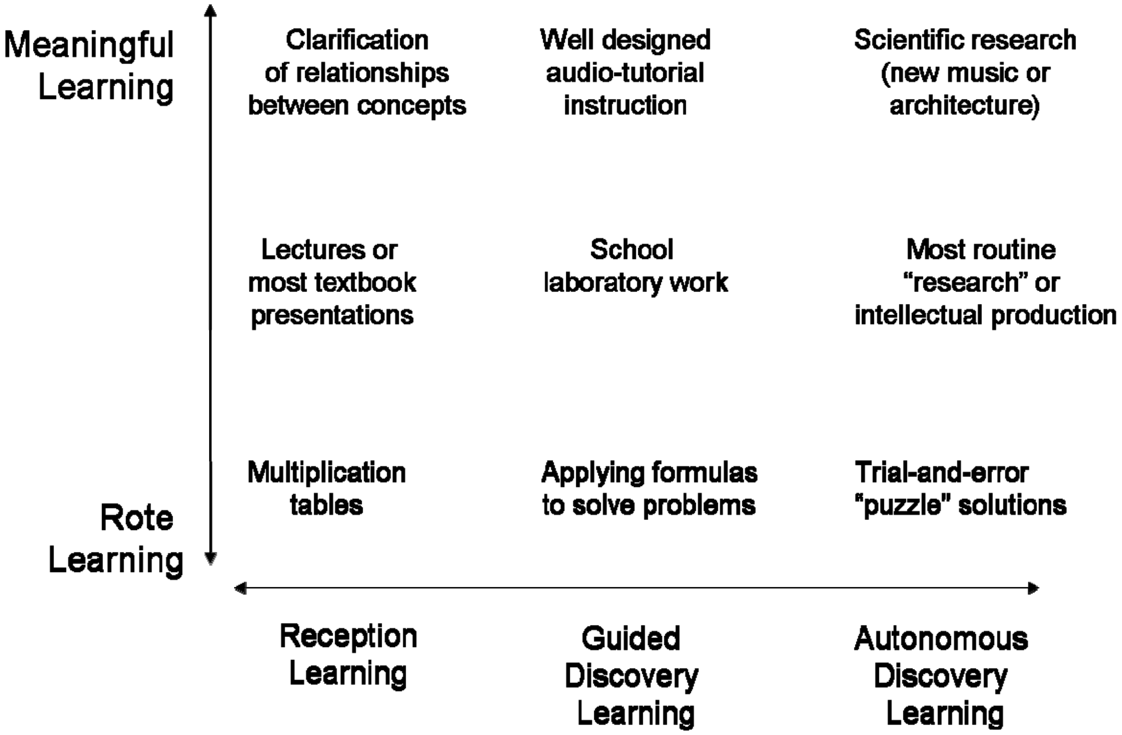
Ausubel was among the first to relate cognitive psychology to instruction, and is well known for developing advanced organizers (Ausubel, 1960). Nevertheless Driscoll (2000) proposes that his main contribution to learning was his development of the theory of meaningful verbal learning (Ausubel, 1963). According to Ausubel, “meaningful learning” takes place when the learner chooses to relate new information to prior knowledge, as opposed to rote learning, which is simply memorization (Ausubel, 1963; Novak & Godwin, 1984).

Ausubel’s development of the theory of meaningful learning was in part, a response to programmed instruction or behaviorism, which tends to promote rote performance over deeper levels of understanding (Ausubel & Fitzgerald, 1961). However, Ausubel advises us that both expository and discovery teaching techniques can promote rote learning, and expresses that there is widespread confusion given expository learning:

> This confusion is partly responsible for the wide spread but unwarranted twin beliefs that reception learning is invariably rote and that discovery learning is inherently and necessarily meaningful. Both assumptions, of course, are related to the longstanding doctrine that the only knowledge one really possesses and understands is knowledge one discovers by oneself. Actually each distinction constitutes an entirely independent dimension of learning (Ausubel, 1963, p 18).

Ausubel is quite clear on his position of discovery learning and suggests those supporting discovery learning “confuse the reception-discovery dimension of the learning process with the rote-meaningful dimension" (Ausubel, 1963, p 18). So Ausubel proposed both expository and discovery instruction promotes rote or meaningful learning.

Novak and Godwin (1978) provide a graphic illustration explaining Ausubel's ideas (See Figure 8). In this graphic Novak and Godwin (1984) elaborated on the work of Bruner and Ausubel, to present Ausubel's ideas as two continuums, the expository-discovery dimension, and the meaningful-rote dimension.



	Reception Learning	Guided Discovery Learning	Autonomous Discovery Learning
Meaningful Learning	Clarification of relationships between concepts	Well designed audio-tutorial instruction	Scientific research (new music or architecture)
	Lectures or most textbook presentations	School laboratory work	Most routine "research" or intellectual production
Rote Learning	Multiplication tables	Applying formulas to solve problems	Trial-and-error "puzzle" solutions

Figure 8. Typical forms of learning

Note: Adapted from "*Learning how to learn*," by Novak, J. D., & Gowin, D. B., 1984, New York, NY: Cambridge University Press. p. 8

This illustration shows that neither form of instruction is as Ausubel explained purely rote or purely meaningful. Neither discovery nor expository instruction is a panacea, each has its advantages and disadvantages. However, Ausubel was not the only well-known critic of discovery learning.

Guidance During Problem Solving

In his classic text, *The Conditions of Learning*, Gagné (1965) states "The discovery method is liable to gross misinterpretation in practical learning situations" (Gagné, 1965, p.165) and explains that proponents of this technique argue for using a minimal amount of instruction, and unfortunately fall into the trap of providing problems "without perquisite knowledge of principles and without guidance" (Gagné, 1965, p.165).

Gagné (1966) described discovery as being different given associative, concept, or principle learning. In addition, Gagné (1966) was one of the first to consider discovery during problem solving. He states that it involves "(1) a process of search, and (2) a process of selection, each of which takes place within the learner's nervous system" (Gagné, 1966, p. 136).

While developing cognitive load theory, Sweller also considered problem solving search, and suggests that if a learner is required to solve problems, while learning, they may spend many hours searching for a problem solution and still not be engaged in schema acquisition (Kirschner, Sweller, & Clark, 2006; Sweller, 1988). In other words, even though these learners may be actively searching for a problem solution, they may not be learning (Sweller, 1988).

Gagné (1965) describes eight forms of learning, of which problem solving is the most complex. Like many authors he says learners may discover principles or problem solutions, but recommends that instructors provide guidance during problem solving. However, at some point, we must teach learners to teach themselves. Problem solving of course, requires the learner to solve problems on their own. This dichotomy of self-guidance and instructor guidance is probably the most important reason for this controversy.

Cognitive load researchers are not against teaching learners to teach themselves, however they are concerned with how one introduces novices to problems and problem solving. Their somewhat counterintuitive solution is to introduce learners to problem solving, by first providing a demonstration or worked examples. Later, as learners develop their skills they suggest allowing learners to practice.

Given the expertise reversal effect (Kalyuga, Ayres, Chandler, & Sweller, 2003), direct instruction is only useful during the earliest stages of learning, during early schema acquisition. Therefore cognitive load theorists suggest fading worked examples, to allow more advanced learners time to practice and automate their skills during problem solving (Renkl, Atkinson, & Maier, 2000; Renkl, Atkinson, Maier, & Staley, 2002). Thus, it is not practice or discovery that cognitive load researchers are against. It is the timing of that practice which is under scrutiny.

Discovery Learning and Constructivism

Discovery learning has morphed and changed over the decades since Bruner's article. Mayer (2004) suggests it is still with us in the writings and practices of constructivism. During the 1990s, many American educators adopted a constructivist epistemology toward teaching and learning. This epistemology suggests the active construction of knowledge (Dewey, 1916; Duffy & Cunningham, 1996; Wittrock, 1974).

Constructivism is primarily a philosophical position, but has implications for instructional design. It suggests we perceive information from the environment, and that our mental models of the environment help us to construct our own unique version of reality (Jonassen, 1991). This relativist epistemology extends into a philosophy of instructional design.

Jonassen (1991) suggests instructional designers should focus less on “prescribing a single best sequence of learning (p.12)” and allow learners to negotiate their own learning. Thus, many constructivists design what they describe as “ill-structured learning” environments, because they feel the learner will construct his or her own interpretation of that environment, and must be allowed to do so.

Jonassen (2002) has updated constructivism to describe “learning as activity.” He and his colleagues are now attempting to integrate activity theory and Ecological Psychology into a constructivist philosophy of learning. In terms of instructional design recommendations, this “learning as activity” mantra becomes “learning by doing.”

Strong Criticism of “Learning by doing”

“Learning by doing” has been a popular approach toward the design of instruction, but many educational psychologists and instructional design researchers have begun to question the efficacy of this approach.

Mayer (2004) describes constructivist instructional design recommendations, as relying on “the constructivist teaching fallacy” (Mayer, 2004, p.15). Specifically, he explains that many constructivists prescribe active learning techniques, which require learners to be behaviorally active. Rather than being behaviorally active, he suggests learners be cognitively active. Mayer puts it best when he says “Activity may help promote meaningful learning, but instead of behavioral activity per se (e.g., hands-on activity, discussion, and free exploration), the kind of activity that really promotes meaningful learning is cognitive activity (e.g., selecting, organizing, and integrating knowledge)” (Mayer, 2004, p.17). While this article stops short of condemning pure discovery or constructivism, he concludes “The research in this brief review shows that the formula constructivism = hands-on activity is a formula for educational disaster” (Mayer, 2004, p. 17).

Kirschner, Sweller, and Clark (2006) followed Mayer, by being critical of constructivist teaching techniques. However they took it a step further, and in a bold move, argued that constructivist, discovery, problem-based, experiential, and inquiry based teaching have been a failure. They describe these instructional design prescriptions as “unguided” or “minimally guided” instruction; echoing Gagné’s earlier argument (that discovery learning does not provide guidance) (Gagné, 1965).

So there is strong criticism of constructivism and discovery learning from those who promote cognitive load theory. However, as Paas, Renkl, & Sweller (2004) warn, this epistemology has a strong following in American education, but “despite a long history, evidence for the effectiveness of discovery learning from controlled studies is very sparse” (Paas, Renkl, & Sweller, 2004, p. 6).

Kirschner, Sweller, and Clark’s criticisms have not gone unheard, and shortly after publishing this paper, several members of the constructivist community responded (e.g. Hmelo-Silver, Duncan, & Chinn, 2007). Hmelo-Silver et al.’s main argument is that problem-based and inquiry learning are not “minimally guided” because these forms of instruction are “scaffolded inquiry.”

The Nature of the Debate

It seems the real problem, given this debate, is more one of scale and a failure to communicate. When debating the nature of learning, researchers must be specific about which learners, and what one means by “learning.” As Gagné (1965) described there are many types of learning. In addition, recall that earlier in this chapter neurological researchers found that procedural and declarative learning are even processed by different portions of the brain (Squire & Zola, 1996). Finally, as this literature review has shown learner expertise is a continuum, extending from novices with little to no prior experience, to those with decades of experience.

Given the entire breadth and depth of “learning,” it is quite possible that both groups of researchers (the critics and proponents of discovery learning) maybe correct for different audiences. Discovery during problem solving may very well be important for more experienced learners, but detrimental for novices. The next section considers audience as a factor in this debate, to ponder scaffolding and guidance in ill-structured learning problems and learning environments.

The Appropriate Environment for the Audience

In 1993, Jonassen, Mayes, & McAleese used the term constructivist learning environments (CLEs). They describe these environments as being for more advanced learners, and they expected more structured approaches for novices (Jonassen, et al., 1993). However, they state:

> We believe that constructivistic learning environments may be used during the latter stages of knowledge acquisition and that they represent rich and meaningful environments for initial knowledge learners. However constructivistic environments are more reliably and consistently applied to support the advanced knowledge acquisition phase. (Jonassen, Mayes, & McAleese, 1993, p.232)

In addition to the above discussion, Jonassen et al. (1993) presented a continuum (See Figure 9) which suggests ill-structured domains should mainly be used during the later stages of the learning process.

Figure 8 even proposes that more structured learning domains, like skill-based or procedure-based learning, are appropriately handled by well-structured instruction. This is an important point because this article shows that some of the most outspoken advocates for constructivism believe constructivist learning environments have their limitations. However an important distinction should be made at this point. While it is important to consider constructivist learning activities in the later stages of learning, most instruction is developed for novices.

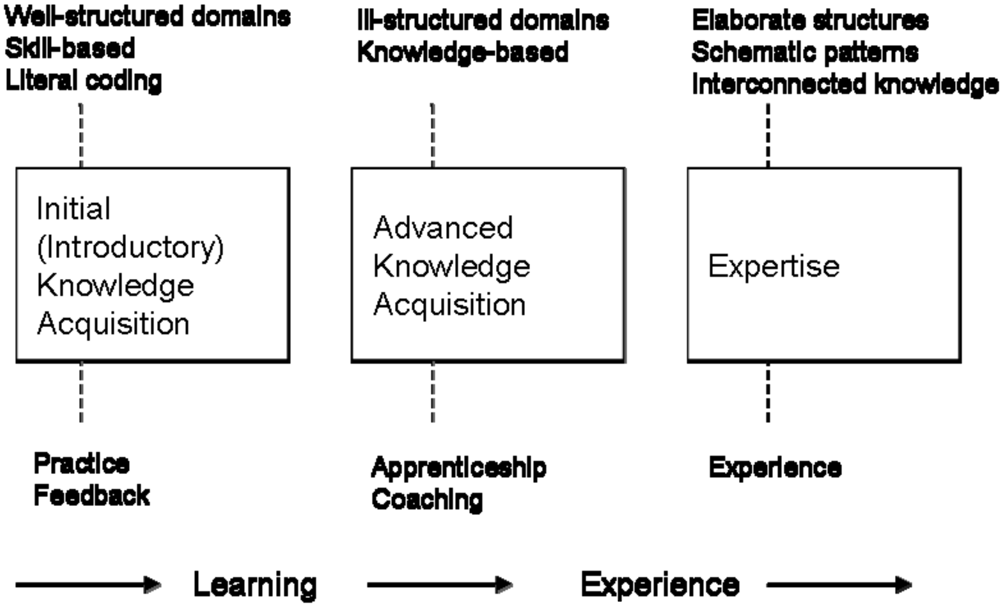


Figure 9. Structured and ill-structured domains

Note. Adapted from "A manifesto for a constructivist approach to uses of technology in higher education," by D. Jonassen, T. Mayes, & R. McAleese, R., 1993, In *T.M. Duffy, J. Lowyck, & D.H. Jonassen (Eds.), Designing environments for constructive learning*. p. 232

Several years later, Jonassen suggested that instructional designers should develop well-structured environments for novices, and specifically refers these designers to Sweller's work with worked examples (Jonassen, 1997).

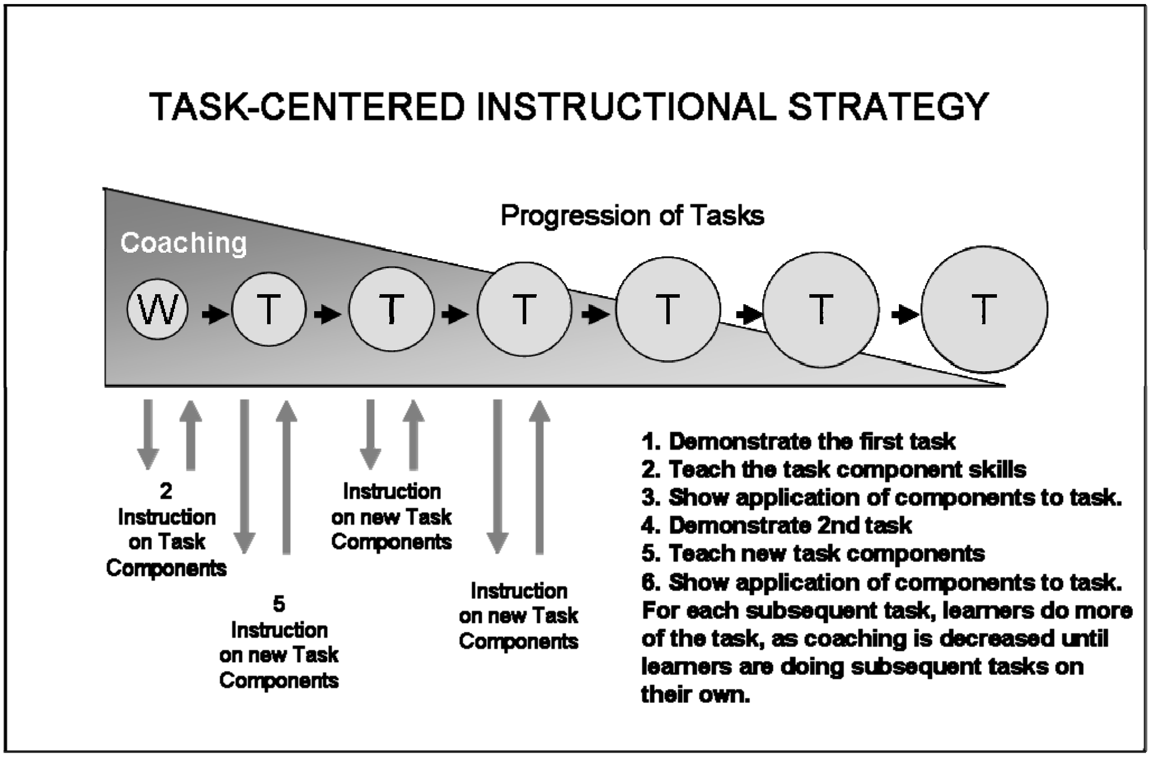
The primary purpose of instruction is to provide learners with a means of learning new material. Guided instruction means providing learners with the underlying schema which amounts to well-structured information (instruction), but providing learners with less information, usually always means providing them with less guidance.

Kirschner, Sweller, and Clark, state "Most learners of all ages know how to construct knowledge when given adequate information and there is no evidence that presenting them with partial information enhances their ability to construct a representation more than giving them full information" (Kirschner, Sweller, & Clark, 2006).

While educators cannot teach a learner everything they need to know, providing novice learners with less, in an effort to allow them to discover it on their own, is probably irresponsible. It's very important that instructors live up to their responsibility, to provide learners with the guidance and instruction that they need during early schema acquisition.

Merrill's Task Centered Strategy

Veteran researcher David Merrill synthesized the literature concerning expository and discovery-based instruction, to produce what he describes as a "task centered strategy" (Merrill, 2007) (See Figure 10). This instructional design model considers the needs of novices and those with more expertise. It begins with well-structured problems and ends with ill-structured learning environments to suit the needs of all learners.



1. Demonstrate the first task
2. Teach the task component skills
3. Show application of components to task.
4. Demonstrate 2nd task
5. Teach new task components
6. Show application of components to task.

For each subsequent task, learners do more of the task, as coaching is decreased until learners are doing subsequent tasks on their own.

Figure 10. Merrill's Task-Centered Instructional Strategy

Note: Adapted from “Levels of instructional strategy,” by Merrill. D.M., 2006, *Educational Technology 46*(4) p.8

Merrill’s “Task-centered Instructional Strategy” is an instructional design model, which is the culmination of years of work, and a practical synthesis of many different learning theories (Merrill, 2007). It begins with a demonstration or worked example and through a series of problems, adds complexity to simplified tasks by adding components of authentic problems. This design strategy guides learners to eventually lead them to solve complex tasks, on their own, without coaching. It suggests demonstrations and worked examples early in the process and later as the learner gains expertise, coaching and guidance is faded, to allow the learner to discover how to use previously learned concepts and principles, to solve authentic problems.

As Merrill and many others have explained, novices need to be guided during the earliest stages of learning. This is the underlying idea of this dissertation, that instructors and instructional materials must guide learners, during early schema acquisition. Later, after learners acquire the underlying schemas and become more experienced, then they may be allowed to discover and practice authentic problems, as suggested by Anderson's ACT-R framework (Anderson, 1993). It is then that discovery learning techniques may be useful and only then, that they should be taught with ill-structured learning environments (Jonassen, Mayes, & McAleese, 1993).

Even though the view of many educators is to support active learning or discovery-based problem solving, this review has shown that this view is not justified by the literature, and not in the best interest of novices. However, good empirical studies have never lost favor with educational researchers, so before closing this discussion of discovery learning, it would be prudent to review the Tuovinen and Sweller (1999) article. This article is important, because it actually compares learners who studied worked examples with those using discovery-practice, during early schema acquisition.

Worked Examples versus Discovery Learning

Sweller (1988) proposed that discovering a problem solution constitutes a dual-task, requiring the learner to search for a problem solution, while trying to learn the underlying problem schema. Studying worked examples is a way to eliminate the second task (problem solving search) because studying worked examples only demonstrates the problem schema (Sweller et al., 1998).

Tuovinen and Sweller (1999) compared the performance of those learners who were given worked examples and those who discovered problem solutions by solving problems on their own. In addition, Tuovinen and Sweller also compared the performance of those with and without experience. This experiment was carried out over three consecutive weeks.

During the first week, learners were asked to fill out and initial survey and then were introduced to FileMaker Pro (a Macintosh database program) via a series of HyperCard stacks (a series of printed electronic slides). During week two, learners were randomly assigned into groups: an exploration group and a worked examples group. The exploration group was given the following text-based instructions:

> Try out the functions in each of the lessons in situations you create-yourself, saving your files on the floppy disk provided. You may use any of the databases on the floppy disk if you wish. You will be asked to solve problems similar to the one shown in the lessons, in the test on this work. So direct your exploration towards gaining adequate mastery of the program to deal with such questions. (Tuovinen & Sweller, 1999, p.337)

The worked example group was asked to read through a worked example that consisted of "a problem statement related to calculation or field construction or use and then an annotated step-by-step example of the way the problem could be solved with computer-screen views seen by the operator working to obtain the solution" (Tuovinen & Sweller, 1999, p.337). These learners were subsequently asked to practice what they had learned on a similar problem.

During the third week, all learners were tested with a paper-based test composed of items similar to those taught in their lessons. Each learner was provided with a series of questions and required to create database files based on those questions. Test scores were analyzed with a 2x2 ANOVA (higher or lower levels of experience) X (worked-example or exploration group). As expected, the results showed a significant main effect with respect to levels of experience, with those with prior experience performing significantly better than those with less experience. Although the main effect for groups was not significant, there was a significant interaction between these variables (See Figure 11). When they compared the mean test scores for those without experience, they found participants in the worked examples group performed significantly better than those in the exploration group (discovery practice). Thus, they confirmed the worked-example effect. As expected, they found means scores for the groups with prior experience were not significantly different, but that worked examples were not as beneficial for learners in this group (further evidence of the expertise reversal effect).

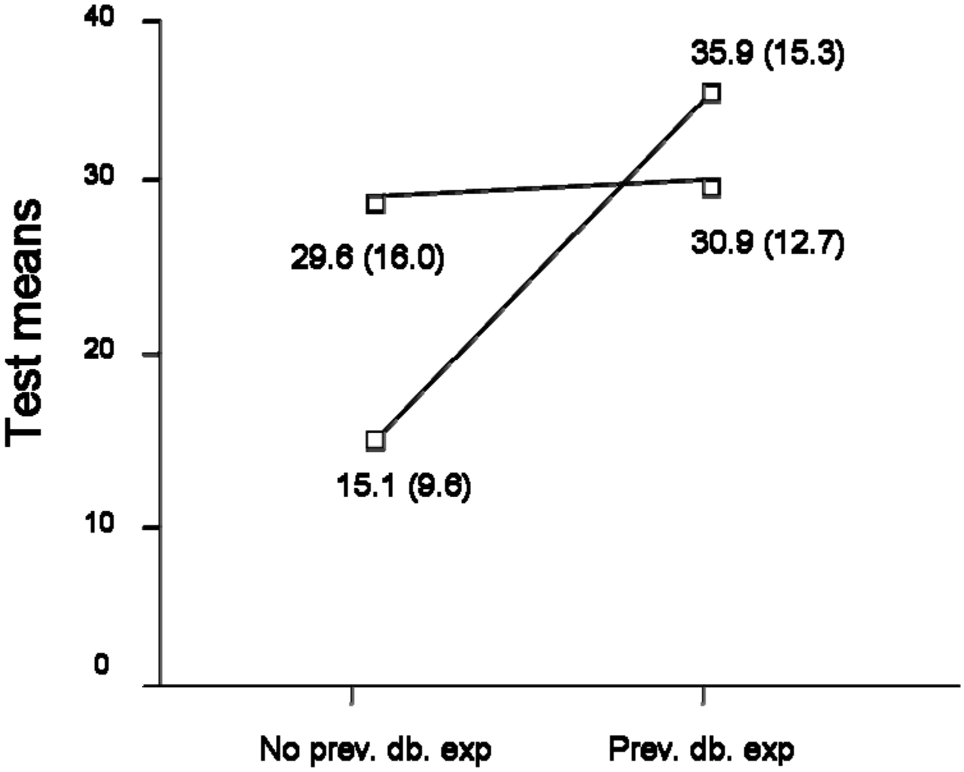


Figure 11. Mean test performance

Note. Adapted from “A Comparison of Cognitive Load Associated With Discovery Learning and Worked Examples,” by J.E. Tuovinen & J. Sweller, 1999, *Journal of Educational Psychology*. 91(2) p.338

While these results are positive for cognitive load theorists, this was just one of many studies, comparing learners who studied worked examples versus those who solved problems. Many, studies have confirmed these results, showing overwhelming evidence in favor of worked examples, as opposed to problem solving (Carroll, 1994; Paas, 1992; Paas & van Merriënboer, 1994; Quilici & Mayer, 1996; Sweller, 1988, Zhu & Simon, 1987). On the other hand, according to Paas, Renkl, and Sweller (2004), there is comparatively little evidence, demonstrating the efficacy of discovery learning during initial skill acquisition.

Tuovinen and Sweller (1999) while confident of their results closed their article with one caveat:

> It can, of course, be argued that exploration practice may be superior to worked examples, even for novices, if measures other than those of the present experiment are used. For example, exploration may favor long-term retention. Although this question must remain open until tested, it should be noted that in the present case, students with no previous database experience who learned by exploration, achieved such low test scores that minimal knowledge was available for long-term retention. (Tuovinen & Sweller, 1999, p.340)

This assertion, that long-term retention may favor exploration is an important point. It is eerily familiar, given the discussion of the animated demonstration literature, for it echoes Palmiter's concerns of an animation deficit (Lipps et al., 1998; Palmiter, 1991; Palmiter et al. 1993).

Thus this dissertation intends to measure learner performance, a week after initial instruction. Given this is the case the next order of business is to discuss how that might be accomplished.

How Has Cognitive Load Been Measured?

Our technology is just beginning to be able to peer inside the working brain to measure changes in brain function, thus it has been very difficult to measure cognitive load. However, humans have been very imaginative and developed physiological, computational, and self report estimates of cognitive load. This section describes the various cognitive load measures developed to date, but begins by describing the predecessors of these measures.

Human Factors and Cognitive Load Research

The measurement of cognitive load has several origins. It may be linked to cognitive psychology, or physiology, but perhaps cognitive load research has its closest ties to ergonomic or human factors (the parent field of human computer interaction, HCI). Consequently, some of the earliest cognitive load articles were published in the journal *Human Factors* (e.g. Paas & van Merriënboer, 1993). The field of human factors studies how people interact with their environment, and more recently has begun to concentrate on the computer interface (Bailey, 1996).

Usability Engineering (or simply Usability) became an important theme in HCI research. Even though researchers of the 1980s considered usability as a multidimensional construct (e.g. Bethke, Dean, Kaiser, Ort, & Pessin, 1981), they mainly described it in terms of “ease of use” or user satisfaction. Later in the 1990s, researchers began to consider other attributes of usability. Nielsen (1993) defines usability by describing five subcomponents (learnability, efficiency, memorability, errors, and satisfaction).

Soloway, Guzdial, and Hay (1994) called for Norman's "user-centered" design philosophy to be more "learner-centered." Nielsen's (1993) definition of learnability "How easy is it for users to accomplish basic tasks" is truly a subjective measure of "perceived usability," rather than a more objective comparison. To his defense, Nielsen (1993) was merely trying to describe a set of heuristic guidelines to help software programmers begin to think about usability. Nielsen (2001) even suggests that we should consider the user's opinions and suggestions only after watching them actually work with the software.

What is needed is a new way of producing design guidelines for software programmers and instructional designers. We need an objective method of evaluating software products based upon Soloway's learner-centered design philosophy (Soloway et al., 1994). Cognitive task analysis has fulfilled this role to date and studies cognitive tasks, but is it able to improve human performance? This may be possible, if we evaluate the instructional strategies that are the most efficient and effective,

Paul Merrill proposed instructional designers should use an information processing approach to task analysis, as they design procedure-based instruction (Merrill, 1971; Merrill 1976; Merrill, 1980). So given Merrill's cognitive perspective, researchers could influence human performance, by objectively comparing learner performance of complex cognitive tasks, and using these observations, improve the instructional strategies used to present these tasks.

Cognitive task analysis requires an objective approach to instructional design research, and in doing so, can refine instructional materials to produce effective efficient instruction. This then is the basis of an objective instructional science.

Instructional science has evolved over the past several decades and as it evolved so did its methods of inquiry. Cognitive load measures are a recent development of instructional science and human factors research. However, before these measures were developed, there were several predecessors that led the way.

Predecessors of Cognitive Load Measurements

NASA (the National Aeronautics and Space Administration), has tested human endurance and the limitations of human ability for several decades. So it should not be surprising that NASA researchers developed several important measures of "mental workload." One in particular, the NASA-TLX (task load index) (Hart & Staveland, 1988) is used to measure the load a person endures, during a task performance.

Hart and Staveland (1988) described their task-load index as being multidimensional, for it combines six subscales (mental demand, physical demand, temporal demand, performance, effort, and frustration level). They used this index in several studies to determine load conditions during several experimental tasks including simple cognitive tasks, manual control tasks, complex laboratory tasks, supervisory control tasks, and aircraft simulation.

Prinzel, Pope, Freeman, Scerbo and Mikulka, (2001) also reported using the NASA-TLX with Electroencephalogram (EEG) and Event-Related Potentials (ERPs) to build "adaptive automation technology." These computer-based systems automatically adapt to the limited capacities of human operators, when the operator is under high load conditions. Prinzel et al (2001) stated that their intention was to build adaptive systems that automate less critical tasks, to efficiently reduce an operator's workload. Their hope is to develop systems that limit disasters, like the challenger space shuttle and three-mile island accidents.

The NASA-TLX is just one example of several workload measures that have been developed. Other authors have developed similar measures, the Cooper-Harper Scale (Cooper & Harper, 1969) and the SWAT, Subjective Assessment Technique (Reid & Nygren, 1988). In all of the above assessments mental load was considered to be a multidimensional construct. The multidimensional nature of mental workload will be described further in later sections.

Before continuing, it should be noted there are two main differences between these measures and those used during cognitive load research, the audience and conditions. Cognitive load researchers typically only measure the performance of novices during learning, where the above measures were much more general and developed for other audiences and circumstances.

Objectivity and Cognitive Load

Brünken, Plass, and Leutner (2003) discuss the measurement of cognitive load in some detail, and even develop a classification scheme to describe cognitive load assessments. They classify cognitive load measurements along two basic dimensions: objectivity and causal relation. In their classification scheme, assessments along the causal relation dimension can be described as either direct or indirect, while those in the objectivity dimension are either objective or subjective (See Table 2).

Table 2

Methods for measuring cognitive load

	Causal Relationship	
Objectivity	*Indirect*	*Direct*
Subjective	Self-reported invested mental effort	Self-reported stress level Self-reported difficulty of materials
Objective	Physiological measures Behavioral measures Learning outcome measures	Brain activity measures (e.g., fMRI) Dual-task performance

Note. Adapted from "Direct measurement of cognitive load in multimedia learning," by R. Brünken, J. L. Plass, & D. Leutner, 2003, *Educational Psychologist*, 38(1), p.55

All of the measures employed by cognitive load researchers have their advantages and disadvantages (Brünken et al., 2003). Tuovinen and Paas (2004) revealed that most studies measuring cognitive load, typically make use of self-reported mental effort ratings developed in the early 1990s. Nevertheless, Brünken, Plass, and Luetner (2003) state:

> Although this technique, which is frequently used in current cognitive load research (See Paas, Tuovinen, Tabbers, &Van Gerven, 2003), appears to be able to assess the subjective perception of invested effort reliably, it remains unclear how this mental effort relates to actual cognitive load (Brünken, Plass, & Luetner, 2003, p.56).

While learners may be able to self report their own levels of cognitive load, this measure is not objective. On the other hand, functional magnetic resonance imaging (fMRI) provides more direct objective cognitive load measurements, but this method is expensive and difficult for educational researchers to use with large populations. Even though this is the case, brain imaging studies are being used to study learners during computer-based problem solving (e.g. Anderson, Albert, & Fincham, 2005).

Given the constraints of this study, it is not possible to use fMRI. Instead, this project used indirect objective methods, specifically behavioral observation, as well as indirect subjective measures like those proposed by Paas and van Merriënboer (1993). However, before explaining these methods, it is necessary to explain the reasoning and derivation of these methodologies.

The Efficiency Perspective of Cognitive Load

Paas and van Merriënboer (1993) base their methodology on an "efficiency perspective" of cognitive load, a slightly different view from Sweller's element interactivity perspective. However, Sweller, van Merriënboer, and Paas (1998) synthesized these views a decade ago and most cognitive load studies since the mid 1990s have used this perspective (Tuovinen & Paas, 2004).

This "efficiency perspective" of cognitive load, dates back before Sweller's seminal article describing cognitive load theory (Sweller, 1988), since Ahern and Beatty previously published an article in the journal *Science* (Ahern & Beatty, 1979). Their article studied the human eye during cognitive processing. By this stage, it had already been determined that the pupil dilates during increased cognitive activity (Janisse, 1977). In addition, it was known that pupil dilation varied based on the momentary cognitive demands of task performance (Beatty, 1977). Ahern and Beatty (1979) took this idea a step further; to hypothesize that higher ability learners should have more efficient cognitive structures (better formed schemata) and that their pupil dilation would reflect this more efficient cognitive processing. They tested this hypothesis and found evidence which supported their case. So, they were able to provide physiological data (further evidence) that greater expertise resulted in higher mental efficiency. Recall that Chase and Simon (1973a) had found that chess experts have the ability to manipulate more information, in a shorter period of time, because they recognize patterns. Thus an expert does not have a larger memory capacity; they are just more efficient with their limited capacity, because they have well formed schemata.

Ahern and Beatty (1979) also found physiological evidence of task automation. Specifically, they state that the "pupillary response during information processing might reflect the effects of prior practice at cognitive tasks" (Ahern & Beatty, 1979, p. 1291). Thus their results support those of Schneider & Shiffrin (1977), who first proposed the idea that experience or practice promotes the automation of learned skills. In Ahern and Beatty's words "the effect of further practice is to make access to the items more automatic and thereby to decrease the processing load associated with item retrieval" (Ahern & Beatty, 1979, p. 1291).

Ahern and Beatty's "processing load" is of course, what Sweller later described as cognitive load (Sweller, 1988). Thus a learner with more expertise has a reduced cognitive load, because they have acquired schemata, which better describes task performance. So a learner with more experience does not have an increased cognitive capacity, but they are more efficient with their limited working memory. This is what Paas & van Merriënboer (1993) later described as the efficiency perspective of cognitive load. So according to Paas and van Merriënboer's efficiency perspective, learners are more efficient during an instructional condition, if their performance is greater than expected, and their invested mental effort is lower than expected (Paas & van

Merriënboer, 1993). Paas and van Merriënboer (1993) developed this perspective to produce a hybrid objective/subjective cognitive efficiency construct (E).

E is for Efficiency

Paas drew upon the field of human factors to develop the efficiency view of cognitive load. In a 1992 article, he states “Cognitive load is a multidimensional concept” (Paas, 1992 p.429) but also says “the intensity of effort is considered to be an index of cognitive load” (Paas, 1992, p. 429). Paas and van Merriënboer (1993) further developed this, to derive a multidimensional construct which they called “relative condition efficiency (E)” (Paas & van Merriënboer, 1993, p.737). Constructs (or latent variables) are not directly observable (Schumaker & Lomax, 2004), but allow researchers to study less tangible concepts, like cognitive load. Schumaker and Lomax (2004) describe latent variables as being inferred from two or more measured indicator variables. Paas and van Merriënboer’s relative condition efficiency (E) construct (Equation 1) is composed of standardized mental effort ratings and performance scores.

$$E = \frac{|Z_{Performance} - Z_{MentalEffort}|}{\sqrt{2}} \tag{1}$$

This formula has recently been revised to remove the absolute value symbols to be mathematically equivalent and somewhat simpler, as in Equation 2 (Clark, Nguyen, & Sweller, 2006a; Paas, Tuovinen, Tabbers, & van Gerven, 2003; Tuovinen & Paas, 2004).

$$Relative\ condition\ efficiency = \frac{Z_{Performance} - Z_{MentalEffort}}{\sqrt{2}} \tag{2}$$

Paas and van Merriënboer (1993) derived their construct (Equation 1) from the point-line distance formula (See Equation 3):

$$\frac{|ax+by+c|}{\sqrt{a^2+b^2}}=0 \tag{3}$$

The point-line distance formula is a geometric formula, for finding the perpendicular distance between a point and a line (Weisstein, 2008). Paas and van Merriënboer used this formula, to combine two sets of z-scores, in order to graph the resulting equation.

In the case of relative condition efficiency (E), the point is a standardized, group mean score, for an instructional condition, given two variables (mental effort and performance). These two variables are plotted relative to one another on a two dimensional graph, with mental effort on the x axis and performance on the y axis (See Figure 12). The denominator of equation, $\sqrt{2}$ rotates the combined scores 45 degrees from either axis to form the efficiency, E=0 line.

Figure 13 is a generalized efficiency graph. Scores above the E=0 base line, in the upper left hand corner of the graph, are expected to have a greater relative condition efficiency, because they have a better performance with decreased mental effort. Paas and van Merriënboer's study is a good example to explain the use of this metric.

Paas and van Merriënboer compared conventional problem solving, worked examples, and completion problems. To compare group mean scores, they conducted a one-way ANOVA and revealed a significant difference between groups, F (2, 42) = 24.76, $p < 0.001$. They found in post hoc comparisons (a Fischer's test), that the conventional problem solving condition (E=-1.15) was significantly less efficient than the other conditions (worked example and problem completion groups), which were not significantly different.

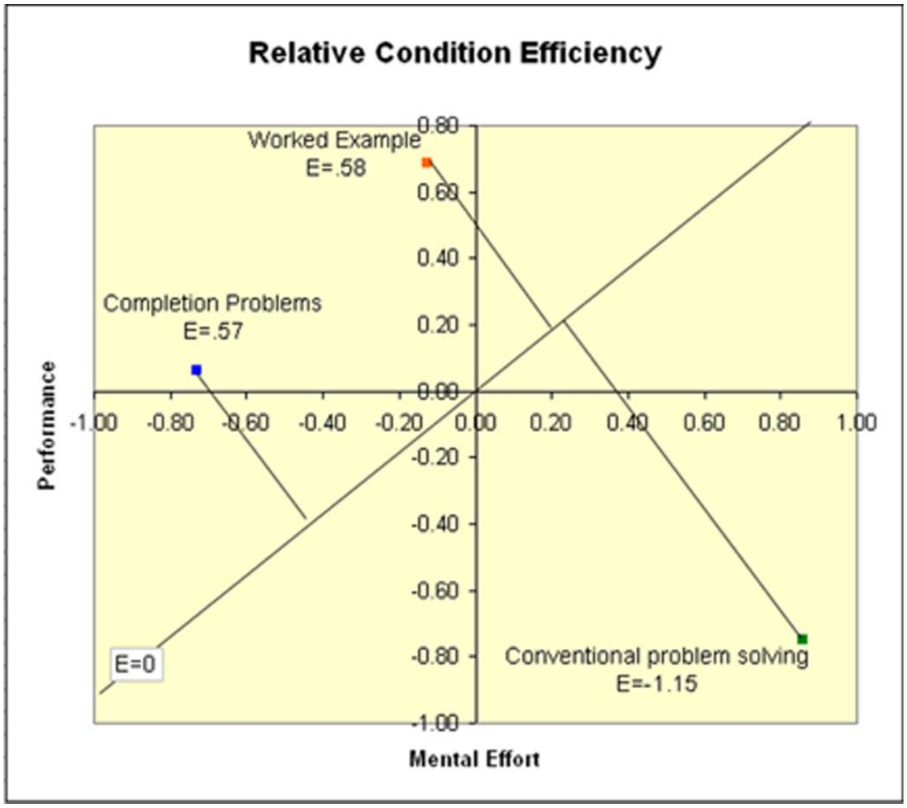


Figure 12. A graph of relative condition efficiency.

Note. Adapted from “The efficiency of instructional conditions: An approach to combine mental effort and performance measures,” by F.G.W.C. Paas and J.J.G. van Merriënboer, 1993, *Human factors*, *35*(4), p.741

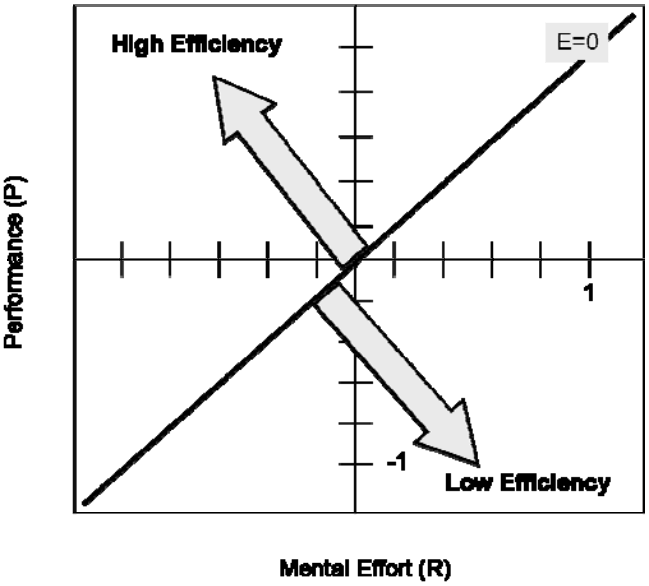


Figure 13. Generalized efficiency graph

Note: Adapted from “Cognitive Load Measurement as a Means to Advance Cognitive Load Theory,” by Paas, F., Tuovinen, J.E., Tabbers, H. & Van Gerven, P. W. M., 2003, *Educational Psychologist*, 38(1) p.68

Paas and van Merriënboer (1993) were explicit and suggested that these combined mental effort/performance-based cognitive load measurements should be qualified with performance data. Therefore, even though the worked example and problem completion group scores did not differ significantly, the mean performance scores for these groups were: *M* = 78.57 for the worked example group, *M* = 67.22 for the problem completion group, and finally *M* = 51.60 for the conventional discovery problem-solving group. Finally it should be stated that relative condition efficiency (E) assumes a linear relationship between perceived mental effort and performance (Paas & Merrienboer, 1993).

Variations on a Theme

Ten years after the original relative condition efficiency article, Paas, Tuovinen, Tabbers, and Van Gerven (2003) reported that researchers had required learners to provide mental effort ratings at different times.

It seems Paas and Van Merriënboer (1993) had used mental effort estimates from the test phase, while Sweller and his associates used mental effort estimates from the learning phase (See Table 3).

To clarify matters, Paas et al (2003) suggested dividing these into two separate efficiency metrics and that researchers use the terms “relative condition efficiency” (Paas & van Merriënboer, 1993, p.739) and “learning efficiency” (Paas et al., 2003, p.69).

Table 3

Relative condition efficiency and learning efficiency

Relative condition efficiency		
	Learning phase	Test phase
Performance score		X
Mental effort estimate		X
Learning efficiency		
	Learning phase	Test phase
Performance score		X
Mental effort estimate	X	

Note. Adapted from "A Multidimensional Approach to the Mental Efficiency of Instructional Conditions" by F. Paas, 2007, retrieved June 5, 2007 from http://www.ou.nl/Docs/Expertise/OTEC/Projecten/onderzoeksvoorstellen%20PDF/Paasproject34%5B1%5D.pdf

Tindall-Ford, Chandler, and Sweller (1997) were perhaps the first to measure what came to be called "Learning efficiency." That is they conducted their mental effort ratings after the learning phase, but before the test phase. As Table 4 shows many cognitive load researchers followed their lead and also measured what came to be called "learning efficiency."

Tuovinen and Paas (2004) explained that each of these metrics signifies different aspects of the learning-testing process. Paas et al (2003) suggests "relative condition efficiency" is mostly concerned with mental effort expended during a test performance, and may be more related to transfer.

The main benefit of this metric is that it helps instructional designer researchers measure the relative efficiency of instructional conditions (Tuovinen & Paas, 2004). On the other hand learning efficiency is concerned with the mental effort expended during training.

Table 4

The timing of mental effort ratings

Studies	Learning phase	Test phase
Paas and Van Merriënboer (1993)		ME, P
Marcus, Cooper and Sweller (1996)		ME, P
Tindall-Ford, Chandler, and Sweller (1997)	ME	P
Yeung, Jin, and Sweller (1997)		ME, P
Kalyuga, Chandler, and Sweller (1998)	ME	P
Kalyuga, Chandler, and Sweller (1999)	ME	P
Yeung (1999)		ME, P
Tuovinen and Sweller (1999)	ME	P
Kalyuga, Chandler, and Sweller (2000)	ME	P
Camp, Paas, Rikers, and Van Merriënboer (2001)		ME, P
Kalyuga, Chandler, and Sweller (2001)	ME	P
Kalyuga, Chandler, Tuovinen, and Sweller (2001)	ME	P
Pollock, Chandler, and Sweller (2002)	ME	P
Van Gerven, Paas, Van Merriënboer, and Schmidt (2002)	ME	P
Van Merriënboer, Schuurman, De Croock, and Paas (2002)	ME	P

Note: adapted from "Exploring multidimensional approaches to the efficiency of instructional conditions" by Tuovinen, J. E. & Paas, F. G. W. C., 2004, *Instructional Science 32*(1-2) p. 136

3 Dimensional Approaches

Tuovinen and Paas (2004) defend the original metric because they say it is important to measure the relative importance of mental effort during a test or performance:

> It is quite feasible for two people to receive the same performance scores, while one of them needs to work laboriously through a very effortful process to arrive at

> the same number of correct answers, whereas the other person reaches the same answers with a minimum of effort (Tuovinen & Paas, 2004, p.140).

Thus, Tuovinen and Paas began to consider new metrics that would consider both the test phase and the learning phase, simultaneously. In their search for this new metric, Tuovinen and Paas (2004) created some new terminology. They refer to both relative condition efficiency and learning efficiency as 2 dimensional (2D) measures because they include one mental effort measure and one performance measure.

Tuovinen and Paas (2004) also developed a 3D (or 3 dimensional measure) by combining the factors of the 2D measures; to do so, they factored in learning effort (E_L), test effort (E_T), and performance (P) in the following formula:

$$\text{3D Efficiency} = \frac{\mathrm{P} - E_{\mathrm{L}} - E_T}{\sqrt{3}} \tag{4}$$

Like 2D efficiency metrics, this 3D measure is also graphed, but in three dimensional space. Tuovinen and Paas (2004) claim this combines the best features of both metrics.

However, Salden, Paas, Broers, and van Merriënboer (2004) produced yet another 3D metric that makes use of total training time. To do so, Salden et al (2004) combined performance (P), mental effort (ME) and total training time (TT) in the following formula:

$$\text{Training Efficiency} = \frac{\mathrm{P - ME - TT}}{\sqrt{3}} \tag{5}$$

This metric makes more sense because it factors in time. Certainly cognitive load is concerned with the amount of time involved during training. This metric can also be graphed in three dimensional space with performance, mental effort and time on the three axes. In addition, other studies since Tuovinen and Paas' article have begun to use 4D metrics which include motivation (Nadolski, Kirschner & van Merriënboer, 2005). However 3D & 4D metrics will not be used in this study.

Performance Efficiency

Consider equation 6. Is mental effort a necessity for an efficiency metric? Also, recall that Brünken, Plass, and Leutner (2003) categorized cognitive load measures based upon direct and indirect, or by subjective and objective methods. As they categorized these measures, they explained that they were uncertain how perceived mental effort is related to cognitive load. Even though it can be argued that perceived mental effort is an indicator of cognitive load, performance time is more objective, and a better indicator of the efficiency of a learner's performance.

This study intends to derive a new formula based completely on the objective measures of performance time (PT) and performance (P), but does not include a subjective mental effort rating. This will be described as "performance efficiency" (PE) (See Equation 6)**:**

$$\text{Performance efficiency} = \frac{Z_{Performance} - Z_{PerformanceTime}}{\sqrt{2}} \tag{6}$$

As with many of the other efficiency formulas, performance time and performance are standardized with Z-scores as in Paas and van Merriënboer's 1993 article. Performance efficiency may also be represented in abbreviated form like the other efficiency metrics, with P representing the Z-score of performance and PT representing the Z-score of performance time (See Equation 7).

$$\text{Performance Efficiency} = \frac{\text{P - PT}}{\sqrt{2}} \tag{7}$$

Finally, this metric is also graphed, and the denominator $\sqrt{2}$ is derived in the same way, from the point-line distance formula. These values are also plotted on a two dimensional biplot, but with performance time on the X axis, and contrasted with performance, on the Y axis (this is discussed in more detail in later chapters).

While this new metric does not represent mental effort or mental efficiency, it is an objective efficiency measure. Also like relative condition efficiency, performance efficiency may be used to compare instructional conditions, to describe a group's performance and relate the performance of groups to one another.

Before the Paas and van Merriënboer (1993) article, performance measures were used exclusively. Performance efficiency provides a simple objective way to express group performance versus time. Like relative condition efficiency, performance efficiency allows one to compare group performance from a graphical perspective.

However, it should be stated that like relative condition efficiency, performance efficiency also assumes a linear relationship between its two factors, in this case, between performance and performance time. This relationship (or slope of the E=0 line), may vary depending on problem complexity and the environment in which problems are solved.

In summary, there is little doubt that Paas and van Merriënboer's efficiency metric has had a dramatic effect on the cognitive load literature. It has helped researchers produce an estimate of cognitive load. The current study uses the relative condition efficiency metric described by Paas and van Merriënboer (1993), but also intends to implement a similar measure, performance efficiency (See Equation 6 or 7).

Rationale

In short, there are two broad instructional strategies compared in this dissertation. Like Tuovinen and Sweller (1999), this dissertation compares Bruner's two modes of instruction, animated worked examples (the expository mode) with discovery learning (the hypothetical mode). However, the literature review has revealed four main areas of inquiry which should also be considered in this study.

The first area of inquiry revolves around Tuovinen and Sweller's closing remarks; they had some reservations about worked examples and retention (Tuovinen & Sweller, 1999). They asked future researchers to consider retention, and implied that retention may not be as durable with worked examples, as it is with discovery problem solving. Given these reservations, there may be a case for Palmiter's animation deficit, and it could be a legitimate concern given animated demonstrations. So, in terms of the animated demonstration literature, would those who studied animated demonstrations have a performance decrement a week later (Palmiter's animation deficit)?

Secondly, the literature review found the worked-example effect was apparent by those who had studied solved problems (Cooper & Sweller, 1987; Sweller & Cooper, 1985; Sweller & Chandler, 1991). Lewis (2005) proposed animated demonstrations are a similar presentation form. Therefore will learners using this form of instruction also have and an increased performance over their problem solving peers (the worked-example effect)?

Third, Paas and van Merriënboer (1994) found that learners, who studied varied-context worked examples, outperformed those that solved problems, and described this as the variability effect (Paas & van Merriënboer, 1994). Given the claims of Lewis (2005) that animated demonstrations act as worked examples, will the variability effect also be apparent given animated demonstrations?

Finally, a fourth area of inquiry is the methodologies used by cognitive load researchers. It seems some members of the cognitive load community have questioned the relationship of mental effort ratings and cognitive load (Brünken et al., 2003). Brünken et al. (2003) also compared cognitive load measures on two dimensions, objectivity and causal relation. Since no one measure was found to be advantageous, it was decided that the best way to understand this relationship was to triangulate multiple measures, because it is currently necessary to use a combination of both objective and subjective measures. To help qualify the subjective nature of mental effort ratings, a new measure, performance efficiency was developed. Therefore this study intends to test performance efficiency, in order to help qualify the results of cognitive load research.

Given the methodologies of cognitive load research, it will be necessary to measure several variables: perceived mental effort, performance time and accuracy. These variables may be addressed on their own, or in combination via constructs like relative condition efficiency or performance efficiency.

Therefore in order to address each of the areas of inquiry proposed above, the following research questions are presented:

Question 1: Is there a significant difference among the instructional strategies, relative to performance time?

Question 2: Is there a significant difference among the instructional strategies, relative to accuracy?

Question 3: Is there a significant difference among the instructional strategies, relative to “relative condition efficiency?”

Question 4: Is there a significant difference among the instructional strategies, relative to “performance efficiency?”

Operational Definitions

Several of the dependent variables in the above research questions are self evident, but Palmiter’s animation deficit and the cognitive load learning effects must be described as a combination of these variables. Therefore this section provides explicit operational definitions.

Palmiter’s Animation Deficit

Lipps, Trafton and Gray (1998) described Palmiter’s animation deficit as “poorer retention despite faster learning following animation training” (Lipps et al., 1998, p. 1). In terms of this study and its dependent variables, an operational definition of Palmiter’s animation deficit, would be a significant increase in performance time with a simultaneous decrease in accuracy, in a delayed performance, one week after initial instruction (given animated demonstrations as an instructional strategy).

The Worked-example Effect

The worked-example effect is often defined as an improvement in learner performance given worked examples. Sweller and Cooper’s early studies were the first to describe this effect (Cooper & Sweller, 1987; Sweller & Cooper, 1985). They described this effect by saying a “decreased solution time was accompanied by a decrease in the number of mathematical errors” (Sweller & Cooper, 1985, p.59). The dependent variables here are solution time and "a reduction of errors", or in terms of this study and the animated demonstration literature, performance time and accuracy (Palmiter, 1991).

Therefore, for the purposes of this dissertation, the worked-example effect will be operationally defined as a significant reduction in performance time and a simultaneous significant increase in accuracy.

The Variability Effect

Paas and van Merriënboer (1994) found the variability effect. They had studied the learner performance of those who had studied "varied context" worked examples. These researchers had developed relative condition efficiency (RCE) and compared the learner performance of learners under a variety of conditions. This dissertation will also use RCE to consider the variability effect.

RCE includes a performance variable and a perceived mental effort rating. In the current study, performance is represented by accuracy. Thus to measure RCE, it was necessary to measure perceived mental effort. In keeping with Paas and van Merriënboer (1994), this study will define the variability effect, as a significant increase in relative condition efficiency, for a high variability instructional condition relative to other conditions (as described in Equation 8).

$$\textit{Relative condition efficiency} = \frac{Z_{Accuracy} - Z_{MentalEffort}}{\sqrt{2}} \qquad (8)$$

This leads us to the methodology of the dissertation (Chapter three).

CHAPTER THREE - METHODS

This chapter outlines the methodology employed during the study. The chapter is broken into several sections to clarify concepts, methods, and instruments. It begins with a section which identifies the participants. This is followed by sections discussing the research design, the materials, dependent variables, constructs, analysis, reliability, and finally the chapter concludes with a section devoted to the pilot study.

Participants

The participants of this study were pre-service teachers. These learners were undergraduates enrolled in an introductory instructional technology course at a large southeastern university. This audience was chosen because it was expected that they would be primarily novices with the procedures presented, although some variability was expected in the population.

An *a priori* power analysis for a four group MANOVA produced a sample size of n=115 participants. This number of participants is necessary to arrive at a power of 0.80, with a small effect size η^2= 0.125, given α =0.05 (α=0.05 is used throughout this study, unless stated otherwise) (Stevens, 2002).

Research Design

This section provides a brief overview of the design of the study; each of these ideas is described in detail in later sections. The literature review suggested contrasting instructional strategies. Since the study compares the performance of learners in four independent groups, this is known as a “between groups” design (Gall, Borg, & Gall, 1996; Mook, 2001).Therefore the design of this study was experimental because it randomly assigned learners to one of four instructional conditions.

Worked-example based instruction typically involves both an example and some level of practice (Sweller, 2006). So combinations of these strategies were analyzed, in order to study animated demonstrations. Since Palmiter felt learners who used animated demonstrations mimicked the instruction (Palmiter, 1991), two conditions were compared, one using an identical problem (the mimic condition) and one using a different problem than that demonstrated. Also, a demonstration-only group (demo) was included to contrast learning under this limited set of circumstances. Finally, since one of the goals of this dissertation was to contrast discovery practice with animated demonstration, as in the Tuovinen and Sweller study (Tuovinen & Sweller, 1999), a fourth practice only condition was included. Consequently, this study compared a total of four instructional conditions:

1. “demo” - animated demonstration only;
2. “demo+practice” - an animated demonstration, plus practice with the demonstrated task;
3. “demo2+practice” - a second animated demonstration, plus practice with a task (different from that demonstrated);
4. “practice” – discovery-based, practice

   (each instructional conditions is discussed in detail, later in this chapter).

Data analysis required a series of univariate and multivariate statistical procedures. Two separate multivariate analysis of variance (MANOVA) tests were conducted, because there were multiple outcome variables (performance time and accuracy) over two sessions. In addition, two efficiency constructs (relative condition efficiency and performance efficiency) were analyzed.

Performance data were gathered during two separate phases, to test the retention of procedural learning and a potential animation deficit (Lipps et al., 1998; Palmiter, 1991). Therefore, learner retention was assessed during a more immediate performance, during an *acquisition phase* (week 1), and then longer-term retention was assessed a week later, during the *retention phase* (week 2). The purpose of the acquisition phase was to introduce learners to the subject matter, and contrast immediate performance. The purpose of the retention phase was to contrast learner performance given the instructional conditions, one week after initial instruction.

Materials

This section of the dissertation describes the materials and overall sequence of events during the study. Subsequent sections describe each instrument in detail. Learners interacted with all instruments and instructional materials via IBM compatible computers, using Windows XP (service pack 2) and Internet Explorer 6.0. These computers had 2GHz AMD (Advanced Micro Devices) Athlon 2400 processors, with 480 MB of RAM. Computers were arranged in a classroom setting. Before learners entered the environment, TechSmith Morae Recorder (screen capture software) was executed and allowed to record individual learners once they interacted with the computer. In addition, a completed project was projected on a screen at the front of the classroom.

A Synopsis of the Acquisition Phase

This section briefly outlines the sequence of events and instruments used during the acquisition phase (See Figure 15 for a flowchart). The recording process, instruments, instructional conditions, and variables are all discussed in later sections of the chapter. Before learners entered the learning environment computers were prepared, specifically Morae Recorder was allowed to record learner interaction on all computers.

The acquisition phase began when all learners were presented with an initial web-based survey, survey 1. Following this demographic survey, they were presented with a brief overview, which introduced the subject matter. After viewing the overview, a JavaScript randomly assigned learners to one of four instructional conditions. Learners in the demo+practice, demo2+ practice, and practice conditions were asked to assemble the Mr. Potato head document (See Figure 14). Rather than interacting with this document, learners from the demo condition were asked to continue. Finally, the acquisition phase concluded by asking all learners to complete a post treatment survey Week 1 survey #2.

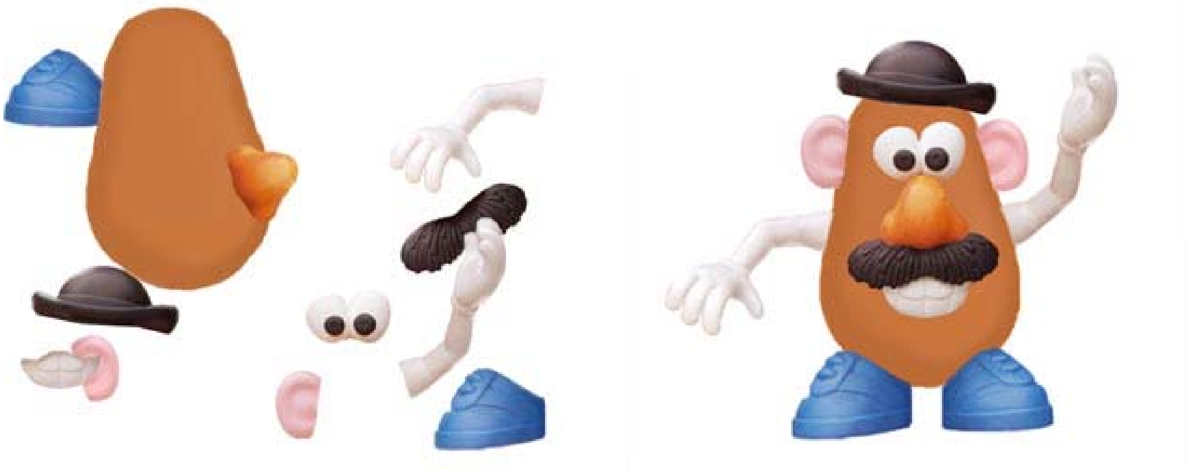


Figure 14. Week one - the “Mr. Potato head” problem

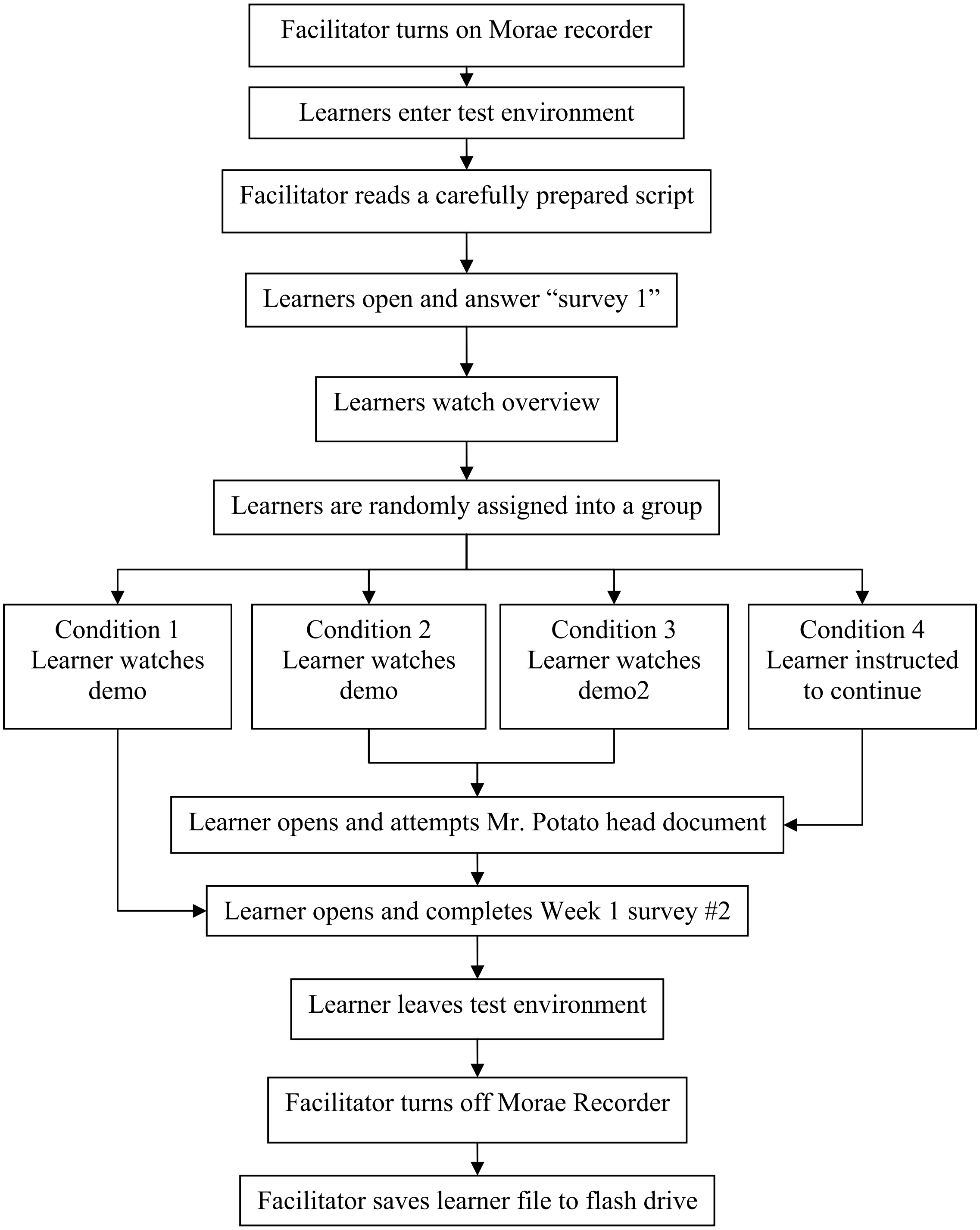

Facilitator turns on Morae recorder
Learners enter test environment
Facilitator reads a carefully prepared script
Learners open and answer "survey 1"
Learners watch overview
Learners are randomly assigned into a group
Condition 1 Learner watches demo
Condition 2 Learner watches demo
Condition 3 Learner watches demo2
Condition 4 Learner instructed to continue
Learner opens and attempts Mr. Potato head document
Learner opens and completes Week 1 survey #2
Learner leaves test environment
Facilitator turns off Morae Recorder
Facilitator saves learner file to flash drive

Figure 15. The materials and processes used during the acquisition phase

A Synopsis of the Retention Phase

During week two, the retention phase, all learners were presented with the different performance problem (the picnic problem) (See Figure 16). Once they had attempted the performance problem, a post-performance survey was administered. The next few sections describe each of the instruments.

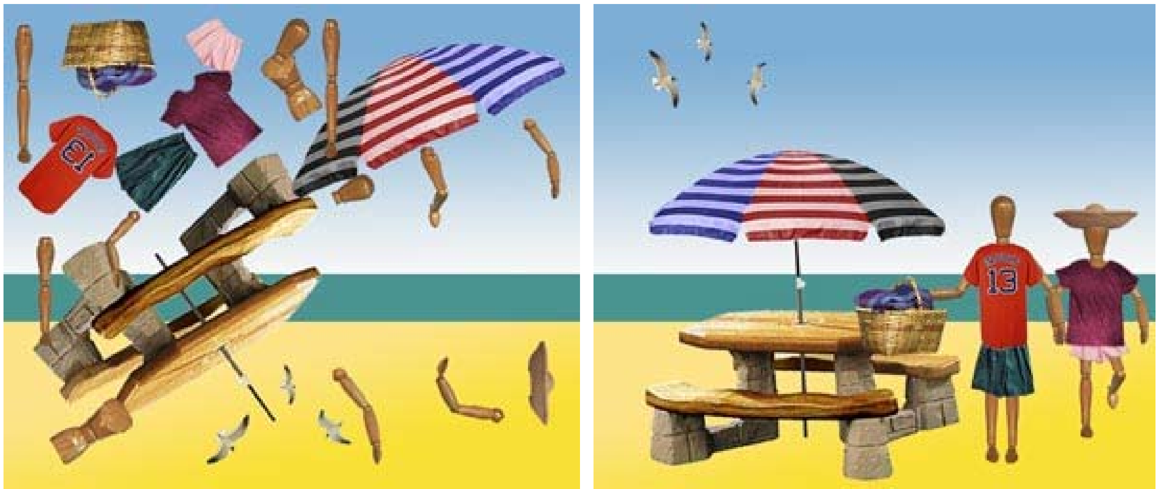
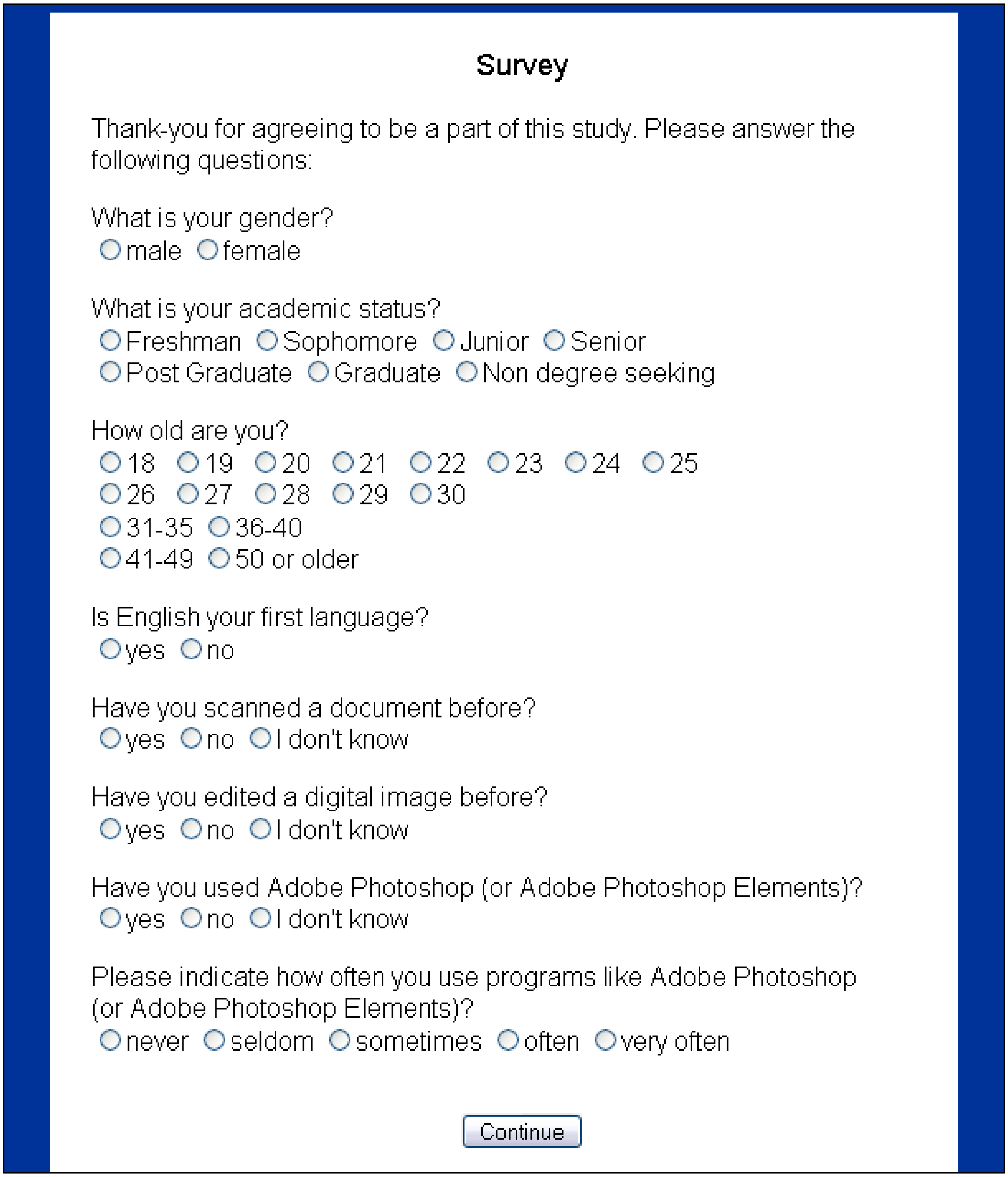


Figure 16. Retention phase - the Picnic problem

Pretreatment Survey (survey 1)

It was important to gather various forms of demographic data. Therefore a pretreatment survey (Survey 1) (See Figure 17) was developed. Survey 1 was a web-based survey developed with Microsoft FrontPage 2003 (Microsoft, 2003a). Once learners filled out the survey, the act of survey submission automatically forwarded the learner to the introductory overview.



Survey

Thank-you for agreeing to be a part of this study. Please answer the following questions:

What is your gender?
○ male ○ female

What is your academic status?
○ Freshman ○ Sophomore ○ Junior ○ Senior
○ Post Graduate ○ Graduate ○ Non degree seeking

How old are you?
○ 18 ○ 19 ○ 20 ○ 21 ○ 22 ○ 23 ○ 24 ○ 25
○ 26 ○ 27 ○ 28 ○ 29 ○ 30
○ 31-35 ○ 36-40
○ 41-49 ○ 50 or older

Is English your first language?
○ yes ○ no

Have you scanned a document before?
○ yes ○ no ○ I don't know

Have you edited a digital image before?
○ yes ○ no ○ I don't know

Have you used Adobe Photoshop (or Adobe Photoshop Elements)?
○ yes ○ no ○ I don't know

Please indicate how often you use programs like Adobe Photoshop (or Adobe Photoshop Elements)?
○ never ○ seldom ○ sometimes ○ often ○ very often

Figure 17. Pretreatment Survey (Survey 1)

Certainly Survey 1 allowed for the collection of basic demographic data, but in addition, it was used to screen second language learners. Several authors who have studied the cognitive load of second language learners (Grace, 1998; Krashen, 1982; Plass & Jones, 2005), proposed these learners may be under an additional load, because of the need to translate instruction into their native language. For optimal learning these authors suggest instruction should be translated into their native language. Since the language of all instruction was in English, those who answered no, to the fourth question "Is English your first language?" were allowed to participate, but their results were removed from the dataset.

Introductory Overview

Paas (1992) and Tuovinen and Sweller (1999) provided their learners with an introductory overview. This gave learners some context for the instructional conditions. The introductory overview used in this study was a short narrated web-based presentation (~ 2 minutes) developed with TechSmith Camtasia 4.0 (TechSmith, 2006). It provided learners with an introduction to graphic design and digital image editing. In addition to describing the field of graphic design, this narrated frame-based (non-animated) presentation, presented all learners with screenshots of Adobe Photoshop Elements 2.0 (Adobe Systems, 2002).

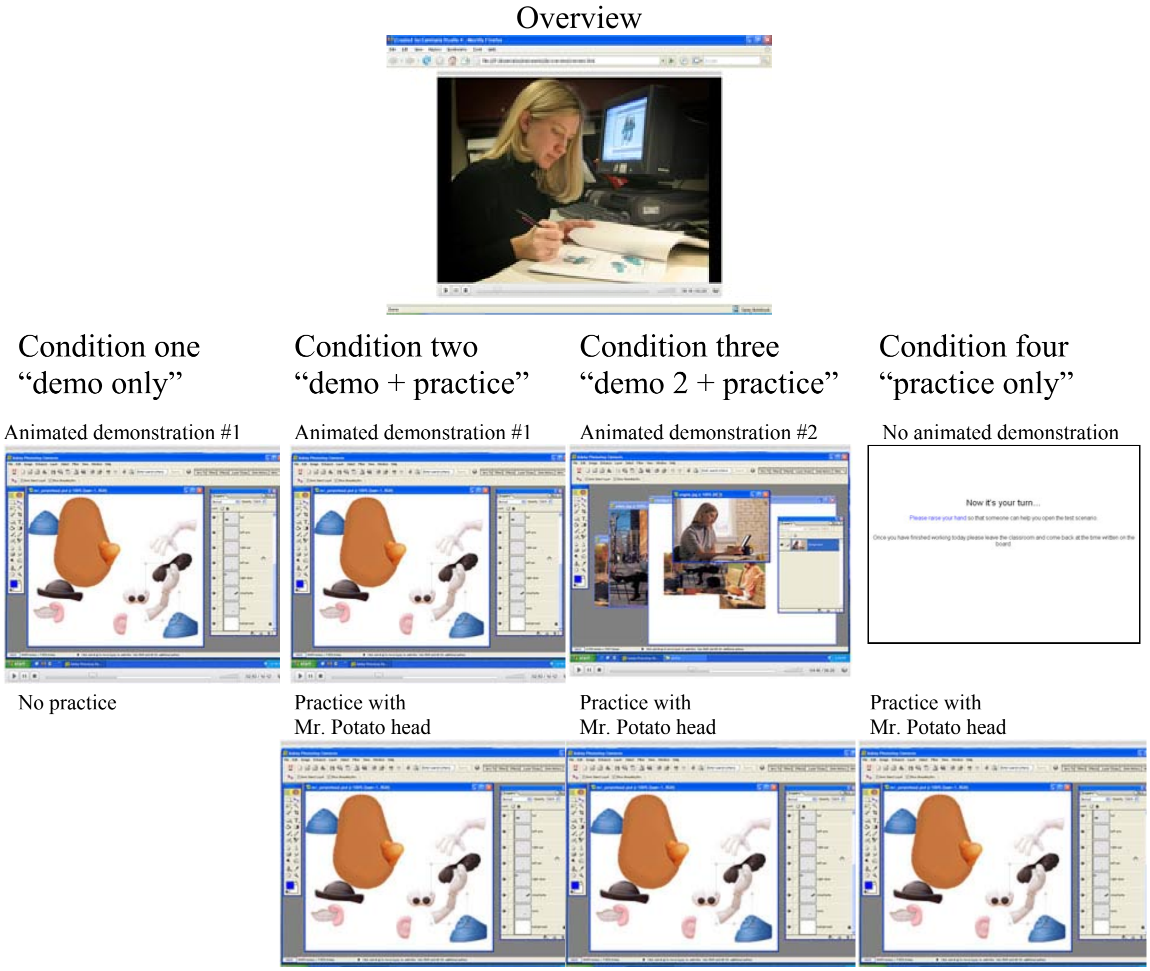
Once the overview concluded, a JavaScript randomly divided learners into four instructional conditions, by forwarding them to a new web page. The practice condition web page suggested learners raise their hand to get assistance from the facilitator (who opened the Mr. Potato head document for them). The facilitator also opened the Mr.

Potato head document for members of groups 2 and 3, once they had completed the remaining components of their instructional conditions.

Instructional Conditions

The four experimental groups make use of two animated demonstrations and practice (See Figure 18). The two animated demonstrations were developed with Techsmith Camtasia Studio 4.0 (Techsmith, 2006). The four instructional conditions were:

- Demo (Condition 1) – The instructional condition is a brief animated demonstration that shows the learner how to put together a Mr. Potato head document with Adobe Photoshop Elements. This condition only demonstrates a series of Photoshop procedures. Learners in this condition did *not* practice the demonstrated procedures.
- Demo + practice (Condition 2) – Learners in this condition viewed the same demonstration as those in condition 1, but also used Adobe Photoshop Elements 2.0, to put together the Mr. Potato head problem (the problem demonstrated).
- Demo 2+ practice (Condition 3) - Learners in this condition watch a different animated demonstration, which demonstrates the same underlying skills as demo 1, but puts together a photo collage rather than Mr. Potato head. After watching this collage demonstration, learners were asked to work with the Mr. Potato head problem.
- Practice (Condition 4) – Learners in this condition received no additional instruction other than the overview, but were asked to put together the Mr. Potato head problem.



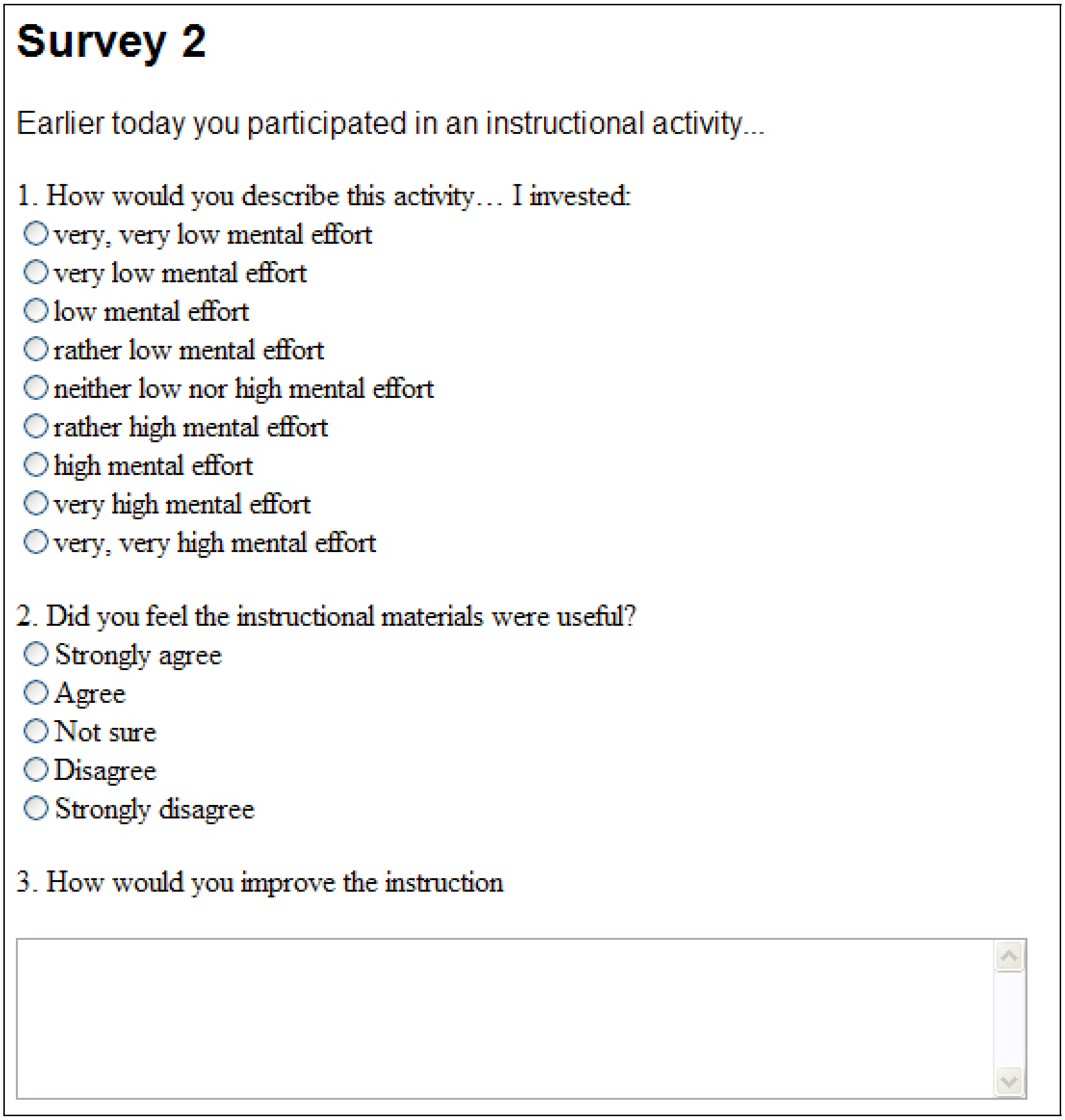
Now it's your turn...

Please raise your hand so that someone can help you open the test scenario.

Once you have finished working today please leave the classroom and come back at the time written on the board.

Figure 18. Instructional conditions

The initial problem scenario, the Mr. Potato head document, is an assembly task that requires assembly within Adobe Photoshop Elements 2.0 (Adobe Systems, 2002). The performance objectives of the Mr. Potato head learning activity required learners to select, move, rotate, and flip Photoshop layers to produce the Mr. Potato head product.



Survey 2

Earlier today you participated in an instructional activity...

1. How would you describe this activity… I invested:

○ very, very low mental effort
○ very low mental effort
○ low mental effort
○ rather low mental effort
○ neither low nor high mental effort
○ rather high mental effort
○ high mental effort
○ very high mental effort
○ very, very high mental effort

2. Did you feel the instructional materials were useful?

○ Strongly agree
○ Agree
○ Not sure
○ Disagree
○ Strongly disagree

3. How would you improve the instruction

Figure 19. Post-treatment survey (Week 1 survey #2)

Once learners had finished the post-treatment survey they were thanked for their participation, asked not to discuss their instruction with others, and finally not to use Adobe Photoshop Elements 2.0 before the next session.

The Retention Phase (the Picnic Problem)

One of the goals of this project was to understand how well learners would remember and be able to apply what they had learned from animated demonstrations. So one week after initial instruction (week two) learners put together another Adobe

Photoshop Elements document that required learners to recall what they had learned one week prior.

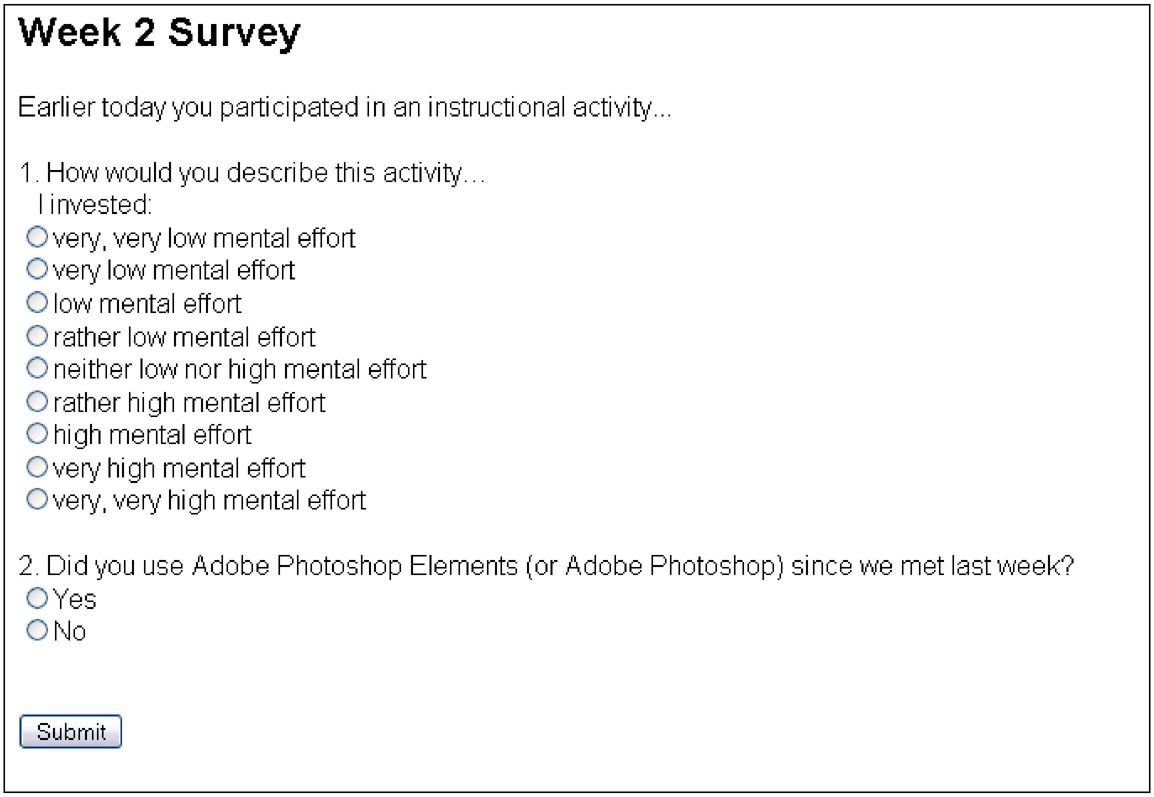
The picnic problem (See Figure 16) is a construction task that requires learners to put together two stick figures within a scene. This picnic problem is somewhat more complicated than the week one practice problem (the Mr. Potato head document) because it is composed of multiple disassembled figures within a scene, along with several other objects (a picnic basket, picnic table, an umbrella, and birds) but it still only used the skills acquired in the Mr. Potato head problem.

The picnic problem requires the same skills as the Mr. Potato head problem, but their newly learned skills must be used to reconstruct a more complex graphic. To complete the picnic problem learners only needed to select, move, rotate, and flip Photoshop layers.

Post-Performance Survey (week 2 survey)

It was also necessary to obtain mental effort ratings during the retention phase (week two) to provide evidence for research question three. So following their performance with the picnic problem, learners filled out the “week 2 survey” (See Figure 20). This survey was similar to the post-treatment survey because it also included a mental effort question. This question is identical to the one used in the Paas and van Merriënboer (1993) study, and the post treatment survey used during the previous week (See Figure 19).

The “week 2 survey” had a question aimed at determining if the student had used Adobe Photoshop Elements (or Adobe Photoshop) in the week since the initial instruction. Results from the learners who answered yes (had used Adobe Photoshop Elements between sessions) were removed from the dataset.



Week 2 Survey

Earlier today you participated in an instructional activity...

1. How would you describe this activity...
I invested:
○ very, very low mental effort
○ very low mental effort
○ low mental effort
○ rather low mental effort
○ neither low nor high mental effort
○ rather high mental effort
○ high mental effort
○ very high mental effort
○ very, very high mental effort

2. Did you use Adobe Photoshop Elements (or Adobe Photoshop) since we met last week?
○ Yes
○ No

Figure 20. Week 2 survey

The Dependent Variables and Constructs

Dependent Variables

Gagné (1964) describes two general categories of dependent variables which are often associated with problem-solving studies. He suggests most researchers are concerned with (1) “the rate of attainment of some criterion performance” and (2) “the degree of correctness of this performance” (Gagné, 1964, p.295).

Gagné’s “rate of attainment” is easily measured as performance time in seconds. “Performance time” was recorded and measured with TechSmith Morae 1.01

(TechSmith, 2004). Recording and measurement procedures are described in later sections.

Gagné's second category, "the degree of correctness" is not as easily defined. Gagné (1964) mentions single problem solutions are usually measured as either pass or fail. In other words, either the learner correctly solves the problem (attains the problem goal) or not. While problem completion is an important part of any problem solving study, this pass/fail measurement oversimplifies the learner's attempt during complex cognitive tasks. Gagné mentions one may score a learner's performance as "partially correct" (Gagné, 1964, p.299) in which case, a partial solution suggests some learning has occurred. This dependent variable will be described as accuracy in keeping with Palmiter's (1991) nomenclature.

Even though Gagné is very well respected in the instructional design community, perhaps it would be prudent to consult the cognitive load literature too. As it turns out Cognitive load theorists are in complete agreement with Gagné and have also found these variables important. Sweller et al (1998) describe three major categories of mental effort measurement techniques (subjective techniques, physiological techniques, and task- & performance-based techniques). As they discuss performance based techniques they describe the variables of the current study when they say:

> These techniques use objective task characteristics (e.g., number of elements that need to be considered such as the number of if-then conditions in a propositional reasoning task) and performance levels (e.g., differential learning times, errors) to obtain information on mental effort. (Sweller, van Merriënboer, & Paas, 1998, p. 267).

Sweller’s variable differential learning times and errors, are comparable to Gagné’s dependent variables of “rate of attainment” and “the degree of correctness.” As stated above, in this study these dependent variables will be referred to as performance time and accuracy, and are the object of research questions one and two.

As discussed earlier, both of questions one and two will be analyzed with a MANOVA because there are multiple outcome variables that differ between groups. Detailed measurement procedures for these variables are described in later sections.

Relative Condition Efficiency

The reader may recall from the literature review that relative condition efficiency (RCE) is a construct which was described by Paas and van Merriënboer (1993). Relative condition efficiency combines performance scores and measures mental effort gathered during the test phase (See Table 5 and Equation 9).

$$\textit{Relative condition efficiency} = \frac{|Z_{\textit{Performance}} - Z_{\textit{MentalEffort}}|}{\sqrt{2}} \qquad (9)$$

Table 5

Relative condition efficiency (RCE)

	Learning phase	Test phase
Performance score		X
Mental effort rating		X

Note. Adapted from “A Multidimensional Approach to the Mental Efficiency of Instructional Conditions,” by F. Paas, 2007, retrieved June 5, 2007 from http://www.ou.nl/Docs/Expertise/OTEC/Projecten/onderzoeksvoorstellen%20PDF/Paasproject34%5B1%5D.pdf

The relative condition efficiency formula has recently been revised to remove the absolute value symbols (Paas, Tuovinen, Tabbers, & van Gerven, 2003).

$$\text{Relative condition efficiency} = \frac{Z_{Performance} - Z_{MentalEffort}}{\sqrt{2}} \qquad (10)$$

Table 6 is a worked example of a relative condition efficiency problem. In this example two instructional conditions are being compared (Conditions 1 & 2). To compare conditions, researchers first gather their raw data, in this case test scores. Each individual test score is standardized, that is it is converted to a z-score.

Next, a researcher must consider the group's mental effort ratings (answers to survey questions). Each individual's mental effort rating is then standardized. A group average Z-score is then calculated for each condition. This average Z-score is then used in the relative condition efficiency formula (See Table 6).

The E score, Relative condition efficiency is calculated by adding together the group mean test score to the group mean mental effort score. The E score for that group is subsequently plotted on a graph, with the group mean mental effort score on the X-axis and the group mean performance score on the Y-axis (See Figure 21).

Recall that E is the perpendicular distance from the group mean score, to the E=0 line (note right angles are shown in blue). Finally group Z-scores are compared in an ANOVA.

Table 6

RCE Example

Condition 1	**Test score**	**Z-Score**	**Mental effort**	**Z-Score**
Student A	70	0.36	2.3	-1.25
Student B	80	1.30	3.4	-0.22
Student C	75	0.81	2.9	-0.68
Average Z-score		**0.82**		**- 0.72**
Plot Values: Performance =0.82 and Mental Effort =-0.72				
Condition 2				
Student D	64	-0.18	4.3	0.63
Student E	52	-1.26	3.6	-0.02
Student F	55	-0.99	5.3	1.5
Average Z-score		**- 0.81**		**0.70**
Plot Values: Performance =-0.81 and Mental Effort =0.70				
Grand Mean	66		3.63	
Std dev	11.08		1.06	

Condition 1

$$E = \frac{0.82 - -0.72}{\sqrt{2}} = 1.09$$

Condition 2

$$E = \frac{-0.81 - 0.70}{\sqrt{2}} = -1.07$$

Note: Adapted from *Efficiency in learning: evidence-based guidelines to manage cognitive load,* by R.C. Clark, F. Nguyen, and J. Sweller, 2006a, San Francisco: Pfeiffer. p 335

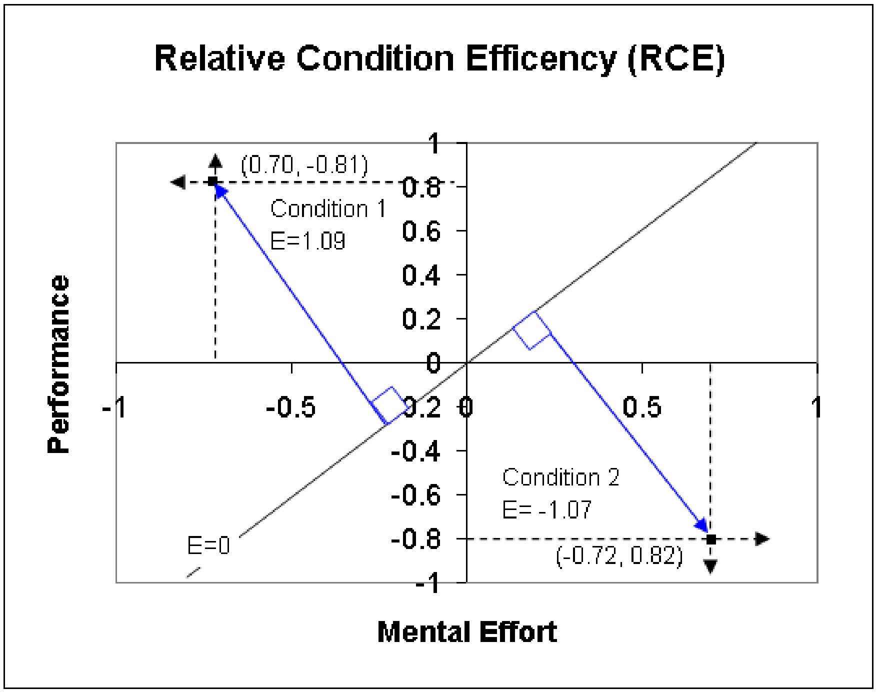


Figure 21. RCE example graph

Note: Adapted from *Efficiency in learning: evidence-based guidelines to manage cognitive load,* by R.C. Clark, F. Nguyen, and J. Sweller, 2006a, San Francisco: Pfeiffer. p 335

Performance Efficiency (PE)

"Performance efficiency" (See Equations 9 & 10) is a slightly modified version of the relative condition efficiency metric shown above, but rather than using a subjective mental effort rating, this measure uses performance time. This altered construct only relies on objective measures. Performance efficiency in this study is:

$$\text{Performance efficiency} = \frac{Z_{Accuracy} - Z_{PerformanceTime}}{\sqrt{2}} \qquad (11)$$

While this dissertation used a single performance problem, and the performance score was the accuracy score, this technique may be generalized to other studies, to use other performance scores (e.g., the total number of problems correct) as in the Paas and van Merriënboer (1993) study. A generalized formula for performance efficiency is:

$$\text{Performance efficiency} = \frac{Z_{Performance} - Z_{PerformanceTime}}{\sqrt{2}} \qquad (12)$$

Detailed procedures for the subcomponents of this construct are described in the data analysis section.

Procedure

This section outlines the data collection procedures. In particular, it discusses preparation of the learning environment, the acquisition and retention phases and concludes with a brief discussion of the software recording procedures.

Preparation of the Learning Environment

A facilitator worked with university technical support, to ensure several programs were installed on the computers used during this study. This study required the following software to be installed or available on each learner's station: a web browser (in this case

Internet Explorer 6.0) (Microsoft, 1995-2004), the Adobe Flash player (version 7.0) (Adobe, 1996-2007), Adobe Photoshop Elements 2.0 (Adobe Systems, 1990-2002), and TechSmith Morae Recorder (Techsmith, 2004).

Prior to data collection, the facilitator prepared the environment. To do so, the facilitator first confirmed that all necessary software had been installed. Next, a folder containing several items was placed on the desktops of all computers. This folder contained the Photoshop document for that session and *.url* files (desktop short-cuts to survey 1 and the day's final survey). Following this, Photoshop was launched and pallet locations were reset. A set of earphones was plugged into each system and the volume levels were checked on all computers. A labeled note card was placed at each station. These note cards were labeled with the date and computer number. Finally, just before learners entered the room the facilitator went to each computer station to start Morae Recorder (the recording software).

Software Recording Procedures

TechSmith Morae, a usability program, was used as the primary tool for data collection (TechSmith, 2004). This usability software is composed of two components, Morae Recorder and Morae Manager.

Morae Recorder acts like a video camera to record a learner's interaction with a computer, and produces a proprietary movie file. This coded movie file is a visual record of a learner's onscreen actions, but, in addition, Morae Recorder encodes a database of all user actions (mouse clicks, keyboard entries, & window events) into the file. This software was installed on lab computers, and turned on before a learners sat down to interact with the computer. Morae was hidden from the learner, making it a non-reactive measure (Campbell, 1957), allowing for the observation of learner behavior, while not being intrusive, or changing the nature of the behavior.

Finally, the data provided by the software recording provided evidence for the research questions, and was gathered during two separate phases (the acquisition phase and retention phase).

Acquisition Phase (Week One)

Once learners enter the environment, it was explained that they were being asked to volunteer to participate in a research study. They were asked to sit at an appropriate computer (one with earphones). Only those stations with the required software had earphones plugged in. Participants were asked to move if they sat at an inappropriate station. Once learners were all seated, they were handed the Institutional research board (IRB) documentation and asked to read and sign it. In addition, learners were asked to print their name on the note card placed at their station.

Once this paperwork was signed, learners were instructed to put their earphones on. These provided learners with an individualized learning experience (free from audio distractions). Participants wore earphones to insure that they did not hear instruction or audio feedback from other computers.

Once the above conditions were met, the acquisition phase began. This began with a scripted introduction. This facilitator explained to the learners that:

- they were taking part in a research study;
- this study was conducted during two sessions (the acquisition and retention phases);

- at some point during the project they would be required to use the computer to work through a problem scenario;
- they could not be helped, that they would have to figure out the problem on their own;
- and finally that all on-screen behavior was being recorded.

Next all learners were told to open a folder on the desktop of their computers and to double-click on the "start" icon, a short cut which led them to survey 1. Once learners answered all questions and submitted the survey, they were forwarded to the introductory overview. A JavaScript randomly assigned each learner to an instructional condition (which may have included an initial assessment, the Mr. Potato head document). After taking part in the instructional condition, all learners concluded activities by completing the post-treatment survey. As learners left the room they were thanked for their participation.

Once all learners had left, the facilitator went to each computer station to save the recordings for later analysis. These recordings were saved according to the computer number and section number [e.g. "001-17.rdg" for section 001 station 17 — .rdg is the Morae, 3-letter file extension]. In addition, it was confirmed that the note cards information at each computer, matched the recording file name. Finally, it was important to ensure all week one files were deleted.

Retention Phase (Week Two)

The retention phase was conducted one week after initial instruction. Additional data concerning the dependent variables were collected during this delayed assessment. The learning environment was prepared in a similar manner, as during the previous week, however, a week two folder was distributed. This folder included the picnic problem and a desktop short-cut to the post-performance survey (week 2 survey). In addition, Adobe Photoshop Elements pallets were reset, and Morae Recorder was also turned on before learners entered the test environment. Finally learners were let into the environment.

During this second meeting with learners, it was reiterated that they were participating in a research project, that they were being recorded, and should complete the picnic problem and post-performance survey before leaving. Once the survey was completed, they were asked to leave and thanked for their participation. Recordings were saved as files, in a similar manner as during the acquisition phase, but kept in a separate folder (labeled week two).

Analysis

This section is structured around the four research questions. Each question is first stated, then introduced in terms of the variables measured, followed by a hypothesis, an expectation, and then finally the analysis procedures are explained. Because of the multivariate nature of questions one and two, these will be discussed together; whereas, questions three and four are discussed separately.

Questions One & Two

Question one: Is there a significant difference among the instructional strategies, relative to performance time?

Question two: Is there a significant difference among the instructional strategies, relative to accuracy?

Measurement of Performance Time.

Earlier in this chapter it was explained that TechSmith Morae (screen capture software) was installed on the learner's computer station, and then allowed to record a learner's interaction, with the computer. This software was also used to analyze this interaction to measure both performance time and accuracy. TechSmith Morae has a second component (Morae Manager) which allows a researcher to analyze the recorded movie files, and document learner interaction days or months later.

Learner interaction was coded, by labeling actions with a series of markers (small flags on the video timeline). The performance time began when the learner first opened the document and was coded with a researcher defined marker (*in point*). The performance ended when a learner completed the greatest number of subtasks required to solve the problem (*out point*). This produces what Morae Manager describes as a *segment*. The duration of a segment is the performance time. Morae displays the duration of these segments in seconds. This duration was logged in an Excel spreadsheet for later analysis.

The *in point* was operationally defined as the point on the timeline when the learner first had the ability to move the cursor (when the cursor changes from an hourglass to an arrow). This position on the timeline was labeled as the "in point" — the beginning of the performance (and performance time).

The end of the performance, the *out point* is a bit more complicated. To find an out point, a researcher must watch the video. Only the time toward correct assembly was counted toward an individual performance time. Therefore, the out point — the end of the performance time, was operationally defined as the point at which the greatest number of pieces within the problem were in proper alignment. In practice, accuracy and performance time were measured simultaneously.

Measurement of Accuracy.

A researcher viewed the recorded video files of each learner's on-screen action, and scored learner interaction using a rubric specifically developed for the problem. Thus two separate accuracy rubrics were developed for use in the study, one for accuracy given the Mr. Potato head problem during the acquisition phase (See Table 7), and a separate rubric for the picnic problem during the retention phase (See Table 8). Both rubrics were based on the problem solving operators required to solve the problem.

Table 7

The Mr. Potato head accuracy rubric

flip	**layer**	**rotate**	**move**	**item**
***	***	***		Right arm
***	***	***		left shoe
***	***			nose
***	***	***		body
***		***		teeth
***		***		hat
***	***	***		left arm
***	***	***		right ear
***	***	***		left ear
	***	***		right shoe
***				moustache
***		***		eyes
0	0	0	0	0

These rubrics are examples of behavior analysis data forms. Behavioral analysis data forms are generally organized into a tabular format (Hinde, 1973; Lehner, 1996). For ease of use Microsoft Excel 2003 spreadsheets (Microsoft, 2003b) were developed for each learner performance, and stored separately as a file.

Three instructional conditions (the demo+practice, demo2+practice, and practice groups) reassembled the Mr. Potato head problem during the acquisition phase. To document their performance, accuracy was measured with the rubric shown in table 7. During week two, the retention phase, all learners reassembled the picnic problem. This performance was also documented with the rubric shown in table 8.

The rubrics used during this study were based on the same underlying point structure. Each learner was granted 1 point for correctly moving a layer, and 1 point for correctly rotating a layer within the scene. In addition, since the main objective of the instructional conditions was for learners to learn how to manipulate Photoshop layers, 2 points were granted for raising or lowering a layer correctly (relative to other layers), and an additional 2 points were granted for flipping a layer horizontally. The cells within each rubric with "***", received no points. Because each rubric was a spreadsheet columns were totaled and them summed with an excel formula to produce a final accuracy score.

To receive credit for an object, it must be visible, in the correct location, correct rotation, and correct layer. However partial credit was given. For example, if the learner had only moved the table to the correct location within the picnic problem, they were given one point for correct piece placement, but they would receive no credit for rotation, unless the table was rotated correctly. If the table was generally in the correct location credit was given. However, researcher judgment was involved and this was not an exact science, as all learners were not held to a strict centimeter by centimeter standard. So for instance, given picnic table placement, learners were given credit if they had placed the table on the left side of the screen in the lower quadrant of the screen. They were also given credit if they rotated the table correctly.

Table 8

Picnic problem accuracy rubric

flip	layer	rotate	move	item	
	***			umbrella	
	***			tshirt	
***	***			head	
***	***	***		right leg	
***	***			head 2	
***	***			purple shirt	
***				hat	
***	***	***		s left leg	
***	***	***		bent right leg	
***	***	***		left leg	
***				green shorts	
***	***	***		arm 2	
***				pink shorts	
***	***	***		left arm	
***	***			body	
***	***			picnic basket	
***	***	***		arm	
***	***			right arm	
***	***			torso	
***	***			table	
***	***			bird3	
***	***	***		bird2	
***	***			bird1	
0	0	0	0		0

In table 8, the learner was given credit for the umbrella if they moved the umbrella in to the correct location and rotated it relative to the picnic table. Finally, learners were given credit if they flip the umbrella so that the pattern was like that of the model.

Many learners continued to interact with the software interface, long after they had “most correctly assembled the scene,” usually in an attempt to complete subtasks that they did not know how to complete. Since a video file was used to document learner interaction, it was possible to detect if a learner disassembled pieces of the scene before the end of the video. Thus, if a learner correctly assembled the scene, and then moved pieces, the point on the timeline, when they had most correctly assembled the problem was deemed to be the end of the performance.

Question One & Two: Hypotheses

$$H_0 = \begin{bmatrix} \mu_{1PT} \\ \mu_{1A} \end{bmatrix} = \begin{bmatrix} \mu_{2PT} \\ \mu_{2A} \end{bmatrix} = \begin{bmatrix} \mu_{3PT} \\ \mu_{3A} \end{bmatrix} = \begin{bmatrix} \mu_{4PT} \\ \mu_{4A} \end{bmatrix} \qquad (13)$$

H_o = There is not a significant difference in performance given the instructional strategy.

$$H_a = \begin{bmatrix} \mu_{1PT} \\ \mu_{1A} \end{bmatrix} \neq \begin{bmatrix} \mu_{2PT} \\ \mu_{2A} \end{bmatrix} \neq \begin{bmatrix} \mu_{3PT} \\ \mu_{3A} \end{bmatrix} \neq \begin{bmatrix} \mu_{4PT} \\ \mu_{4A} \end{bmatrix} \qquad (14)$$

H_a = There is a significant difference in performance given different instructional strategies.

Questions One & Two: Expectation.

Sweller and Cooper (1985) found that learners who studied worked examples took significantly less time to solve problems (performance time) with fewer errors (accuracy). This “worked-example effect” is the main precedence for the current project. While Tarmizi and Sweller (1988) reported that there are some circumstances when worked example-based instruction is not as effective as solving problems, it was assumed that this was not the case given the current project.

Given Sweller and Cooper’s initial findings (Sweller & Cooper, 1985), it was expected that learners who studied animated demonstrations [animated worked examples, according to Lewis (2005)] would take less time to solve problems (performance time) with fewer errors (accuracy), than learners who learned through discovery problem solving.

However, since this dissertation studied two separate performances, the Mr. Potato head problem, during the acquisition phase, and the Picnic problem, during the retention phase, the expectations for these outcomes could differ.

Given Sweller and Cooper's results (Sweller & Cooper, 1985), it was expected that during both the initial assessment and the delayed assessment, that learners in the animated demonstration conditions would out perform their peers in the practice condition. Given the multiple outcome variables involved in this assertion, research questions one and two were answered with a MANOVA (See Equations 13 & 14). The results of all research questions are discussed in detail, in Chapter four.

Question Three

Question 3: Is there a significant difference among the instructional strategies, relative to "relative condition efficiency?"

Relative Condition Efficiency

Paas and van Merriënboer (1993) described the original efficiency metric. "Relative condition efficiency" is a metric for measuring the relative efficiency of instructional conditions. This construct is based upon a combination of performance scores (accuracy in the current study) and mental effort ratings (See Table 9).

Learners produced mental effort ratings by filling out a survey question (a 9-point mental effort rating) following their performance, with the picnic problem. This survey question is identical to the one use in the Paas and van Merriënboer (1993) study. Paas and van Merriënboer used test scores from a statistics test (originally reported in Paas, 1992) with the percentage correct as their "raw score" (p.429).

Table 9

Relative condition efficiency

	Learning phase	Test phase
Performance score		X
Mental effort rating		X

Note. Adapted from "A Multidimensional Approach to the Mental Efficiency of Instructional Conditions," by F. Paas, 2007, retrieved June 5, 2007 from http://www.ou.nl/Docs/Expertise/OTEC/Projecten/onderzoeksvoorstellen%20PDF/Paasproject34%5B1%5D.pdf

Since this dissertation used two performance problems (one in each phase), two measures of relative efficiency were calculated (RCE1 & RCE2). In each case, the raw performance score for that phase was obtained from an accuracy rubric. Mental effort ratings and accuracy scores were standardized, to produce performance and mental effort z-scores for each individual. Once a list of Z-scores was developed, group E scores were computed from the relative efficiency formula (See Equation 15). Next, each group score was graphed. Finally, an ANOVA of the Z-scores was used to determine if they were significantly different.

$$\textit{Relative condition efficiency} = \frac{Z_{Performance} - Z_{MentalEffort}}{\sqrt{2}} \tag{15}$$

Question Three: Hypotheses.

$$\mathrm{H}_o = \mu_{1\mathrm{RCE}} = \mu_{2\mathrm{RCE}} = \mu_{3\mathrm{RCE}} = \mu_{4\mathrm{RCE}} \tag{16}$$

H_o = There is not a significant difference in relative condition efficiency (RCE) given different instructional strategies.

$$H_a = \mu_{1RCE} \neq \mu_{2RCE} \neq \mu_{3RCE} \neq \mu_{4RCE} \tag{17}$$

H_a = There is a significant difference in relative condition efficiency (RCE) given different instructional strategies.

Question Three: Expectation.

Paas and van Merriënboer (1993) studied a similar set of instructional conditions and found that learners, who studied worked examples, significantly out-performed those who solved problems. Given this precedence with relative condition efficiency, it was expected that those learners who studied animated demonstrations would out-perform those who solved problems.

Question Four

Question 4: Is there a significant difference among the instructional strategies, relative to "performance efficiency?"

Performance Efficiency

Performance efficiency is a new construct which was developed during this study. This metric is a variant of the methodology first proposed by Paas and van Merriënboer (1993) in that it uses z-scores, and graphs its results in much the same manner, but it only relies on objective measures. Like relative condition efficiency, one begins by standardizing performance time and performance scores (accuracy in this study). Like relative condition efficiency, performance efficiency scores can then be analyzed with the following formula:

$$\mathit{Performance\ Efficiency} = \frac{Z_{Performance} - Z_{PerformanceTime}}{\sqrt{2}} \tag{18}$$

Next this metric is graphed as in the Paas and van Merriënboer (1993) article. In addition, a one-way ANOVA is used to compare groups, and may be followed by post hoc comparisons to determine significant differences.

Question Four: Hypotheses.

$$H_o = \mu_{1PE} = \mu_{2PE} = \mu_{3PE} = \mu_{4PE} \tag{19}$$

H_o = There is not a significant difference in performance efficiency given the type of instruction.

$$H_a = \mu_{1PE} \neq \mu_{2PE} \neq \mu_{3PE} \neq \mu_{4PE} \tag{20}$$

H_a = There is a significant difference in performance efficiency given different instructional conditions.

Question Four: Expectation.

Because this is the first use of this metric, there is no precedence for this type of study. However, Tuovinen and Paas (2004) calculated a similar metric, their 3D efficiency metric, and found no significant differences between learners who studied worked examples, versus learners who learned through discovery-practice. Given this precedence, it is expected that there will be no significant differences in performance efficiency between the instructional conditions of the present study.

Reliability

Summer and Fall Participants

Because the power analysis suggested a sample size of 115 participants, it was necessary to collect data across two semesters, given the size of the classes. Unfortunately several months passed between semesters and since learners were sampled over the summer and fall semesters and from two different classrooms data from these groups may differ. Thus, it was necessary to see if these potential differences influenced the dataset. A MANOVA was used to see if these groups differed significantly with respect to performance time and accuracy.

Inter-observer Reliability

Observational data has its advantages and disadvantages. While it may be a more direct method of observing behavior, with less conceptual interference from tests or questionnaires, this type of data has its own issues, like coding errors and observer drift (Knupfer & McLellan, 1996; Talpin & Reid, 1973). So this study checked the reliability of the data by using inter-observer reliability estimates.

Cohen (1960) developed a scale for observational agreement, and describes it as:

$$\mathrm{K} = \frac{\pi_0 - \pi_e}{1 - \pi_e} \tag{21}$$

Kappa (K) has two subcomponents, π_0 is the proportion of rater pairs exhibiting agreement, and π_e is the proportion expected to exhibit agreement by chance alone (Cantor, 1996). Given the above one would expect a kappa K=1, if the raters were in perfect agreement. However, this is rarely the case, so agreement must be rated given a range of varying strengths of agreement. Please see table 10, a table describing the strength of agreement given Cohen’s Kappa based on table provided by Landis and Koch (1977).

These estimates were made for a randomly selected group of participant data files (*n*=20). Finally, inter-observer reliability estimates were only conducted on measures of performance time and accuracy given the delayed assessment in the retention phase.

Table 10

Strength of agreement

Kappa Statistic	Strength of agreement
< 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Note. Adapted from "The measurement of observer agreement for categorical data." by J.R. Landis, and G.G. Koch, 1977, *Biometrics*, 33 (1), 159-174.

Pilot Studies

Fall 2006

An initial pilot study was conducted in December 2006. The purpose of this study was to confirm that the data collection procedures were feasible. To accomplish this goal, TechSmith Morae Recorder (the recording software) was installed on a set of lab computers. Then during a single class meeting, recordings of learner interaction were performed while learners interacted with a nascent form of the Mr. Potato head document.

Spring 2007

Another pilot study was conducted during the spring of 2007. This pilot study was conducted over two consecutive weeks to test all procedures and revised versions of the instructional materials. The purpose of this pilot was to test the follow-up procedures over a two week period.

During the spring pilot, 3x5 note cards were numbered and placed at each computer station. Once learners entered the test environment, they were asked to write their name on the note card to indicate their presence during week 1. During week 2, these same cards were placed at the same computer stations, in order to track learners between sessions. It was found that note cards may be used to describe the context of each computer station, regardless of learner presence. So, for instance, if a learner was not present the following week, this was noted.

CHAPTER FOUR - RESULTS

The purpose of this chapter is to present the results of the dissertation. The chapter considers data preparation, preparatory data analysis, reliability analysis, results of the research questions, and then concludes with the limitations of these results.

Data Preparation

Data were collected with a series of web-based surveys and recordings made with TechSmith Morae 1.01 (TechSmith, 2004). The data were logged within a set of Excel spreadsheets (Microsoft, 2003b). Some minor calculations were made within these spreadsheets, but for the most part, statistical calculations were made with SAS 9.1.3 Service Pack 2 for Windows (SAS, 2002-2003). In addition, data were analyzed and represented graphically using bivariate plots prepared with a series of SAS macro programs (%MULTNORM) (SAS, 2007b), (ELLIPSES) (Friendly, 2007b), (OUTLIER) (Friendly, 2007c), (CQPLOT) (Friendly, 2007d).

Sample Preparation

As with any study, it was necessary to process the data before statistical analysis could be conducted. Appendix B describes the decision rules for sample preparation, but a summary is provided here. In short, even though a total of 215 students participated in this study, not all learners followed the instructions or completed both phases of the study. Participants were removed from the dataset because of various reasons: 25 did not return the second week and were lost due to attrition, 25 learners were removed because they did not complete the surveys required by the study or did not follow instructions, 13 used Adobe Photoshop Elements (or Adobe Photoshop) between sessions, 13 learners had technical difficulties, 9 were removed because they were second language learners (discussed in Chapter three), and finally 8 learners were removed because they used the Adobe Photoshop help system or "surfed" the web looking for help. This provided a sample size of n=122. As described in Chapter two, this sample size is sufficient to arrive at a power of 0.80, with a small effect size η^2= 0.125, given α =0.05.

Sample

A sample size of n=122 learners followed the instructions, completed all surveys, and attempted the required performances. Table 11 describes this sample according to the demographic variables gathered with survey 1.

Table 11

Sample by instructional condition

Group	*n*	Gender		Level				Age
		Female	Male	Fresh	Soph	Junior	Senior	*M*
demo	33	23	10	5	14	11	3	19.97
demo+practice	29	23	6	4	12	12	1	20.48
demo2+practice	36	29	7	1	19	15	1	21.78
practice	24	18	6	4	15	4	1	19.75
total	122	93	29	14	60	42	6	

Preparatory Data Analysis

Fidell and Tabachnick (2003) describe preparatory data analysis as being "conducted before a main analysis to assess the fit between the data and the assumptions of that main analysis" (Fidell & Tabachnick, 2003, p. 115). As an important first step to any analysis, this section of the chapter assesses the fit of the dataset to the assumptions of a multivariate analysis of variance (MANOVA). Specifically, it considers the fit of a pooled data set solution to the assumptions of a MANOVA.

A Pooled-semester Solution for Data Analysis

An *a priori* power analysis suggested a sample size of *n* = 115 participants, in order to detect a small effect size. A sample size of this magnitude required data to be collected across two semesters (the summer and fall semesters of 2007). Therefore it was important to question if this pooled dataset would affect statistical tests. To answer this question, an analysis was conducted to determine if a pooled-semester dataset was a viable solution for analysis.

The reader may recall that the demo group did not assemble the week one problem (the Mr. Potato head problem). Therefore the week two performance (the picnic problem) was chosen to compare semester subgroups, because it was the only performance in which all participants were involved. Thus a MANOVA of dependent variables, week two performance time (PT2) and week two accuracy (AC2), was used to compare semester subgroups.

A MANOVA makes several assumptions (assumptions of independence, normality and homoscedasticity) (Stevens, 2002; Tabachnick & Fidell, 2001). Stevens (2002), provides a general procedure for assessing each of these assumptions. The assumptions for this MANOVA are discussed in detail in Appendix B. In short, the independence assumption was met (Glass & Hopkins, 1984). The normality assumption was violated, since the %MULTNORM macro program revealed non-normality. This violation was primarily due to a series of multivariate outliers, so potential outliers were removed and transformations were implemented to test the "homoscedasticity" assumption. Box's M test (Box, 1954) was performed and X^2 (3, N=88) =4.50, p=0.21, φ=0.23, therefore the variance-covariance matrices were not found to be significantly different, so there was no evidence that the homoscedasticity assumption was violated. Therefore a MANOVA was conducted.

A MANOVA was used to compare the two semester subgroups (the summer and fall subgroups). The MANOVA indicated that there was not a significant difference between the two semester subsets, since Wilks' Λ =0.95, F (2, 95) = 2.47, p = 0.09, η^2=0.05.

Given the MANOVA did not find significant differences between the two semester subgroups, the use of a pooled data set was found to be a viable solution for analysis. For a detailed account of this analysis, consider Appendix C, Table 12, and Figures 22.

Table 12

Comparison of summer and fall semesters

	Summer semester	Fall semester
n	28	60
Transformed accuracy (TAC2)		
M	6.56	6.53
SD	0.24	0.23
Transformed performance time (TPT2)		
M	31.54	33.54
SD	3.56	4.52

Reliability Analysis

An analysis of inter-observer agreement was performed to assess the consistency of the researcher's assessments. A single researcher analyzed the data for this study. A later analysis by the same researcher was used to judge the consistency of assessments. Performance time and accuracy measurements were compared given 20 learner data files from the week two performance. The 20 learner data files were chosen at random using a random number generator in Microsoft Excel (Microsoft, 2003b). Cohen's *κ* was used to compare inter-observer agreement, and resulted in accuracy (AC2), *κ* = 0.29 (fair agreement) and performance time (PT2), *κ* = 0.47 (moderate agreement).

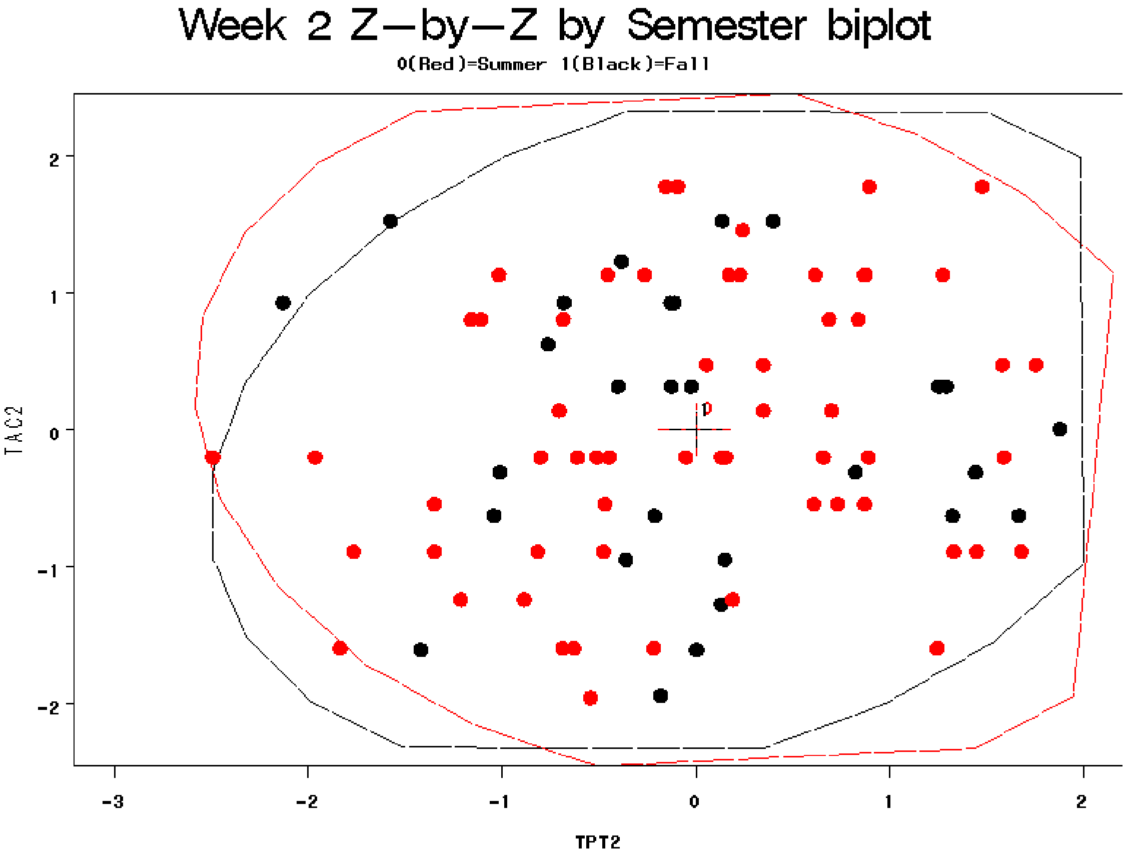


Figure 22. Week two Z-by-Z semester comparison

Results

Next the chapter turns its attention to the results of the research questions. The overall structure of this section is based upon these questions, but recall that there were two phases of the overall experiment during which there were two performances (Note Table 13). Table 13 explains the terminology and time table for this chapter.

Table 13

Research question by phase matrix

Phase	Question 1	Question 2	Question 3	Question 4
Acquisition Phase (Week one)	Acquisition Phase MANOVA		Relative Condition Efficiency (RCE1)	Performance Efficiency (PE1)
Retention Phase (Week two)	Retention Phase MANOVA		Relative Condition Efficiency (RCE2)	Performance Efficiency (PE2)

Questions one and two were evaluated with two separate MANOVAs, one for each phase or week, of the experiment. Thus the week one analysis became the acquisition phase MANOVA and week two the retention phase MANOVA. In addition, questions three and four were also analyzed over two weeks, so relative condition efficiency was described as RCE1 and RCE2. The same naming convention was used for performance efficiency (PE1 & PE2).

Questions One & Two

Questions one and two considered multiple outcome variables (performance time and accuracy) so they were analyzed as a MANOVA, therefore the results of these two questions are discussed together.

The Acquisition Phase MANOVA

Research questions one and two investigated group differences given the two dependent variables, performance time and accuracy, for the four instructional conditions (demo, demo+practice, demo2+practice, and practice). The purpose of the acquisition phase (week one) was to introduce all learners to the subject matter, but recall that the design of this experiment required the demo group (*n*=23) to refrain from practicing during this phase, so they did not assemble the Mr. Potato head problem during week one. Therefore only three groups of learners (demo+practice, demo2+practice, and practice) had a performance during week one (See Figure 23).

In addition, during preparatory data analysis, a series of individuals had to be removed from the data set because these observations were potential multivariate outliers. After these outliers were removed from the initial sample of *N*=122 participants, the total number of practicing learners in the acquisition phase was reduced (*n* = 69). This number represents both the outliers removed from the overall data set and a loss of the demo group learners, who did not practice during the acquisition phase. Thus the group composition of practicing learners in the acquisition phase was demo+practice group (*n* =21), demo2+practice group (*n* = 31), and practice group (*n* = 17) (See Figure 23).

Assumptions of the MANOVA

A MANOVA makes several assumptions (assumptions of independence, normality and homoscedasticity) (Stevens, 2002; Tabachnick & Fidell, 2001). This analysis is discussed in detail in Appendix C; however a brief presentation of this analysis is described in this section of the chapter.

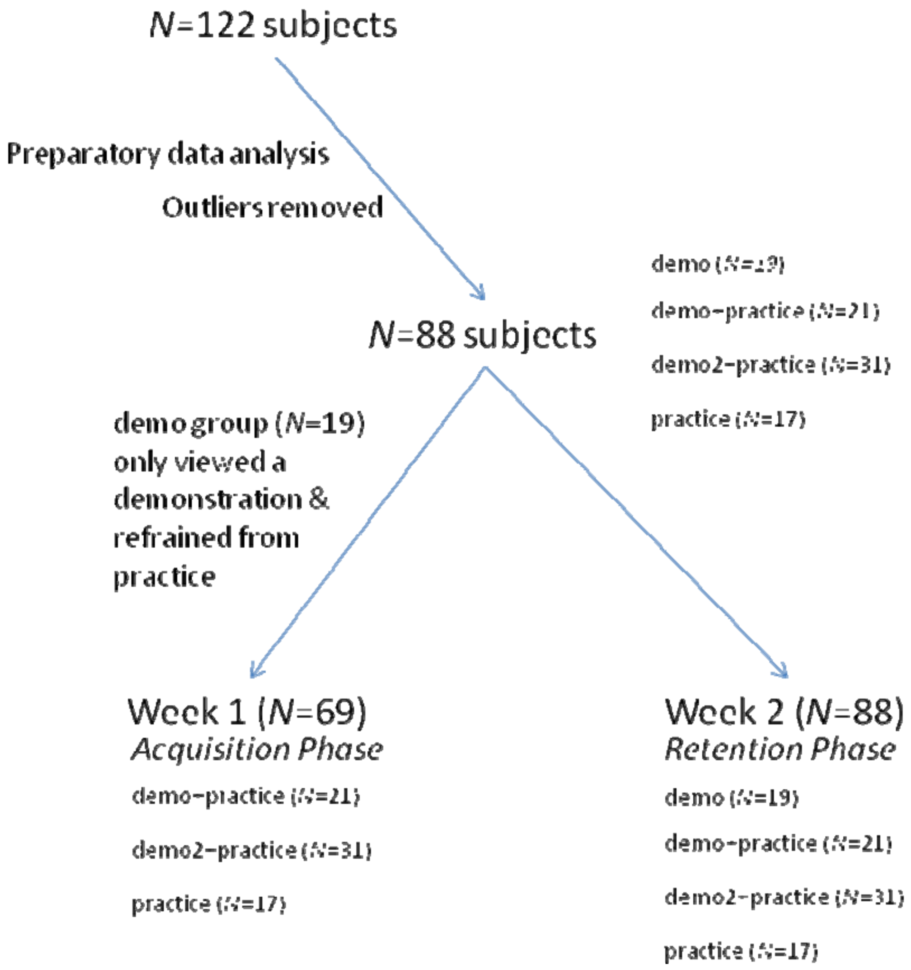


Figure 23. Flowchart of the reduction process

According to Glass and Hopkins (1984) this data met the independence assumption. However, the %MULTNORM macro revealed that the data was non-normal, violating the normality assumption. Analysis of the data set with the OUTLIER macro (Friendly, 2007b) revealed multivariate outliers, but these outliers were retained in order to maintain power. Tabachnick and Fidell (2001) recommend that researchers who retain outliers transform their data, therefore transformations were performed. Next Box's M test (Box, 1954) was performed, and since $X^2(6, N = 69) = 7.97$, p=0.24, φ=0.34 the groups were found to be homogeneous, suggesting there was no evidence the homoscedasticity assumption was violated. Thus even though multivariate outliers were retained, it was reasonable to proceed with the MANOVA.

The Acquisition Phase MANOVA

The acquisition phase MANOVA found there was a significant difference between groups, because Wilks' Λ = 0.68, *F* (2, 68) =6.83, *p* <0.0001, η^2=0.32 (See Figures 24 & 25). The F tests for performance time and accuracy were also statistically significant, as *F* (2, 68) = 3.19, *p*=0.0478 for accuracy (AC1), and *F* (2, 68) =7.84 *p*=0.0009 for performance time (PT1).

Table 14 details the acquisition phase dependent variables, by group. Post hoc comparisons with Scheffé's test (*p*<0.025) revealed that learners of both the demo+practice and demo2+practice groups assembled the Mr. Potato head problem, in significantly less time than the practice group. However, no significant differences between groups were found given accuracy (AC1) with Scheffé's test (*p*<0.025).

Table 14

Acquisition phase dependent variables by group

	demo	demo+practice	demo2+practice	practice
n		21	31	17
Transformed Performance time (TPT1)				
M	NA	19.66	22.40	28.62
SD	NA	6.35	6.28	9.01
Transformed Accuracy (TAC1)				
M	NA	0.56	0.99	1.44
SD	NA	0.79	1.99	1.13

Figure 26 is a graphic representation of this dataset. Group colors are demo+practice=red, demo2+practice=green, practice=black. This bivariate plot of the acquisition phase dataset includes transformed performance time and accuracy scores, and is shown by group. Since accuracy was transformed with a TAC1=log (25-AC1) transformation, the most accurate performances are at the bottom of the graph. This also applies to Figures 24 through 25.

In addition, given Figure 26 there seems to be a ceiling effect. Notice how the transformed accuracy scores are all near the bottom of the graph. The consequences of this ceiling effect are fully described in Chapter five.

Finally, Tabachnick and Fidell (2001) mention two other important aspects of a MANOVA, the effect size, and correlation between the dependent variables. The effects size for this MANOVA was η^2=0.32, therefore this combination of variables accounts for 32%, a reasonable proportion of the total variance (Tabachnick & Fidell, 2001).

Finally these performance time and accuracy were negatively correlated since *r* (67) =-0.13, p = 0.29. This was expected since performance time increases as accuracy decreases.

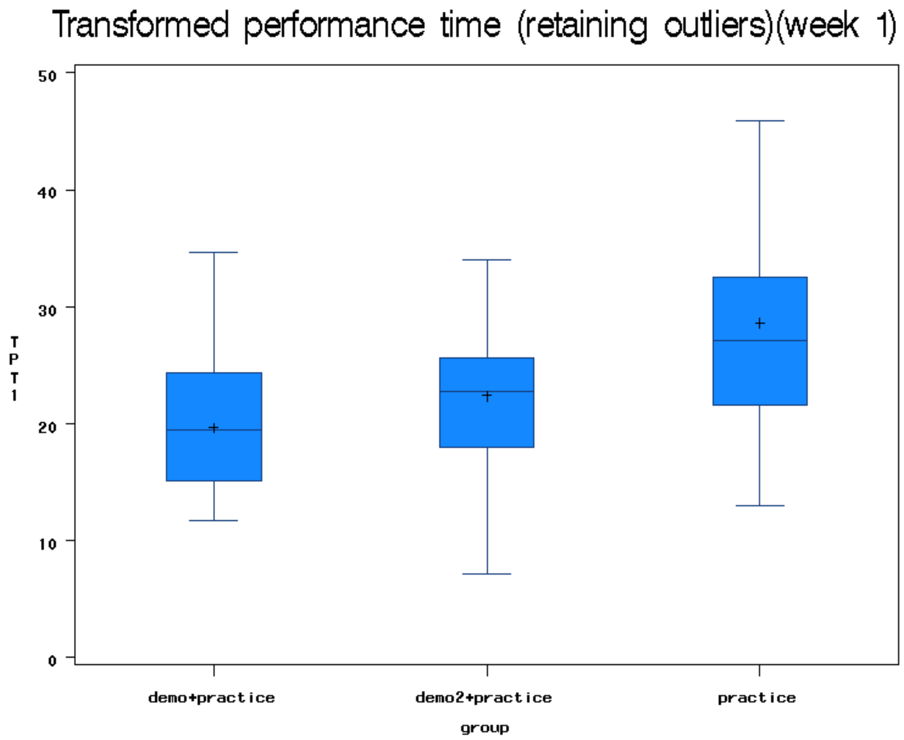


Figure 24. Acquisition phase transformed performance time (retaining outliers)

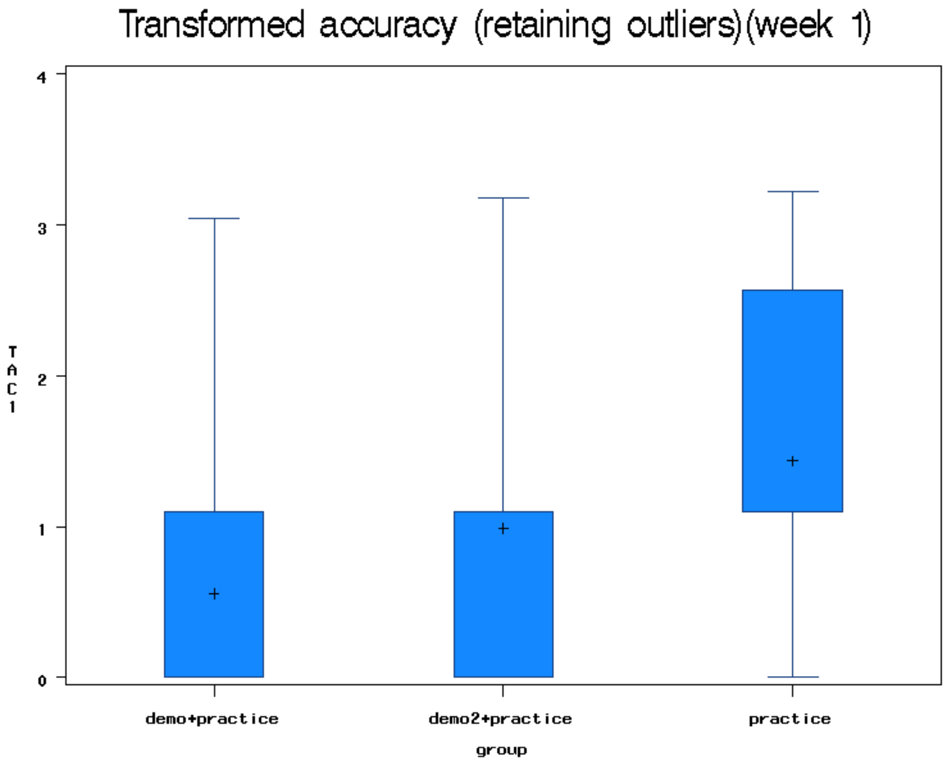


Figure 25. Acquisition phase transformed accuracy (retaining outliers)

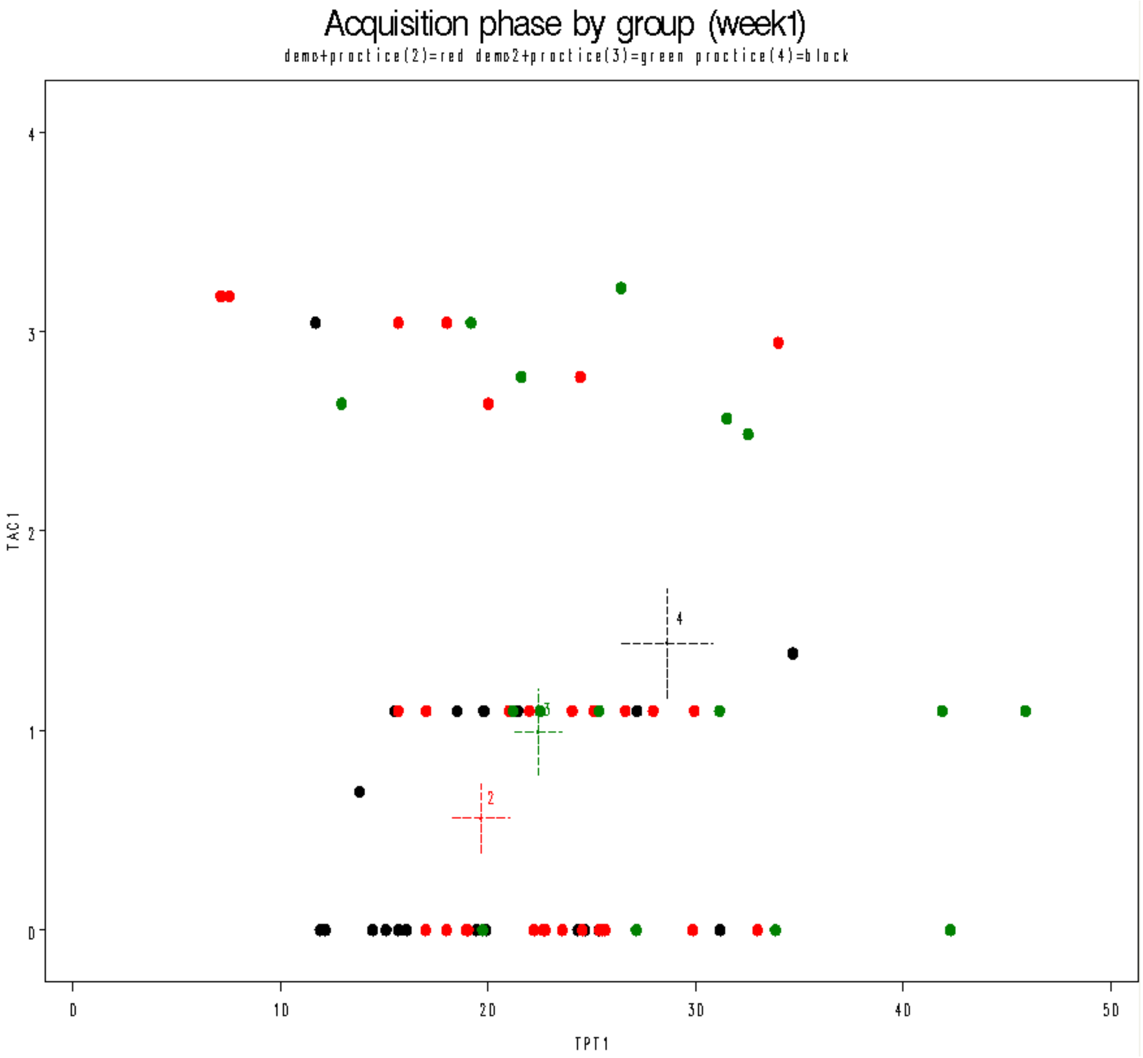


Figure 26. Solution two: retaining potential outliers

The Retention Phase MANOVA

The reader may recall that the week two performance was described as the retention phase. This week two performance was analyzed with a MANOVA to determine group differences a week after initial instruction.

Earlier in this chapter during the preparatory data analysis section, a MANOVA of the week two dataset was considered, but this MANOVA compared the performance of the two semester subsets (used semester as the grouping variable). On the other hand, the purpose of the retention phase MANOVA was to analyze group differences one week after initial instruction. Therefore further discussion of this dataset must be considered (with group as the grouping variable), because research questions one and two require an analysis of this dataset to determine the differences in group performance given the four different instructional conditions.

Unlike learners in the acquisition phase, all groups of learners in the retention phase assemble the problem scenario (the picnic problem). Also recall that during preparatory data analysis 34 multivariate outliers were removed from the initial sample. Therefore this same group composition must be retained, so the group composition of the retention phase was demo ($n = 19$), demo+practice ($n = 21$), demo2+practice ($n = 31$), and practice ($n = 17$), for an overall $n = 88$.

The Retention Phase Assumptions

As with all forms of analysis in this chapter, the assumptions of the test were analyzed first. A detailed analysis of the retention phase MANOVA is described in Appendix D.

According to Glass and Hopkins (1984) learners in this sample met the independence assumption, but the %MULTNORM macro program (SAS, 2007b) revealed non-normality. These outliers were removed and transformations were implemented. Later, Box's M test was conducted and it found the variance-covariance matrices were homogeneous, since $X^2(9, N = 88) = 4.43$, p=0.88, φ=0.22. This finding showed that there was no evidence that the transformed dataset violated the homoscedasticity assumption, thus it was reasonable to consider a retention phase MANOVA.

The Retention Phase MANOVA

The overall goal of the retention phase MANOVA was to determine if group differences existed a week after initial instruction. It was hypothesized that learners in the demonstration conditions would out-perform those in the practice condition. However, the results of the MANOVA found that there was not a significant difference given learner performance one week after initial instruction, since Wilks' Λ =0.96, F (3, 87) =0.64, p =0.70, η^2=0.04. Table 15 lists the group means for each of the dependent variables transformed performance time (TPT2) and transformed accuracy (TAC2).

Tabachnick and Fidell (2001) suggest researchers consider two other important aspects of a MANOVA, the effect size and correlation between the dependent variables. The effects size for this MANOVA was η^2=0.04, therefore this combination of variables accounts for only 4% of the total variance (Tabachnick & Fidell, 2001). As for the correlation between the dependent variables, r (120) =-0.14, p = 0.12. Finally, Tabachnick and Fidell (2001) explain that it is better to have uncorrelated dependent variables, because this way, they measure separate aspects of the independent variables.

Table 15

Transformed performance time (TPT2) and accuracy (TAC2) by group

	demo	demo+practice	demo2+practice	practice
n	19	21	31	17
Transformed performance time (TPT2)				
M	34.10	31.92	33.29	32.09
SD	3.78	4.93	4.57	3.44
Transformed accuracy (TAC2)				
M	6.55	6.55	6.54	6.50
SD	0.26	0.25	0.22	0.21

Research Question Three

Research question three was concerned with relative condition efficiency (RCE) (See Equation 22) (Paas and van Merriënboer, 1993). Relative condition efficiency is a construct, a combination of observable measurements (Kerlinger, 1986) in this case a performance measure and a mental effort rating.

$$\textit{Relative condition efficiency} = \frac{Z_{Performance} - Z_{MentalEffort}}{\sqrt{2}} \qquad (22)$$

Week One Relative Condition Efficiency (RCE1)

The reader may recall that Paas and van Merriënboer (1994) found significant results when they compared groups of learners who studied either high or low variability worked examples, versus those who solved high or low variability problems. Specifically, they found that those learners who studied varied context examples invested less time and mental effort during practice (the variability effect). This study examines this effect using animated demonstrations.

To consider Relative condition efficiency (RCE) accuracy scores (AC1) were measured with a rubric (See Table 7). In addition acquisition phase mental effort ratings (AME) were measured following the construction of the Mr. Potato head problem. Mental effort was measured with the first question on the post treatment survey (week 1 survey 2): "I invested:" with nine possible responses, from "very, very low mental effort" to "very, very high mental effort." This is the question Paas and van Merriënboer (1993) used in their study. Given Paas and van Merriënboer results, it was hypothesized that learners in the animated demonstration conditions (demo+practice and demo2+practice) would out-perform learners in the practice condition.

During the acquisition phase, standardized week one accuracy scores (AC1) were combined with standardized acquisition phase mental effort ratings (AME) to provide week one relative condition efficiency (RCE1) (See Equation 23). In addition, relative condition efficiency was analyzed for the retention phase. This provided week two relative condition efficiency (RCE2), a combination of standardized retention phase mental effort ratings (RME) and standardized week two accuracy scores (AC2) (See Equation 24).:

$$RCE1 = \frac{Z_{AC1} - Z_{AME}}{\sqrt{2}} \tag{23}$$

$$RCE2 = \frac{Z_{AC2} - Z_{RME}}{\sqrt{2}} \tag{24}$$

The general procedure for analyzing relative condition efficiency (Paas & van Merriënboer, 1993) was used to compare group scores. Group relative condition efficiency scores z-scores were compared with an ANOVA. The performances of three groups were compared, group composition was demo+practice group (*n*=21), demo2+practice group (*n*=31), and practice group (*n*=17). The assumptions of this ANOVA were analyzed. An analysis of these assumptions is presented in Appendix E.

According to Glass and Hopkins (1984) learners in this data set met the independence assumption, but a Kolmogorov-Smirnov test revealed non-normality. Transformations were implemented. Later a Levene's test compared the transformed means to find they were not significantly different, *F* (2, 68) =2.26, *p*=0.11. This finding showed that there was no evidence that the transformed dataset violated the homoscedasticity assumption, thus it was reasonable to consider an ANOVA. The

ANOVA was conducted and it revealed that there were significant differences between groups since *F* (2, 68) = 3.69, *p*=0.03 (See Figure 27 & Table 16). Even though these groups were significantly different, post hoc comparisons with Scheffé's test (*p*<0.05) found no significant differences between groups. Table 16 lists group means for RCE1.

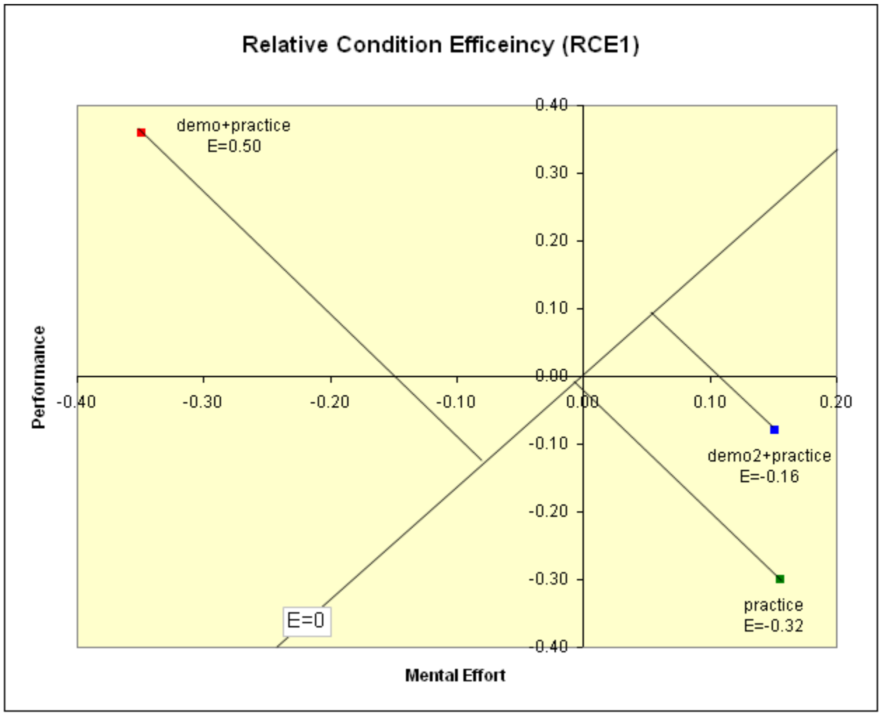


Figure 27. Week one relative condition efficiency (RCE1)

Table 16

Week one relative condition efficiency (RCE1) by group

	demo	demo+practice	demo2+practice	practice
n	NA	21	31	17
Relative condition efficiency (RCE1)				
M	NA	0.50	-0.16	-0.32
SD	NA	0.63	1.11	1.25

Week Two Relative Condition Efficiency (RCE2)

During week 2 (the retention phase), relative condition efficiency was analyzed by combining accuracy scores (AC2) with retention phase mental effort ratings (RME) (See Equation 23). Accuracy scores (AC2) were measured with a rubric (See Table 8). This rubric measured the learner's performance given the picnic problem. Retention mental effort ratings (RME) were also measured following the week two performance. Group relative condition efficiency scores z-scores for the retention phase (RCE2) were compared with an ANOVA. A detail analysis of these assumptions of this ANOVA is presented in Appendix E.

According to Glass and Hopkins (1984) learners in this data set met the independence assumption, and a Kolmogorov-Smirnov test revealed a normal distribution. Later a Levene's test compared the means to find they were not significantly different, F (3, 87) =0.56, p=0.64. This finding showed that there was no evidence that the data set violated the homoscedasticity assumption, thus it was reasonable to consider an ANOVA.

It was hypothesized that learners in the animated demonstration conditions would out-perform learners in the practice condition. Group z-scores were tested with an ANOVA and revealed that there were no significant differences between groups as F (3, 87) = 0.38, p=0.77 (See Figure 28 & Table 17). Relative condition efficiency is a combination of week two accuracy (AC2) and retention mental effort. Tables 18 and 19 provide the data for these two components of relative condition efficiency.

Table 17

Week two relative condition efficiency (RCE2) by group

	demo	demo+practice	demo2+practice	practice
n	19	21	31	17
RCE2				
M	-0.14	0.17	0.00	-0.05
SD	0.81	0.95	0.92	1.10

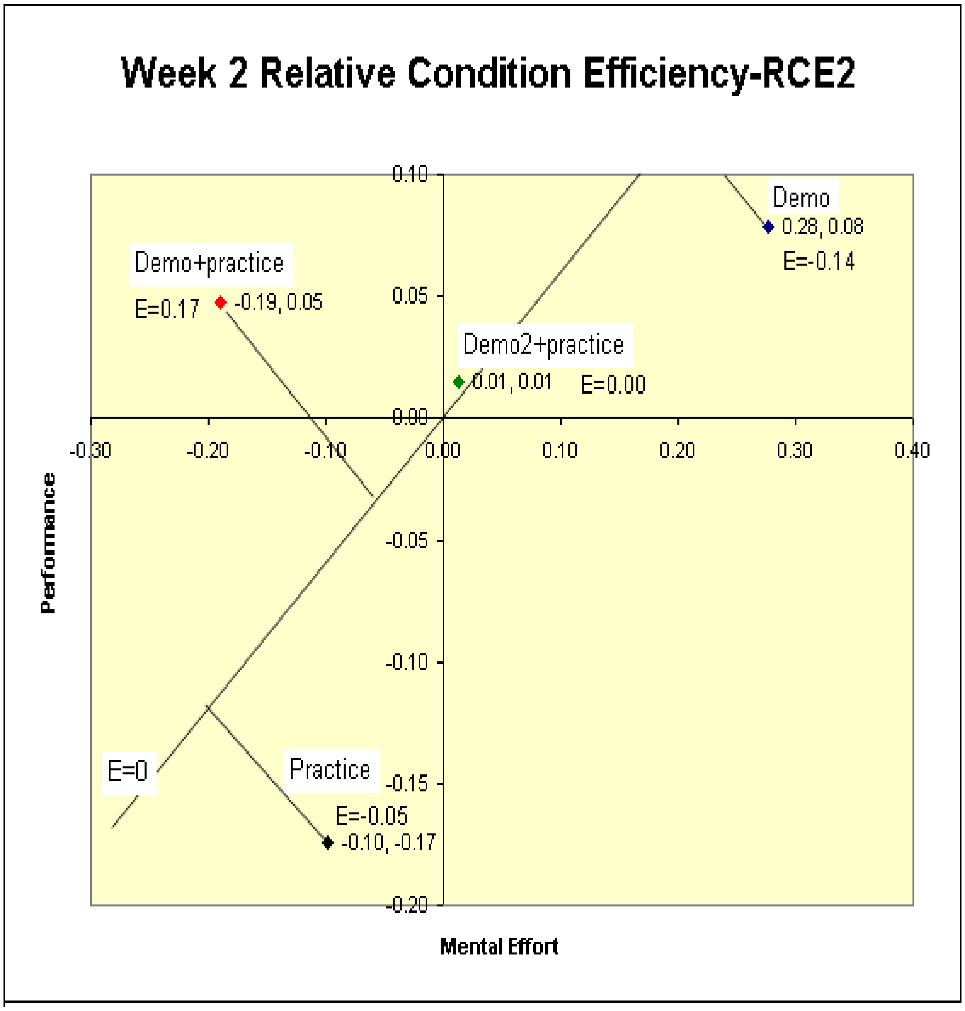


Figure 28. Week two relative condition efficiency (RCE2)

Table 18

Retention accuracy (AC2) by group

	demo	demo+practice	demo2+practice	practice
n	19	21	31	17
Performance time (PT2)				
M	0.08	0.05	0.02	-0.17
SD	1.14	1.09	0.93	0.92

Table 19.

Retention mental effort by group

	demo	demo+practice	demo2+practice	practice
n	19	21	31	17
Retention mental effort (RME)				
M	0.28	-0.19	0.01	-0.10
SD	0.84	1.10	1.01	1.03

Finally, an ANOVA was conducted to compare retention mental effort ratings (RME). The results of this ANOVA indicated that there were no significant differences between groups as F (3, 87) = 0.38, p=0.77.

Research Question Four

This question dealt with performance efficiency, a new metric developed in this study. As described in the literature review, there has been some discussion in the cognitive load literature concerning the objective/subjective nature of cognitive load measurements (Brünken, Plass, & Leutner, 2003). Performance efficiency was developed to help researchers objectively compare and contrast their measurements. For this reason, performance efficiency (See Equation 25) only includes objective measures, a performance score and performance time.

The reader may recall that Chapter three included a discussion of problem solving dependent variables. Gagné (1964) reported that researchers are primarily concerned with performance time and the degree of correctness (accuracy). Gagné’s proposal was the impetus for the subcomponents of performance efficiency.

Since these two dependent variables are the most commonly gathered problem solving variables, it made sense to develop a metric based upon the needs of these researchers.

$$Performance\ efficiency = \frac{Z_{Performanc\,e} - Z_{Performanc\,eTime}}{\sqrt{2}} \qquad (25)$$

The process of calculating performance efficiency is very similar to that of relative condition efficiency. However, the subcomponents of performance efficiency are somewhat different, because it only includes an objective performance measure (accuracy in the current study) and performance time. Equation 25 is a generalized formula for performance efficiency which may be used in any study.

Acquisition Phase Performance Efficiency (PE1)

Week one performance efficiency (PE1) was calculated by standardizing performance scores, in this case, week one accuracy (AC1) and performance time (PT1). Group z-scores were then analyzed with the formula in Equation 26, and graphed as in Figure 29. Next an ANOVA is used to compare group performance efficiency scores. This may be followed by post hoc comparisons to determine significant differences.

$$PE1 = \frac{Z_{AC1} - Z_{PT1}}{\sqrt{2}} \qquad (26)$$

Before conducting the ANOVA, the assumptions of that ANAOVA were analyzed. A detailed analysis of these assumptions is presented in Appendix F. However a brief explanation is made here. According to Glass and Hopkins (1984) learners in this data set met the independence assumption, and a Kolmogorov-Smirnov test revealed non-normality. The distribution was subsequently transformed. Later, a Levene's test compared the means to find they were not significantly different, *F* (3, 87) =0.03, *p*=0.97. This finding showed that there was no evidence that the data set violated the homoscedasticity assumption, thus it was reasonable to consider an ANOVA.

Since no precedence for this metric exists, it was hypothesized that a finding of no significant difference would be found for these conditions. This expectation was not found to be the case, since the ANOVA for performance efficiency (PE1) revealed significant differences among group means, as *F* (2, 68) = 13.95, *p*<0.0001 (See Table 20 and Figure 29).

Post hoc comparisons with Scheffé's test (*p*<0.05), revealed significant differences between all groups, and the demonstration groups (demo+practice and demo2+practice) had more efficient performances, than the practice group. The demonstration groups were not found to be significantly different from one another.

Table 20

Week one performance efficiency (PE2) by group

	demo	demo+practice	demo2+practice	practice
n	NA	21	31	17
Performance efficiency (PE1)				
M	NA	0.55	0.04	-0.75
SD	NA	0.61	0.77	0.99

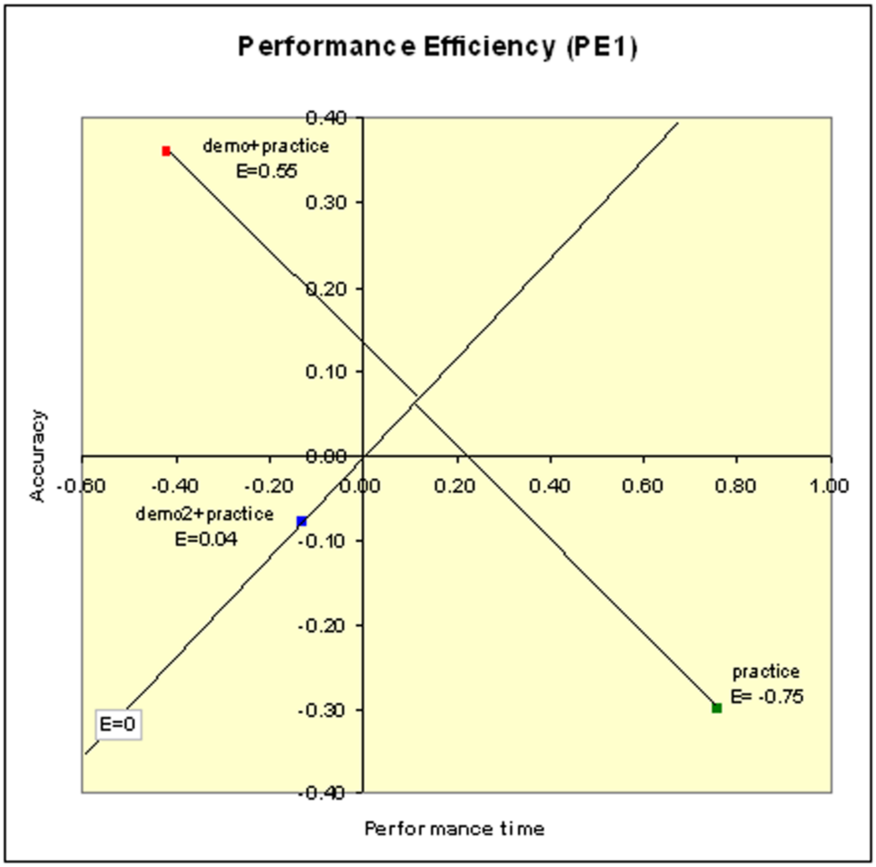


Figure 29. Week one performance efficiency (PE1)

Retention Phase Performance Efficiency (PE2)

Week two performance efficiency (PE2) was also calculated in a similar manner as the acquisition phase. Week two accuracy scores (AC2) and performance times (PT2) were standardized (See Equation 27), and graphed (See Figure 30). As in the acquisition phase an ANOVA of the group means was performed.

$$PE2 = \frac{Z_{AC2} - Z_{PT2}}{\sqrt{2}} \tag{27}$$

Before conducting the ANOVA, the assumptions of that ANAOVA were analyzed. A detailed analysis of these assumptions is presented in Appendix F. However a brief explanation is made here. According to Glass and Hopkins (1984) learners in this data set met the independence assumption, and a Kolmogorov-Smirnov test found a normal distribution. A Levene's test compared the means to analyze the variance-covariance matrices. They were not significantly different, as *F*=0.56 (3, 87), *p*=0.64. This finding showed that there was no evidence that the data set violated the homoscedasticity assumption, thus it was reasonable to consider an ANOVA.

Since no precedence for this metric exists, it was hypothesized that group means would not differ a week after initial instruction. This expectation was found to be the case, since an ANOVA for performance efficiency (PE2) revealed no significant differences in group means, because *F* (3, 87) = 0.42, *p*=0.74 (See Table 21 & Figure 30).

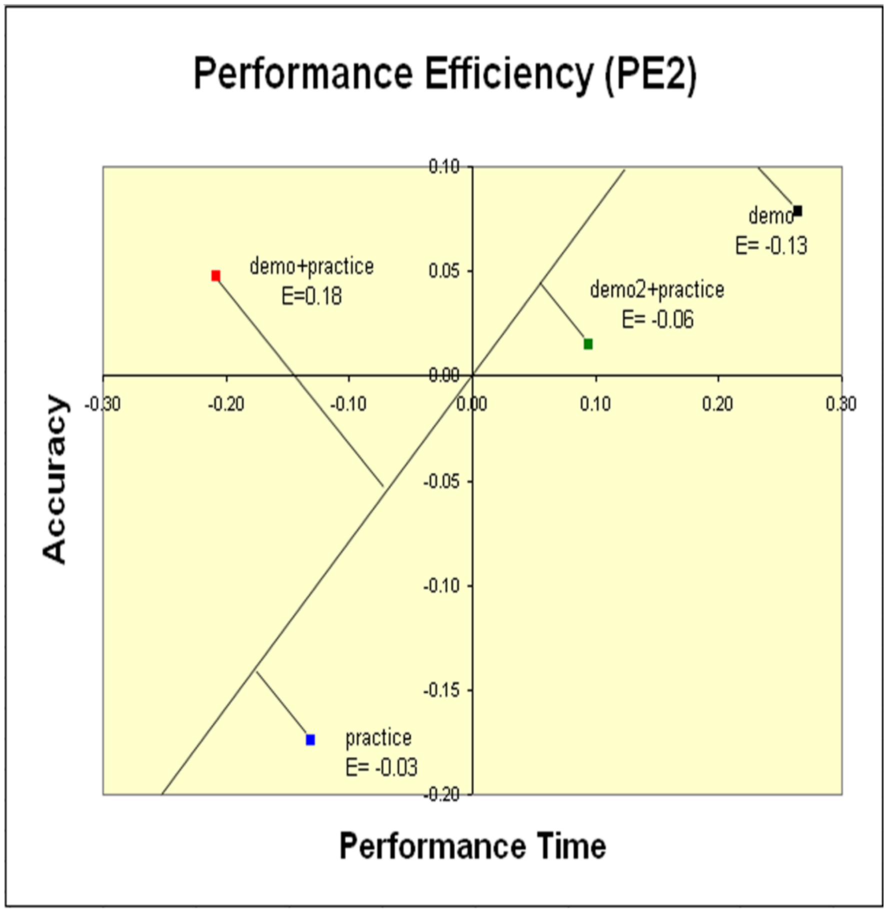


Figure 30. Retention phase performance efficiency (PE2)

Table 21

Week two performance efficiency (PE2) by group

	demo	demo+practice	demo2+practice	practice
n	19	21	31	17
Performance efficiency (PE2)				
M	-0.13	0.18	-0.06	0.03
SD	0.95	1.05	0.97	0.76

Limitations of these Results

The results of any study are limited by various types of error. In some cases error is unavoidable and the general linear model (Y=μ + α + ε), even assumes that there will be some error (ε) within any ANOVA or MANOVA (Keppel, 1991; Stevens, 2002). However, it is the responsibility of the researcher to minimize error. Mitchell and Jolley (2004) have proposed three sources of error (participant errors, observer errors, and administration errors). In addition to these sources of errors, there are two main types of measurement error, systematic and random errors (Mitchell & Jolley, 2004; Pedhazur & Schmelkin, 1991). This section discusses these errors in relation to this dissertation.

Observer & Administration Errors

A systematic observer error is one in which the observer repeatedly makes errors, because of instrumentation or bias (Pedhazur & Schmelkin, 1991). One important criticism of the current study is that the primary researcher served as the only rater. This situation allows for a type of systematic error called observer bias. However, Mitchell and Jolley (2004) explain that observer bias may be avoided, if the researcher is “blind to the conditions” that is, they are unaware of which condition that they are rating. This was the case in this dissertation, since files were blindly rated and then later categorized.

While multiple raters may have strengthened the results of this study, this did not occur. A potential solution for this issue is for future researchers to replicate the study with multiple raters.

It is important to realize that although systematic errors may provide consistent results, these results may be consistently incorrect, and thus systematic errors reduce the validity of the measurement (Pedhazur & Schmelkin, 1991) therefore it is important to reduce systematic error.

One way researchers may reduce systematic observer errors is to refine their instrumentation (Pedhazur & Schmelkin, 1991). For instance, the rubrics used to produce the accuracy variable in this study could be improved. Future researchers could refine the scoring of accuracy to perform a GOMS level of analysis. GOMS is an acronym (Goals, Operators, Methods, and Selection rules). Card, Moran, and Newell (1983) developed this process of analyzing computer interaction. A GOMS level analysis or another more modern HCI/usability analysis could further define learner actions, to categorize and represent learner actions more precisely. So for instance, rather than simply stating that an action was completed, these actions could be thoroughly defined and each problem solving operator could be scored individually.

An important criticism of this study is that it had a number of outliers (See Table 22). As with any study that has outliers, the results are less generalizable because the outliers were removed. These outliers may be a result of measurement errors, execution faults, or intrinsic variability (Barnett, 1978). See Appendix B for a detailed analysis of how and why these outliers were removed.

Table 22

Multivariate outliers

Obs	ID	Group	PT2	AC2	DSQ	prob
1	45	2	1700	48	6.065	0.0482
2	23	1	1811	41	6.088	0.04764
3	61	3	576	48	6.535	0.03811
4	40	2	571	48	6.6	0.03689
5	25	1	1694	38	6.874	0.03216
6	27	1	629	36	6.901	0.03173
7	15	1	781	35	7.246	0.02671
8	29	1	1852	48	8.154	0.01696
9	72	3	979	34	8.303	0.01574
10	20	1	1187	34	8.553	0.01389
11	98	4	1983	46	8.957	0.01135
12	12	1	2017	42	9.371	0.00923
13	46	2	1406	34	9.854	0.00725
14	43	2	1206	33	10.635	0.00491
15	110	4	788	33	10.846	0.00441
16	96	4	2275	40	16.613	0.00025
17	118	4	582	31	16.784	0.00023
18	97	4	827	30	17.742	0.00014
19	111	4	1518	31	17.793	0.00014
20	91	3	1225	30	18.04	0.00012
21	2	1	1715	30	23.236	9E-06
22	28	1	1330	28	24.692	4E-06
23	13	1	2316	34	26.665	2E-06
24	48	2	2625	39	28.251	1E-06
25	39	2	2682	41	28.395	1E-06
26	62	3	1531	25	37.43	0
27	19	1	1036	20	55.317	0
28	53	2	869	18	65.183	0
29	36	2	1223	18	66.467	0
30	22	1	152	16	81.725	0
31	18	1	608	13	94.623	0
32	104	4	2101	14	105.327	0
33	122	3	242	4	162.198	0
34	123	1	242	0	196.277	0

Outliers were removed from the initial sample during preparatory data analysis. The group composition of the outliers is demo =14, demo+practice=8, demo2+practice=5, and finally practice=7. The number of outliers in this study is troubling, because this lead to an unequal reduction from groups.

These extreme cases should be studied further. Why were there more demo learners who were removed from the final analysis? Were their scores extreme because they performed well, or poorly? Why were there fewer demo2+practice learners removed? Did the demo2+practice group feel more confident than other groups? Did the practice learners quit, because they felt unprepared? Without further study or testimonial from these learners, this is all speculation. Perhaps future studies will consider learner motivation. This then brings us to the next source of error, learner error.

Learner Errors

Mitchell and Jolley (2004) also describe participants or learners as a potential source of error. According to these authors, learner error can be either systematic or random. The fact learners were told that they must “figure out the problem scenario on their own” may have frustrated, or even motivated some learners. In addition, learners may attempt to figure out the hypothesis of the study. Either of these situations could create participant bias (Mitchell & Jolley, 2004). A potential solution to limit participant bias, is to inform all participants that their responses will be anonymous (Mitchell & Jolley 2004). In an effort to limit participant bias, learners in this study were told their responses would be anonymous.

Mitchell and Jolley (2004) also describe participants as having random error during a study, so learner behavior may be variable. However it is important to realize that learner errors are actually a field of study, because learner errors are a part of the learning process (Nielsen, 1993; Reason, 1990). This will be discussed in more detail in Chapter five.

Summary of the Results

Although the results of this study are quite interesting, a detailed discussion of these findings will not be made until chapter five. The purpose of this section is merely to summarize the results of the study and to provide some closure for Chapter four. The section is structured according to the research questions, but also considers the results according to the phases of the study.

Table 23 summarizes the results of the study. Significant results are marked by one or more asterisks, a single asterisk (*) represents significant results, while multiple asterisks represent highly significant results (***). Non significant results are represented by an abbreviation (NS).

Table 23

Results by phase matrix

<table>
<tr><th>Phase</th><th>Question 1</th><th>Question 2</th><th>Question 3</th><th>Question 4</th></tr>
<tr><td>Acquisition Phase
(Week one)</td><td colspan="2">Acquisition Phase
MANOVA
***</td><td>Relative Condition
Efficiency (RCE1)
*</td><td>Performance
Efficiency (PE1)
***</td></tr>
<tr><td>Retention Phase
(Week two)</td><td colspan="2">Retention Phase
MANOVA
(NS)</td><td>Relative Condition
Efficiency (RCE2)
(NS)</td><td>Performance
Efficiency (PE2)
(NS)</td></tr>
</table>

Research Questions One and Two

Research questions one and two considered performance time and accuracy. There were two phases of the study, the acquisition phase and retention phase, and two performances. During the acquisition phase (week one) the demo group did not assemble the Mr. Potato head problem. However, the other three groups worked with this problem, the demo+practice, demo2+practice, and practice conditions. So an acquisition phase MANOVA of performance time and accuracy, was used to compare group performances.

During the acquisition phase, it was hypothesized that learners in the animated demonstration conditions (demo+practice and demo2+practice) would out-perform learners in the practice condition. It was found that there was a significant difference between the groups, since Wilks' Λ=0.68, F (2, 68) = 6.83, p <0.0001, η^2=0.32. Post hoc comparisons with Scheffé's test (p<0.05) revealed that learners of both the demo+practice and demo2+practice groups assembled the acquisition phase problem, in significantly less time than the practice group. The retention phase MANOVA found no differences between groups a week after initial instruction.

Research Questions Three and Four

Research questions three and four dealt with the two efficiency constructs (relative condition efficiency and performance efficiency). The results for these metrics varied given condition. During the acquisition phase, significant differences between conditions were revealed given week one relative condition efficiency (RCE1) since F (2, 68) = 3.69, p=0.03. However, post hoc comparisons with Scheffé's test (p<0.05) found no significant differences between groups.

Also during the acquisition phase, performance efficiency (PE1) was found to be significantly different, because *F* (2, 68) = 13.95, *p*<0.0001. In addition, significant differences were also revealed during post hoc comparisons, with Scheffé's test (*p*<0.05), given week one performance efficiency (PE1). This analysis revealed significant differences during the acquisition phase. More specifically, it revealed that the demonstration groups (demo+practice and demo2+practice) had more efficient performances, than the practice group.

During the retention phase (week two) the four instructional conditions (demo, demo+practice, demo2+practice, and practice) were not found to differ, given the efficiency metrics (performance efficiency or relative condition efficiency), or their subcomponents. A complete discussion of each of these measures is made in Chapter five.

CHAPTER FIVE - DISCUSSION

The purpose of this chapter is to discuss the results of this study. This chapter will systematically consider all of the results, but more importantly discusses these results in the context of instructional design theory. It first considers the results according to each research question, then offers a discussion describing the implications of the results, and concludes by considering future research.

Before discussing the results of the study, it is important to recall the purpose of this dissertation. As stated in Chapter one, the purpose of this dissertation has been to assess initial skill acquisition, using animated demonstrations and practice. The two main goals of the dissertation were to: (1) consider the worked example and variability effects using animated demonstrations (Paas & van Merriënboer, 1994; Sweller & Chandler, 1991); and (2) determine if learners would exhibit a delayed performance decrement, known as Palmiter's animation deficit (Palmiter, 1993; Lipps et al., 1998). To address these goals, four research questions were developed, in order to consider skill acquisition from an HCI and cognitive load perspective. The results of these questions will be discussed and the implications of this research are considered.

Research Questions One and Two

Research questions one and two were developed to determine if learners using animated demonstrations would exhibit the worked-example effect. The results of these questions will be reviewed and discussed in relation to this effect and cognitive load theory. In addition, these questions were used to test the durability of worked example based instruction. Since there were multiple outcome variables (performance time and accuracy), multivariate statistics were necessary. Also since there were two performances, two MANOVAs were conducted, one for the acquisition phase (week one) and one for the retention phase (week two). Each phase will be discussed separately.

The Acquisition Phase MANOVA

From an instructional perspective, the purpose of the acquisition phase was to introduce learners to the subject matter, but in terms of the overall study, the main purpose of this phase was to gather performance data. During the acquisition phase, only three groups assembled the problem scenario (the Mr. Potato head problem). The demo group, *n*=23, were asked to refrain from practicing during week one, in order to measure retention during week two (the retention phase). Once outliers were removed, the number of practicing learners during week one was *n*=69.

The Results in Terms of the Worked-example Effect

Sweller and Cooper reported that those learners who studied worked examples during early schema acquisition, significantly out-performed their peers, who had learned the same procedures through active problem solving (Cooper & Sweller, 1987; Sweller & Cooper, 1985). They described this effect by saying a "decreased solution time was accompanied by a decrease in the number of mathematical errors" (Sweller & Cooper, 1985, p.59). Therefore this study put forth the hypothesis that the demonstration learners (demo+practice & demo2+practice) would outperform their peers who learned through problem solving. It was proposed the animated demonstrations would act as worked examples to promote skill acquisition, resulting in improved learner performance. The acquisition phase MANOVA found that this expectation was the case, because there was a significant difference between groups, given both performance time and accuracy, since Wilks' Λ=0.68, F (2, 68) = 6.83, p<0.0001, η^2=0.32. Performance time and accuracy were also both found to be statistically significant given α=0.05, since F (2, 68) = 3.19, p=0.048 for accuracy (AC1), and F (2, 68) = 7.84, p=0.0009 for performance time (PT1) (See Table 24). This is the result predicted by the worked-example effect.

Table 24

Acquisition phase dependent variables

	demo	demo+practice	demo2+practice	practice
n		21	31	17
Transformed Performance time (TPT1)				
M	NA	19.66	22.40	28.62
SD	NA	6.35	6.28	9.01
Transformed Accuracy (TAC1)	NA	0.56	0.99	1.44
M	NA	0.79	1.99	1.13
SD				

However, it should be clearly stated that although the acquisition phase MANOVA found significant differences between groups, post hoc comparisons with Scheffé's test (p<0.025) found group differences for accuracy (TAC1) were not significantly different. Nevertheless, significant group differences for performance time (TPT1) were revealed in post hoc comparisons with Scheffé's test (p<0.025). More specifically, post hoc comparisons found the demonstration groups (demo+practice & demo2+practice) assembled the Mr. Potato head problem, in significantly less time than the practice groups.

The worked-example effect has been described by many authors (Sweller, 2006), but has rarely been defined in terms of performance variables, like those in this study. According to Rourke and Sweller, “The worked-example effect occurs when learners presented worked examples to study, during a learning phase, solve test problems more effectively, than learners presented the equivalent problems to solve during the learning phase” (Rourke and Sweller, in press, p.1). This somewhat vague definition makes mention of the learning phase, but does not define the effect in terms of performance variables.

So while this study can claim that it has found positive evidence that the animated demonstration learners solved problems more effectively, than their peers who learned through problem solving, this dissertation will not claim that these learners demonstrated the worked-example effect. It intends to hold this effect to a more stringent operational definition, in which learners must have both a decreased performance time and an increased accuracy, in a manner similar to that described by Sweller and Cooper (1985). While this study came very close to finding a worked-example effect given the instructional conditions, again accuracy was not found to be significantly different in post hoc comparisons with Scheffé’s test (p<0.025).

Why was Accuracy Not Significantly Different?

Since accuracy did not differ during the acquisition phase (week one), it cannot be stated that learners exhibited the worked-example effect. This result is contrary to the expectations of this research. After reviewing the accuracy results of the acquisition phase (See Figure 31) one can see evidence of a ceiling effect during week one. In this figure, the accuracy scores of the demonstration groups (demo+practice and demo2+practice) are clumped near the bottom of the graph (signifying more accurate performances).

The reader may recall that scores were transformed to reduce the influence of outliers and make the skewed data more suitable for Box's M test. Even though transformations altered the distribution of the variable (to help Box's M test resolve homoscedasticity) it did not change the fact that these scores were originally very high (creating a ceiling effect) (Alliger, Ranges, & Alexander, 1988; Lord, 1955).

The expectation of an ANOVA or MANOVA, is that group scores exhibit a normal distribution, but in some settings, groups may score very high on some scales (Lord, 1955). Gall, Borg, and Gall (1996) state "A ceiling effect occurs when the range of difficulty of the test items is limited, and therefore scores at the higher end of the possible score continuum are artificially restricted" (Gall, Borg, & Gall, 1996, p.533).

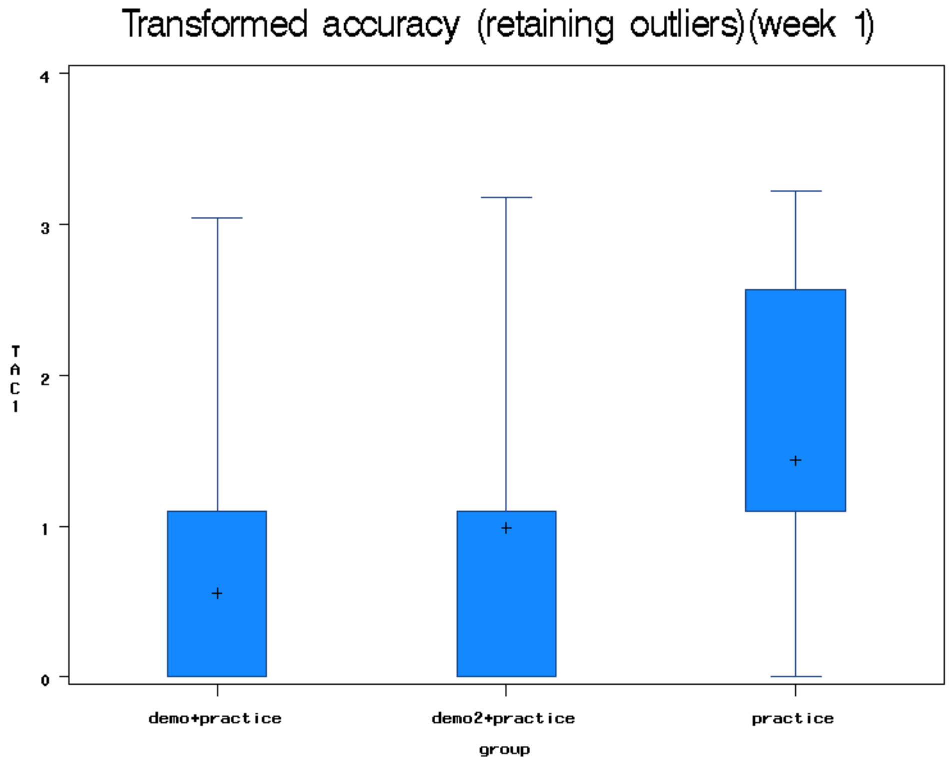


Figure 31. Acquisition phase transformed accuracy (retaining outliers)

Essentially, a ceiling effect tends to restrict variability, making it difficult to determine group differences (Alliger, Ranges, & Alexander, 1988). Since an analysis of variance is dependent upon some variability in the sample, this result is contrary to a MANOVA's assumption of normality.

While the Mr. Potato head problem may have been somewhat simple, it was developed to teach learners the required skills, and simple enough to allow at least some learners to complete the task. Had the Mr. Potato head problem been more difficult, differences in variability may have been easier to detect, but making training more difficult, to make research easier is not ethical.

In addition to being the subject of this research, the Mr. Potato head problem had another important role, to teach novice learners how to use Adobe Photoshop Elements. It is important to remember that the purpose of the acquisition phase was that novices be allowed to practice their new skills, and learn how to use this software. Finally, good instructional design ensures that learners learn.

The Retention Phase MANOVA

Background

As Tuovinen and Sweller (1999) concluded their article, they remarked that "…exploration may favor long-term retention. Although this question must remain open until tested…" (Tuovinen & Sweller, 1999, p. 340). Thus the purpose of the retention phase MANOVA was to test this idea. For this reason, it analyzed group differences one week after initial instruction. This retention interval was also chosen in order to consider the results in relation to Palmiter's animated demonstration study (Palmiter, 1991). Each of these researchers had reservations concerning the durability of learning given worked examples. Palmiter considered learning by animated demonstration as mimicking the demonstrated procedures. In addition she noted decreased performance. Here is how she explained her results:

> These demonstration users were not as proficient at remembering the procedures they had learned during training when tested at the 7-day delay session. It appeared that they simply mimicked the tasks that they had seen during training and had not encoded them so that they could remember the tasks later for faster performance… The demonstration users on the other hand, had difficulty transferring the knowledge to a new situation. They spent more than double the time spent in training (a significant increase) to perform the similar task a week later. (Palmiter, 1993, p.74).

This quote is based on Palmiter's dissertation study (Palmiter, 1991). She studied three groups of learners those that practiced after having studied text-based job aids, animated demonstration, and hybrid animated demonstrations with text. Palmiter's results are in direct contrast with the cognitive load literature, which suggests learning via this form of animated worked example.

Given the worked-example effect, it would be expected that Palmiter's demonstration learners would at least do as well as their peers, who studied text-based job aids. Also, including text-based instruction with animation has the potential to produce the split-attention effect (Chandler & Sweller, 1992; Moreno & Mayer, 1999a; Sweller & Chandler, 1991; Tarmizi & Sweller, 1988; Ward & Sweller, 1990). So given the split-attention effect, Palmiter's demonstration learners should be expected to do better than those that studied hybrid animated demonstrations, which included text-based instructions. Her findings did not support this hypothesis, since both groups had similar results. In addition it may be expected that animated demonstration learners would do better than their peers who used text based instruction, since this presentation form is more concrete than the abstract text-based instruction. Thus the current study is in part, a replication (using Palmiter's dependent variables), but it also takes into account the learners' cognitive load.

Results of the Retention Phase

Given Sweller and Cooper's initial findings (Sweller & Cooper, 1985), it was expected that during the retention phase, those who learned with animated demonstrations would take less time to solve problems (performance time) with fewer errors (accuracy), as compared with learners who learned through problem solving (practice). These expectations were not met, since the week two results found that there was not a significant difference between groups, since Wilks' Λ =0.96, F (3, 87) =0.64, p =0.70, η^2=0.04. Table 25 lists the group means for each of the dependent variables, transformed performance time (TPT2) and transformed accuracy (TAC2).

Table 25

Retention phase MANOVA by group

	demo	demo+practice	demo2+practice	practice
n	19	21	31	17
Transformed perf time (TPT2)				
M	34.10	31.92	33.29	32.09
SD	3.78	4.93	4.57	3.44
Transformed accuracy (TAC2)				
M	6.55	6.55	6.54	6.50
SD	0.26	0.25	0.22	0.21

The Results is Terms of Palmiter's Animation Deficit

Lipps et al. (1998) described Palmiter's animation deficit as a short term performance gain by learners using animated demonstrations (during early skill acquisition), but a significant loss in long term retention. Specifically, Palmiter describe the phenomenon this way: "the demonstration groups became significantly slower between the training and delay test session...accuracy between sessions decreased significantly for the demonstration groups and increased significantly for the text-only group" (Palmiter & Elkerton, 1991b, p. 260).

The findings of the current study do not support this animation deficit. They are more like those of Waterson and O'Malley (1992) or Lipps et al. (1998), who both found no evidence of Palmiter's animation deficit. It should be noted that Waterson and O'Malley (1992) added narration to their animated demonstrations. In each case these researchers found no evidence of an animation deficit. Further research concerning an animation deficit is suggested, and more evidence should be collected before this claim can be either further refuted, or justified. Yet given the results of this study, it seems this retention deficit, is not a concern, even given retention intervals as long as a week. So, there is no evidence for the idea proposed by Tuovinen and Sweller (1999) who suggested that "…exploration may favor long-term retention" (p. 340).

Finally, during week two all three demonstration conditions, competed equally well with the practice condition (See Table 25). This finding is somewhat surprising, for it also includes the demo group, who had not practice during week one, and in the end were not statistically different from the other groups. Therefore, it seems an animated demonstration alone, was sufficient for schema acquisition.

Why Are These Results Different From Palmiter's?

There are several reasons why the results of this study are different from those by Palmiter. First and foremost, Palmiter's instructional conditions were not like those in the present study. Palmiter chose to study students who learned individual discrete tasks. Her articles describe these tasks (e.g. copy button or copy field), but describes them as being in isolation, and not in the context of an overall problem. Whereas the current study studied learners *in situ*, that is, learner performance was studied given an authentic context, as they used their skills as a part of a larger project. This was necessary to gather data concerning the learner's cognitive load during problem solving, but also to measure learning in an ecologically valid manner.

The animated demonstrations presented in this study were just over ten minutes long, whereas Palmiter describes the tasks in her animated demonstrations as "deleting a field with only three procedural steps to more complex tasks such as creating a hierarchical pop-up button with 12 steps" (Palmiter & Elkerton, 1991b, p. 259). These HyperCard tasks take far less time to complete than the tasks of this study.

In her defense, Palmiter's study was implemented 17 years before this study. The technology at her disposal was considerably limited. It is little wonder that she did not study web-based narrated animated demonstrations. Web-based, animated demonstrations could not have been developed in 1991. The web as we know it did not exist then. The World Wide Web, as it was known then, did not support graphics, let alone animation or audio.

The importance of adding narration to an animated demonstration should not be underestimated. It promotes what Mayer (2001) describes as multimedia learning.

Narration provided the demonstration learners with a verbal narrative and could direct the learner's attention during the presentation, while Palmiter's narrative was only presented during the text-only condition. Palmiter may be correct that only providing learners with an animation produces mimicry of the animated demonstration, what Ausubel described as rote learning (Ausubel, 1963).

Self-explanation of non-narrated animated demonstrations may not be sufficient to produce meaningful learning. The addition of narration creates a meaningful learning environment for novices, because the instructor provides guidance, and an explanation of the procedures being demonstrated. Instructor explanation may produce the guidance necessary for schema acquisition, whereas Palmiter's non-narrated animated demonstrations could only produce rote learning.

Therefore, given all of these differences, it is not unexpected that the findings in this study are quite different from those of Palmiter's. The findings of this study are certainly not the last word given retention and animated demonstration, but given these results, the evidence does not support Palmiter's (1993) mimicry model.

Just "Too Easy"

Critics of this study may suggest the reason for a finding of no significant differences during the retention phase, was that the performance problem was "just too easy." Before coming to this conclusion, please consider Figure 32. These results were for a single problem, in which partial credit was given. If one were to take a more conservative approach, to only consider those learners who actually solved the problem, they would find that 58% of the demo learners solved the problem, 44% of the demo2+practice group, and 38% of the practice group solved the problem with no errors.

In addition, the purpose of the retention phase problem was (1) to assess retention in all groups; and (2) allow learners additional practice at a level appropriate for the audience. The intended audience for this assessment was a group of novices, those who likely only had the prior week's introduction to Adobe Photoshop Elements 2.0. From a learner's perspective, the retention phase problem (the picnic problem) was neither too easy, nor too hard. As evidence for this claim, consider the retention phase mental effort ratings (RME), *N*= 122, suggested that it was neither too easy nor too hard, since *M*= 5.00, *SD*=1.56, (5.0 is "neither low nor high mental effort"). Thus, this group as a whole felt it was neither too easy nor too hard, quite the contrary, they found it to be "just right" (See Figure 32).

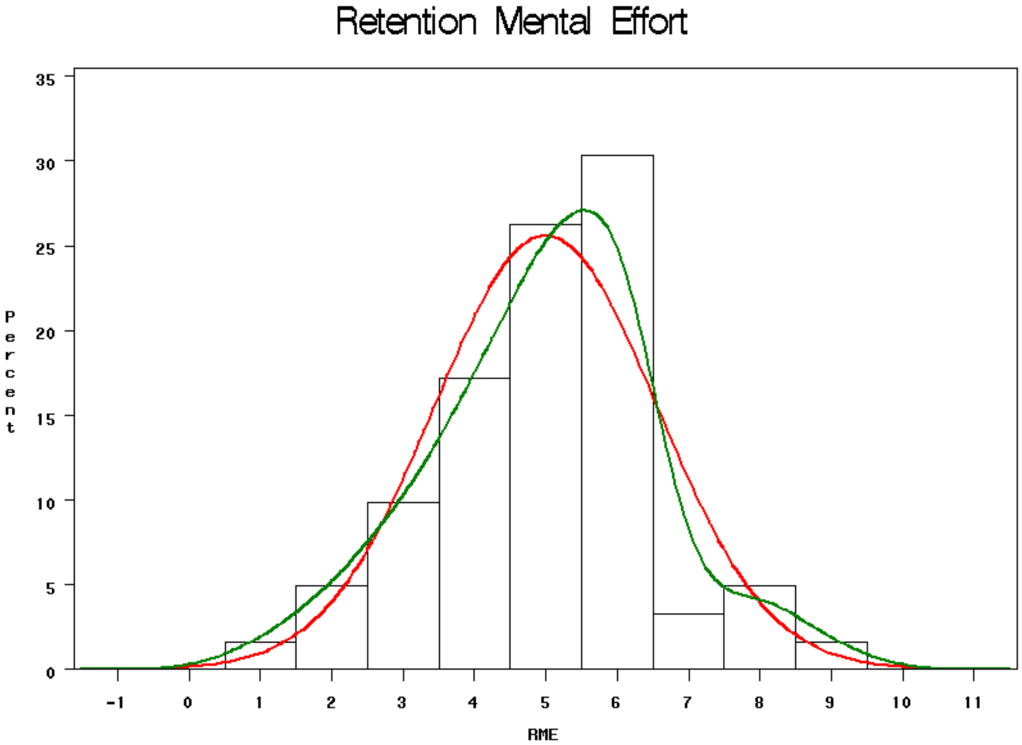


Figure 32. Retention mental effort histogram

After reviewing Figure 32, one can see that even though the overall group mean for the retention mental effort (RME) rating, was *M* = 5.00, suggesting learners invested "neither low nor high mental effort," but the group distribution was negatively skewed, Kolmogorov-Smirnov *D* (3, 87) = 0.16, $p < 0.01$. This is evidence that the retention phase problem (the picnic problem) was not "too easy." Finally, as this histogram shows the greatest number of individuals chose a six out of nine, which means they felt they invested "rather high mental effort" while solving this problem.

Research Questions Three and Four

Research questions three and four were developed to consider animated demonstrations from a cognitive load perspective. This was accomplished by analyzing animated demonstrations with two efficiency metrics (relative condition efficiency and performance efficiency).

Research Question Three - Relative Condition Efficiency

Research question three was concerned with the relative condition efficiency (RCE) of the instructional conditions. This metric was developed as an approach to compare instructional conditions given mental effort and performance measures (Paas & van Merriënboer, 1993).

In many ways the current study was modeled after a study by Tuovinen and Sweller (1999), which used relative condition efficiency to compare the learner performance of those who learned via worked examples or discovery problem solving. They found that novice learners who studied worked examples scored significantly higher on pencil and paper tests, than their peers who learned through discovery problem solving.

The current study compared similar conditions to those used in the Tuovinen and Sweller study, but used animated demonstrations. The expectation was that the animated demonstration conditions would out perform their problem solving peers. The week one results were consistent with Tuovinen and Sweller's results, since there was a significant difference between groups scores, *F* (2, 68) = 3.93, *p* = 0.03 (See Table 26 & Figure 33). However, post hoc comparisons with Scheffé's test (*p*<0.05) revealed no significant differences.

Table 26

Week one relative condition efficiency by group

	demo	demo+practice	demo2+practice	practice
n	NA	21	31	17
RCE1				
M	NA	0.50	-0.16	-0.32
SD	NA	0.63	1.11	1.25

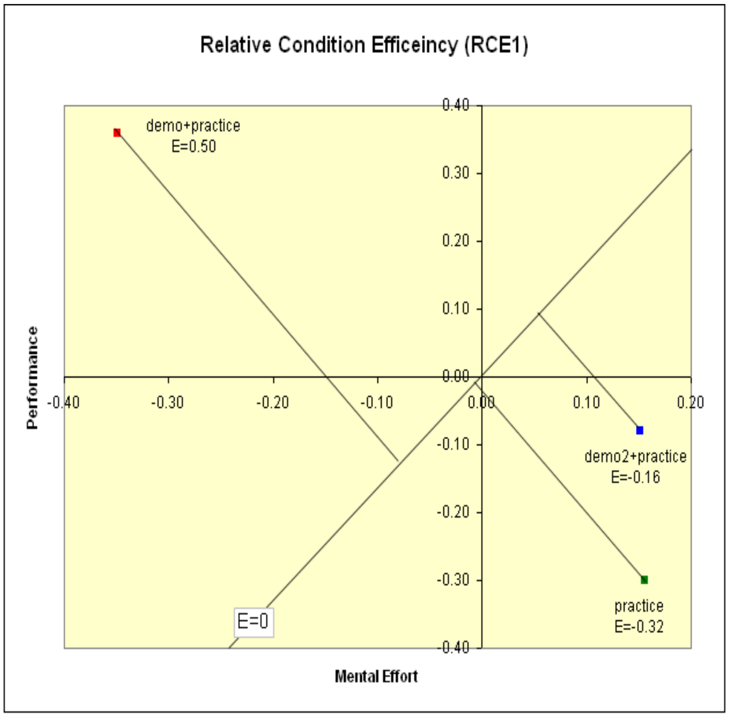


Figure 33. Week one relative condition efficiency (RCE1)

The week two results found that there was no difference between group scores, since an ANOVA revealed an *F* (3, 87) = 0.38, *p* = 0.77 (See Table 27 & Figure 34).

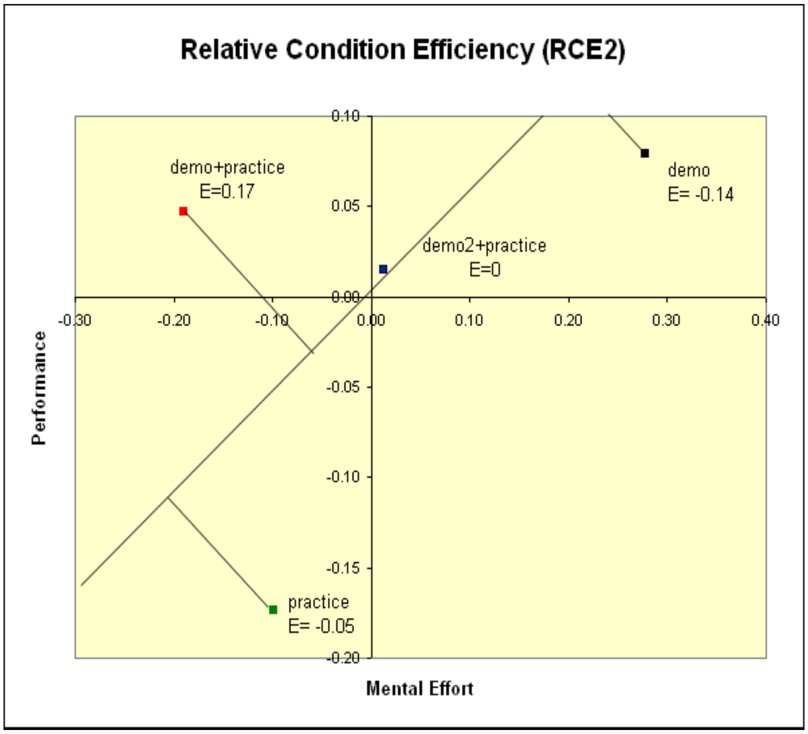


Figure 34. Week two relative condition efficiency (RCE2)

Table 27

Week two relative condition efficiency (RCE2) by group

	demo	demo+practice	demo2+practice	practice
n	19	21	31	17
RCE2				
M	-0.14	0.17	0.00	-0.05
SD	0.81	0.95	0.92	1.10

The Results in Terms of Variability Effect

Paas and van Merriënboer (1994) compared high and low variability instructional conditions. They found that learners gained most from varied context examples. Specifically, they invested less time and mental effort, which they later described as the “variability effect” (Sweller, van Merriënboer, & Paas, 1998).

The development of the second animated demonstration (demo2) was an attempt to determine if varying the context of animated demonstrations would induce the variability effect. The inclusion of a second animated demonstration (the photo collage demonstration) produced a “varied context example” as described by Clark, Nguyen, and Sweller (2006a), and in turn generated the demo2+practice group.

As predicted by the variability effect or cognitive load theory, those who studied a varied context demonstration before practice (demo2+practice), significantly out-performed those who learned through problem solving (the practice condition). This is an important finding, because it shows animated demonstrations are useful as authentic instruction, and may significantly improve learner performance. It extends the use of animated demonstrations from only being used in a similar context, to different problem scenarios, in which the learner must focus on the underlying problem structure, to grasp the problem schema.

These results do not support Palmiter’s mimicry model (Palmiter, 1993), for it shows learners who study varied context animated demonstrations are able to learn an underlying problem schema, to later reconstruct that schema from memory. Therefore, those who study animated demonstrations do not mimic the actions of the instructor, for they are interpreting the new problem, in terms of the problem solving operators, and are using a problem schema to solve the problem.

As stated above the week one results are positive evidence of the variability effect given animated demonstrations, and although significant differences were found at the $p=0.03$ level, group differences could not be detected in post hoc comparisons with Scheffé’s test ($p<0.05$). Therefore, it should not be stated that the subjects in this dissertation exhibited the variability effect. It seems the accuracy ceiling effect, has again caused some difficulties, this time accuracy (AC1) made it difficult for the RCE1 ANOVA to detect group variability.

Research Question Four - Performance efficiency

Research question four considered performance efficiency (PE), a new metric developed during this study. Performance efficiency is a construct which is a combination of Z-scores, in this case performance time (PT2) and accuracy (AC2). Performance efficiency was calculated in both the acquisition and retention phases.

There was no precedence for this metric, so it was hypothesized that a result of no significant differences would be found. The week one performance efficiency metric was calculated, but significant differences were found since F (2, 68) = 12.95, p<0.0001. Post hoc comparisons with Scheffé's test (p<0.05) found that both the demo+practice and demo2+practice conditions had significantly more efficient performances than the practice condition (See Table 28 & Figure 35).

Table 28

Week one performance efficiency (PE1) by group

	demo	demo+practice	demo2+practice	practice
n	NA	21	31	17
Relative condition efficiency (PE2)				
M	NA	0.55	0.04	-0.75
SD	NA	0.61	0.77	0.99

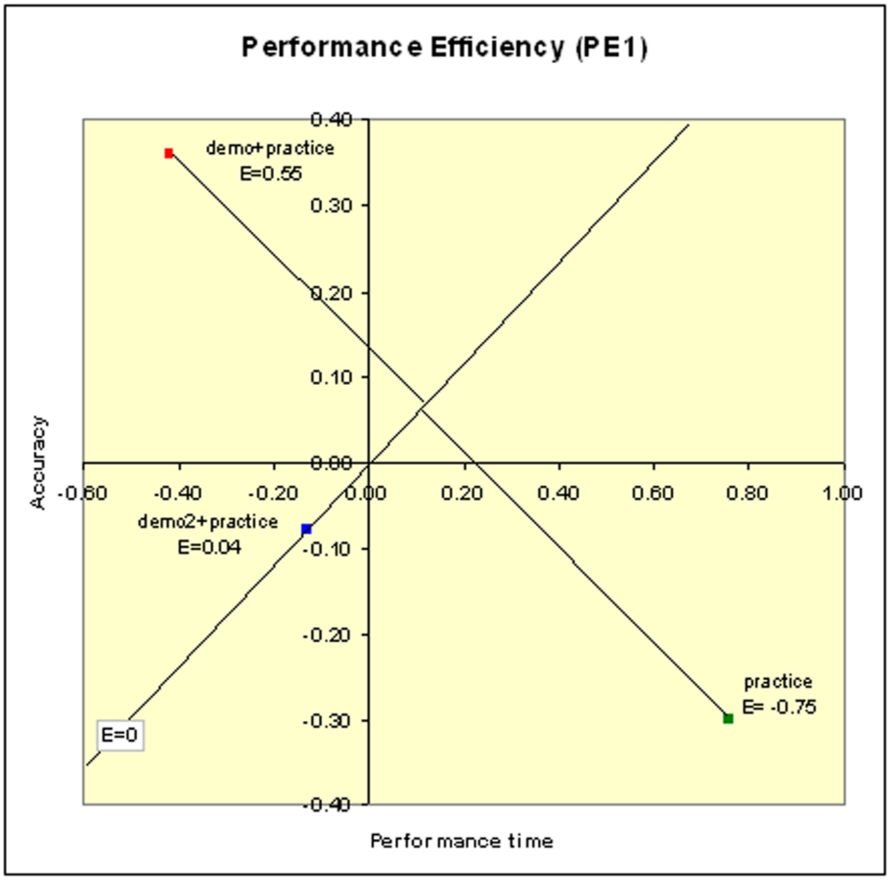


Figure 35. Week one performance efficiency (PE1)

Like the other efficiency metrics, performance efficiency (PE) is read from the diagonal line in the center of the graph, where E=0. Because less time is more efficient and a greater accuracy score is worth more points, conditions in the upper left quadrant of the graph (above the E=0 line) are the most efficient, with greater quantities of E indicating a greater performance efficiency. E in this case is from performance efficiency, and is the perpendicular distance from E=0 to the group mean score.

Performance efficiency was also measured during week two (PE2) the retention phase. The expectation was that there would be no significant differences between groups. This expectation was found to be the case, during the retention phase, since an

ANOVA for performance efficiency revealed no significant differences between group means, because *F* (3, 87) = 0.42, *p*=0.74 (See Figure 36 & Table 29).

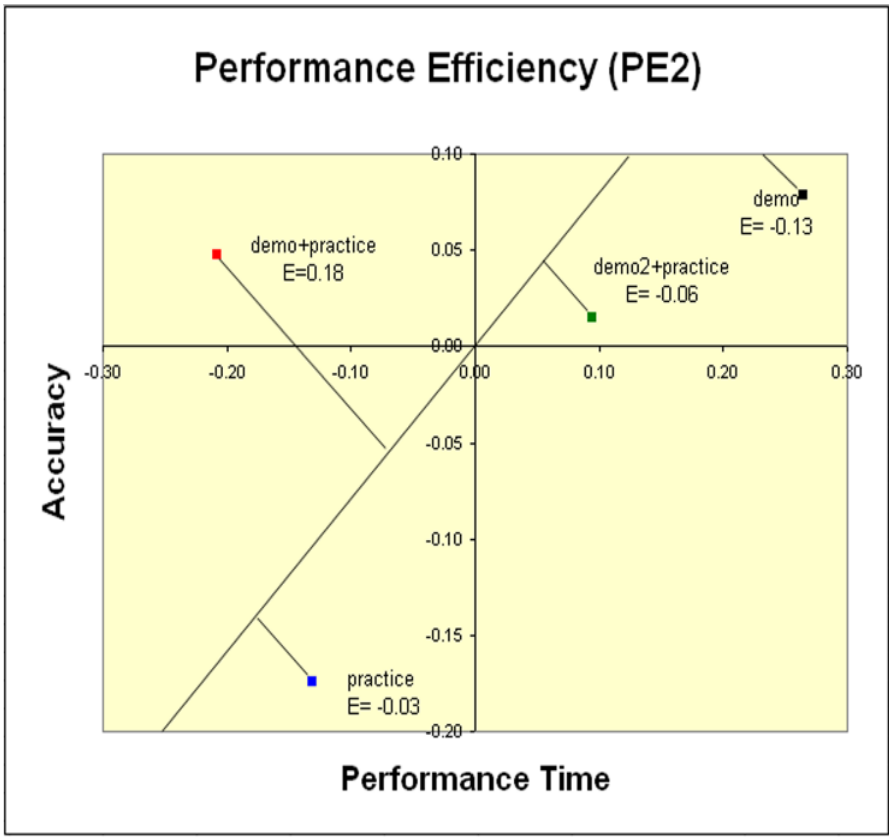


Figure 36. Performance efficiency (PE2)

Table 29

Performance efficiency by group

	demo	demo+practice	demo2+practice	practice
n	19	21	31	17
Performance efficiency (PE)				
M	-0.13	0.18	-0.06	0.03
SD	0.95	1.05	0.97	0.76

The reason this metric was developed was because there has been some discussion concerning the subjectivity of cognitive load measurements (e.g., Brünken, Plass, & Luetner, 2003). So this metric was devised to bridge these hybrid subjective/objective measures with a metric that was completely based on objective performance measures.

The reason cognitive load measures like relative condition efficiency are so useful, is that they allow a researcher to contrast two or more instructional conditions on two scales simultaneously. Performance efficiency allows one to graph performance, in this case accuracy, versus performance time. It is useful to contrast performance efficiency (PE1) (See Figure 35) to relative condition efficiency (RCE1) (See Figure 33) to see the relative contributions of each variable.

Finally performance efficiency is very generalizable, for it is conceivable that any performance measure could be contrasted with its performance time, to be graph and analyzed, in this manner. Therefore, this measure may be used outside of the cognitive load literature.

Why Were the Worked Example or Variability Effects Not Evident?

Even though animated demonstrations act as animated worked examples (Lewis, 2005), it is quite possible that they may not exhibit the worked example or variability effects. As discussed in Chapter two, Sweller and his associates have encountered this phenomenon with other worked examples (Tarmizi & Sweller, 1989; Ward & Sweller, 1990). Both of these studies found that if learners using worked examples had to integrate multiple sources of information, the worked examples would be no more effective than problem solving, and in some cases, may even be less effective.

Ward and Sweller (1990) proposed that in these cases, when learners were required to integrate multiple sources of information that learners may become overloaded and subsequently not exhibit the worked-example effect.

In the current study, demonstration learners significantly outperformed their problem solving peers, but their performance did not meet the operational definition of the worked example or variability effect. If one were to speculate why demonstration learners did not meet these effects, it could be stated that these animated demonstrations were somewhat lengthy, over 10 minutes. This has a potential for overload, and may have decreased the ability of the animated demonstrations to communicate its message. However, learner performance was not significantly different a week later, so, although these animated demonstrations did not technically exhibit the worked example or variability effects, this presentation form was not detrimental to learning. Quite the contrary, during the acquisition phase, learners using this form of instruction significantly outperformed their peers given performance efficiency and performance time.

The Implications of this Study

The context of this dissertation study

Computers have become very important in our knowledge worker society (Cortada, 1998). Lewis (2005) describes animated demonstration as a presentation form that is generalizable to all computer-based procedures. So an efficient method of instruction that applies to all computer-based procedures is very valuable. In the past few years, animated demonstrations have become increasingly common, and now are used by both education and industry.

Well-known companies, like Bank of America, Amazon.com and Microsoft, are all using animated demonstrations as a way to teach clients how to use their online services. Microsoft has even begun to incorporate “demos” (animated demonstrations) into its Office® products, as training and support. They also offer this training through a separate website called the "Office Demo Showcase" (Microsoft, 2007). Finally, organizational training groups like Element K are also beginning to offer animated demonstrations as an "online training service."

Even though these well-known companies are developing materials, educators debate if we should use these forms of instruction. The argument against animated demonstrations, are that they are a passive form of instruction.

Implications for Instructional Designers

If enough evidence-based literature begins to guide the actions of instructional designers, we will begin to produce efficient effective instruction. Unfortunately, the current literature is rarely based on empirical evidence. Mayer (2004) describes this literature as often about the "fuzzy and unproductive world of educational ideology" (p.18). When considering initial skill acquisition, educators all too often immediately employ discovery learning or what Mayer (2004) describes as the "constructivist teaching fallacy" (p.15) or "learning by doing" (p. 17). As this dissertation has shown, this "learning by doing" philosophy dates back to the 1960s and the discovery learning movement. It was then that Bruner (1961) proposed two approaches toward instruction.

One view, the expository approach, is that learners should be guided during early instruction. The alternative perspective was that learners should be allowed to discover problem solutions on their own (discovery learning). In short, Bruner (1961) says "Practice in discovering for oneself teaches one to acquire information in a way that makes that information more readily viable in problem solving. So goes the hypothesis" (Bruner, 1961, p.26).

This hypothesis was tested in this dissertation, and this study, like many other worked example studies (e.g. Tuovinen and Sweller, 1999), has found that those learners who studied worked examples (the expository mode) performed significantly better than their peers who learned through discovery problem solving (the hypothetical mode).

Klahr and Nigam (2004) describe why this is the case, when they propose those learning "in discovery situations are more likely than those receiving direct instruction to encounter inconsistent or misleading feedback, to make encoding errors and causal misattributions, and to experience inadequate practice and elaboration" (Klahr & Nigam, 2004, p.661). According to Kirschner, Sweller, and Clark (2006) those using direct methods of instruction are better able to acquire a problem schema. However, with enough practice, it may be possible that those learning through discovery learning may eventually "catch up," but at what cost? Sweller would suggest these learners may be overloaded, and could even encounter enough extraneous cognitive load that they would be unable to solve problems, perhaps never to learn the desired procedure, simply because of the instructional strategies employed.

In the current study, it was found that learners using direct methods of instruction had improved performance during early schema acquisition. So why should we allow learners to encounter misleading feedback? Or as Sweller (1988) describes it, spend their time in problem solving search without truly learning? The learner's time is important. Why should educators through their inaction, allow learners to wander, perhaps aimlessly, in an attempt to solve problems?

It is much more ethical for educators to take action, and guide learners. As this and many other studies have shown, strong guidance through direct instruction has a clear advantage for leading novices to a better understanding of the problems they are trying to solve (Kirschner, Sweller, & Clark, 2006).

While educators are entitled to their own opinions, the empirical evidence, now including the results of this study, shows animated demonstrations promote improved learner performance and are the most efficient means of teaching procedural skills. This time savings should be used to allow novices to simply learn more, and be more productive. Therefore, it is the recommendation of this study that instructional designers use this effective and evidence-based instructional strategy, to provide learners with an efficient means of accomplishing procedure-based learning.

Implications for Researchers

As the previous section described, the results of this study are important to instructional designers, but this study has some important implications for educational researchers. This is because recordings are not just an efficient means of conveying content to learners, but also a practical tool to allow researchers to review, categorize and analyze learner behavior.

Recall that Nielsen (1993) defined usability in terms of five attributes (learnability, efficiency, memorability, errors, & satisfaction). Given this dissertation, one can see how these attributes are related to instructional design given e-learning environments, cognitive load theory, and the methodology of this study. Thus this methodology was successful, because it showed that using software to record learner on-screen action is an effective means of evaluating e-learning environments, from a usability, or learnability perspective.

An important implication of this study is that it demonstrated a new approach to instructional design research, for it provided a means of evaluating learning as it occurs. By recording learner on screen actions, it was possible to document how novices behave and react when tasked with an unfamiliar learning environment. Recordings documented learner errors, the problem solving operators they employed, and the problems that they solved.

Researchers of course, are human and therefore they too, are constrained by working memory. This methodology decreases the cognitive load of the researcher (researcher cognitive load), by freeing them from the constraints of time. Therefore researchers do not have to record behaviors as they occur, because this methodology allows them to document learner actions weeks or months after the actual behavior. So perhaps the most important benefit of this methodology is that it allows researchers to document multiple outcome variables that may occur simultaneously.

Finally, the most important implication of this research is that this methodology may be generalized to any e-learning environment. Recordings of learner on-screen actions allow educational researchers to review learner behavior repeatedly if necessary, to document multiple aspects of that behavior (Martin & Bateson, 1993). This allows researchers to easily collect several variables that may be occurring simultaneously.

Conclusions

The importance of this study's results should not be underestimated. The results of this study are further evidence of the worked-example effect, but now given animated demonstrations. Since animated demonstrations are increasingly being used, these results provide further support for their use. Therefore this study has established that animated demonstrations are indeed an effective and efficient form of instruction.

Clark (1994) described this idea well, during the Clark-Kozma debates, when he said "The designer can and must choose the less expensive and most cognitively efficient way to represent and deliver instruction" (Clark, 1994, p.22). Clark (2001) drew attention to cognitive load research and suggested that it was a promising area. Cognitive load research ensures that learners are able to learn, and in the most efficient manner possible.

In this study learners were given the opportunity to learn in a variety of ways. Some would hold that experience is the best teacher, but this position diminishes the role of the instructor. Instructors have purpose in any learning environment, they provide guidance and support.

However, in an e-learning environment that role may be reduced because of an inability to communicate with "anytime anywhere" learners, but animated demonstrations allow researchers to overcome the obstacle of time and place because it allows the instructor's guidance to be there "just in time" for that "e-learner."

It's important to note, that although this dissertation used media and made several comparisons, it did not compare different forms of media, it compared different instructional strategies, given a learner-centric view (Jonassen, Campbell, & Davidson, 1994; Mayer, 1997). So unlike previous animated demonstration researchers (e.g. Palmiter, 1991; Waterson & O'Malley, 1993) this study compared different instructional strategies. In doing so, it found direct instructional strategies are more effective and efficient.

This dissertation was successful on several levels. In short this study:

- found positive evidence of both the worked example and variability effects given animated demonstrations;
- demonstrated the durability of worked example based instruction;
- found recording learner on-screen actions is a practical means of documenting the learnability of several instructional strategies;
- investigated the utility of a new metric called "performance efficiency," and used this metric to objectively compare several instructional conditions, to analyzed the relative efficiency of learner performance;
- and finally, found further evidence that Palmiter's animation deficit is not a concern given narrated animated demonstrations.

This study is not the final word given animated demonstrations and cognitive load. Thus, as with many research projects, this dissertation generated more questions than it answered, and therefore recommends future research.

Future Research

This section discusses some of these unanswered questions, and poses them in a form that future researchers may find useful.

The Length of an Animated Demonstration

Sweller (1994) proposed that element interactivity is a source of intrinsic cognitive load. Certainly animated demonstrations have element interactivity, or an inherent complexity associated with them. As this study showed, learners who are exposed to animated demonstrations were no different from those who practiced, one week after initial instruction. It was proposed that the reason these conditions did not exhibit the worked-example effect, was because these demonstrations were somewhat lengthy. Would this be the case if the length of the instruction were different?

Pollock et al. (2002) proposed the element interactivity effect. In doing so, they provided evidence that it is not the amount of information that matters, but the number of interacting elements within the instruction. However, as the length of the instruction increases, the probability of interacting elements also increases. Therefore there is a potential for a longer animated demonstration to become less useful.

Given this is the case, what is a good guideline for the length of an animated demonstration? Should they be 2-5 minutes or should they be as long as 10-20 minutes? Does learner performance deteriorate as a function of the length of the animated demonstration? These are all good questions for future researchers.

Length of the Retention Interval

While this study found retention was no less durable given animated demonstrations, it did so with a fairly short retention interval. One week may not be long enough to find any differences. Future researchers should consider similar work with longer retention intervals.

How would learner performance be affected given animated demonstration and longer retention intervals? Would the demo group be as productive given a two week interval? Are Tuovinen and Sweller's concerns founded given three to four weeks?

In a related line of reasoning, Lewis (2005) proposed learners could explore a series of animated demonstrations (a demobank) to learn computer-based procedures. The issue here is that learners do not know how to accomplish their tasks nor do they know the underlying moves associated with a task, therefore will providing a series of animated demonstrations impair or improve learning?

Learner Errors

Earlier in this chapter, the methodology of this study was described as taking a learnability perspective. This methodology allows an educational researcher to analyze why a learner makes errors. A variety of different types of learner errors could be detected when viewing recordings of learner on-screen action. These errors were related to the various aspects of the problems presented.

Some learners clearly had not learned, or had forgotten how to rotate objects within a scene a week after initial instruction. In this case, a learner's final product would have each piece of the problem placed correctly, but not rotated correctly. In addition, many learners had difficulty remembering how to flip layers in the scene. This was perhaps the most common error. Finally learners often had difficulty relocating layers relative to one another.

In each of these cases, learners had difficulties with the underlying skills of the presentation. This is as opposed to the situation during week one, when a majority of learners succeeded. Again this was not associated with any one instructional condition and was something that occurred in all conditions. Recall that the groups were not significantly different a week after instruction.

Note the columns of the rubric in Table 30. The columns in this problem are identical to the columns in week one.

Table 30

Picnic problem accuracy rubric

flip	**layer**	**rotate**	**move**	**item**
	***			umbrella
	***			tshirt
***	***			head
***	***	***		right leg
***	***			head 2
***	***			purple shirt
***				hat
***	***	***		s left leg
***	***	***		bent right leg
***	***	***		left leg
***				green shorts
***	***	***		arm 2
***				pink shorts
***	***	***		left arm
***	***			body
***	***			picnic basket
***	***	***		arm
***	***			right arm
***	***			torso
***	***			table
***	***			bird3
***	***	***		bird2
***	***			bird1
0	0	0	0	0

These were the underlying skills of the lesson, and perhaps is the best way to categorize learner errors. Remedial work with additional demonstrations or feedback, could help to alleviate these learner errors. Future researcher should refine the methodologies of this study, to consider more refined methods of categorizing learner error. It is hoped that future researchers will use techniques like those employed in this study, to evaluate instructional materials, or take this learnability perspective toward instructional materials, to make them more "learnable."

This learnability perspective acts as an extension to cognitive load theory, for it allows cognitive load researchers, to objectively document learner errors. It also allows researchers to analyze instructional design strategies, to consider how these strategies affect the performance of complex cognitive tasks. Since these methods are generalizable, this objective approach toward instructional design research may be used by any educational researcher, to evaluate and refine procedural instruction, to produce more efficient and effective instructional materials.

Are There Other Useful Metrics?

Instructional science is still evolving, and cognitive load theory is just one aspect of this growing field. One of the goals of this dissertation was to synthesize cognitive load theory with human computer interaction (HCI) research. Performance efficiency is a tangible result of this synthesis. Its application to instructional design research was expected, but this metric also has applications in other related fields, perhaps as a research tool in human factor's research.

Now the question becomes: Are there other useful metrics like performance efficiency? Perhaps cognitive load measurements could measure performance over time, and allow researchers to develop "learning rate" metrics. Figure 37 is a graphic illustration showing how this hypothetical "learning rate" metric, may look if it compared two instructional strategies over time. Perhaps someday instructional designers will be able to improve the rate at which learners learn, in other words improve "the learning curve," to produce an expert level performance quicker.

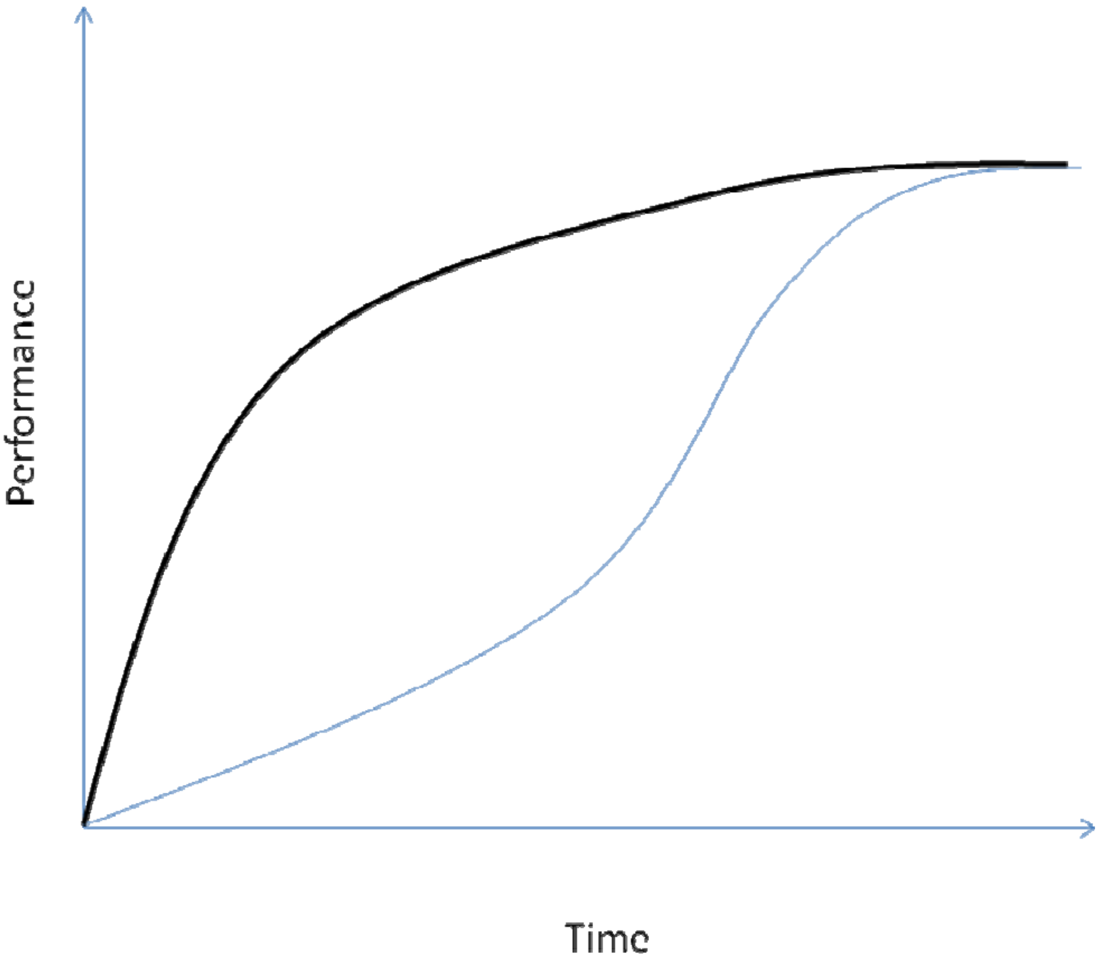


Figure 37. A hypothetical learning rate metric

Finally, future researchers will probably study learner performance and cognitive load, given more objective methods, perhaps in a HCI context. No matter how Instructional Science continues to grow and evolve, it’s important that we attempt to answer useful, practical questions about learning and instruction.

REFERENCES

Adobe Systems. (1996-2007). *Adobe Flash Player* [Computer program] Mountain View, CA

Adobe Systems. (1990-2002). *Adobe Elements 2.0* [Computer program] Mountain View, CA

Alliger, G.M., Ranges, P.J., & Alexander, R.A. (1988). A method for correcting parameter estimates in samples subject to a ceiling. *Psychological Bulletin, 103* (3) 424-430.

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*(4) 369-406

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*(3), 313-341.

Anderson, J. R., Albert, M. V., Fincham, J.M. (2005). Tracing problem solving in real time: HMRI analysis of the subject-paced tower of Hanoi. *Journal of cognitive neuroscience, 17* 1261-1274.

Anderson, J.R. and Fincham J.M. (1994). Acquisition of procedural skills from examples. *Journal of experimental psychology: Learning, memory, and cognition. 20* (6) 1322-1340

Anderson, J.R., Fincham, J.M. & Douglass, S. (1997). The role of examples and rules in the acquisition of a cognitive skill. *Journal of experimental psychology: Learning, memory, and cognition*, 23(4) p.932-945

Anderson, J. R. & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Atkinson, R.C. & Shiffrin, R.M. (1968). Human memory: a proposed system and its control processes. *In K.W. Spence (ed.), The psychology of learning and motivation: Advances in research and theory, Vol. 2* (pp. 89–195). New York: Academic Press.

Atkinson, R.C. & Shiffrin, R.M. (1971). The control of short term memory. *Scientific american*, 225(2) 82-90.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. W. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*, 181–214.

Ausubel, D.P. (1963). *The psychology of meaningful verbal learning; an introduction to school learning*. New York: Grune and Stratton.

Ausebel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart, & Winston.

Ausubel, D. P., & Fitzgerald, D. (1961). Growth, Development, and Learning. *Review of Educational Research, 31*(5) 500-510.

Baddeley, A. D. (1986) *Working memory*. Oxford: The Oxford University Press

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

Baddeley, A. (2006) Working memory: an overview. *In Working memory and education*. *Pickering, S. J. Ed.* (pp. 3-26) London: Academic Press.

Baddeley, A.D. & Hitch, G.J. (1974). Working memory. *In G.A. Bower (ed.), Recent Advances in Learning and Motivation, Vol. 8* (pp. 47–89). New York: Academic Press.

Bailey, R.W. (1996) *Human Performance Engineering*, (3rd ed.), Englewood Cliffs, NJ: Prentice Hall

Bakeman, R. & Gottman, J.M. (1986). *Observing interaction: an introduction to sequential analysis*. Cambridge, England: Cambridge University Press.

Barnett, V. (1978). The study of outliers: purpose and model. *Applied Statistics*. 27(3) 242-250

Bartlett, F.C. (1932). *Remembering*. Cambridge: The Cambridge University Press

Bartlett, F.C. (1958). *Thinking*. New York: Basic Books

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the royal society of London. Series A, mathematical and physical sciences, 160* (901) 268-282.

Bear, M., Connors, B., & Paradiso, M. (2001). *Neuroscience: exploring the brain.* Baltimore, MD: Lippincott, Williams & Wilkins.

Beatty. J. (1977). Activation and attention. In M. C. Wittrock, J. Beatty, J. E. Bogen, M. S. Gazzaniga, H. J. Jerison, S. D. Krashen. R. D. Nebes, & T. Teyler (Eds.). The human brain. (pp.63-86) Englewood Cliffs, New Jersey: Prentice-Hall.

Bethke, F.J., Dean, W.M., Kaiser, P.H., Ort, E. and Pessin, F. H. (1981). Improving the usability of programming publications. *IBM Systems Journal, 20* (3) 306-320

Bobrow, D.G., & Norman, D. A. (1975). Some principles of memory schemata. In D.G. Bobrow & A. M. Collins (Eds.), *Representation and understanding: Studies in cognitive science*. New York: Academic Press

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. *Annals of Mathematical Statistics*, *25,* 290-302.

Bonwell, C. & Eison, J. (1991). *Active learning: Creating excitement in the classroom*. AEHE-ERIC Higher Education Report No.1. Washington, D.C.: Jossey-Bass.

Breierova, L. & Choudhari, M. (2001). *An introduction to sensitivity analysis*. retrieved Febrary 16, 2008 from http://sysdyn.clexchange.org/sdep/Roadmaps/RM8/D-4526-2.pdf

Brown, M.B., Forsythe, A.B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346) 364-367.

Bruner, J. S. (1961). The act of discovery. *Harvard Educational Review. 31*(1) 21–32.

Brünken,R., Steinbacher,S., Plass, J.L. & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology. 49*(2) 109-119.

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning, *Educational Psychologist, 38*(1), 53–61.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin 54* (4), 297-312.

Cantor, A.B. (1996). Sample-size calculations for Cohen's kappa. *Psychological Methods 1*(2) 150-153.

Carroll, W. (1994).Using worked examples as an instructional support in the algebra classroom. *Journal of Educational Psychology, 86*(3) 360–367.

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*. *8*(4), 293-332.

Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, *62*, 233-246.

Chase, W.G., & Simon, H.A. (1973a). Perception in chess. *Cognitive Psychology, 4*(1) 55-81

Chase, W.G., & Simon, H.A. (1973b). The mind's eye in chess. In W.G. Chase (Ed.), *Visual information processing* (pp. 215--281). New York: Academic Press.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5* 121-152.

Chi M.T.H., Bassok M., Lewis M., Reimann P., Glaser R. 1989. Self-explanations: how students study and use examples in learning to solve problems. *Cognitive Science*. 13, 145– 82

Clark, J. M. & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review 3* (3) 149-170.

Clark, R. E. (1994). Media Will Never Influence Learning. *Educational Technology Research and Development, 42* (2) 21-29.

Clark, R.E. (1983). Reconsidering research on learning from media. Review of Educational Research 53(4) 445-459.

Clark, R. E. (2001). What is next in the media and methods debate? In R. E. Clark (Ed.), *Learning from Media* (pp. 327–337). Greenwich CT: Information Age Publishers Inc.

Clark, R.C. (1999). *Developing technical training*. Silver Spring, MD: International Society for Performance Improvement

Clark, R.C., Nguyen, F., and Sweller, J. (2006a). *Efficiency in learning: evidence-based guidelines to manage cognitive load.* San Francisco: Pfeiffer.

Clark, R.C., Nguyen, F., and Sweller, J. (2006b). *Efficiency in learning: evidence-based guidelines to manage cognitive load [CD].* San Francisco: Pfeiffer.

Cohen, N.J. & Squire, L.R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. *Science 210* (4466) 207-210.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20* (1) 37-46.

Cole, D.A., Maxwell, S.E., Arvey, R., and Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin. 115*(3) 465-474.

Cooper, G.E., & Harper, R.P. (1969). *The use of pilot ratings in the evaluation of aircraft handling qualities* (Report No. NASA TN-D-5153). Moffett Field, CA: NASA Ames Research Center. retrieved April 23, 2008 from http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19690013177_1969013177.pdf

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*. 79(4), 347-362.

Cornett, L. L. (1993). *Animated demonstrations versus text: a comparison of training methods*. Unpublished Masters Thesis. Rice University

Cortada, J. W. (Ed.) (1998). Where did the knowledge worker come from? In *Rise of the Knowledge Worker.* (pp. 3-22) Boston: Butterworth-Heinemann

Curtis, N.A. (2007). Are histograms giving you fits? New SAS software for analyzing distributions. retrieved December 31, 2007 from http://support.sas.com/rnd/app/papers/distributionanalysis.pdf

De Groot, A. D. (1965). *Thought and choice in chess*. The Hague: Mouton Publishers.

D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature*. 378(6554) 279-281.

D’Esposito, M., Aguirre, G.K., Zarahn, E., Ballard, D., Shin, R.K., Lease, J. (1998). Functional MRI studies of spatial and nonspatial working memory. *Cognitive Brain Research. 7*(1) 1–13.

Devlin, S.J., Gnanadesikan,R., Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62(3), 531-545.

Dewey, J. (1916/1997). *Democracy and education*. New York: Simon and Schuster.

Driscoll, M. P. (2000). *Psychology of learning for instruction* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Duffy, T.M., & Cunningham, D.J. (1996). Constructivism: Implications for the design and delivery of instruction. *In D. H. Jonassen (Ed.), Handbook of research for educational communications and technology* (pp. 177-198). New York: Simon & Schuster Macmillan.

Finch, F. (2005) Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated. *Methodology 1*(1) 27–38.

Fidell, L. S., & Tabachnick, B. G. (2003). Preparatory data analysis. *In J. A. Schinka & W. F. Velicer (Eds.), Handbook of psychology: Research methods in psychology* (Vol. 2, pp. 115-141). New York: John Wiley & Sons.

Fleming,M. L., and Levie,W. H. (1993). *Instructional message design: principles from the behavioral and cognitive sciences*. (2nd Ed) Educational Technology Publications, Englewood Cliffs, NJ.

Friendly, M. (1991). Statistical Graphics for Multivariate Data. *Proceedings of the SAS SUGI 16 Conference*, retrieved December 29, 2007 from http://www.math.yorku.ca/SCS/sugi/sugi16-paper.html

Friendly, M. (2006). Data ellipses, he plots and reduced-rank displays for multivariate linear models: SAS software and examples. *Journal of Statistical Software, 17*(6) 1-43.

Friendly M (2007a). HE plots for multivariate general linear models. *Journal of Computational and Graphical Statistics, 16*(2) 421–444.

Friendly, M. (2007b). SAS macro programs: ellipses. retrieved November 4, 2007 from http://www.math.yorku.ca/SCS/sasmac/ellipses.html

Friendly, M. (2007c). SAS macro programs for statistical graphics: OUTLIER. retrieved November 4, 2007 from http://www.math.yorku.ca/SCS/sssg/outlier.html

Friendly, M. (2007d). SAS macro programs: cqplot. retrieved November 4, 2007 from http://www.math.yorku.ca/SCS/sasmac/cqplot.html

Gagné, R. M. (1964). Problem solving. *In A.W. Melton (Ed.), Categories of human learning*. New York: Academic Press.

Gagné, R. M. (1965). *The conditions of learning*. New York: Holt, Rinehart & Winston

Gagné, R. M. (1966). Varieties of learning and the concept of discovery (pp.135-150) In Shulman, L. S. and Keislar, E. R. (Eds) *Learning by discovery: A critical appraisal.* Chicago: Rand McNally and Co.

Gagné, R. M. (1968). Learning hierarchies. *Educational psychologist*, *6*(1), 1-9.

Gagne, R. M. (1972). Domains of learning. *Interchange,3*, 1-8.

Gagné, R. M. & Paradise, N. E. (1961). Abilities and learning sets in knowledge acquisition. *Psychology Monographs. 75* (14) 1-23.

Gall, M. D.; Borg, W. R.; & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). New York: Longman Publishers.

Gerjets, P. Scheiter, K. and Catrambone, R. (2004).Designing instructional examples to reduce intrinsic cognitive load: molar versus modular presentation of solution procedures. *Instructional Science. 32*(1) 33–58

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn & Bacon

Glaser, R. (1966). Variables in discovery learning. *In L. S. Shulman, & E. R. Keislar, (Eds.), Learning by discovery: A critical appraisal* (pp. 13-26). Chicago: Rand McNally.

Gnanadesikan, R. & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics, 28* (1) 81-124.

Grace, C. A. (1998). Retention of Word Meanings Inferred from Context and Sentence-Level Translations: Implications for the Design of Beginning-Level CALL Software. *The Modern Language Journal, 82*(4) 533-544.

Grafton, S.T., Mazziotta,J.C., Presty, S., Friston,K.J., Frackowiak,R.S.J. and Phelpsis, M.E. (1992). Functional anatomy of human procedural learning determined with regional cerebral blood flow and PET. *The Journal of Neuroscience, 12* (7) 2542-2548.

Harrison, S. M. (1995). A comparison of still, animated, or nonillustrated on-line help with written or spoken instructions in a graphical user interface. In *I. R. Katz, R. Mack, L. Marks, M. B. Rosson, & J. Nielsen (Eds.), Proceedings of the ACM conference on human factors in computing systems* (pp. 82-89). Denver, CO: Association for Computing Machinery.

Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In *P.A. Hancock & N. Meshkati (Eds.), Human mental workload* (pp. 139–183). Amsterdam: Elsevier.

Hegarty,M., Kriz,S., & Cate, C. (2003). The roles of mental animations and external animations in understanding mechanical systems. *Cognition and instruction, 21*(4), 325–360.

Hilbert, T. S., & Renkl, A. (2007). Learning how to Learn by Concept Mapping: A Worked-Example Effect. *Paper presentation at the 12th Biennial Conference EARLI 2007 in Budapest, Hungary*

Hinde, R.A. (1973). On the design of check-sheets. *Primates, 14*(4) 393-406.

Hinsley, D., Hayes, J., & Simon, H. (1976). From words to equations: Meaning and representation in algebra word problems. In *P. Carpenter & M. Just (Eds.), Cognitive processes in comprehension*. Hillsdale, NJ: Erlbaum.

Janisse, M.P. (1977). *Pupillometry*. Washington: Hemisphere Publishing Company.

James, W. (1890/1950). *The Principles of Psychology*. Dover: New York.

Jeung, H. &Chandler, P., & Sweller (1997). The role of visual indicators in dual sensory mode instruction. *Educational Psychology*, 17(3), 329-344.

Jonassen, D. (1991). Objectivism versus constructivism: Do we need a new philosophical paradigm? *Educational Technology Research and Development, 39*(3), 5-14.

Jonassen, D.H. (2002). Learning as activity. *Educational Technology, 42*(2), 45-51.

Jonassen, D. H., Campbell, J. P., & Davidson, M. E. (1994). Learning with media: Restructuring the debate. *Educational Technology Research and Development, 42*(2), 31-39.

Jonassen, D., Mayes, T., & McAleese, R. (1993). A manifesto for a constructivist approach to uses of technology in higher education. In *T.M. Duffy, J. Lowyck, & D.H. Jonassen (Eds.), Designing environments for constructive learning* (pp. 231-247). Heidelberg: Springer-Verlag.

Jonides, J., Smith, E. E., Koeppe, R. A., Awh, E., Minoshima, S., & Mintun, M. A. (1993). Spatial working memory in humans as revealed by PET. *Nature*, 363, 623–625.

Kalyuga,S., Chandler,P., & Sweller, J (1997). Levels of expertise and user-adapted formats of instructional presentations: a cognitive load approach*. In Anthony Jameson, Cécile Paris, and Carlo Tasso (Eds.), User Modeling: Proceedings of the Sixth International Conference, UM97. Vienna*, New York: Springer Wien New York.

Kalyuga,S., Chandler,P., Tuovinen,J.,& Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*(3) 579-588.

Kalyuga, S. Ayres,P., Chandler,P. and Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1) 23–31.

Kalyuga, S., Chandler, P., & Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors, 40* (1) 1-17.

Keppel G. 1991. *Design and Analysis: A Researcher's Handbook.* (3rd ed.) Englewood Cliffs: Prentice Hall.

Kerlinger, F. N. (1986). *Foundations of Behavioral Research*, (3rd ed.) New York: Holt, Rinehart, and Winston

Keselman, H. J. (2005). Multivariate normality tests. In *B. S. Everitt & D. C. Howell (Eds.), Encyclopedia of statistics in behavioral science* (Volume 3, 1373-1379). New York:Wiley

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist.* 41(2) 75-86.

Klahr, D. and Nigam, M. (2004). The Equivalence of Learning Paths in Early Science Instruction. *Psychological Science, 15*(10) 661-667.

Knupfer, N. N. & McLellan, H. (1996). Descriptive research methodologies. *In* D. H. Jonassen (Ed), *Handbook of Research for Educational Communications and Technology*, 1196-1212. New York: Macmillan.

Kozma, R. (1991). Learning with media. *Review of Educational Research, 61*(2), 179-212.

Kozma, R. (1991). Will media influence learning? Reframing the debate. Review of Educational Technology Research and Development, 42(2), 1042-1629.

Krashen, Stephen D. (1982). *Principles and practice in second language acquisition*. Oxford; New York: Pergamon

Landis,J.R., and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33* (1), 159-174.

Larkin, J., McDermott, J., Simon, D., & Simon H. (1980). Expert and Novice Performance in Solving Physics Problems. *Science, 208* (4450) 1335-1342.

Lefevre, J. & Dixon, P. (1986). Do Written Instructions Need Examples? *Cognition and Instruction. 3*(1), 1-30

Lehner, S. 1996. *Handbook of ethological methods*, (2nd ed.) Cambridge University Press, Cambridge, United Kingdom.

Lewis, D. (2005). Demobank: a method of presenting just-in-time online learning. *In the Proceedings of the Association for Educational Communications and Technology (AECT) Annual International Convention (vol 2, p. 371-375) October 2005, Orlando, FL.*

Lipps, A. W., Trafton, J. G., & Gray, W. D. (1998). Animation as documentation: A replication with reinterpretation. retrieved June 12, 2007 from http://www.stc.org/confproceed/1998/PDFs/00006.PDF

Locke, L. F., Spirduso, W.W., Silverman, S. J. (2000). *Proposals that work: A Guide for Planning Dissertations and Grant Proposals*. Thousand Oaks, CA: Sage Publications

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, 7(4), 560-572.

Loftus, E.F. & Hoffman, H. G. (1989). Misinformation and memory: the creation of new memories. *Journal of Experimental Psychology: General*, 118(1), 100-104.

Lord, F. M. (1955). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement*. *15* (4) 383-389.

Magill, R.A. & Hall K.G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science 9*(3), 241-289.

Marcus, N., Cooper, M. & Sweller, J. (1996) Understanding instructions. *Journal of Educational Psychology*. 88(1) 49–63

Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. *Biometrika*, *58*(1) 105-121.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3)519-530.

Mardia, K.V. (1975). Assessment of multinormality and the robustness of Hotelling's T-squared test, *Applied Statistics, 24*(2) 163-171

Marsh, P. O. (1979). The instructional message: A theoretical perspective. *Educational Communication and Technology Journal*, 27(4) 303-18

Martin, P., & Bateson, P. (1993). *Measuring behaviour: an introductory guide*. (2^{nd} ed) Cambridge: Cambridge University Press

Mayer, R. (1997). Multimedia Learning: Are We Asking the Right Questions? *Educational Psychologist, 32*(1), 1-19

Mayer, R. (2001). *Multimedia Learning*. Cambridge: Cambridge University Press

Mayer, R. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist. 59*(1) 14-19.

Mayer, R. E., & Anderson, R.B. (1991).Animations need narrations: an experimental test of a dual-coding hypothesis. *Journal of Educational Psychology, 83*(4) 484-490

Mayer, R.E., Bove, W., Bryman, A., Mars, R. and Tapangco, L. (1996). When less is more: meaningful learning from visual and verbal summaries of science textbook lessons. *Journal of Educational Psychology. 88*(1) 64-73

Mayer, R. E. & Moreno, R. (1998). A cognitive theory of multimedia learning: implications for design principles. *Paper presented at the annual meeting of the ACM SIGCHI Conference on Human Factors in Computing Systems. Los Angeles, CA.*

Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology*, 84, 389–401.

Maxwell, S. (2001). Methodological and statistical concerns of the experimental behavioral researcher. *Journal of Consumer Psychology, 10*(1/2) 29-30.

Melton, A.W. (1963). Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavio*r, *2* 1-21

Merrill, M.D. (1983). Component Display Theory. In C. Reigeluth (ed.), *Instructional Design Theories and Models*. Hillsdale, NJ: Erlbaum Associates.

Merrill, M. D. (2007). A task-centered instructional strategy. *Journal of Research on Technology in Education, 40*(1) 33-50.

Merrill, P.F. (1971). Task analysis: an information processing approach. CAI Center Tech Memo Number 27. (ERIC Document Reproduction Service No. ED050554)

Merrill. P. F. (1976). Task Analysis ─ An information processing approach. *National Society for Performance and Instruction Journal, 15*, (2). 7-11.

Merrill, P.F. (1980). Analysis of a procedural task. National Society for Performance and Instruction Journal, 17(2), 11-26.

Microsoft (1995-2004). *Microsoft Internet Explorer* [Computer program] Redmond, WA

Microsoft (2003a). *Microsoft FrontPage 2003* [Computer program] Redmond, WA

Microsoft (2007). *Office demo showcase*. Retrieved August 13, 2007 from http://office.microsoft.com/en-us/FX010804491033.aspx

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

Mishkin, M. (1978). Memory in monkeys severely impaired by combined but not by separate removal of amygdala and hippocampus. *Nature, 273*, 297 – 298.

Miyake, A., & Shah, P. (1999). Toward unified theories of working memory. *In Miyake, A., & Shah, P. (Eds.) Models of working memory: Mechanisms of active maintenance and executive control* (pp. 442-481). New York: Cambridge University Press.

Molenda, M, Reigeluth, C.M. y Nelson, L. M. (2003). Instructional design. *In L. Nadel (Ed.). Encyclopedia of Cognitive Science, London*, Nature Publishing Group, Vol. 2, 574-578. retrieved January 30, 2008 from http://www.indiana.edu/~molpage/ID_Cog%20Sci.pdf

Moore, M.G. (1989). Three types of interaction. *The American Journal of Distance Education*. 3(2) retrieved January 7, 2006 from http://www.ajde.com/Contents/vol3_2.htm

Mook D.G. (2001). Psychological research: the ideas behind the methods. London: Norton

Moreno, R., & Mayer, R. E. (1999a). Multimedia-supported metaphors for meaning making in mathematics. *Cognition and Instruction, 17*(3), 215–248.

Moreno, R., & Mayer, R. E. (1999b). Visual presentations in multimedia learning: Conditions that overload visual working memory. In D. P. Huijsmans & A. W. M. Smeulders (Eds.), *Lecture notes in computer science: Visual information and information systems* (pp. 793–800). Berlin: Springer.

Mousavi, S.Y., Low, R. and Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology*, 87(2) 319-334.

Nadolski, R. J., Kirschner, P. A., & van Merriënboer, J. J. G. (2005). Optimising the number of steps in learning tasks for complex skills. *British Journal of Educational Psychology, 75*, 223–237.

Newell, A., Shaw, J.C., Simon, H.A. (1958a). Elements of a theory of human problem solving. *Psychological Review, 65*, 151-166.

Nielsen, J. (1993). *Usability engineering*. Boston: Academic Press.

Nielsen, J. (2001) *First Rule of Usability? Don't Listen to Users.* retrieved 02-11-06 from http://www.useit.com/alertbox/20010805.html

Olson, C. L. (1974). Comparative Robustness of Six Tests in Multivariate Analysis of Variance. *Journal of the American Statistical Association, 69* (348) 894-908.

Owen, E. and Sweller, J. (1985). What Do Students Learn while Solving Mathematics Problems? Journal of Educational Psychology, 77(3) 272-284.

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal Educational Psychology. 84* (4), 429-434.

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

Paas, F. G. W. C., Renkl, A. & Sweller, J. (2004). Cognitive load theory: instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science, 32*, 1–8.

Paas, F., Tuovinen, J. E., Tabbers, H. K., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38* (1), 63–71.

Paas, F. G. W. C., and van Merrienboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental-effort and performance measures. *Human Factors, 35*(4), 737-743.

Paas, F. G. W. C., and van Merrienboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem solving skills: A cognitive load approach. *Journal Educational Psychology, 86*(1) 122-133.

Paas, F. G. W. C. (2007). A Multidimensional Approach to the Mental Efficiency of Instructional Conditions. retrieved June 5, 2007 from http://www.ou.nl/Docs/Expertise/OTEC/Projecten/onderzoeksvoorstellen%20PDF/Paasproject34%5B1%5D.pdf

Paivio, A. (1971). *Imagery and Verbal Processes*. New York: Holt, Rinehart & Winston.

Paivio, A. (1978). Mental comparisons involving abstract attributes. *Memory and Cognition, 6*, 199-208.

Palmiter, S. (1991). *Animated Demonstrations for learning procedural, computer-based tasks*. Doctoral Dissertation. University of Michigan, Ann Arbor, (Proquest No. 9124076)

Palmiter, S. and Elkerton, J. (1991a). Animated demonstrations versus written instructions for learning procedural tasks: A preliminary study. *International Journal of Man-Machine Studies. 34*, 687-701.

Palmiter, S. & Elkerton, J. (1991b). An evaluation of animated demonstrations of learning computer-based tasks. In the *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*. New Orleans, Louisiana, United States. p.257 - 263

Palmiter, S., Elkerton, J. & Baggett, P. (1991). Animated demonstrations vs. written instructions for learning procedural tasks: A preliminary investigation, *International Journal of Man-Machine Studies, 34*, 687-701.

Parkin, A. (1998). The central executive does not exist. *Journal of the International Neuropsychological Society, 4*, 518–522.

Paulesu, E., Frith, C. D., & Frackowiak, R. J. (1993). The neural correlates of the verbal component of working memory. *Nature, 362*, 342–345

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: an integrated approach*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Penney, C.G. (1989). Modality effects and the structure of short-term memory. *Memory and Cognition 17*(4) 398–442.

Pershing, J. A., Warren, S. J., & Rowe, D. T. (2006). Observation methods for HPT. *In J. A. Pershing (Ed.), The Handbook of Human Performance Technology: Principles, Practices, and Potential* (3rd ed.). San Francisco, CA: Pfeiffer.

Peterson, L. R. & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology 58*(3), 193–198

Peter L. Pirolli and John R. Anderson (1985).The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology, 39*(2), 240-272

Plass, J., & Jones, L. (2005). Multimedia learning in second language acquisition. *In The Cambridge handbook of multimedia learning, R. Mayer, Ed.* (pp. 467-488). New York: Cambridge University Press.

Poldrack, R. A. & Gabrieli, J. D. E. (2001). Characterizing the neural basis of skill learning and repetition priming: Evidence from mirror-reading. *Brain, 124*(1) 67-82.

Pollock, E. Chandler, P. and Sweller, J. (2002). Assimilating complex information. *Learning and Instruction. 12*(1), 61-86.

Postman, L. 1963. Does interference theory predict too much forgetting? *Journal of Verbal Learning and Verbal Behavior 2*(1) 40-48.

Price, E. & Driscoll, M. (1997). An inquiry into the spontaneous transfer of problem-solving skill. *Contemporary Educational Psychology 22*(4), 472–494

Prinzel, L. III, Pope, A., Freeman,F, Scerbo, M. and Mikulka, P. (2001). *Empirical analysis of EEG and ERPS for psychophysiological adaptive task allocation.* (Report No. NASA/TM-2001-211016) Hampton, VA: National Aeronautics and Space Administration. retrieved December 10, 2003 from http://techreports.larc.nasa.gov/ltrs/PDF/2001/tm/NASA-2001-tm211016.pdf

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, 88, 144–161.

Rabitz, H. (1989). System analysis at molecular scale. *Science 246*(4927), 221–226.

Reason J.(1990). *Human error*. New York: Cambridge University Press

Reimann, P. & Neubert, C. (2000). The role of self-explanation in learning to use a spreadsheet through examples. *Journal of Computer Assisted Learning, 16*(4), 316-325

Renkl, A., Atkinson, R. K., & Maier, U. H. (2000). From studying examples to solving problems: Fading worked-out solution steps helps learning. In *L. Gleitman & A. K. Joshi (Eds.), Proceeding of the 22nd Annual Conference of the Cognitive Science Society* (pp. 393–398). Mahwah, NJ: Erlbaum.

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *Journal of Experimental Education, 70* (4), 293–315.

Reid, G.B., & Nygren, T.E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *In P.A. Hancock & N. Meshkati (Eds.), Human mental workload* (pp. 185–218). Amsterdam: Elsevier.

Rieber, L. P. (1990). Animation in computer-based instruction. *Educational Technology Research & Development, 38*(1), 77-86.

Rieber, L. P. (2000). *Chapter 6--Review of Instructional Visual Research: Animated Visuals.* retrieved January 3, 2004 from http://www.nowhereroad.com/cgl/chapter6/

Rieber, L., & Parmley, M. W. (1992). Effects of animated computer simulations on inductive learning with adults: a preliminary report. *In: Proceedings of Selected Research and Development Presentations at the Convention of the Association for Educational Communications and Technology*. (ERIC Document Reproduction Service No. ED 348 019).

Rieber, L. P.,& Parmley, M. W. (1995). To teach or not to teach? Comparing the use of computer-based simulations in deductive versus inductive approaches to learning with adults in science. *Journal of Educational Computing Research. 13*(4) 359–374.

Rossett, A. & Gautier-Downes, J. (1991). *A handbook of job aids*. San Diego: Pfeiffer.

Rourke, A. & Sweller, J. (in press). The worked-example effect using ill-defined problems: Learning to recognize designers' styles. *Learning and Instruction* retrieved November 9, 2008 from http://www.science-direct.com/science/article/B6VFW-4SHN0C2-1/2/2936a72b8a9e3e1e28af287c61cb30ed

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. E. Brewer (Eds.), *Theoretical issues in reading comprehension*. Hillsdale, NJ: Lawrence Erlbaum.

Rumelhart, D. E. and A. Ortony. (1977). The representation of knowledge in memory. In R. D. Anderson, R. J. Sprio and W. E. Montague (Eds.) *Schooling and the acquisition of knowledge*. Hillsdale, NJ Lawrence Erlbaum Associates. 99-135

Rumelhart, D. & Norman, D. (1978). Accretion, tuning and restructuring: Three modes of learning. In. J.W. Cotton & R. Klatzky (eds.), *Semantic Factors in Cognition*. Hillsdale, NJ: Erlbaum.

Russell, T. L. (1999). *The no significant difference phenomenon*. Chapel Hill: Office of Instructional Telecommunications, University of North Carolina.

Rummel, R. J. (1970). *Applied factor analysis*. Evanston, ILL: Northwestern University Press.

Salden, R.J.C.M., Paas, F., Broers, N.J. & Van Merriënboer, J.J.G. (2004) Mental efficiency as a determinant for the dynamic selection of learning tasks in aviation training, *Instructional Science 32*(1–2) 153–172.

SAS (2002-2003). SAS 9.1.3 Service Pack 2 for Windows [Computer program] Cary, NC: SAS Institute Inc.

SAS (2007a). *SAS Macro Programs for Statistical Graphics*. Retrieved November 4, 2007 from http://ftp.sas.com/samples/A56143

SAS (2007b). *Macro to test multivariate normality*. Retrieved November 4, 2007 from http://support.sas.com/kb/24/983.html

Schacter, D. L. (1987). Implicit memory: history and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(3), 501-518.

Scheiter, K, Gerjets, P. & Schuh, J. (2004). The impact of example comparisons on schema acquisition: do learners really need multiple examples? In Y. B., Kafai, B. Sandoval, N. Enyedy, A. S. Nixon & F. Herrera (Eds.), In the *Proceedings of the 6th International Conference of the Learning Sciences*. Mahwah, NJ: Erlbaum

Schoenfeld, A. H. (2004). The math wars. *Educational Policy, 18*(1) 253-286.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: Detection, search, and attention. *Psychological Review, 84*(1) 1-66

Schumaker, R.E., & Lomax, R.G. (2004) *A Beginner's Guide to Structural Equation Modeling*. Mahwah , NJ: Lawrence Erlbaum Associates.

Schunn, C. D. & Anderson, J. R. (2001). Acquiring expertise in science: Explorations of what, when, and how. In K. Crowley, C. D. Schunn, & T. Okada (Eds.) *Designing for science: Implications from everyday, classroom, and professional settings*, (pp. 83-114). Mahwah, NJ: Lawrence Erlbaum Associates.

Scoville, W.B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, 20, 11-21.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (Complete Samples). *Biometrika 52*(3/4) 591-611.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review, 84*(1), 127-190.

Skinner, B. F. (1958). Teaching Machines. *Science, 128* (3330) 969-977.

Simon, H. A., & Gilmartin, K. (1973). A simulation of memory for chess positions' *Cognitive Psychology 5* (1), 29-46.

Squire, L.R. (1986). Mechanisms of Memory. *Science, 232*, (4758) 1612-1619.

Squire, L.R. (1992). In the public domain memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological Review.99*, (2) 195-231.

Squire L. (1993). Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychological Review. 99* (2) 195-231.

Squire, L. R., & Zola, S. M. (1996). Structure and function of declarative and nondeclarative memory systems. *Proceedings of the National Academy of Sciences, 93*, 13515-13522.

Soloway, E., Guzdial, M., & Hay, K. H. (1994). Learner-centered design: The challenge for HCI in the 21st century. *Interactions, 1*(2), 36-48.

Sternberg, R.J. (2002). *Cognitive Psychology*. (3rd ed.) Belmont, CA: Wadsworth Publishing

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (4th Ed.). Hillsdale, NJ: Erlbaum

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257-285.

Sweller, J. (1993). Some cognitive processes and their consequences for the organisation and presentation of information. *Australian Journal of Psychology 45*(1) 1-8

Sweller, J. (2002). Visualisation and Instructional Design. *In the Proceedings of the International Workshop on Dynamic Visualizations and Learning*, Tübingen, Germany, retrieved November 9, 2008 from http://www.iwm-kmrc.de/workshops/visualization/sweller.pdf

Sweller, J. (2003). Evolution of human cognitive architecture. *In B. Ross (Ed.), The Psychology of Learning and Motivation.* San Diego: Academic Press.

Sweller, J. (2006). The worked-example effect and human cognition. *Learning and Instruction. 16*(2) 165-169.

Sweller, J. & Chandler, P. (1991). Evidence for Cognitive Load Theory. *Cognition and Instruction 8*(4) 351-362.

Sweller, J. and Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12*(3) 185-233.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2(1), 59-89.

Sweller, J. (1993). Some cognitive processes and their consequences for the organisation and presentation of information. *Australian Journal of Psychology. 45*(1) 1-8

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, 12(3), 185-233.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2(1), 59-89.

Sweller, J., Van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10, 251-296.

Tabachnick, B. G., Fidell, L. S. (2001). *Using multivariate statistics.* (4th ed.). Boston, MA: Allyn and Bacon.

Taplin, P. S., and Reid, J. B. Effects of instructional set and experimenter influence on observer reliability. *Child Development, 44*(3) 547-554.

Tarmizi, R.A. and Sweller, J. (1988). Guidance during mathematical problem solving. *Journal of Educational Psychology*, 80 (4) 424-436

Techsmith (2006). *TechSmith Camtasia Studio 4.0* [Computer program] Okemos, MI

Techsmith (2004). *TechSmith Morae 1.0.1* [Computer program] Okemos, MI

Thompson R, Kim J. (1996). Memory systems in the brain and localization of memory. *Proceedings of the National Academy of Sciences of the United States of America*. 93, 13438-13444

Tuovinen, J. E. & Paas, F. G. W. C. (2004). Exploring multidimensional approaches to the efficiency of instructional conditions. *Instructional Science 32*(1-2) 133–152.

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology*, 91(2), 334-341.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Van Lehn, K. (1996). Cognitive skill acquisition. *Annual Review of Psychology, 47*, 513 539

van Gog T., Paas F., van Merriënboer, J.J.G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction 16*(2) 154-164

Van Merriënboer, J.J.G. & Paas, F. G. W. C. (1990). Automation and schema acquisition in learning elementary computer programming: Implications for the design of practice. *Computers in Human Behavior, 6*(3) 273-289

Van Merriënboer, J. J. G., & de Croock, M. B. M. (1992). Strategies for computer-based programming instruction: Program completion vs. program generation. *Journal of Educational Computing Research, 8*(3) 365-394

Van Merriënboer, J. J. G., Schuurman, J. G., de Croock, M. B. M., & Paas, F. G. W. C. (2002). Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training efficiency. *Learning and Instruction 12* (1) 11-37.

Van Merriënboer, J.J.G. and Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17* (2), 147-177

Van Merrienboer, J. (1997). *Training complex cognitive skills: A four-component instructional design model for technical training*. Englewood Cliff, New Jersey. Educational Technology Publications.

Wagner, E. D. (1994). In support of a function definition of interaction. *The American Journal of Distance Education. 8 (*2), p. 6–29.

Ward, M. and Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*(1), 1-39.

Waterson, P. E. and O'Malley, C. E. (1993): Using Animated Demonstrations in Multimedia Applications: Some Suggestions Based upon Experimental Evidence. *In: Proceedings of the Fifth International Conference on Human-Computer Interaction*. pp. 543-548.

Weisstein, E. W. (2008). Point-Line Distance--2-Dimensional. retrieved April 22, 2008 from http://mathworld.wolfram.com/Point-LineDistance2-Dimensional.html

Wilson, B.G. & Cole, P. (1996). Cognitive Teaching Models, *In D. H. Jonassen (Ed.), Handbook of research for educational communications and technology* (pp. 601-621). New York: Simon & Schuster Macmillan.

Wittrock, M. C. The learning by discovery hypothesis. *In L. S. Shulman and E. R. Keislar (Eds.), Learning by discovery: a critical appraisal*. Chicago: Rand McNally, 1966, 33-76.

Wittrock, M. C. (1974). A generative model of mathematics learning. *Journal for Research in Mathematics Education 5*(4) 181-196.

Zhu, X., & Simon, H.A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*(3), 137-166.

APPENDICES

APPENDIX A: ANIMATED DEMONSTRATION STUDIES

Authors (# subjects)	Instructional conditions Test session variables	Results
Palmiter & Elkerton (1991) (N=48) Subject matter: HyperCard 12 procedures	-Text only, -Animated demonstration, -Animated demonstration w/text Immediate test, & delayed test (1 week later) Performance time, accuracy, retention, & transfer	*Speed (Performance time)*: a significant session x media interaction F (2, 42) =7.06, p<0.003 with both demonstration groups completing tasks in significantly less time than text group at the initial test. *Accuracy*: There was a significant session x media interaction F (2, 42), p < 0.001 with the demonstration groups completing significantly more correct trials than the text group at the initial test. Retention: There was no significant difference between the groups in performance time a week later. *Transfer*: The demonstration groups completed similar task in less time during the initial session than the text group. A significant Session x Media interaction F (2, 42) = 3.64, p < 0.04 was found. A significant increase in time for the demonstration groups between sessions.
Waterson & O'Malley (1993) (N=30) Subject matter: Cricket Graph 6 procedures	-Text only, -Animated demonstration, - Combination group (Narrated demonstration) Performance time, task type (identical, similar, different)	*Performance time*: The combination group completed tasks in significantly less time than text group given identical tasks F (2, 54) =14.08, p<0.01, and similar tasks F (2, 54) = 9.85, p<0.01, but not significantly different given different tasks (p=0.07)
Lipps, Trafton, & Gray (1998) (N=64) Subject matter: Microsoft Excel (12 procedure)	-Text only, immediate test -Animated demonstration, no immediate test, All delayed test (1 week later) Accuracy, Performance time	Accuracy: F (1, 60) = 9.56, p < .005, MSE =0.01, the demonstration group significantly more accurately at the than text group acquisition session Performance time: F (1, 60) =15.88, p <.001, MSE = 0.13 the animation group performed tasks in significantly less than time than the text group (but not at delayed test)

APPENDIX B: THE POOLED-SEMESTER SOLUTION

An *a priori* power analysis suggested a sample size of *n* = 115 subjects, in order to detect a small effect size. A sample size of this magnitude required data to be collected across two semesters (the summer and fall semesters of 2007). Therefore it was important to question if this pooled dataset would affect statistical tests. To answer this question, an analysis was conducted to determine if a pooled-semester dataset was a viable solution for data collection.

The reader may recall that the demo group did not assemble the week one problem (the Mr. Potato head problem). Therefore the week two performance (the picnic problem) was chosen to compare semester subgroups, because it was the only performance in which all participants were involved. Thus a MANOVA of week two dependent variables, performance time (PT2) and accuracy (AC2), was used to compare semester subgroups.

A MANOVA makes several assumptions (assumptions of independence, normality and homoscedasticity) (Tabachnick & Fidell, 2001; Stevens, 2002). According to Stevens (2002), each of these assumptions should be considered before proceeding with a MANOVA, and Stevens provides a general procedure for assessing each of these assumptions. The next few subsections are arranged according to Stevens' general procedure. Finally, this section concludes with an assessment of the pooled-semester solution for data analysis.

Independence Assumption - Are the Observations Independent?

The first step in Steven's general procedure is concerned with the independence assumption (Stevens, 2002). Specifically, it questions if the observations are independent of one another? Stevens (2002) lists this assumption first and emphasizes that violations of this assumption are serious.

According to Glass and Hopkins (1984) “Whenever the treatment is individually administered, observations are independent. But where treatments involve interaction among persons, such as discussion method or group counseling, the observations may influence each other (p.353).” In this study, the treatments were individually administered, so according to Glass and Hopkins, it may be said that this assumption has been met, since learners were required to work alone during each performance.

The Normality Assumption

As with ANOVA, normality is an important consideration, but given multiple outcome variables multivariate normality must be assumed. Stevens (2002) stated “a statistical test for multivariate normality is still not available on SAS” (p.263). However, it has been several years since Stevens’ published this text and now a SAS macro program, the %MULTNORM macro program is available from the SAS website (SAS, 2007b).

The %MULTNORM macro allows researchers to test for multivariate and univariate normality (SAS, 2007b). It provides a Shapiro Wilk’s test (Shapiro & Wilk, 1965) for each of the variables. Shapiro and Wilk developed the *W* statistic to test for univariate normality. However, SAS (2007b) provides this statistic to help researchers to make decisions about multivariate normality. If the Shapiro Wilk’s test rejects univariate normality, this is a good indication that the sample may not be multivariate normal. Stevens (2002) also makes this point as he describes his general procedure for checking the assumptions of a MANOVA. Finally the Shapiro Wilk’s test rejects univariate normality, when the null hypothesis is rejected (when the p-value of W is found to be less than 0.05).

In addition, to the Shapiro Wilk's test, the %MULTNORM macro provides two multivariate statistics, Mardia skewness β_1p and Mardia kurtosis β_2p. These statistics are based on several articles by Mardia (Mardia, 1970; Mardia 1975). When considering skewness or kurtosis, if the null hypothesis, is rejected (in this case at the p=0.05 level), then this suggests that the data set is multivariate non-normal (Keselman, 2005).

This %MULTNORM macro and its associated tests were implemented for both week two accuracy (AC2) and week two performance time (PT2) given performance on the picnic problem. Again this problem was chosen because all participants were involved. The %MULTNORM macro program revealed non-normality (violating the normality assumption) when the Shapiro-Wilks' W= 0.76, p<0.0001 for accuracy (AC2), and for performance time (PT2) the Shapiro-Wilks' was W = 0.95, p=0.0015. Mardia skewness was found to be β_1p= 146.5, p<0.0001 and Mardia kurtosis was β_2p=12.01, p<0.0001. Violations of the normality assumption have effects on power and type I error (Stevens, 2002). However, Stevens (2002) describes a MANOVA as being robust to violations of the normality assumption, with respect to type I error. Stevens also discusses skewness and kurtosis in relation to this assumption. He explains that multivariate skewness has negligible effects on power, but Olson (1974) found platykurtosis has a substantial effect on power.

The level of platykurtosis in this sample is cause for concern, for it attenuates power (Stevens, 2002), but there may be a good reason for the platykurtosis and non-normality in this dataset. Both platykurtosis and normality are affected by multivariate outliers (Tabachnick & Fidell, 2001).

Are there Multivariate Outliers?

Outliers are individual observations that differ dramatically from the rest of the observations (Glass & Hopkins, 1996). Multivariate outliers differ from the rest of the observations on two or more scales (Stevens, 2002). Figure 38 is a graphic representation of the week two dataset. This figure was generated with a SAS macro (ELLIPSES) (Friendly, 2007b).

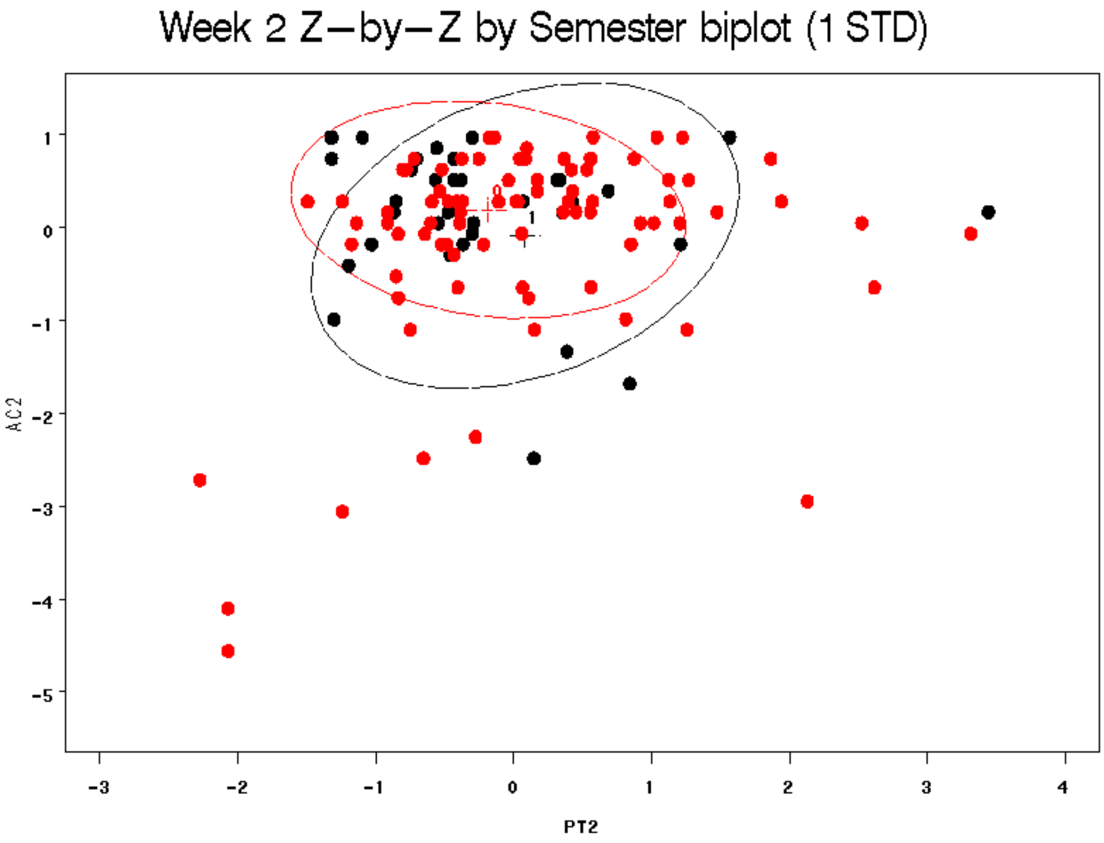


Figure 38. A bivariate plot of the week two performance time and accuracy Z-Scores

Week two accuracy (AC2) and performance time (PT2) Z-scores are compared in this plot, 0 (red) = “Summer,” and 1 (black) = “Fall.” The ellipses represent a single standard deviation (σ=1) from the mean of each group. Group means (the center point within the ellipses) are labeled with cross marks. Figure 38 allows the reader to see that some values are more extreme, relative to those that are more normal. Notice how several points lie far outside the ellipses. Some of these points (the extreme ones) are probably multivariate outliers.

Potential Multivariate Outliers

Graphics like those in Figure 38 are useful for visualizing the overall data set, but more can be done to analyze the data for outliers. The OUTLIER macro (Friendly, 2007c) was used to analyze the week two data set to detect multivariate outliers given the two semesters sub-groups. The OUTLIER macro uses "multivariate trimming," a procedure first described by Gnanadesikan and Kettenring (1972). Multivariate trimming trims potential multivariate outliers in a series of iterative passes. The OUTLIER macro made seven passes to trim potential outliers. To arrive at a correct number of passes, the researcher chooses a relatively low number, and then that number is increased by one, until no new outliers are found (Friendly, 2007c).

Table 31 lists the potential outliers. The OUTLIER macro isolated these observations because the probability of their squared Mahalanobis distances, D^2 (DSQ) was less than 0.05 (Friendly, 1991). Researchers may remove outliers from a sample, but if these observations are deemed to be a part of the population, the researcher may retain those values (Tabachnick and Fidell, 2001). However, if the researcher chooses to retain outliers, Tabachnick and Fidell advise them to reduce the impact of the outliers, by transforming the dataset.

Table 31

Potential multivariate outliers

Obs	ID	Group	PT2	AC2	DSQ	prob
1	45	2	1700	48	6.065	0.0482
2	23	1	1811	41	6.088	0.04764
3	61	3	576	48	6.535	0.03811
4	40	2	571	48	6.6	0.03689
5	25	1	1694	38	6.874	0.03216
6	27	1	629	36	6.901	0.03173
7	15	1	781	35	7.246	0.02671
8	29	1	1852	48	8.154	0.01696
9	72	3	979	34	8.303	0.01574
10	20	1	1187	34	8.553	0.01389
11	98	4	1983	46	8.957	0.01135
12	12	1	2017	42	9.371	0.00923
13	46	2	1406	34	9.854	0.00725
14	43	2	1206	33	10.635	0.00491
15	110	4	788	33	10.846	0.00441
16	96	4	2275	40	16.613	0.00025
17	118	4	582	31	16.784	0.00023
18	97	4	827	30	17.742	0.00014
19	111	4	1518	31	17.793	0.00014
20	91	3	1225	30	18.04	0.00012
21	2	1	1715	30	23.236	9E-06
22	28	1	1330	28	24.692	4E-06
23	13	1	2316	34	26.665	2E-06
24	48	2	2625	39	28.251	1E-06
25	39	2	2682	41	28.395	1E-06
26	62	3	1531	25	37.43	0
27	19	1	1036	20	55.317	0
28	53	2	869	18	65.183	0
29	36	2	1223	18	66.467	0
30	22	1	152	16	81.725	0
31	18	1	608	13	94.623	0
32	104	4	2101	14	105.327	0
33	122	3	242	4	162.198	0
34	123	1	242	0	196.277	0

Stevens' (2002) procedure allows for outliers to be retained in the dataset, but he advises researchers to consider transforming the dataset in order to protect Box's M test (the next step in the process).

Transformations

Transformations were performed (See Figure 39). Compare the upper and lower panels. As this figure shows performance time was positively skewed, and accuracy was negatively skewed. The reader may recall that Mardia skewness was found to be β_1p= 97.65, p<.0001 and Mardia kurtosis was β_2p=9.00, p<.0001.

Stevens (2002) advises researchers to transform positively skewed data, by using an x = $\sqrt{x}$ transformation, in this case TPT2 = $\sqrt{PT2}$ (where TPT2=transformed week two performance time, and PT2= performance during week two).

Rummel (1970) gave several data transformations for negatively skewed data (like the accuracy data). Rummel's suggestion of adding a constant (c) to the log of the variable, x =log (x+c), was found to be the best solution. So, Rummel's transformation became TAC2 = log (50-AC2), where, TAC2 = the transformed week two accuracy score during week two and AC2 = the accuracy score. Again these transformations were made in order to protect Box's M test from non-normality.

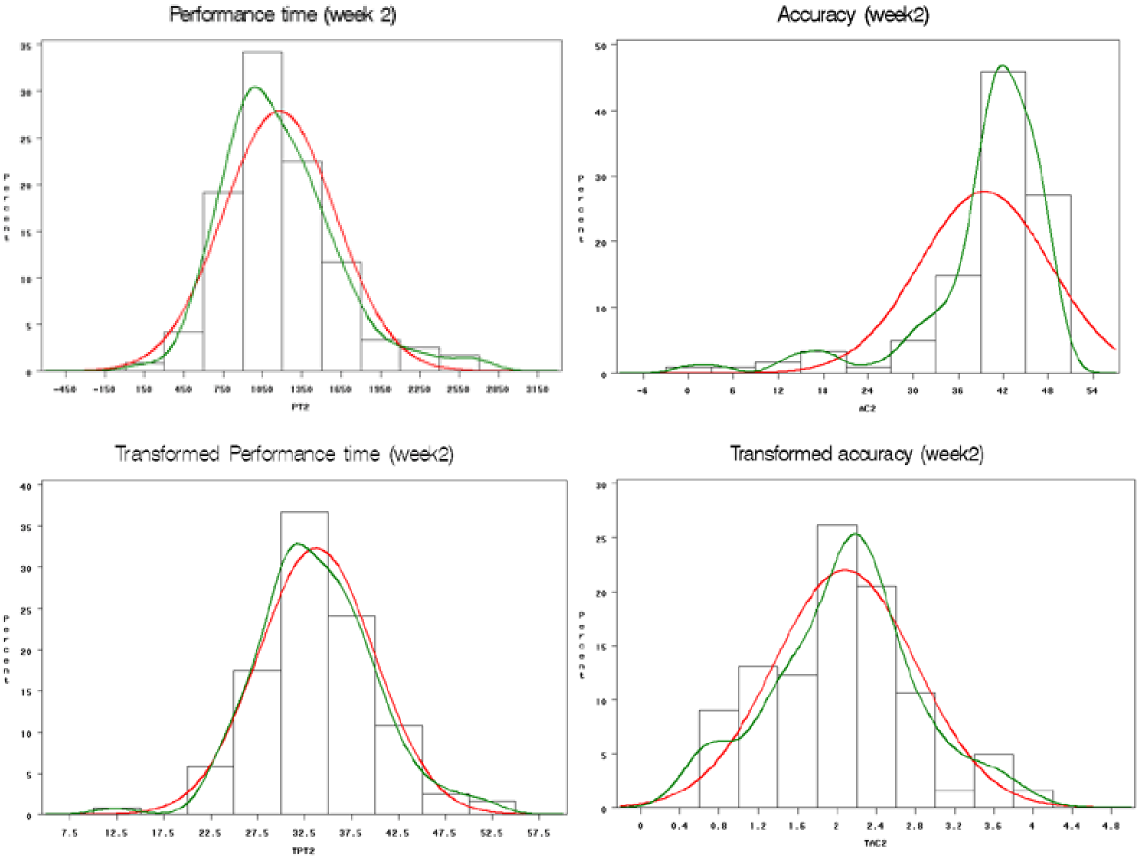


Figure 39. Week two histograms demonstrating the effects of transformations

Normality Following Transformation

The %MULTNORM macro program was run once again following transformations. Although transformations made a difference this macro again revealed non-normality when the Shapiro-Wilks' *W*= 0.95, *p*=0.0012 for accuracy (AC2), and for performance time (PT2) the Shapiro-Wilks' was *W* = 0.98, *p*=0.42. Mardia skewness was found to be β_1p= 21.11, *p*=0.0003 and Mardia kurtosis was β_2p=3.22, *p*=0.0013. Even though this macro revealed the normality assumption had been violated, a MANOVA is robust to violations of this assumption (Stevens, 2002). So this analysis continued to investigate the assumptions of the MANOVA, to consider this dataset with Box's M test.

The Homoscedasticity Assumption - Are the Matrices Homogeneous?

Box's M test is used to test the assumption of "homoscedasticity" (Box, 1954) and is a generalized version of Bartlett's test (Bartlett, 1937). When Box's M test is significant or heterogeneous, the group variance-covariance matrices differ (e.g. $\Sigma_1 \neq \Sigma_2$) violating the homoscedasticity assumption. So for a sample to meet the homoscedasticity assumption, the variance-covariance matrices should not be significantly different (e.g. $\Sigma_1 = \Sigma_2$).

Box's M test was performed, and it was found that the variance-covariance matrices were significantly different or heterogeneous, as $X^2(3, N=122) = 8.31$, $p=0.04$, $\varphi=0.26$. As a result, this overall dataset failed to meet the homoscedasticity assumption even after transforming the data. This leaves the analysis with little choice but to remove the outliers (listed in Table 31) from the dataset. This decision was not made lightly. The homoscedasticity assumption is a necessary requirement of a MANOVA. Therefore a dataset without the multivariate outliers in Table 32 is used for the remainder of the study.

Removal of the Outliers

The OUTLIER macro (Friendly, 2007c) was again used to identify multivariate outliers. Next the output from this macro was used in a SAS data step, to actually remove them from the dataset. The new dataset, $n=88$, included 28 observations from the summer semester, and 60 from the fall semester. The group composition of these outliers is demo =14, demo+practice=8, demo2+practice=5, and finally practice=7. This was somewhat troubling as it made for an unequal reduction in groups.

APPENDIX B (CONTINUED)

Table 32

Multivariate outliers

Obs	ID	Group	PT2	AC2	DSQ	prob
1	45	2	1700	48	6.065	0.0482
2	23	1	1811	41	6.088	0.04764
3	61	3	576	48	6.535	0.03811
4	40	2	571	48	6.6	0.03689
5	25	1	1694	38	6.874	0.03216
6	27	1	629	36	6.901	0.03173
7	15	1	781	35	7.246	0.02671
8	29	1	1852	48	8.154	0.01696
9	72	3	979	34	8.303	0.01574
10	20	1	1187	34	8.553	0.01389
11	98	4	1983	46	8.957	0.01135
12	12	1	2017	42	9.371	0.00923
13	46	2	1406	34	9.854	0.00725
14	43	2	1206	33	10.635	0.00491
15	110	4	788	33	10.846	0.00441
16	96	4	2275	40	16.613	0.00025
17	118	4	582	31	16.784	0.00023
18	97	4	827	30	17.742	0.00014
19	111	4	1518	31	17.793	0.00014
20	91	3	1225	30	18.04	0.00012
21	2	1	1715	30	23.236	9E-06
22	28	1	1330	28	24.692	4E-06
23	13	1	2316	34	26.665	2E-06
24	48	2	2625	39	28.251	1E-06
25	39	2	2682	41	28.395	1E-06
26	62	3	1531	25	37.43	0
27	19	1	1036	20	55.317	0
28	53	2	869	18	65.183	0
29	36	2	1223	18	66.467	0
30	22	1	152	16	81.725	0
31	18	1	608	13	94.623	0
32	104	4	2101	14	105.327	0
33	122	3	242	4	162.198	0
34	123	1	242	0	196.277	0

In addition to providing a table of potential outliers, the OUTLIER macro calculates Mahalanobis distances (D_i) for each item in the dataset, then plots them as a squared distance (DSQ) relative to the χ^2 Quantile (SAS, 2007a) (See Figure 40). Outliers in this plot are substantially above the blue line (Friendly, 1991).

However a chi square plot is also subject to the effects of outliers (Friendly, 1991). The dotted blue line (the expected value of the χ^2 Quantile) is not level. This is because this line is being influenced by the outliers in the upper right-hand corner of the plot (Friendly, 1991). Friendly was aware of this scenario and designed the OUTLIER macro to use "multivariate trimming."

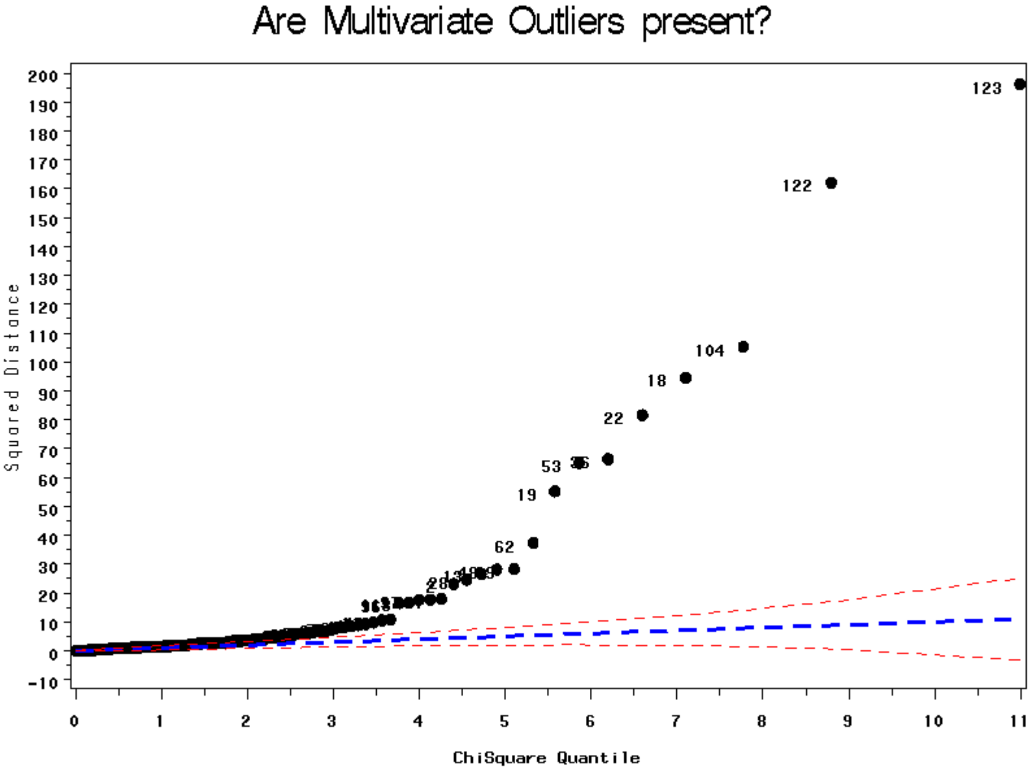


Figure 40. Potential multivariate outliers

The OUTLIER macro trims potential multivariate outliers in a series of iterative passes (7 passes in the current study) (Friendly, 1991). Thus these values were trimmed from the data set using a SAS data step, to produce a similar plot, free from the effects of outliers. Figure 41 is a plot of this dataset, with these values trimmed from the dataset, however the data are represented as a detrended quantile-quantile or QQ plot, (prepared using Friendly's cqplot macro, Friendly, 2007e). Notice how all values now lie within the confidence bands (the dotted red lines).

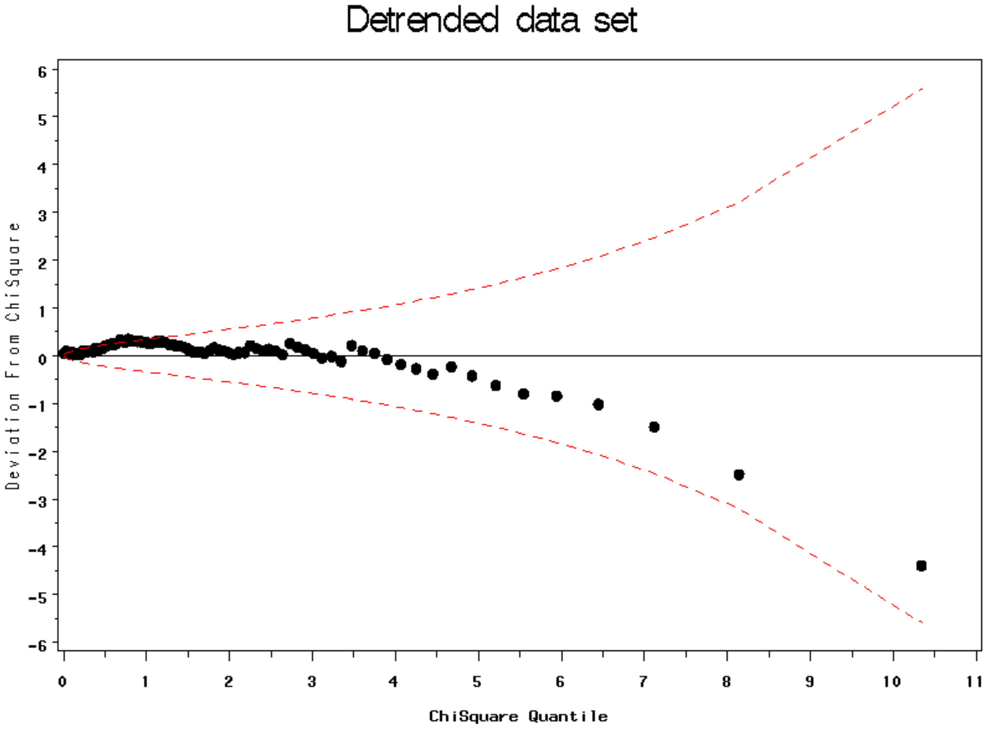


Figure 41. Detrended QQ plot, the dataset after outlier removal

After outliers were removed, the %MULTNORM macro was revisited. This was to test for the normality of the dataset without outliers, and this new dataset revealed a different set of results. The macro revealed non-normality for week two accuracy (AC2) because there was a Shapiro-Wilks' *W*= 0.94, *p*=0.0005, but week two performance time (PT2) was normal since the Shapiro-Wilks' was *W* = 0.96, *p*=0.07. However, skewness and kurtosis were much closer to normality, since Mardia skewness was β_1p= 2.05, p=0.72 and Mardia kurtosis was β_2p=-1.99, p=0.0467 (See Figure 42).

Earlier it was discussed that Olson (1974) found that kurtosis does have an effect on power, and given that this is the case, transformations were implemented. The distribution of accuracy in Figure 42 is positively skewed without the outliers.

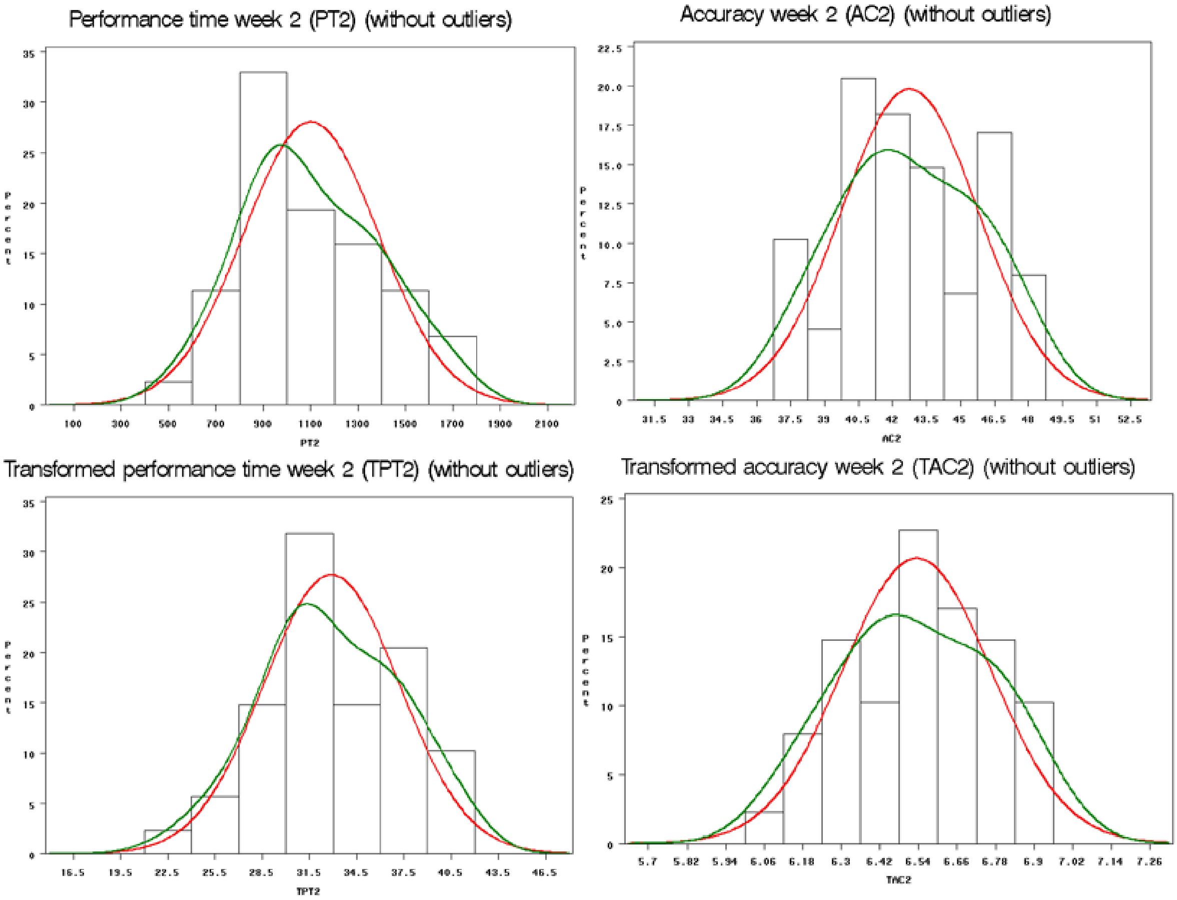


Figure 42. Week two histograms demonstrating the effects of transformations

So given this new dataset both variables required an x = $\sqrt{x}$ transformation, in this case TPT2 = $\sqrt{PT2}$ (where PT2= performance during week two, & TPT2=transformed week two performance time) and TAC2 = $\sqrt{AC2}$ (where AC2=

accuracy during week two, & TAC2=transformed week two accuracy). This new outlier free dataset has a somewhat different distribution.

Normality Following Transformation

The %MULTNORM macro program was run once again following transformations. This macro again revealed univariate non-normality (violating the normality assumption) because the Shapiro-Wilks' W= 0.94, p=0.0005 for transformed accuracy (TAC2), but for transformed performance time (TPT2) the Shapiro-Wilks' was normal because W = 0.97, p=0.23. In addition, Mardia skewness was found to be normal β_1p= 0.86, p=0.93 and Mardia kurtosis was also found to be normal β_2p=-1.90, p=0.06. Even though this macro revealed univariate non-normality, or that the assumption had been violated, a MANOVA is robust to violations of the normality assumption (Stevens, 2002). So this analysis continued to assess the assumptions of this MANOVA, to consider this dataset with Box's M test.

The Homoscedasticity Assumption

Following transformations, Box's M test was performed with this dataset (without the multivariate outliers), and it was found that the variance-covariance matrices were not significantly different, or were homogeneous, since X^2(3, N=88) =4.50, p=0.21, φ=0.23. Since they were found to be homogenous, there is no evidence that the homoscedasticity assumption has been violated given this dataset, so it is reasonable to consider a MANOVA.

The Decision to Use a MANOVA

In summary, this analysis has shown the sample was non-normal, but departures from normality have a limited effect on Type I error (Stevens, 2002, Mardia, 1971). Mardia kurtosis was originally found to be β_2p=12.01, p<.0001. However, this was mainly due to a group of multivariate outliers, which were subsequently removed from the initial dataset, to produce an outlier free dataset n=88. Box's M test was conducted with the outlier free dataset and it was found that the variance-covariance matrices were not significantly different, or homogeneous as $X^2(3, N=88) = 4.50$, $p=0.21$, $\varphi=0.23$. Therefore it was reasonable to continue with a MANOVA of the pooled semester dataset especially given MANOVA is robust to departures from normality (Stevens, 2002).

APPENDIX C: THE ACQUISTION PHASE MANOVA

This MANOVA was analyzed according to Stevens' general procedure for assessing the assumptions of a MANOVA (Stevens, 2002). Therefore the next few sections several questions will address the assumptions of the acquisition phase (week one) MANOVA.

Are the Observations Independent?

When considering a MANOVA, one must first consider the independence assumption (Stevens, 2002). Earlier it was stated that each learner was required to work alone and scores were measured separately, thus according to Glass and Hopkins (1984) this sample met the independence assumption.

Is the Acquisition Phase Dataset from a Normal Population?

The next step in Steven's general procedure is to address the normality assumption (Stevens, 2002). Therefore the %MULTNORM macro was implemented and revealed that the acquisition phase dataset was non-normal (violating the normality assumption). Multivariate non-normality was revealed when the macro revealed a Shapiro-Wilks' W= 0.62, p<0.0001 for accuracy (AC1), and for performance time (PT1) the Shapiro-Wilks' was W = 0.88, p<0.0001. Mardia skewness was found to be β_1p= 66.70, p<0.0001 and Mardia kurtosis was β_2p=3.79, p<0.0001.

Analysis of the dataset with the OUTLIER macro (Friendly, 2007b) revealed an additional 20 potential multivariate outliers in the acquisition phase dataset (See Table 33). Stevens (2002) provides several reasons for finding outliers, he suggests it may be due to recording or entry errors, or an instrumentation error. Stevens also states "If, however, none of these appears to be the case, then one should not drop the outlier, but perhaps report two analyses (one including the outlier and the other excluding it)" (Stevens, 2002, p.17).

Given Stevens' suggestion, and the fact that there were a series of potential outliers in the acquisition dataset, this study will present both prospective solutions: solution one (remove the potential outliers), and solution two (retain the outliers). Both solutions are summarized in the next few sections, and then this section concludes with arguments for solution two, retaining outliers. The next section considers solution one, removing outliers.

Solution one: Removing outliers

Acquisition phase outliers were removed by first using the OUTLIER macro to identify potential multivariate outliers, those with $p<0.05$ (See Table 33). Next a SAS data step used the output from the OUTLIER macro, to remove these values, leaving 49 week one learners. The group composition following outlier removal was demo+practice $n=19$, demo2+practice $n=23$, practice $n=7$.

Solution one normality.

Once the acquisition phase outliers were removed, the normality assumption needed to be tested with the dataset. Normality was tested with the %MULTNORM macro. Multivariate non-normality was revealed when Mardia kurtosis was found to be β_2p=-2.33, p=0.02 and Mardia skewness was found to be β_1p= 2.41, p=0.66. This macro also revealed a Shapiro-Wilks' W= 0.63, $p<0.0001$ for accuracy (AC1), and for performance time (PT1) the Shapiro-Wilks' was W = 0.95, p=0.04.

Table 33

Potential acquisition phase outliers

Observation	ID	group	AC1	PT1	DSQ	probability
1	69	3	24	1088	6.568	0.037474
2	102	4	24	1146	7.793	0.020309
3	37	2	21	1203	11.76	0.002795
4	105	4	22	1755	26.471	0.000002
5	112	4	24	1788	29.103	0
6	106	4	22	2107	43.231	0
7	99	4	13	1059	108.762	0
8	101	4	12	993	130.059	0
9	79	3	11	401	156.189	0
10	119	4	11	168	160.342	0
11	92	3	9	598	209.017	0
12	120	4	9	467	210.438	0
13	84	3	6	1155	306.232	0
14	113	4	4	368	386.521	0
15	93	3	4	325	387.432	0
16	78	3	4	246	389.273	0
17	33	2	4	137	392.167	0
18	77	3	1	57	525.092	0
19	95	3	1	51	525.294	0
20	107	4	0	698	557.854	0

Solution one transformations.

Since this dataset is not normal, data transformations were performed (Stevens, 2002) (See Figure 43). Because performance time was positively skewed an x = $\sqrt{x}$ transformation was used, in this case TPT1 = $\sqrt{PT1}$. Also since accuracy was negatively skewed an x =log (x+C) transformation was used, so given accuracy this transformation became TAC1 = log (25-AC1).

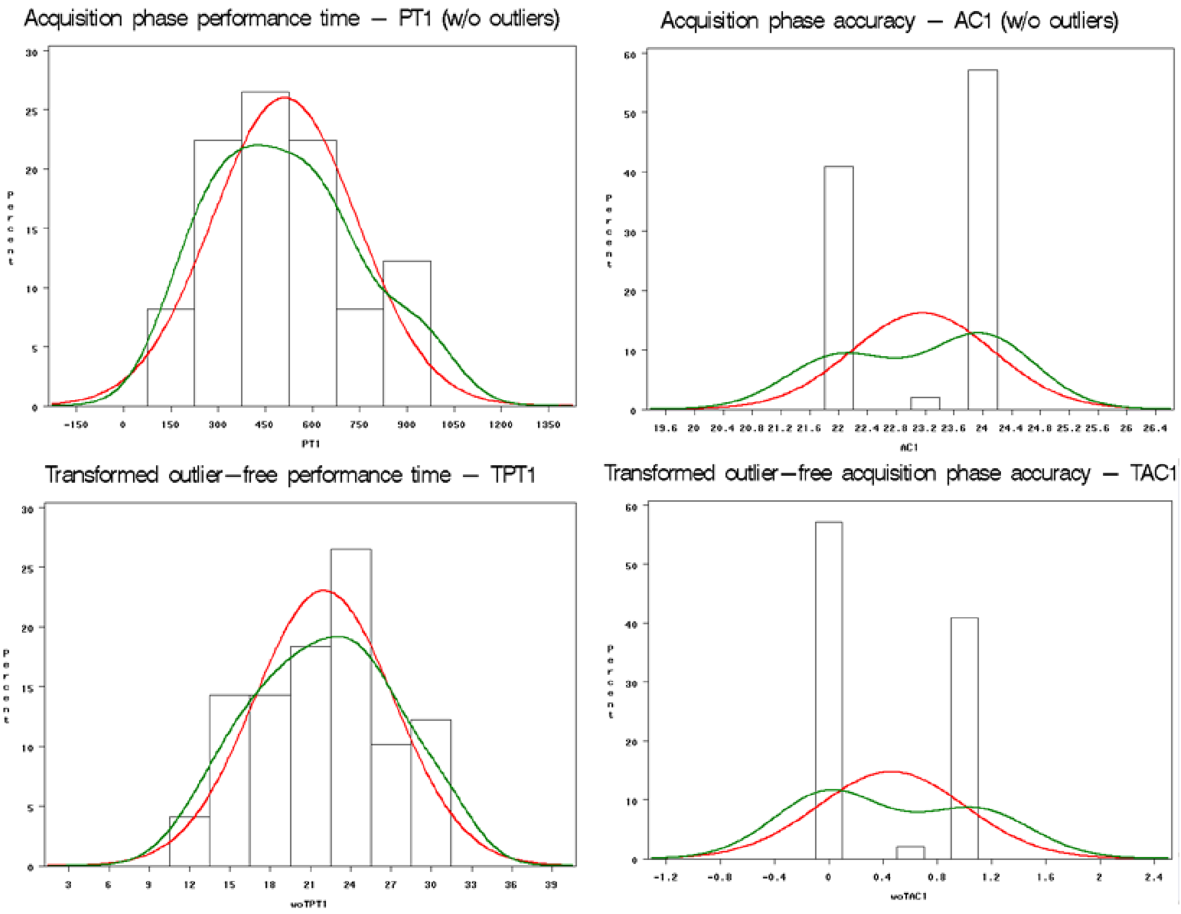


Figure 43. Week 1 histograms demonstrating the effects of transformations

Normality following transformation.

The %MULTNORM macro program was run once again following transformations and revealed univariate non-normality (violating the normality assumption) because the Shapiro-Wilks' W= 0.63, p<0.0001 for transformed accuracy (TAC2), although transformed performance time (TPT2) exhibited normality as Shapiro-Wilks' W = 0.96, p=0.22. Mardia skewness was found to be normal β_1p= 1.06, p=0.90 but Mardia kurtosis was also found to be non-normal β_2p=-2.53, p=0.01. Even though the normality assumption had been violated, a MANOVA is robust to violations of the normality assumption (Stevens, 2002). So this analysis continued to assess the assumptions of this MANOVA, to consider this dataset with Box's M test.

Solution one's Box's M Test.

Once transformations were completed, it was then possible to test this dataset with Box's M test. When Box's M test was implemented it was found that the variance-covariance matrices were not significantly different since $X^2(6, N=48) =1.43, p=0.96, \varphi=0.17$. Given this was the case there was no evidence that the homoscedasticity assumption was violated.

The solution one MANOVA.

Since this dataset met the assumption of homoscedasticity, a MANOVA was conducted. This MANOVA indicated that there was a significant difference between the groups, since Wilks' Λ =0.76, F (2, 48) = 3.28, p = 0.01, η^2=0.24. Post hoc comparisons with Scheffé's test (p<0.025) found the demo+practice group produced the Mr. Potato head problem in significantly less time than either the demo2+practice or practice groups, but it found no significant differences between groups given accuracy (See Table 34, Figures 44, 45, & 46).

Table 34

Descriptive data for the solution one dataset

	demo	demo+practice	demo2+practice	practice
n	NA	19	23	7
Accuracy (AC1)				
M	NA	0.38	0.42	0.78
SD	NA	0.52	0.54	0.54
Performance time (PT1)				
M	NA	19.29	23.24	25.47
SD	NA	5.35	4.23	4.60

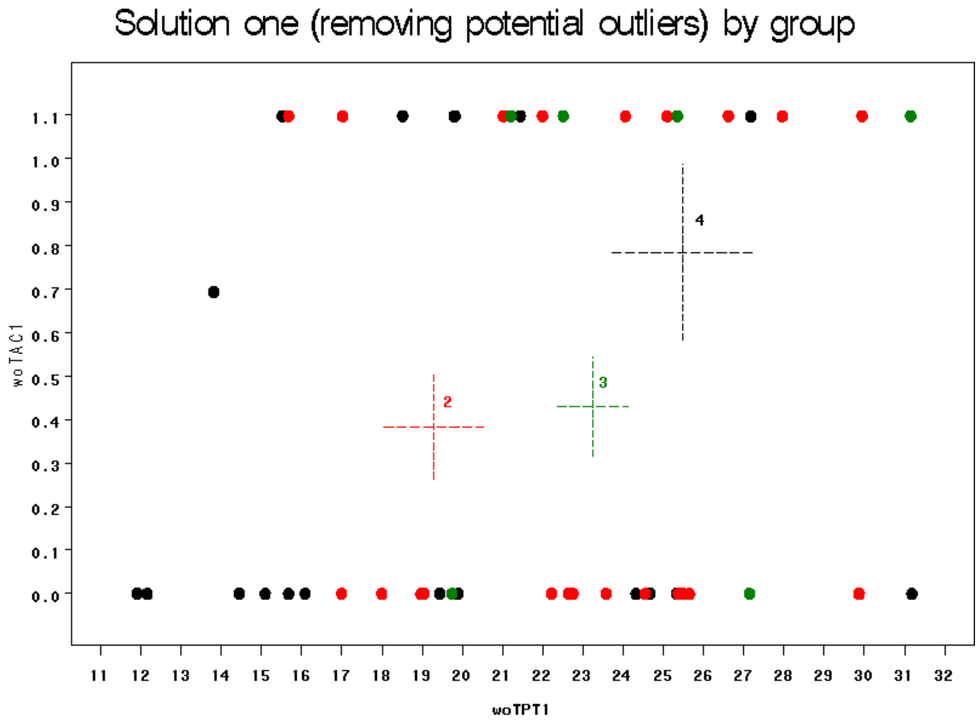


Figure 44. Solution one by group

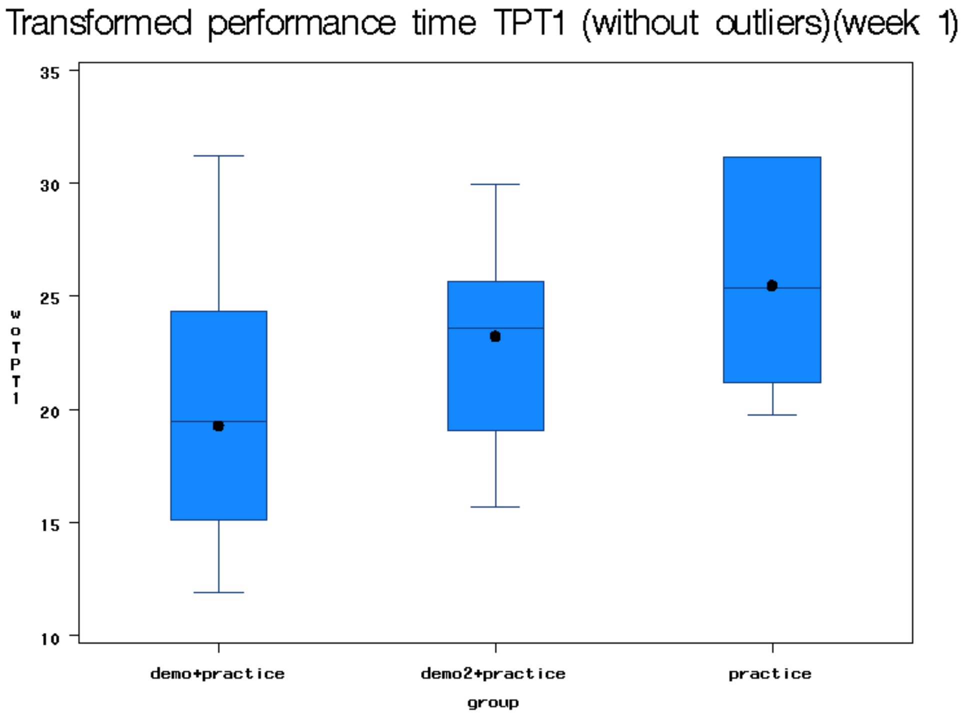


Figure 45. Transformed acquisition phase performance time TPT1 (without outliers)

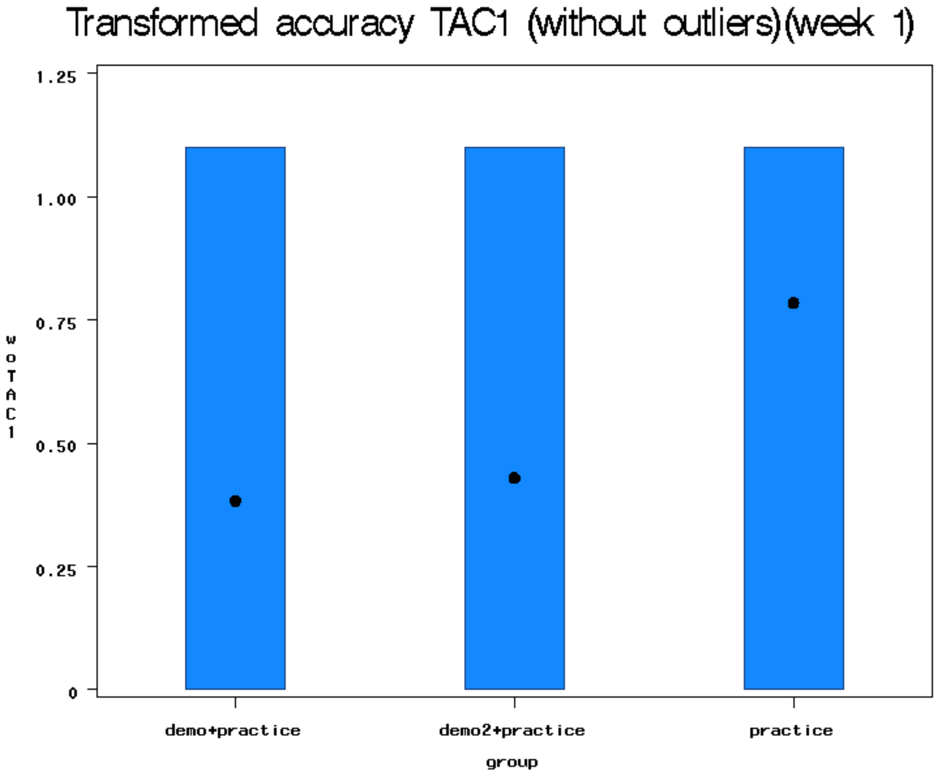


Figure 46. Transformed acquisition phase accuracy AC1 (without outliers)

Recall that earlier in this section, it was suggested that because there were potential outliers, that there were two potential solutions, one was to remove the outliers and proceed with the analysis (solution one above). The second solution was to retain the outliers in an effort to preserve power.

Solution Two: Tansforming the Dataset

This section is based upon the solution two (retaining outliers) and considers the analysis with outliers included in the dataset. Given outliers are retained Tabachnick and Fidell (2001) recommended researchers minimize the influence of outliers by transforming the data. Stevens (2002) advises researchers to transform positively skewed data (like the performance time dataset) by using an x= $\sqrt{x}$ transformation. So the transformation for performance time was TPT1 = $\sqrt{PT1}$, where TPT1 = transformed performance time (week one) and PT1 = performance time (week 1). Figure 33 shows histograms of the solution two transformed dataset.

Negatively skewed data, like the accuracy dataset may use a constant in the transformation, for an x = log (x+C) transformation (Rummel, 1970). In this case the transformation was TAC1 = log (25-AC1).

Figure 47 shows that the acquisition phase performance time (PT1) and accuracy (AC1) scores have some level of skewness and kurtosis. Kurtosis is especially evident in the accuracy data. Skewness and kurtosis was reduced when transformations were applied (compare the upper and lower panels). These transformations were implemented, to protect Box's M test from the influences of non-normality.

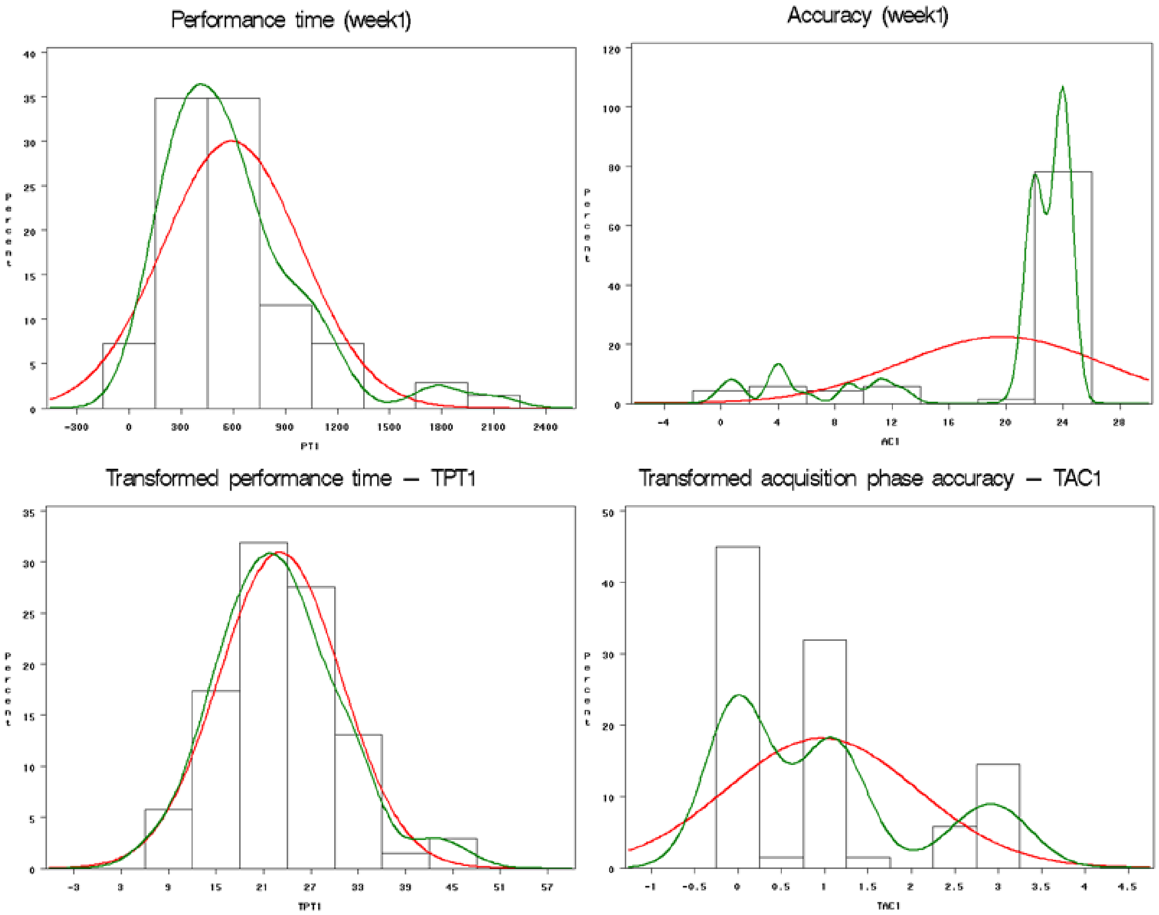


Figure 47. Solution two acquisition phase performance time and accuracy histograms

Normality following transformation.

The %MULTNORM macro program was run once again following transformations and revealed univariate non-normality since the Shapiro-Wilks' *W*= 0.76, *p*<0.0001 for transformed accuracy (TAC2), although transformed performance time (TPT2) exhibited normality as Shapiro-Wilks' *W* = 0.97, *p*=0.31. Mardia skewness was also found to be non-normal β_1p= 17.01, *p*=0.002, but Mardia kurtosis was found to be normal β_2p=0.07, *p*=0.94. Even though the normality assumption had been violated, a MANOVA is robust to violations of the normality assumption (Stevens, 2002). So the analysis continued to assess the assumptions of this MANOVA, to consider this dataset with Box's M test.

Solution two homoscedasticity.

Box's M test was performed, given the dataset which retained outliers. This test made use of the transformed dataset. It was found that the variance-covariance matrices were not significantly different, or homogeneous, $X^2(6, N=69) = 7.97$, p=0.24, φ=0.34. So there was no evidence that this dataset violated the homoscedasticity assumption, suggesting it was reasonable to consider a MANOVA.

The solution two MANOVA.

Like the solution one MANOVA, the solution two MANOVA (retaining the outliers) found that there was a significant difference between the group centroids, since Wilks' Λ=0.68, F (2, 68) =6.83, p <0.0001, η^2=0.32 (See Figures 48 & 49). The F tests for performance time and accuracy were statistically significant, as the F (2, 68) = 3.19, p=0.0478 for accuracy (AC1) and F (2, 68) =7.84 p=0.0009 for performance time (PT1). Table 35 details the acquisition phase dependent variables, by group. However, unlike the results in solution one, if the outliers were retained (solution two) this produced a different set of results, because post hoc comparisons with Scheffé's test (p<0.025) revealed that learners of both the demo+practice and demo2+practice groups assembled the Mr. Potato head problem, in significantly less time than the practice group. Even though this was the case, no significant difference between groups were found given accuracy (AC1) with Scheffé's test (p<0.025).

Table 35

Solution two results for the acquisition phase dependent variables

	demo	demo+practice	demo2+practice	practice
n		21	31	17
Transformed Performance time (TPT1)				
M	NA	19.66	22.40	28.62
SD	NA	6.35	6.28	9.01
Transformed	NA	0.56	0.99	1.44
Accuracy (TAC1)	NA	0.79	1.99	1.13
M				
SD				

The acquisition phase MANOVA.

Figure 51 is a graphic representation of the solution two dataset (retaining outliers). Group colors are demo+practice=red, demo2+practice=green, practice=black. This bivariate plot of the acquisition phase dataset includes transformed performance time and accuracy scores, and is shown by group. Since accuracy was transformed with the TAC1=log (25-AC1) transformation, the most accurate performances are at the bottom of the graph. This same rule applies to Figures 48 through 51. These figures are the same dataset with, and without potential outliers.

There are several arguments against solution. First, consider Table 36, these individual would have to be removed if solution one were chosen. It should be noted that half of the outliers in this table, are from the practice group. Secondly, even though both solutions one and two were statistically viable, given the assumptions of a MANOVA, these values may be transformed.

Table 36

Potential acquisition phase outliers (solution one)

Observation	ID	group	AC1	PT1	DSQ	probability
1	69	3	24	1088	6.568	0.037474
2	102	4	24	1146	7.793	0.020309
3	37	2	21	1203	11.76	0.002795
4	105	4	22	1755	26.471	0.000002
5	112	4	24	1788	29.103	0
6	106	4	22	2107	43.231	0
7	99	4	13	1059	108.762	0
8	101	4	12	993	130.059	0
9	79	3	11	401	156.189	0
10	119	4	11	168	160.342	0
11	92	3	9	598	209.017	0
12	120	4	9	467	210.438	0
13	84	3	6	1155	306.232	0
14	113	4	4	368	386.521	0
15	93	3	4	325	387.432	0
16	78	3	4	246	389.273	0
17	33	2	4	137	392.167	0
18	77	3	1	57	525.092	0
19	95	3	1	51	525.294	0
20	107	4	0	698	557.854	0

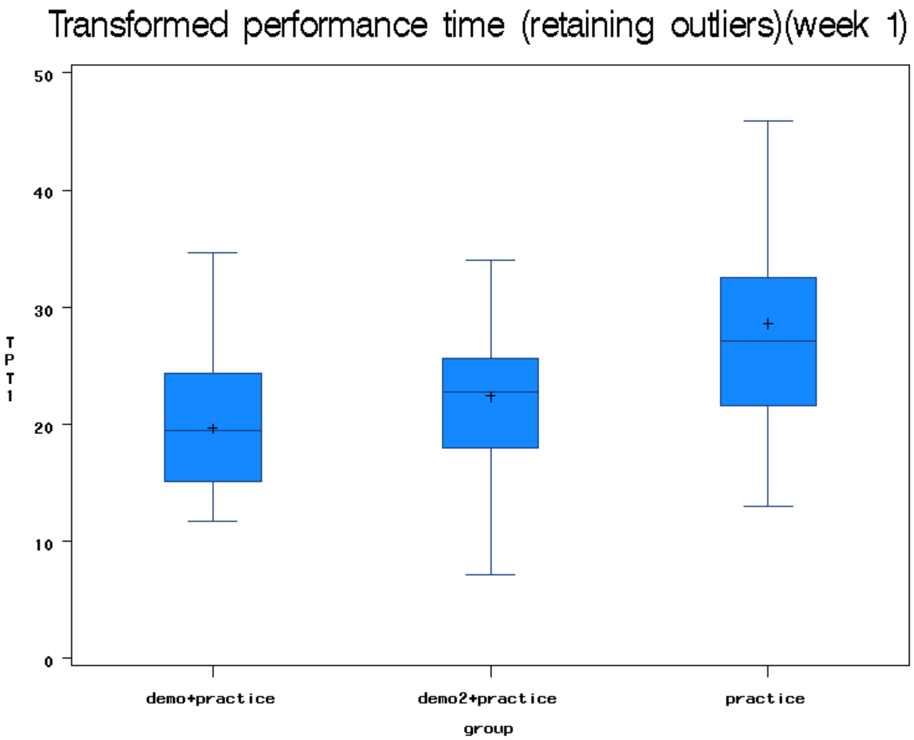


Figure 48. Acquisition phase transformed performance time (retaining outliers)

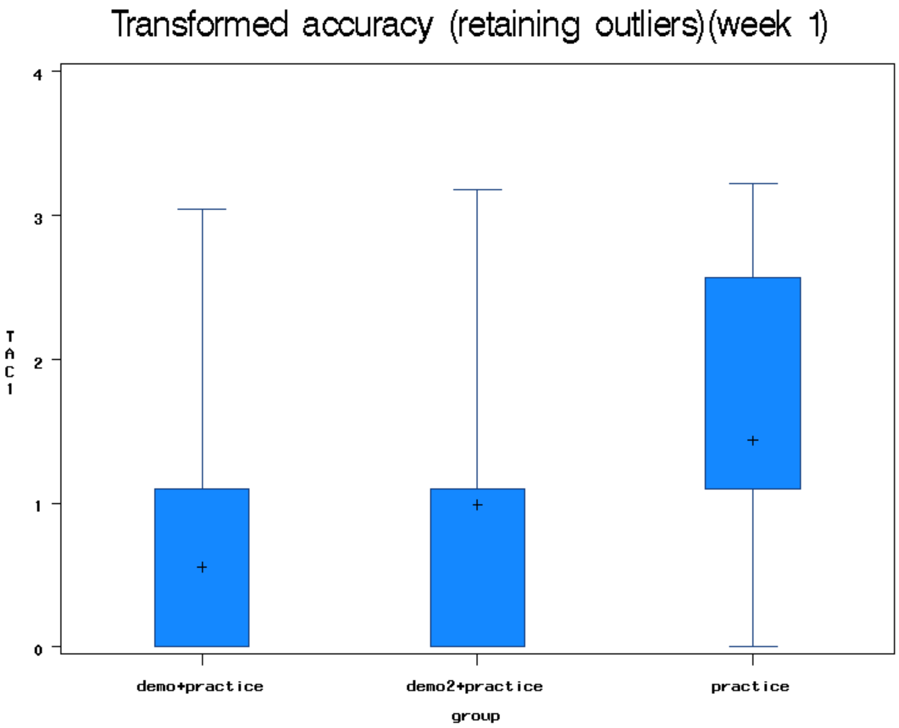


Figure 49. Acquisition phase transformed accuracy (retaining outliers)

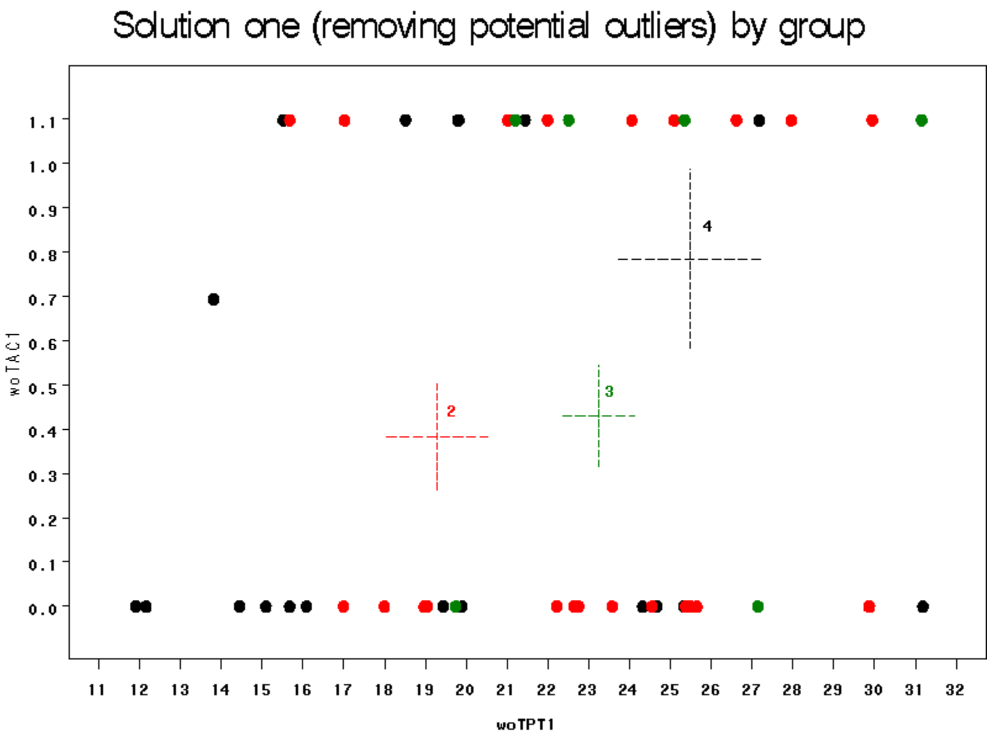


Figure 50. Solution one: without potential outliers

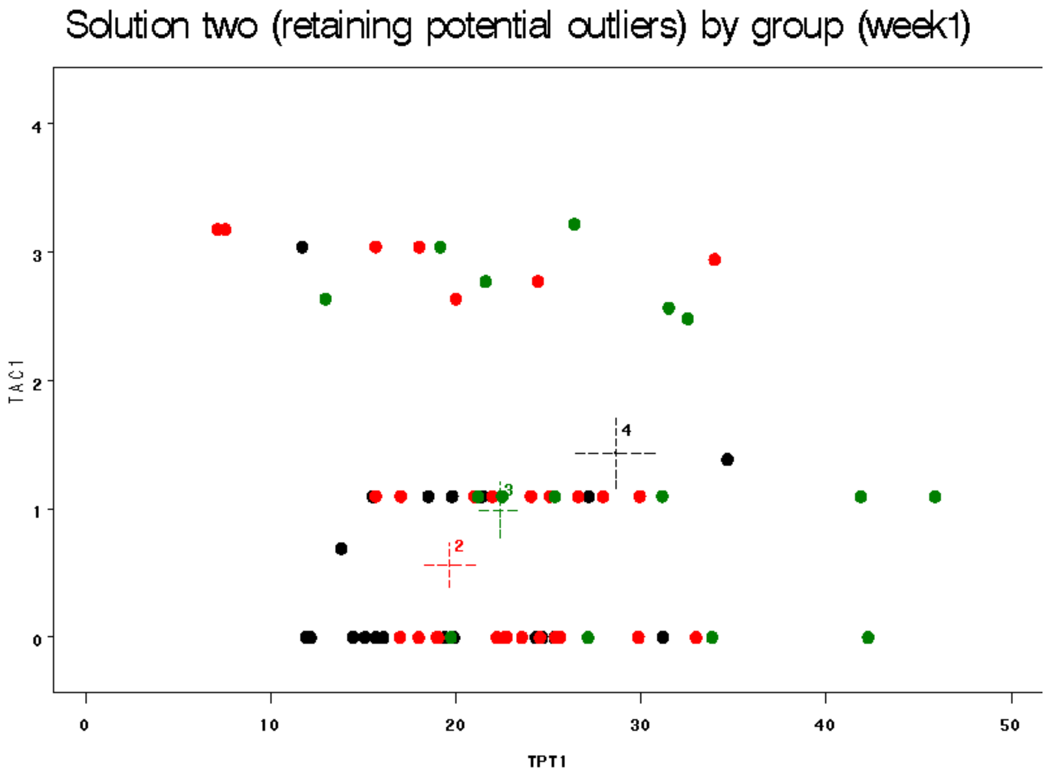


Figure 51. Solution two: retaining potential outliers

Third and perhaps the most important argument against solution one, is that the effect size for solution two was η^2=0.32, as opposed to η^2=0.24 for solution one, therefore the total proportion of variance for solution two is greater (Tabachnick & Fidell, 2001). So, reducing the number of observations ultimately decreases the power of solution one. So given each of these arguments, and even though solution one is statistically viable (given the assumptions of a MANOVA), solution two is the best fit given the dataset. Therefore solution two will be, the solution of choice, and hereafter described as the results of the acquisition phase MANOVA.

APPENDIX D: THE RENTENTION PHASE MANOVA

As with all MANOVAs in this chapter the retention phase MANOVA was analyzed according to Stevens' general procedure for assessing the assumptions of a MANOVA (Stevens, 2002). This procedure begins with the independence assumption. Since learners were required to individually assemble the picnic problem, according to Glass and Hopkins (1984) learners in this sample met the independence assumption.

The Retention Phase Normality Assumption

The %MULTNORM macro program (SAS, 2007b) revealed non-normality given the retention phase data (See Figure 38). This non-normality was revealed when the Shapiro-Wilks' W= 0.94, p=0.0005 for accuracy (AC2), and for performance time (PT2) the Shapiro-Wilks' was W = 0.96, p=0.07. Mardia skewness was found to be β_1p= 2.05, p=0.73 and Mardia kurtosis was β_2p=-1.99, p=0.05.

Since the %MULTNORM macro revealed non-normality, the OUTLIER macro (Friendly, 2007c) was used to test for multivariate outliers. As discussed earlier this OUTLIER macro revealed a total of 34 potential multivariate outliers in the week two dataset (See Table 22). Output from this macro was used to remove these outliers from the dataset.

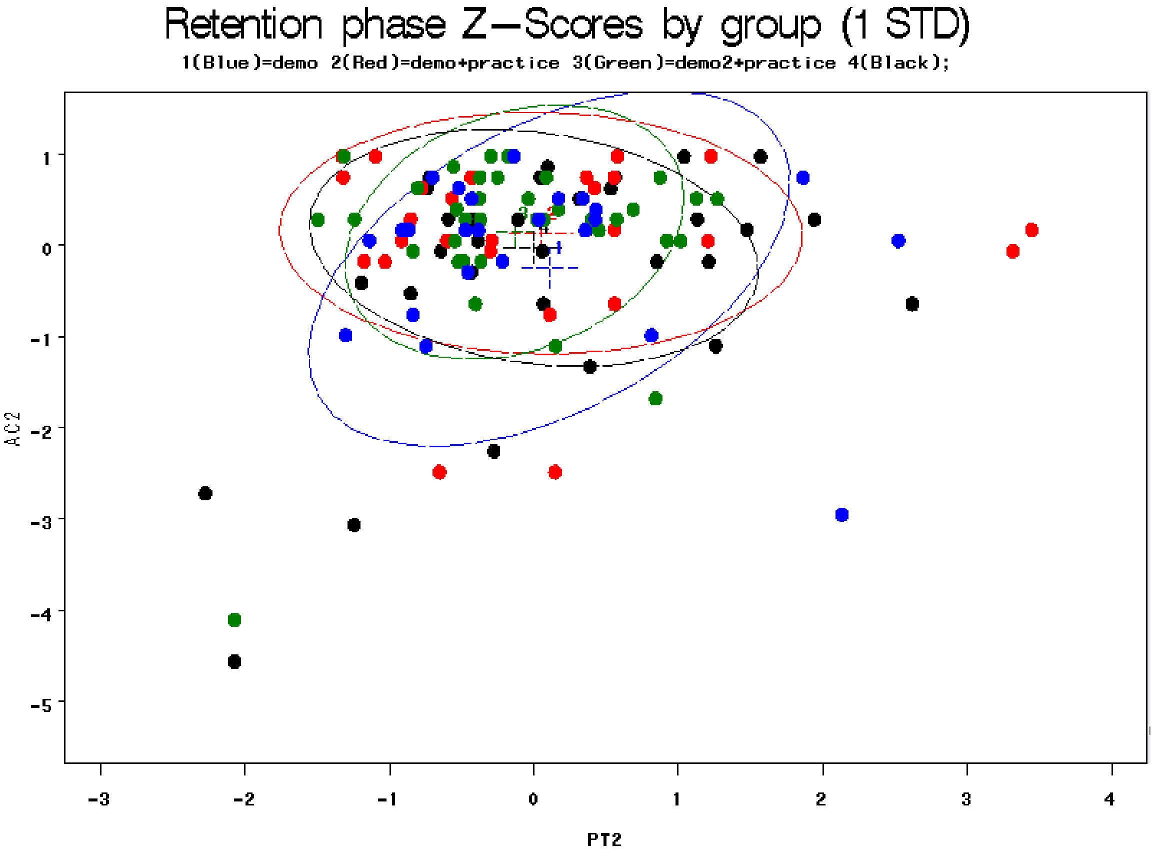
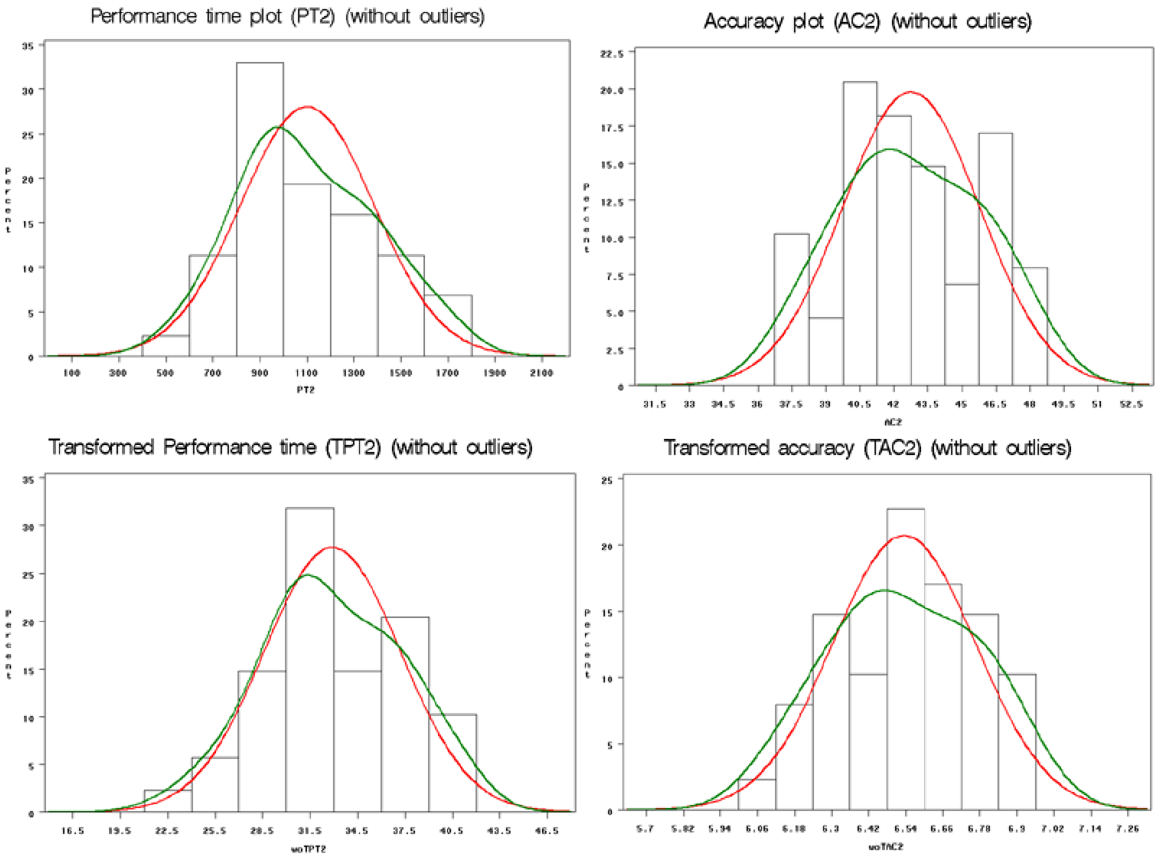


Figure 52. Retention phase Z-scores by group

Data Transformations

Once the 34 potential outliers were removed from the week two dataset, it was found that the resulting dataset was somewhat skewed (See Figure 39). Given this was the case the dependent variables (performance time and accuracy) were transformed. Because both performance time (PT2) and accuracy (AC2) were positively skewed an $x = \sqrt{x}$ transformation was implemented with both variables. These transformations were implemented, to protect Box’s M test from the influences of non-normality.


Performance time plot (PT2) (without outliers)

Accuracy plot (AC2) (without outliers)

Transformed Performance time (TPT2) (without outliers)

Transformed accuracy (TAC2) (without outliers)

Figure 53. Week two histograms demonstrating the effects of transformations

Are the Variance-covariance Matrices Homogeneous?

Once the week two variables, performance time and accuracy, were transformed, Box's M test was conducted to test for the assumption of homogeneity of the variance-covariance matrices. It was found that the matrices were not significantly different, or homogeneous, since $X^2(9, N{=}88) = 4.43$, $p{=}0.88$, $\varphi{=}0.22$. This finding shows that there is no evidence that the transformed dataset violates the homoscedasticity assumption. Given this is the case it is reasonable to consider a MANOVA.

Retention Phase Results

The overall goal of the retention phase MANOVA was to determine if group differences existed a week after initial instruction. It was hypothesized that learners in the animated demonstration conditions would out-perform learners in the practice condition. However, the results of the MANOVA found that there was not a significant difference between the group centroids, as Wilks' Λ =0.96, F (3, 87) =0.64, p =0.70, η^2=0.04 (See Figures 54 & 55). Table 37 lists the group means for each of the dependent variables transformed performance time (TPT2) and transformed accuracy (TAC2).

Table 37

Transformed performance time (TPT2) and accuracy (TAC2) by group

	demo	demo+practice	demo2+practice	practice
n	19	21	31	17
Transformed performance time (TPT2)				
M	34.10	31.92	33.29	32.09
SD	3.78	4.93	4.57	3.44
Transformed accuracy (TAC2)				
M	6.55	6.55	6.54	6.50
SD	0.26	0.25	0.22	0.21

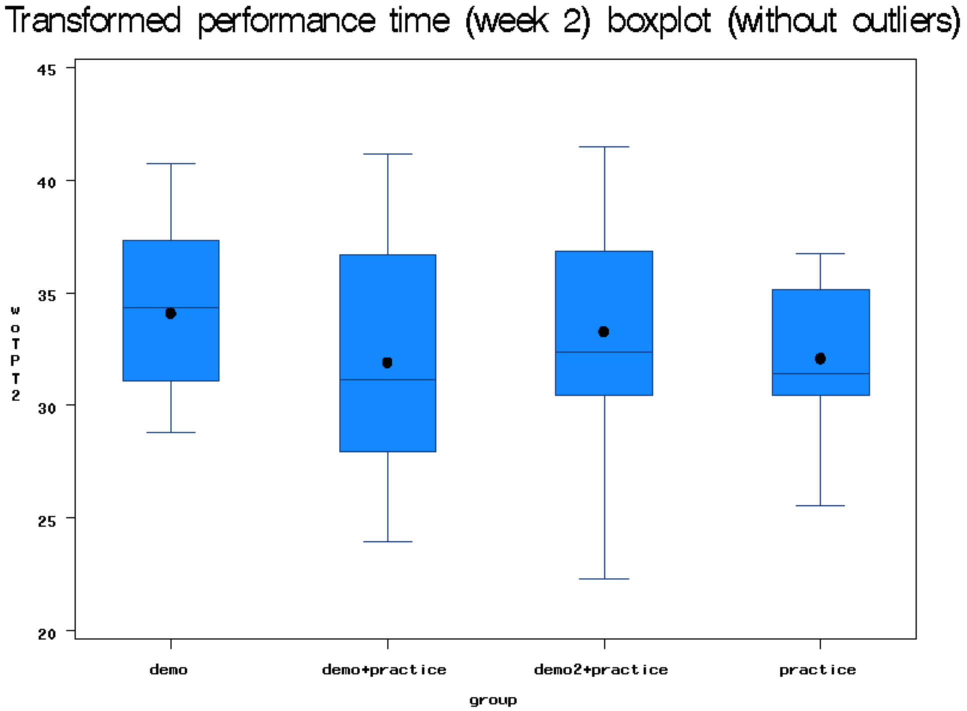


Figure 54. Transformed performance time (TPT2) without outliers

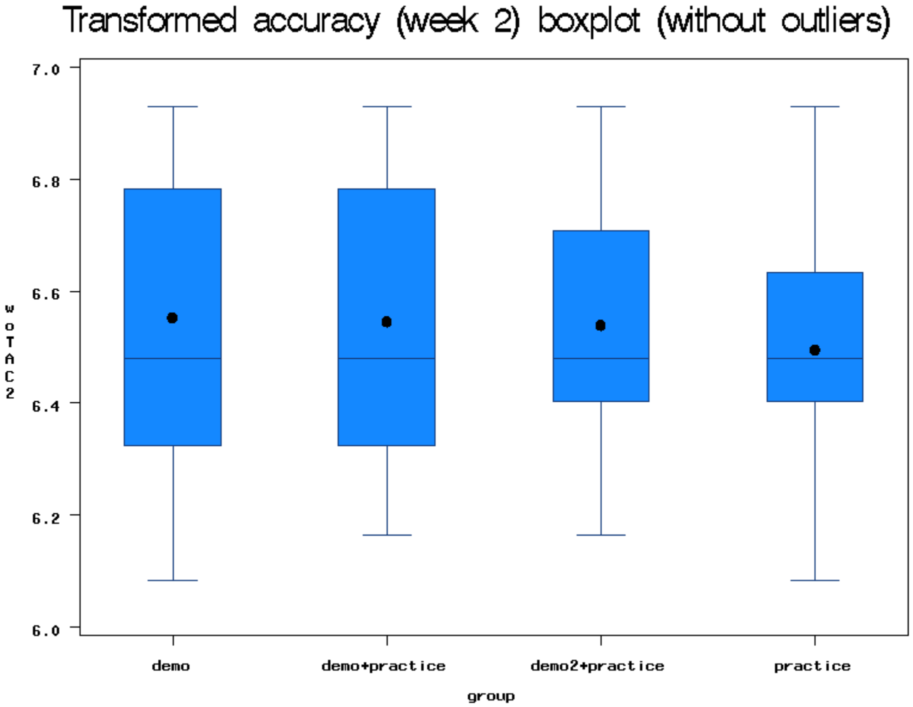


Figure 55. Transformed Accuracy (TAC2) without outliers

APPENDIX E: RELATIVE CONDITION EFFICIENCY

Relative condition efficiency was calculated for both week one (RCE1) and week two (RCE2). During each of these calculations an ANOVA was used to contrast group differences. This Appendix describes the assumptions of each of these ANOVAs.

Week One Relative Condition Efficiency (RCE1)

As with any analysis of variance, one must first consider the independence assumption (Stevens, 2002). In the current study each learner was required to work alone and their scores were measured separately, so according to Glass and Hopkins (1984) the sample met the independence assumption.

Next researchers must consider the normality assumption. In order to assess the normality of RCE1, a Kolmogorov-Smirnov test was implemented and it revealed non-normality for RCE1, as D (2, 68) =0.15, p=0.01.

Given non-normality was found, transformations were implemented. Rummel (1970) provides a series of approaches toward variable transformations. Several of his approaches were tried, where transformed week one relative condition efficiency (TRCE1) involved a constant, TRCE1= $(4\text{-RCE1})^{1/2}$.

Following variable transformations, a Levene's test compared the transformed means to find they were not significantly different, F (2, 68) =2.26, p=0.11. This finding showed that there was no evidence that the transformed dataset violated the homoscedasticity assumption, thus it was reasonable to consider an ANOVA.

Week Two Relative Condition Efficiency (RCE2)

Week two Relative Condition Efficiency (RCE2) also required an ANOVA, so the assumptions of an ANOVA were considered first.

First among these assumptions is the independence assumption (Stevens, 2002). Recall that learners were required to work alone and scores were measured separately, so again according to Glass and Hopkins (1984) this sample can be said to meet the independence assumption.

The next assumption to be considered for the RCE2 ANOVA is the normality assumption. A Kolmogorov-Smirnov test for RCE2 revealed a marginally normal distribution *D*=0.09 (3, 87), *p* = 0.054. Therefore transformations were not necessary for this variable.

Finally Levene’s test compared the means to find that they were not significantly different, *F* (3, 87) = 0.56, *p* = 0.64. This finding showed that there was no evidence that the data set violated the homoscedasticity assumption, thus it was reasonable to consider the RCE2 ANOVA.

APPENDIX F: PERFORMANCE EFFICIENCY

Performance efficiency was calculated for both the week one (PE1) and week two (PE2). During each of these calculations an ANOVA was used to contrast group differences. This Appendix describes the assumptions of each of these ANOVAs.

Week One Performance Efficiency (PE1)

Stevens (2002) advises researchers to consider the assumptions of an ANOVA, prior to running the analysis, therefore the assumptions of the PE1 ANOVA were considered. First was the independence assumption. According to Glass and Hopkins (1984) learners in this data set met this assumption because they were required to work alone and scores were measured separately. Next the normality assumption was considered, and a Kolmogorov-Smirnov test revealed non-normality *D* (2, 68) =0.15, *p*=0.01 (See Figure 56). The green line represents the sample, versus normality, the red line.

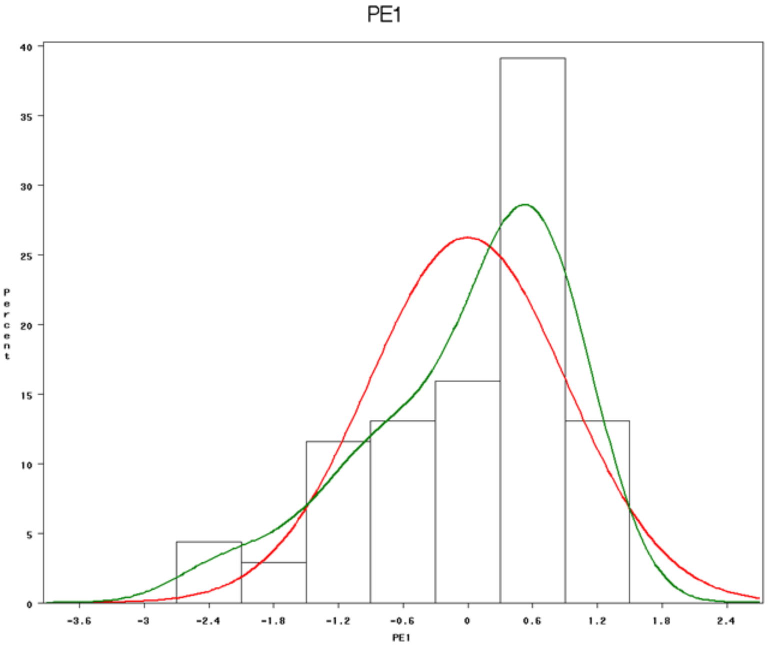


Figure 56. Week one performance efficiency

Given PE1 was negatively skewed, the distribution was subsequently transformed. Both Tabachnick and Fidell (2001) and Rummel (1970) provide several approaches to variable transformation. Many of these approaches were considered and since the distribution was negatively skewed, it was suggested that using a constant and reflecting the variable should be used in this case. In addition, because the distribution was somewhat leptokurtic (See Figure 56), a ratio, or one over 1/X transformation was used in the transformation, providing a transformation of TPE1=1/ (4-PE1) where TPE1 is transformed variable, transformed week one performance efficiency.

Following this transformation, Levene's test was used to compare group means and found that these means were not significantly different, since *F* (2, 68) =0.03, *p*=0.97. This finding provided no evidence that the data set had violated the homoscedasticity assumption. Therefore it was reasonable to consider the PE1 ANOVA.

Week Two Performance Efficiency (PE2)

There are three major assumptions of an ANOVA which need to be considered before analyzing the PE2 ANOVA, these are the independence, normality, and homogeneity of variance assumptions (Stevens, 2002).

According to Glass and Hopkins (1984), the independence assumption requires that observations within groups be independent or not influence one another. Specifically they say "Whenever the treatment is individually administered, observations are independent (Glass and Hopkins, 1984, p353)." So according to this definition, this data set met the independence assumption. This is because treatments were administered to individual learners and observations were made independently of one another.

APPENDIX F (CONTINUED)

The second assumption to be considered for the PE2 ANOVA, is the normality assumption. To test the normality assumption a Kolmogorov-Smirnov test was implemented and found a normal distribution since *D* (3, 87) =0.05, *p*=0.15. So given this normal distribution, no variable transformations were necessary.

Finally the third assumption, the homogeneity of variances assumption was also considered for the PE2 ANOVA. To do so a "proc univariate" procedure was run, using SAS and the Levene's test was used to compare the means. They were not significantly different, since *F* (3, 87) =0.56 *p*=0.64. This no significant difference finding showed that there was no evidence that the data set violated the homoscedasticity assumption, thus it was reasonable to consider an ANOVA for PE2.

ABOUT THE AUTHOR

David Lewis received a bachelor's degree in Biology from the University of South Alabama in Mobile, Alabama. He went on to complete a master's degree in Educational technology at San Diego State University.

While working on his doctorate and this dissertation, David also served as an Instructional designer at the University of South Florida, in Tampa, and at Florida International University, in Miami. While serving as an Instructional Designer he supported students, staff and faculty by designing and developing web-assisted, hybrid and fully-online courses and programs. In addition, he has taught fully-online and campus-based graduate level courses.

Finally, David has supported faculty by professionally designing, developing, and implementing web-based course materials since the spring of 1995.