December 2012

# Rapid Knowledge Assessment (RKA): Assessing Students Content Knowledge Through Rapid, in Class Assessment of Expertise

Erin Margaret O'Connell
*University of Wisconsin-Milwaukee*

# RAPID KNOWLEDGE ASSESSMENT (RKA): ASSESSING STUDENTS CONTENT KNOWLEDGE THROUGH RAPID, IN CLASS ASSESSMENT OF EXPERTISE

by

Erin O'Connell

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Chemistry

at

The University of Wisconsin-Milwaukee

December 2012

# ABSTRACT

RAPID KNOWLEDGE ASSESSMENT (RKA): ASSESSING STUDENTS CONTENT KNOWLEDGE THROUGH RAPID, IN CLASS ASSESSMENT OF EXPERTISE

by

Erin O'Connell

The University of Wisconsin-Milwaukee, 2012

Under the Supervision of Professor Kristen Murphy

Understanding how students go about problem solving in chemistry lends many possible advantages for interventions in teaching strategies for the college classroom. The work presented here is the development of an in-classroom, real-time, formative instrument to assess student expertise in chemistry with the purpose of developing classroom interventions. The development of appropriate interventions requires the understanding of how students go about starting to solve tasks presented to them, what their mental effort (load on working memory) is, and whether or not their performance was accurate. To measure this, the Rapid Knowledge Assessment (RKA) instrument uses clickers (handheld electronic instruments for submitting answers) as a means of data collection. The classroom data was used to develop an algorithm to deliver student assessment scores, which when correlated to external measure of standardized American Chemical Society (ACS) examinations and class score show a significant relationship between the accuracy of knowledge assessment ($p=0.000$). Use of eye-tracking technology and student interviews supports the measurements found in the classroom.

I would like to dedicate this work to my sister, Katharine. Without you this would not have been possible. Thank you for all of your love, guidance, patience, and support.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

I would like to start by thanking my family and friends for all of their love and support.

Thank you Mike for being there for me, even when my stress was causing you stress.

I would like to thank my advisor, Kristen Murphy, for all of her help throughout this process.

I would like to thank Dr. Geissinger, Dr. Aldstadt, and Dr. Knaus for serving on my committee.

Thank you Karen Knaus, Anja Blecking, and Lisa Lanning for being expert raters.

I would also like to thank my group members for your continual support.

Lastly, I would like to thank the Research Growth Initiative of UW-Milwaukee for funding my project.

# Chapter 1: Introduction and Background

## 1. Introduction

Is it possible to learn more about a student's understanding of a concept than whether or not the student can correctly solve a problem or task? What is meant when an instructor says that a student does not understand a concept? Developing a way to understand more about our student's problem solving processes in relationship to their performance would provide valuable insight into instructional methods and interventions where needed. Traditional methods of assessment primarily include performance evaluation. These methods have been used to determine instructional strategy and pace in the classroom, but with advancements in educational psychology these methods are no longer the optimal measurement of knowledge, "…knowledge levels of learners need to be assessed and monitored continuously during instructional episodes to dynamically determine the design of further instruction"(Kalyuga, 2006). Developments in theories such as cognitive load theory, schema learning theory, and adaptive control of thought-rational have led to research on how students learn and integrate information into a domain of knowledge (where a domain is considered any category of related information on a subject) (Anderson, 1996; Marshall, 1995; van Merrienboer & Sweller, 2005). This project draws from all of these theories, however, the project primarily arose from studies done by Kalyuga and Sweller (2004, 2005) involving problem solving in algebra using cognitive load theory.

## 1.2 Theories on Learning

### 1.2.1 Schema

Schemata are mental models made of information treated as a single unit that allow for quick information retrieval if well developed. When learning, new information is incorporated into long-term memory structures known as schemata. Schemata are the mental models of information built by a person over time as new knowledge is made into its own domain or connected into a previously established domain through organization into groups or "chunks" of related concepts to decrease the load on working memory (Chi, Glaser, & Rees, 1982; Larkin, McDermott, Simon, & Simon, 1980). The amount of information that is processed and placed into schema is based on the load that may be handled by working memory. Working memory may process 7±2 pieces of new information at a time (Miller, 1956). A piece of information is any new information that is not currently connected to a schema, or that a person does not recognize as relating to an existing schema. Since the load that working memory can handle is small compared to the amount of information learned in a lecture, it is necessary to understand the efficiency with which students map new information. With the establishment of a measure of efficiency, one can determine how to best approach the development and application of appropriate learning interventions.

In 2004 Kalyuga and Sweller used the idea of schemata to determine if the first step in problem solving by student's related to the efficiency of the students' schema in algebra. They found that first step was a valid measure of schema for algebra, and repeated the study in geometry with the same outcome. The steps chosen by students were directly related to a point in the thought process by the students, as the steps for a math problem could be listed sequentially and followed. The more advanced the process the more advanced the first step in the problem

solving process. In subjects such as chemistry the process is often not a direct step-wise pathway. However, if first steps truly relate to the level of development in schema as is suggested by Kalyuga and Sweller (2004), then first steps in chemistry should also relate to the development of a student's schema for the topic being assessed. Sweller, van Merrenboer, and Paas (1998)and Paas, Tuovinen, Tabbers, and Van Gerven (2003) discussed the use of schema in the role of cognitive load measurement as part of cognitive load theory (CLT).

### 1.2.2 Cognitive Load Theory

Cognitive load is the amount of information that working memory may cognitively process at any given time (Paas & van Merrienboer, 1994a). The limit on working memory of 7±2 pieces of information as discussed earlier has been carried through further study of working memory. Working memory is the cognitive system associated with handling information brought into working memory for the purpose of understanding and interpretation as a means of learning (Baddeley, 1986; Sweller et al., 1998).

Initial studies to investigate the difference between one person's level of knowledge and that of another person found that the difference did not lie in the conscious processing being done, but in the vast knowledge base acquired over exposure to similar or related information (Chase & Simon, 1973). This knowledge base for information is long-term memory. Long-term memory is the storage space for more permanent information that may be used individually or for understanding of more complex matters (Ericsson & Kintsch, 1995), where this information is stored as schemata (discussed in the previous section). A well-developed schema is automated and has efficient "chunking" of information. Working memory is the processing center between information stored in long-term memory and new information or stimuli to be stored in long-term memory from working memory (Baddeley, 1986). The more automated and chunked the

information in a domain within long-term memory is, the less capacity it takes up in working memory, therefore decreasing the load on working memory (Paas et al., 2003). Many studies measuring working memory to aid in the development of instruction have been performed (Chandler & Sweller, 1991, 1996; Mayer & Moreno, 2003; Sweller et al., 1998).

In order to measure load on working memory during learning one must look at what affects cognitive load on learning, or germane cognitive load. Sweller, van Merriënboer, & Paas (1998) published an article in which three types of load were addressed as affecting/determining cognitive load. Intrinsic cognitive load is the load imposed by the material itself. This load cannot be altered, as it is based on the nature of the material being learned. Extraneous cognitive load is the load imposed on cognition by the environment (i.e. instruction, distractions in the surroundings, or superfluous information). Extraneous load may be changed by changing the way in which the material is presented, or altering the environmental settings if not applicable to active learning. Finally there is germane cognitive load. Germane cognitive load is the load imposed on working memory when processing and integrating information into a schema. Changing instruction or task representation affects the amount of germane cognitive load imposed on a learner by creating extraneous load. Sweller (1994) discusses the ability to change germane cognitive load via instruction.

#### 1.2.2.1 Germane Cognitive Load

Germane cognitive load consists of three factors: mental load, mental effort, and performance (Sweller et al., 1998). Mental load represents the load on cognition that stems from the task itself (intrinsic load) and its interaction with a subject's long-term memory. The better the schema in long-term memory, the less controlled processing needs to occur in working memory. Mental effort (ME) is the amount of cognitive resources in working memory directed at task

solution. Mental effort is directly related to the mental load because as the expertise of the learner increases, the germane load occurring from the task decreases and the load on the working memory decreases (as the schema that is called into working memory is more advanced and requires less load on working memory). If mental load decreases it is expected that mental effort would decrease, and vice versa. This is due to schema development and activation, and to mental load. Both the mental load and ME are related to performance. Performance, or how well one succeeds at a task, is measured in mistakes, time on task (TOT), and if an answer is correct or incorrect (Paas, van Merrienboer, & Adam, 1994). The more the automated the processing of long-term memory information becomes, the more mental effort decreases as the information becomes more automated in working memory. Since the three layers of germane load are so interconnected, it is important to understand how and why measuring them is important.

#### 1.2.2.2 Measuring Cognitive Load

Measuring cognitive load has been done using several techniques. The techniques of concern here are the ones involving the measurement of mental effort (Sweller et al., 1998). One of these techniques is subjective, and involves student self-reporting. For subjective reporting there are several methods that involve single or multiple scales for self-reporting (Hendy, Hamilton, & Landry, 1993). In the subjective technique a subject reports his or her mental effort using a provided scale. The scale used by Paas and van Merriënboer (Paas & van Merrienboer, 1994a) involved nine points that looked into very fine detail. The scale ranged from “very, very, very little” to “very, very, very high”. It was later argued that there is little difference between some of the finer points and that a person cannot tell the difference between “very, very, very little” and “very, very little”. Therefore it was suggested that the scale could be trimmed down

to a seven-point scale (Marcus, Cooper, & Sweller, 1996). For determining mental effort in math only a five-point scale was utilized; this was based on the determination that only a general understanding of mental load by the students was needed to be understood for a true measure of mental effort (Kalyuga & Sweller, 2005). This scale has since been modified and validated for use in chemistry (Figure 1)(Knaus et al., 2009)

How much mental effort did you expend on the previous question?

A. Very little

B. Little

C. Moderate amounts

D. Large amounts

E. Very large amounts

**Figure 1. Mental effort scale used in chemistry assessment testing.**

Other methods of measuring mental effort include physiological and task/performance-based techniques (Sweller et al., 1998). Physiological techniques involve measurements of heart rate, brain activity, and eye activity, while task and performance-based techniques involves taking measurements while two tasks are performed concurrently (Sweller et al., 1998). Of these two methods neither of them is practical for large-scale use in the classroom for rapid knowledge instrumentation on single tasks. The ease of use and cost effectiveness of subjective reporting via a validated scale is therefore much more appropriate for large-scale classroom measurements. The only contention one might have is the ability to know students are accurately reporting and gauging their mental effort. However, measurements of pupil size can be done practically in an interview setting lending support to the validity of subjective reporting

of mental effort by students .[1] To do this, an instrument known as an eye-tracker is required (Eye-tracking will be discussed later in this chapter.)

### 1.2.3 Adaptive control of thought- rational (ACT-R)

Another theory of cognition uses production sets. Production sets are a series of rules for matching information in memory to response outcomes. Unlike in cognitive load theory where information is networked, recall in production sets is based on exposure. This idea of producing an outcome through exposure comes from a theory on adaptive control of thought-rational (ACT-R). ACT-R helps to understand how students might use their schema, which suggests productions are used as a representation of knowledge to be accessed for use when problem solving[2]. It is a way to describe how schema may function through deriving information or knowledge through exposure to situations and information, an idea known as abstraction[8]. Abstraction lends to the idea that schema may be generalized and not concrete linkages of information, thereby suggesting that through ACT-R, one gains information by exposure to concrete examples that signal a set of production rules. This then allows for the relationship of new information to other sets of information stored in memory. The more advanced a learner is in a domain (degree of expertise), the more likely their schema are abstracted or generalized giving them greater utility and faster searching.

"According to the ACT-R theory, the power of human cognition depends on the amount of knowledge encoded and the effective deployment of the encoded knowledge" (Anderson, 1996, p. 355).

[1] Eye-tracking studies have shown that pupil size is an accurate measure of mental effort (Beatty, 1982), therefore the assumption is made that if ME is measured through tracking it can be compared to a self-assigned ME for comparison and support of a subjective system.

Therefore ACT-R is concerned with declarative knowledge (the schema mapping from long-term memory, or the "chunking" of information of *what* you need), and the production rules (on *how* to perform a task either physically or cognitively) (Anderson, 1983)[2] of memory. Later studies and expansions on ACT-R lead to further information on long-term memory structures, allowing for the connection in how to determine *when* and *why* to use procedural (knowledge that has become automated) and declarative knowledge, known as conditional knowledge (Figure 2) (Brunning, Schraw, Norby, & Ronning, 2004).

The use of ACT-R to understand how we as humans perform functions of retrieval, input, and use of knowledge is pertinent when contemplating efficiency of a student's thought process and goal state (i.e. conclusion or answer). While the math studies do not specifically measure all three areas of knowledge involved in long-term memory, it does note the importance it plays in the use of schema which is what the instrument sets out to measure. In the RKA, the understanding ACT-R is important for the same reason. The instrument developed here in this thesis is intended to look at a student's efficiency through first step in problem solving, which is a measure of the student's use of schema. By relating the first step to the performance the measurement taken is essentially referencing the use of procedural, declarative, and conditional knowledge through long-term memory.

[2] The original work on ACT-R done by Anderson came from his earlier works on his theory of ACT done in 1983.

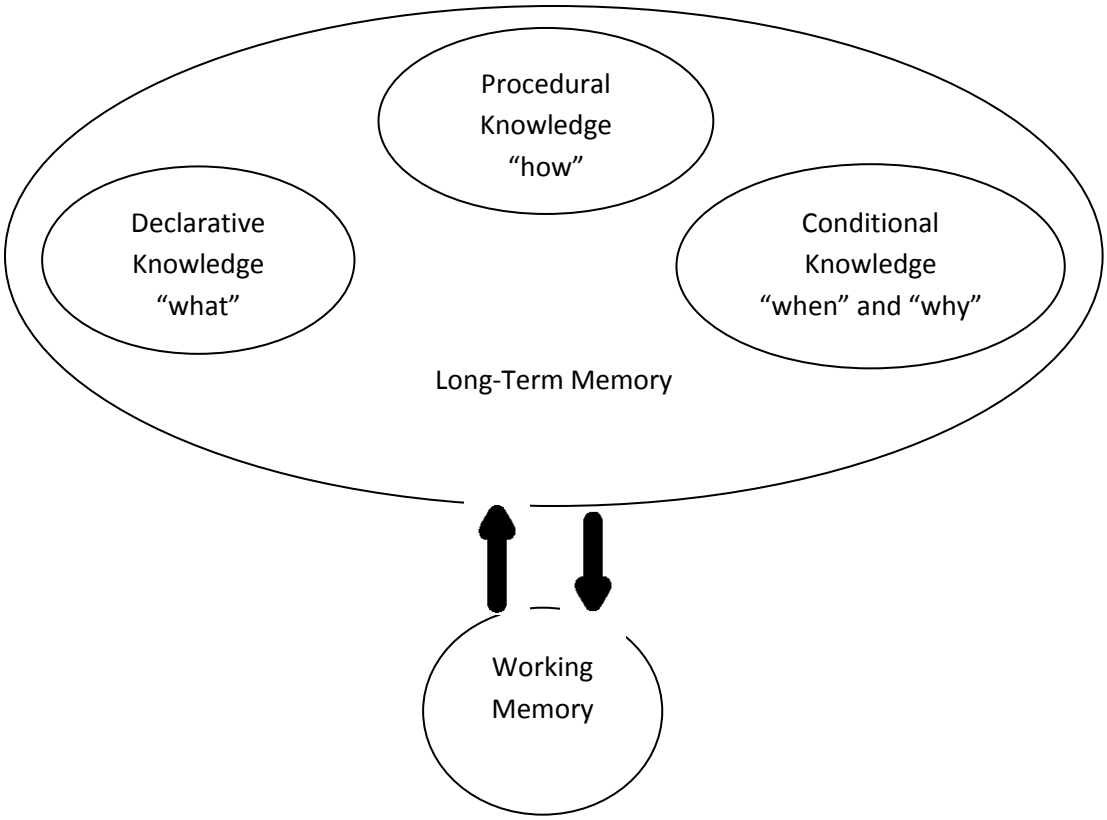



**Figure 2: The relationship between the different types of knowledge in memory.**

The connections of pieces of information(Miller, 1956) into a network of knowledge that a person forms about a topic that can be used when "problem solving" is the basis for schema (Marshall, 1995, Chapter 2). Kalyuga and Sweller developed an instrument to help assess students' expertise and understanding of algebra (Kalyuga & Sweller, 2004, 2005). Expertise is a measure of how well a schema in the domain of study has been developed. Through ranking expertise Kalyuga and Sweller were able to specifically look at a student's increasing or stable level of understanding in the domain of math. They were able to determine expertise by looking at an individual's first step in his or her problem solving process (Kalyuga & Sweller, 2004), and in 2005 they added a mental effort portion to allow for an efficiency evaluation. Students were instructed to report their mental effort using a scale ranging from "extremely easy" to "extremely difficult" (Kalyuga & Sweller, 2005). Efficiency was calculated through a comparison

of mental effort and performance, to determine the degree to which the load on working memory affected performance. Kalyuga and Sweller found that one's expertise in math could be assessed using the difficulty level of the first step used during problem solving. The measurements taken by Kalyuga and Sweller (2004, 2005) show promise for measuring understanding (expertise) of students' individually in math, but math is a linear process and the developed procedure lacks the ability to test for expertise in interdependent domains such as chemistry on a large scale.

The study of problem solving in chemistry will help lead to a larger knowledge base for use in classroom instruction and general conveyance of chemical knowledge in a public forum. In as such, it is necessary to determine how one will appropriately measure a student's current understanding of chemistry. Because the domain of chemistry is vast and contains many sub-domains, the best place to begin collecting knowledge on understanding in chemistry is at the beginning of knowledge development in the domain. For this reason an instrument, Rapid Knowledge Assessment (RKA), was designed to work within the general chemistry curriculum. This study draws from the principles used by Kalyuga and Sweller in their 2004 and 2005 studies, but expands upon them for use in a classroom setting. Development of this instrument also lends to the development of complementary classroom interventions for varying groups of learners and knowledge levels. No intervention will ever be able to benefit an entire group, but if interventions that will target a larger portion of the classroom population based on classwide information are developed, a higher success rate can be achieved. The problem has been finding a way for instructors to determine what types of interventions will be needed in the classroom. The RKA instrument acts as an in-classroom instrument to measure students' knowledge in chemistry through assessing individual and class-wide expertise.

## 1.3 Expertise in Math: why a look at math sheds light on chemistry

The 2004 study by Kalyuga and Sweller involved the development of a system to measure an individual's knowledge in specific areas of math using the person's own schema. The expectation was that if students could be identified at different levels of knowledge in different content areas within algebra, then better instructional strategies could be implemented. The goal was to reduce expertise reversal effect (regression to a less efficient method of problem solving or understanding (Kalyuga, Ayres, Chandler, & Sweller, 2003)) through new instructional strategies, as studies show that this may occur if students are directed towards only one method of thinking about problem solving (Kalyuga, 2007). This is typically identified through lower performance. Expertise reversal effect occurs when instructional procedures do not work well for both lower-level novices and higher-level novices (Kalyuga et al., 2003). If an instructional procedure does not teach to every knowledge level of learner, but caters towards teaching only the lower level knowledge students, there exists the possibility that the higher knowledge level students may stop progressing and reverse to a lower level function of their knowledge. The theory behind this is based on cognitive load theory, where the processing capacity of working memory may be benefited or hindered by instruction and the student's current schema (van Merrienboer & Sweller, 2005). This is because external information competing with information needed for concept processing may become integrated into the schema being developed by the student, and become part of the normal recall of information in the future.

Another theory includes the use of production sets. Production sets are a series of rules for matching information in memory. The study also contains theory from Adaptive Control of Thought- Rational to understand how students might use their schema, which suggests

production sets are used as a representation of knowledge to be accessed for use when problem solving (Anderson, 1996). It is a way to describe how schema may function through the idea of abstraction (Anderson & Lebiere, 1998). By developing a measurement technique using the student's own schema, Kalyuga and Sweller therefore intended to measure the student's level of knowledge and assign students a level of expertise that would help the instructor better design instructional presentations within the classroom.

To begin, Kalyuga and Sweller (Kalyuga & Sweller, 2004) needed to develop a rapid cognitive diagnostic test that measured long-term working memory (schema) and working memory while performing knowledge-driven tasks. Through reviewing the path experts take based on their ability to recall essential elements in the task, Kalyuga and Sweller were able to design four experiments that were used to help determine a student's expertise level. They found tendencies for assessing correct solution paths to be an effective measure of cognitive processes in problem-solution schemas when comparing these first steps identified by the students to the continuum of the problem-solving pathway (a linear pathway for these tasks). For these reasons the four experiments were designed around the idea that the first step in the problem solving process would lead to subtasks within an overall goal. If the subtasks were correctly identified, and if an advanced first step was given and carried out, the learner (student) would have identified with larger chunks of information in their long-term memory and come closer to the correct solution for the task, therefore showing a higher level of expertise with each subtask completed. The first-step approach was used for comparison of the instrument in external validation of the efficiency ratings to exam scores of this concept. The information from these results was then used to develop instructional strategies

Kalyuga and Sweller's (2004) first experiment was designed based on the first-step taken in the problem solving process for finding the solution of a task within algebra. In the tasks set to rapidly measure students' knowledge, the students were asked to give their immediate next step towards finding a solution to the task presented. They were then asked to find the solution for the task provided. This process allows for the students to connect process with solution, and to accurately relate a "step" during the task. It is important to group these two pieces together, so as to have an accurate reporting of the information. If asked to recall later what was done in the beginning, students may not be able to identify with their thought process. Three tasks were given, and the time taken to complete the tasks was recorded. More knowledgeable learners should identify more advanced steps (moves) for the problem states perceived, while less knowledgeable learners should identify less advanced steps, as their perceived problem states would be less efficient. The performance on the tasks was also measured with scoring on a basis of correct or incorrect. The results from the rapid measurement were then correlated to the results from a list of 12 similar tasks. What was found was concurrent validity for the rapid assessment, as demonstrated by the Pearson correlation scores between the traditional and rapid tests, where at the 95% confidence interval a high significance was found; $r(44)=0.92$, $p=0.01$, CI=95% (Kalyuga 2004). Experiment two was used to see if similar results could be seen within a different domain in mathematics, and the same experiment was applied using geometry instead of algebra. Similar results were found for the geometry study. One could argue that while both geometry and algebra are in the domain of math, that both types of math require a separate set of knowledge to draw upon for problem solving. Because separate schema may be activated during problem solving in algebra than in geometry, this lends towards validity of the instrument in other domains of knowledge and subdomains.

In experiment three, Kalyuga and Sweller (2004) studied whether the expertise reversal effect could be measured and adjusted for using the rapid measurement idea to assess the students, and then develop an instructional format to limit such an effect. In this experiment the steps to the task were mapped out in a grading rubric and assigned efficiency levels starting with the most basic part of the task solution as "1" and ending with the final step in the task solution as the highest number. Here students were also asked just as before to solve the problems and give a first step. However, this time the point at which the student started, according to the rubric, was used to indicate the knowledge level of student. This was appropriate, as the higher-knowledge level learners would use more efficient steps as theorized earlier using CLT. This efficiency number summed with the performance score (correct 1, incorrect 0) and the time for the assessment helped to determine the level of knowledge of each student. At this point in the experiment, different instruction was presented for students of different knowledge levels. They found statistically significant correlations ($p<0.01$) between high-level knowledge and low-level knowledge learners to suggest that different levels of knowledge may be measured using the rapid measurement tasks utilizing efficiency ratings. The success of an instrument in one domain supports the development of an instrument in other domains. Indeed, this was specifically addressed by the authors:

> "In more complex domains involving multiple-step problems, students might be able to take many different routes to problem solutions. If all those routes are identifiable, the method still could be used in both paper-based and electronic forms." (Kalyuga and Sweller, 2004, p. 566)

Experiment four then went on to test if what was found in experiments one through three could be used effectively when training students on computers. Because math is a linear

process, the program was set up to walk students through the information in a linear style. There were four stages, and in each stage the students were given less guidance than in the previous stages. By stage four only problem-solving exercises were given. A diagnostic test was given at the beginning of the program to assess where in the computer training process to begin students. Students with the lowest level of knowledge were started at the first stage, while students with a higher level of knowledge were started at the fourth stage. The diagnostic test based students' knowledge on the number of tasks the students correctly solved. To be considered to have a higher knowledge level at least two of the problems had to be solved correctly. It was found that using such a strategy limited the expertise reversal effect for higher knowledge students, along with advancing the expertise of the lower knowledge students. The results in this experiment were also found to be statistically significant for the learner adapted format between the two knowledge level groups ($t(24)=2.26$, $p<.05$). The important aspect of this experiment is that by using computers and having a more controlled environment for each level of learner, it can be seen that using a rapid measure of students' level of expertise based on cognitive processes is functional. Because of this there is reason to believe that an electronic system for a classwide rapid measurement may also be possible. This strategy can help to improve instruction within the learning environment, while aiding different knowledge level students.

Further studies by these researchers included experiments where mental effort (ME) was used to help assign an expertise rating to students (Kalyuga and Sweller, 2005). Mental effort was done via self-reporting by students through use of a seven-point quasi-interval scale ("extremely easy" to "extremely difficult"). The scale was administered as a survey after completion of the task, where the students were asked to provide the difficulty the task posed for them. The mental effort rating was combined with the performance rating to produce an efficiency rating.

From the efficiency rating the treatment and non-treatment group were found to have significantly different average efficiency gains ( $t(14)=1.89$, $p<0.05$)(Kalyuga & Sweller, 2005). The mental effort rating is combined with performance to form efficiency, as it is a measure of expertise. The more “difficult” the task is reported by the students, the more load there is on working memory. Because working memory is the location of information processing from and into schema, the more difficult the task is rated, the less developed a schema the student possesses on a topic, and therefore the less expertise they have. This also shows that the complexity of the information is being reported from a subjective standpoint. The more complex the information is viewed to be, the more load is imposed on working memory. The objective complexity of the task is based on the actual components required for task solution.

## 1.4 Summary

The efficiency rating, in conjunction with the experiments from 2004, helped to develop the basis for the RKA experiment, in which a proxy for the cognitive processes of the working memory can be measured and used to determine expertise in a more complex domain, such as chemistry, using multiple-step problems with multiple solution possibilities. The RKA uses a five-point scale for in classroom activities as well as interviews, and checks the reliability of student self-reporting as a valid method of measurement using measurements of pupil dilation found through eye-tracking, something that was not done in the studies by Kalyuga and Sweller to check reliability of student mental effort.

Because using computers in a classroom is not practical for lecture settings, the RKA needed to be developed for class-wide measurements using alternate techniques. The use of clicker systems allows for computer data analysis while collecting information of a large set of individual users. Students will not be allowed to jump across multiple tasks, but the design of

the instrument is to assess student knowledge to lead classroom instruction. Therefore, it is not being used as a direct tool to teach information, and expertise reversal is minimized.

## 1.5 Using pupil diameter as a measure of mental effort through eye-tracking

Kalyuga and Sweller decided to measure mental effort using a subjective seven-point rating scale ranging from “Extremely easy” to “Extremely difficult” (Kalyuga & Sweller, 2005). While their results tracked with the expectation that higher performance would track with lower ME for higher expertise students, the article does not mention how the study went about assuring validity of ME. For this reason in the development of the RKA eye-tracking technology was utilized to assess validity of subjective reporting of ME by students. Research on the use of pupil size to measure cognitive function has taken place since the late 1800s (Beatty & Wagoner, 1978). Since then great progress has been made in the use of pupil size as a measure of load on working memory (Beatty & Wagoner, 1978; Kahneman & Jacson, 1966; Stone, Lee, Dennis, & Nettelbeck, 2004). Research has found not only links between pupil dilation and mental effort, but those measurements can also be differentiated to show the load on cognition at different points in time during a task (Kahneman & Jacson, 1966). The importance of tracking pupil measurements is therefore paramount in determining if subjective reporting of ME by participants is an accurate measure of ME in research. To carry out this research one must first consider how to track pupil diameter of participants. Earlier methods of tracking used to be through restrictive devices that allowed for measurements to be taken and/or photographs of the eye that were turned into slides for measurement. More recently, computer based technology allowing for video eye-tracking has been developed to record such information using

less restrictive measures (Jacob & Karn, 2003). Recent studies have even utilized mobile eye-tracking systems to measure pupil dilation (Klingner, Kumar, & Hanrahan, 2008).

The importance of eye-tracking can be seen in works such as Keith Rayner's 1998 article following a history of 20 years of eye movement research. This work shows that there is more to eye-tracking than just measurements of pupil diameter. Data on what information in a task is accessed, how long it is accessed (fixations), the path in which information is accessed (saccades), and how often something is accessed are all important pieces of the process that can be collected for analysis. Through using an eye-tracker all of this information and more can be obtained for a single individual in now non-invasive, user-friendly formats. For this reason computer-based eye-tracking was implemented in the RKA interviews to establish validity and reliability of subjective measures of ME by students.

## 1.6 Summary:

With chemistry being such a complex domain that does not follow a linear path, we want to see if the students schema may be used to measure their expertise based on how much information the students can take in and process based on their cognitive load. The mental effort rating they choose will relate to the student's cognitive load and be used as an indicator for his or her expertise level in conjunction with his or her performance and first step, same as in the experiments in math.

Through evaluating Kalyuga and Sweller's studies (Kalyuga & Sweller, 2004, 2005) we expect it is possible to create an instrument that will assess a more complex domain, because there was success determining knowledge level (expertise) of students using linear processes such as in algebra and geometry. The RKA will use the concepts of the math studies (mental effort, first step, and performance) to assign expertise ratings to the students and class as a whole.

# Chapter 2: Instrument Design and Implementation

## 2. Methods

The purpose of this system is to help collect information on how students build a knowledge base, or schema, when learning chemistry. In class the tasks are presented as "spot review" questions. This allows the students to feel more relaxed and view the tool as being there to assist them, rather than as a way to test how quickly they can solve a problem. To accomplish this, a methodology was developed that allows for quick and accurate data collection of several constructs (performance, ME, and first step), using clickers.

### 2.1 Demographic Information

The *Rapid Knowledge Assessment* study underwent development and testing at the University of Wisconsin-Milwaukee. Demographics about the university were obtained from the University of Wisconsin website[2]: 'The University is located in Milwaukee, Wisconsin and is a Research 1, urban, doctoral degree-granting university consisting of 14 schools and colleges that together offer 180 degree programs. UW-Milwaukee is reported to be the most diverse of the University of Wisconsin system universities'.[3] The University has an enrollment of 29,768 students, of which 5,090 are in postgraduate studies[4].

[3] Obtained from UWM website: http://www4.uwm.edu/discover/about.cfm (accessed June, 2012)
[4] Obtained from UWM website: http://www4.uwm.edu/discover/facts.cfm (accessed June, 2012)

Table 1. Institutional Data

| Institution | Type | Size* | Location | Demographic |
|---|---|---|---|---|
| University of Wisconsin-Milwaukee | Doctoral | 29,768 | Milwaukee, WI | 52% Female<br><br>17% Targeted Minority |

*Both undergraduate and graduate students **Targeted minority excludes Asian students

## 2.2 Overview of Classroom Study

The study took place in the Chemistry and Biochemistry Department over eleven semesters, from the spring of 2007 through the fall of 2011. Students from a Preparatory Chemistry course and the first semester of a two part General Chemistry course took part in the study over these semesters. Of these 2188 students, 1521 consented to participate in the study via an approved IRB protocol (IRB #09.047). Only data from consenting students was included in the study. Prerequisites for preparatory chemistry consist of a C grade or better in college algebra. The class meets for three lectures a week for 50 minutes each and one discussion for 50 minutes, without a laboratory section. This adds up to 60 hours over the course of the semester. A professor or academic staff from the chemistry department at the university teaches the lecture, while either graduate or undergraduate teaching assistants teach the discussions. Undergraduate teaching assistants must be chemistry or biochemistry majors and have completed the majority of their required courses for their degree. Preparatory chemistry is not designed for non-science majors. Students required to take a chemistry course(s) for their major must take preparatory chemistry if they are not prepared for their required chemistry

course(s) as determined by a chemistry placement test. The lectures from this course used in the study were taught by one of three instructors trained in use of the instrument. Prerequisites for general chemistry included a grade of C or better in college level algebra and preparatory chemistry. Those placing out of preparatory chemistry or college level algebra through placement testing were not required to have completed these prerequisites before participating in general chemistry. The course meets for 3 lectures (50 minutes each), 1 discussion (50 minutes), and 1 laboratory section (2 hours, 50 minutes) per a week adding up to 105 hours over the semester. A professor or academic staff from the chemistry department at the institution teaches the lecture, while graduate and undergraduate students teach the discussions and labs. The same requirements are set for undergraduate teaching assistants for this course as for preparatory chemistry. General chemistry is designed for students who intend to take higher level chemistry courses for their major. Students required to take a specialized chemistry course such as nursing chemistry do not participate in general chemistry, however some students with attaining a degree that has a specialized chemistry course offered still take general chemistry. An example of this is engineering, where a student may intend on obtaining a certification that requires a more in-depth knowledge of chemistry. Demographic data for the two courses is listed in Table 2.

Table 2. Demographic Characteristics of Students in General Chemistry I and Preparatory Chemistry

| | Accepted IRB (n=1331) | Did not accept IRB (n=595) |
|---|---|---|
| | N (%) | N (%) |
| Gender[a] | | |
| Female | 866 (56.9) | 314 (47) |
| Male | 654 (43) | 354 (53) |
| Unknown | 1 (0.1) | 0 (0.0) |
| | p=0.000 | |
| Year in School | | |
| Freshman | 388 (30.3) | 140 (25.5) |
| Sophomore | 547 (42.7) | 262 (47.6) |
| Junior | 213 (16.6) | 100 (18.2) |
| Senior | 133 (10.4) | 48 (8.7) |
| | p=0.077 | |
| | Mean (SD) | Mean (SD) |
| ACT[b] Composite Score | 22.76 (3.444) | 22.18 (3.685) |
| | p=0.001 | |
| ACT[b] Math Score | 22.92 (4.000) | 22.36 (4.157) |
| | p=0.005 | |
| ACT[b] Science and Reasoning Score | 22.90 (3.531) | 22.46 (3.723) |
| | p=0.012 | |

[a] Gender was obtained through institutional research. Not every participant's gender was identified.
[b] ACT scores were obtained through institutional research and were not available for all students.

## 2.3 Task Design

The principal investigator initially designed tasks for the instrument through collaboration with another departmental staff member for use in the earliest phase of the study in 2007. The tasks were designed based on review information in different content areas within chemistry that were considered important for the understanding of chemistry. The tasks were designed to cover various aspects within the different content areas being tested. The content areas identified for testing initially included properties of matter (formula calculations), stoichiometry,

aqueous reactions, and gases. In time, the instrument was expanded to include general chemical knowledge (unit conversion, temperature conversion, etc.), atomic structure, electronic structure, periodicity, chemical bonding and molecular structure, liquids and solids, reactions, and thermochemistry. The list consisted of fifteen tasks. In 2008, the collaboration of the principle investigator and I began for the purpose of determining task revisions, additions, and deletions. Initial task revisions were carried out based on performance and open-response submissions (2008- 2009). If performance ratings ran too low or too high for the class overall, the tasks were not accurately measuring a range of student abilities. At the same time, if the open-ended responses collected were too vague, continually contained too high a count of complete answers, and/or did not yield a fair number of first steps or had too many first steps for coding the task was not effectively measuring varying mental efforts. Therefore, task revisions were carried out if a general issue or theme was prevalently apparent and could be changed to allow for further testing of the task and its possible use in the final iteration of the task list. To test the tasks, Pearson correlations were performed between internal and external measures of mental effort, efficiency, performance, final examination performance, and percent grade in the class. This allowed for testing of reliability of the information.

## 2.4 Data Collection

### 2.4.1 Open-ended Response Collection

Open-ended responses on the first step in a student's problem solving process were collected on paper during the submission of the task response. Paper sheets were distributed to the students before the projector displayed the task and the timer started for data collection (Appendix A). The top area of the sheet provided a space for the students to record their first step, while the bottom portion of the sheets provided space for the students to work out their

solution to the task. During the collection time for task responses, the top portion containing the first step open-response was torn off and handed in. The bottom portion was retained by the student to allow for review and comparison when the instructor provided a method of task solution.

### 2.4.2 System for Classwide Data Collection

The classwide data was collected with a personal response system (or student "clickers") and the system that was used is a product of eInstruction called Classroom Performance System (CPS) (einstruction, 2007). The system works through using a receiver and software that can accept information sent to the computer by clickers that are compatible with the CPS unit. The system can be set up to accept open responses as numbers or letters and has a timer function that can be used in the data collection phase. Additionally, this system gives feedback to the students through both their clicker and the software (projecting on a screen) that their response has been received by the system. The clickers also have an LCD screen to allow them to view their response before submitting. Students can submit multiple responses while the question is still open with only the last response graded by the system. The students are trained on using the clickers and are expected to purchase, register, and use their clickers as part of their final grade. The data is then exported into excel.

To begin with, each task, or question, is posed to the students using the system and/or on paper. Each task has three areas where data is collected. The first area is performance, the second is mental effort, and the third is the first step in the student's task solving process.

#### *2.4.2.1 Area 1: Performance*

Performance is collected as open response via the clicker system. The students are given 2 to 3 minutes to solve each task. The tasks presented to the students come from one of twelve content areas: General, Atomic Structure, Electronic Structure, Periodicity, Chemical Bonding and Molecular Structure, Liquids and Solids, Properties of Matter- Formula Calculations, Reactions, Stoichiometry, Aqueous Reactions, Gases, and Thermochemistry. Multiple tasks were developed for each content area based on collaboration between the principle investigator and myself. The tasks were written to assess concepts that would be present in most general chemistry courses at many universities, and that would assess various levels of student understanding. The design and order of the tasks focuses on assessing those students with average and slightly less than average understanding of the concepts presented, since this is the group often most affected by interventions.

Once the students have solved the task, they enter their response using their clicker. The answer is coded into the system often with a 5% error range for numerical answers. In the case of an integer response (such as atomic number), no range was allowed.

#### *2.4.2.2Area 2: Mental Effort*

Students are next prompted to report their mental effort, which is collected using five quasi-interval ratings. This quasi-interval scale used here was developed based on the project done by Kalyuga and Sweller(2005) (Figure 3). The students are given a maximum of 30 seconds to enter

in a response. Once the students have gotten used to the order of the system, this part normally takes much less time (approximately 10 seconds).

#### *2.4.2.3 Area 3: First Response*

The students are provided with 30 seconds to 1 minute to complete this segment. The time allowed varies depending on the amount of reading required. The order of the steps was always listed in typed length from shortest to longest. This deterred the students from selecting an answer based on the order it appeared in the list.

It is important that this information was collected *last* so as not to influence the answers to the previous two areas of data collected. It is important to note that the students were only able to answer the question that was open on the screen in front of them. The order in which information was presented to the students in the first phase of the study may be seen in Figure 3. Once an area was closed, the responses for it were not available to the students to prevent back entering information. Normally the amount of time given above in each area was adequate for that given area; however, discretion was applied based on the instructors feeling for how students were moving along.

Once the system was closed so that no more answers could be accepted, the solution for the task was presented to the students. By walking the students through how to solve the task, it allows for them to reflect on their previous responses they gave and gauge their understanding of the concepts involved in solving the task. It also gives the students a chance to ask questions they may have related to the topics being covered. The students were told to focus on the explanation being presented as the task is being solved, and not on copying the solution down. All of the answers were later posted to each question on the web page for the class.

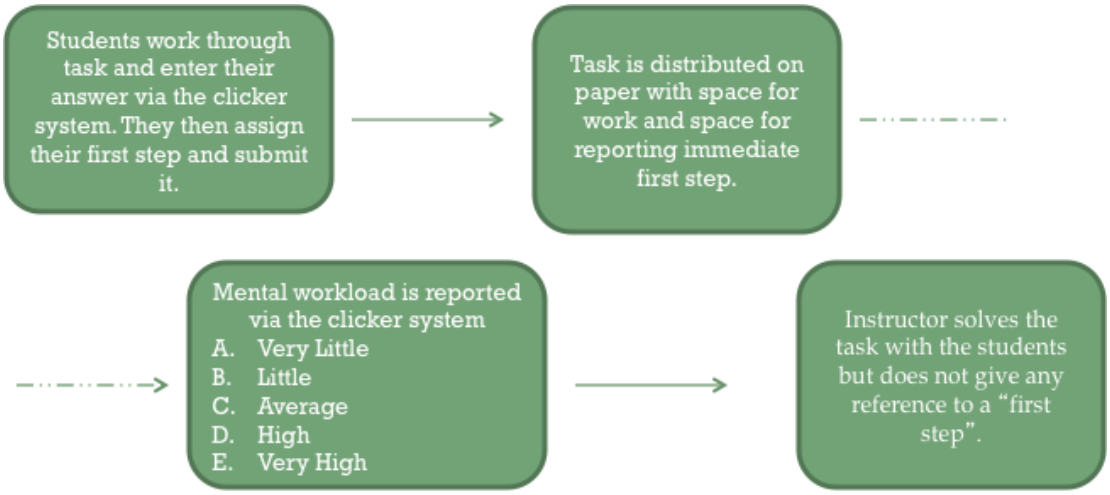



**Figure 3:** Primary phase of the study: Shows the order and time allotted for each area within the system.

## 2.5Instrument Development:

The beginning of the RKA was undertaken in the spring semester of 2007. Information collected in this semester was from three lecture sections of preparatory chemistry students at the University of Wisconsin-Milwaukee (instructor 1, N= 135; instructor 2, N= 229). In this instance, only open responses (OR) were collected from the class. An open response was considered a student-generated first step in the problem solving process undertaken by the student. The purpose of the OR, and the reason why it is referred to here as information instead of data, was to cultivate a basic understanding of student-generated first steps in problem solving. The basis of the instrument, as mentioned earlier, is to create a comparison between several different levels: performance, mental effort, and first step in one's problem solving process. Therefore, in order to create a balanced instrument for classroom use that could be

entirely clicker based, it had to be determined if students overlapped significantly in their first steps.

To cultivate data in this manner, a classwide system for collecting student OR needed to be generated. It was determined that the simplest, fastest, and most accurate way to collect this information would be through the use of a paper system. At the beginning of several classes, students were given sheets of paper on which to record his or her first step in their problem solving process (Appendix A). The students were then presented with a task and given five minutes to work on solving the task. After the task was solved students clicked in their answers to CPS. Students were instructed to record the first step in their problem solving process on the slip of paper and hand it in. Once the OR slips were collected, and the five minutes had passed, the instructor for the lecture would go through the solution for the task. It is important to note that when the solution was presented to the class, there was no inference to any specific correct way to solve the task. This was done to keep students from thinking any one process was more correct than another.

### 2.5.1 Data information and processing

The data collected on the OR forms was taken and coded into a response type. The types of responses were then reviewed for similarities. For example, if response A said “writing the formula for aluminum acetate” and response B said “$Al(CH_3COO)_3$”, both of these responses were grouped together as the same type of response. If this was an appropriate step in the problem solving process, the grouped response was coded into an option for use in the electronic response system (Figure 4). The final step entered into the electronic list was always the “out” response “Reading the exercise, however I am not sure how to start the exercise”. This

allowed for students to present a first step of reviewing the task without selecting a random first step from the list provided to them.

**Exercise: What is the percent composition by mass of oxygen in aluminum acetate?**

| Steps- "For this exercise, my first step is…" | |
|---|---|
| A | writing the chemical formula |
| B | determining the mass of oxygen |
| C | determining the number of moles of oxygen |
| D | determining the molar mass of aluminum acetate |
| E | reading the exercise, however I am not sure how to start the exercise |

Example of OR answers that were used to determine Step A:

"$Al(CH_3COO)_3$"

"$Al(C_2H_3O_2)_3$"

"writing the formula for aluminum acetate"

"writing the chemical formula"

**Figure 4 Example of a task where open response options were coded into response types and entered into the electronic CPS system as an option for multiple choice response.**

The mental effort and task performance data collected via the clicker system were recorded by the CPS software and later exported to excel. The data was organized for analysis in excel, and then loaded into SPSS for analysis. Once in SPSS, Pearson correlations were performed between the task performance, mental effort, final exam scores, and class grades (percentage scores).

### 2.5.2 Continuing RKA development on a larger scale

The project was continued as described above and expanded upon in the fall of 2008. In the fall semester instructor 1 taught both the preparatory chemistry course (N=183) and the general chemistry course (N=150) (there was more than one lecture section of each course during this semester). The data obtained was once again based on the information collected from the OR forms in lecture, with the addition of one important step: mental effort. The students were asked to report mental effort (ME) using a Likert scale[5]. During this semester only a general list of first steps was beginning to be generated. Starting in the fall of 2008 student data was needed to test the reliability and validity of the system. For this reason students were provided with a consent form approved by the internal review board (IRB 09.047). Only students who signed the consent form and authorized the use of their class data were used in the study.

### 2.5.3 Task Development

Development of tasks for further use in the study was carried out through review of earlier semester's class materials to cover varying aspects within each of the twelve noted content areas. These tasks were written through collaboration of the principal investigator, another university instructor (proof of concept phase only), and myself (after the proof of concept phase data collection). The list of tasks may be found in Appendix B. After the tasks were developed, they were tested through classroom application. Students were presented with the tasks during the "spot review". After data collection was completed, the tasks were reviewed based on the amount of time the students' required to complete the task, the overall class performance (did high performing and low performing students in other areas still show a difference in

[5] Explained in section 2.7.4

performance for the given task), and the quality/type of open-responses submitted by the students. If the task took too long to complete and showed overall low performance with OR submissions indicating little understanding of what the task was asking, the task was reviewed. If upon reviewing the task it was apparent that minor changes would allow for continued use of the task (a point of clarification in the task, or a simple rewording), the task was altered for re-testing. However, if the task showed that little information would be able to be determined on the classes understanding of the concept, even with small changes, the task was discarded. Discarded tasks were replaced with new tasks if another task did not test the same concept. If other tasks tested the same concept, no new tasks were created and tested for inclusion in the system.

#### *2.5.3.1 OR for multiple choice development*

The method described on OR above was used to generate first step responses by students over four semesters in eight lectures (N=1331). Individuals submissions from OR were collected and coded, as in Figure 5, and used to create an electronic list of steps for use in collecting clicker data on first step in one's problem solving process.

This coding process of OR to MC was followed for the spring and fall of 2009 semesters. Over these two semesters two classes General Chemistry 1 (N=521) and Preparatory Chemistry N=176) were provided with the secondary phase of the system (Figure 2). In this phase students were asked to provide a written first step (OR) and then later select their first step from a list of multiple-choice options. The other parts of the phase remained the same as in phase 1 of the instrument.

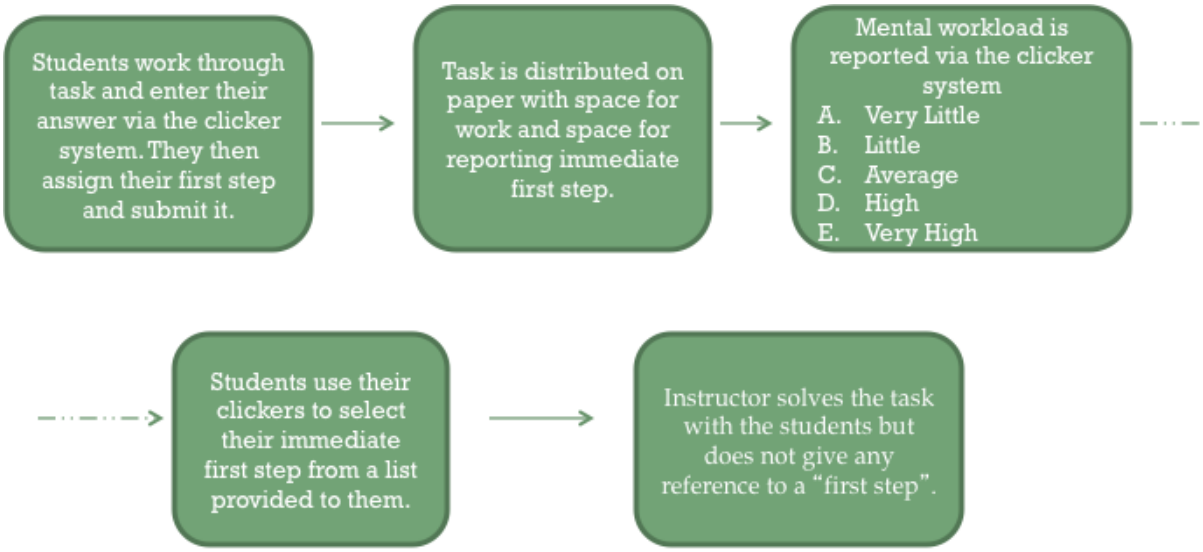



**Figure 5 Secondary Phase of the study: Shows the addition of the multiple choice step in the order and the amount of time allotted for it in the system.**

During the secondary phase of the instrument, the OR and MC responses by students were compared for consistency (Figure 6). Students who selected from the MC list a first step that corresponded with the OR they listed received a score of 1, while those students that did not match received a score of zero.

**Exercise: What is the percent composition by mass of oxygen in aluminum acetate?**

| Steps- "For this exercise, my first step is…" | | | |
|---|---|---|---|
| A | 1 | writing the chemical formula | |
| B | 0 | determining the mass of oxygen | |
| C | 0 | determining the number of moles of oxygen | |
| D | 0 | determining the molar mass of aluminum acetate | |
| E | 0 | Reading the exercise, however I am not sure how to start the exercise | |

Example of OR answers that were used to determine Step A:

"$Al(CH_3COO)_3$"

"$Al(C_2H_3O_2)_3$"

"writing the formula for aluminum acetate"

"writing the chemical formula"

**Figure 6 Example of a task where open response options were matched to multiple choice responses in the CPS system.**

### 2.5.4 Complete electronic testing

After collecting data for comparison of OR to MC in 2009, the study moved on to phase 3 in the spring of 2010 (Figure 7). In phase 3 of the instrument OR was no longer collected. The remainder of the areas of the instrument remained the same. Data collected here in CPS was exported to excel where it was organized for use in SPSS. Once in SPSS that data was analyzed.

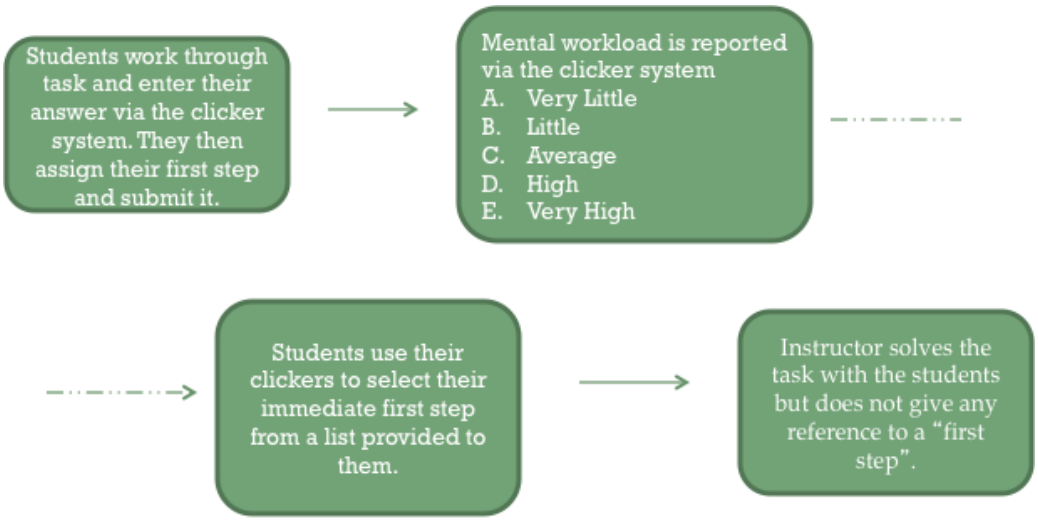



**Figure 7. Phase 3: Electronic data collection**

## 2.6 Efficiency and Complexity Rating System:

Four experts from two universities were asked to review each task in the inventory list for task complexity and efficiency of the first step responses. The complexity rating ratings were done using a method developed by the American Chemical Society Exams Institute (Knaus, Murphy, Blecking, and Holme, 2011). The experts were first provided a list of tasks. Each task was accompanied by instructions for assigning complexity, a response sheet for recording complexity and the rubric to assign complexity. A sample of these instructions, response sheet, and rubric may be found in Appendix C.

The experts were also asked to rate the efficiency of each of the first step responses created from the open responses given by the students (see data information and processing). It was explained to the raters that each answer set would end with the same possible response to allow students to say they do not know how to solve the problem. This was to be rated as "1", as there is no efficiency in problem solving if this step was selected. The rest of the responses were then numbered in order of efficiency for solving the task, with 1 being the least efficient

and the final number being the most efficient. The number of first responses available varies. This is due to the fact that the choices available were generated from student responses.

## 2.7 Eye-tracking

Over four semesters from spring 2009 through Fall 2010, 73 novice and 12 expert interviews were conducted using SensoMotoric Instruments (SMI) eye-tracking technology (Figure 8) (Sensomotoric Instruments, 2011c). Novice interviews consisted of undergraduate students completing the first semester of general chemistry, or who had just completed the first semester of general chemistry and were just beginning general chemistry 2. Expert interviews were conducted for further clarification on possible task improvement, while novice interviews were done to collect information on use of mental effort and first steps in the classroom instrument. A total of 48 tasks from the classroom instrument were coded into webpages using HTML. Their corresponding OR, ME, MC components were also coded into the system using the same method. ME and MC options were coded for selection by participants through use of radial buttons. The interviews were designed to take place in one-hour blocks, in which the participant would complete as many of the tasks and their corresponding components as possible. The participants were encouraged to talk through their thought process while solving the task. Participants who did not complete all the exercises were invited back for another interview. Not all participants returned to finish the set of tasks.

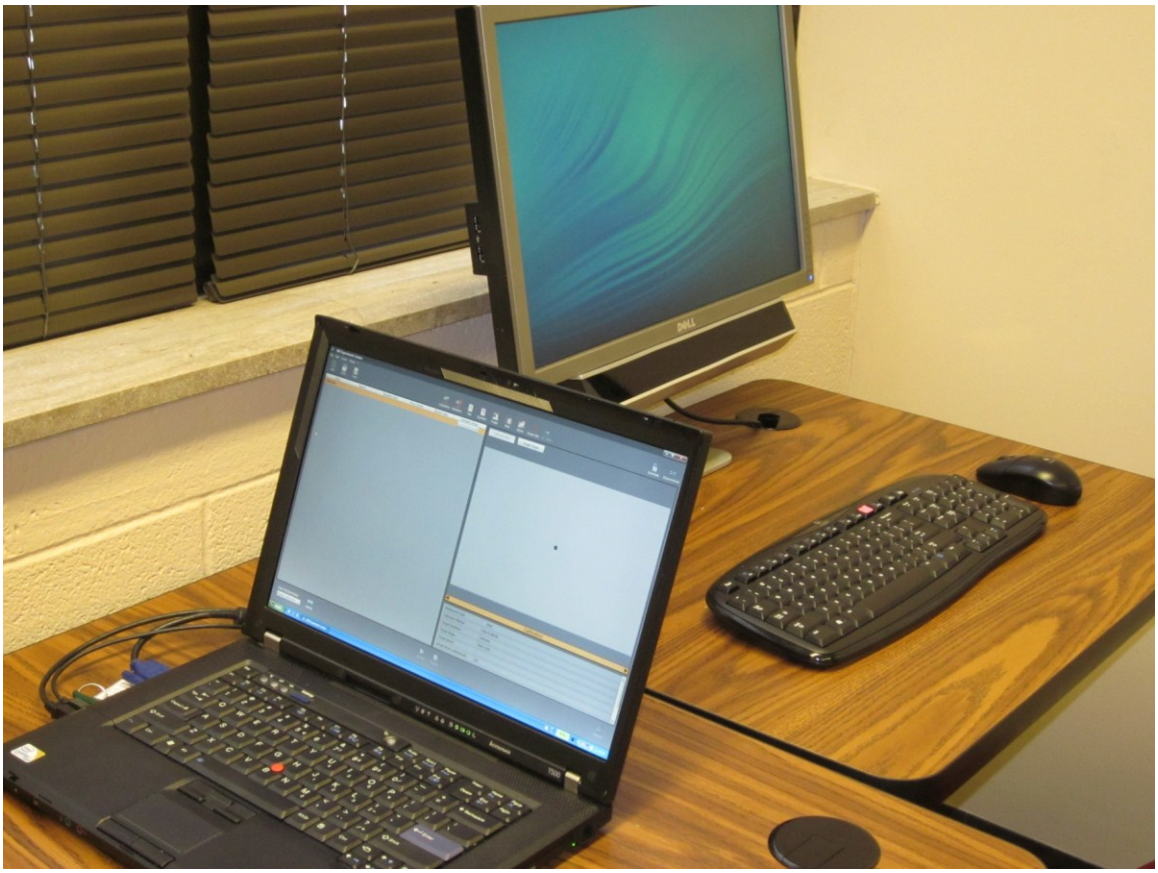


**Figure 8. Eye-tracking computer and RED tracker set-up.**

### 2.7.1 Interview Design

The interviews were set up to follow phase 2 of the classroom study (see open response for multiple choice development). Participants in the interview were presented with a task on screen with an area to type in an answer (Figure 9). Once the participants submitted their response they would click to proceed to the next page. The following page prompted the participants, and presented the participant with an area in which, to provide his or her OR. Upon completing the OR the participant would click a designated key on the keyboard to continue. The participant was then prompted to provide his or her mental effort. After completing this and clicking to move on, the participant was provided a list of electronic first steps with which to choose the first step in his or her own problem solving process (Figure 10).

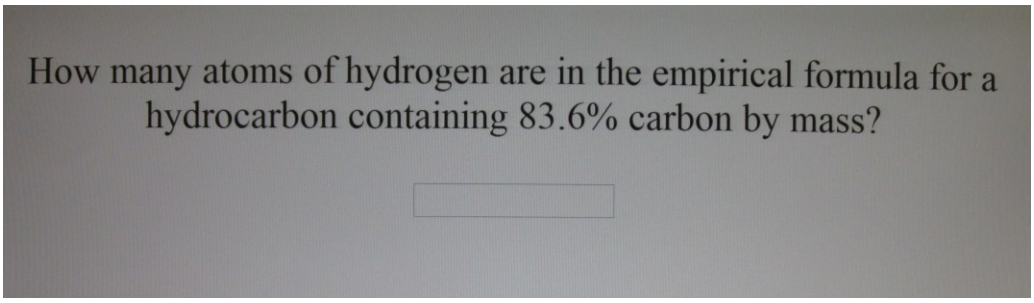



**Figure 9. Screenshot of how the task appears.**

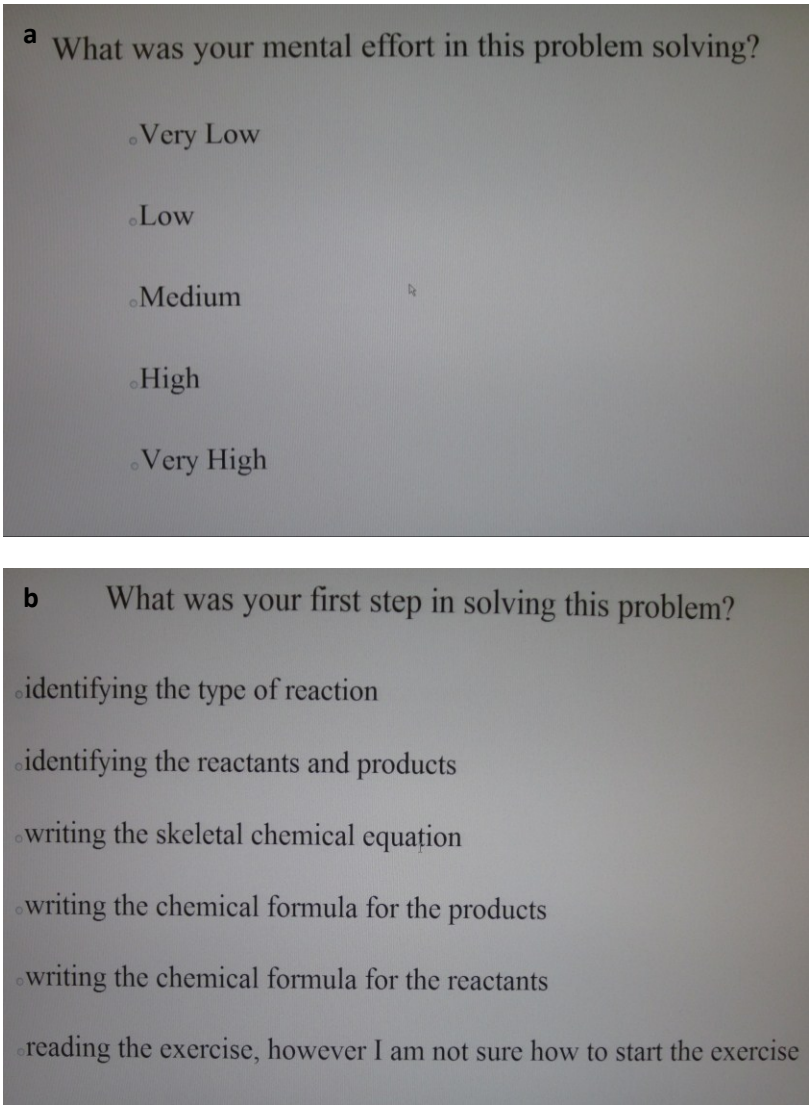



**Figure 10. a) Screenshot of how the ME appears. b) Screenshot of how the MC appears.**

### 2.7.2 Data Collection

The SMI tracker collects data using software programs, iView and an experiment center program (Sensomotoric Instruments, 2011b). The iView software operates and runs the tracker itself, while the experiment center program is used to create the experiment to be run, and connects with the iView software to overlay tracking information onto the experiment file. To allow for accurate tracking of a participant's eye movements in relation to the information presented on the computer screen, a calibration is run at the computer. The calibration functions by providing what appear to the participant as a random point on the screen. The participant follows the point as it changes location on the screen. To adjust calibration and to make sure the participants' eye-movements are recorded properly, the height and distance of the monitor may be adjusted. The way the participant sits may also be adjusted. The system is set so calibration occurs within a set space. The space the participant has to move around during the experiment is 40cmx20cm at 70cm distance from the tracker. The calibration is accurate within 0.4˚ when set to these specifications. Through looking at the laptop monitoring experiment center software, the interviewer can see when tracking is lost and when it is recovered. This allows for the interviewer to let the participant know when to adjust their posture while sitting at the tracker during the interview. When tracking was lost from participants' change in posture, it occurred because the distance from the participants' eyes to the tracker moved outside of the calibrated range. Tracking was also lost of the participants looked away from the screen. The tracker only records eye data when the reflection of the light from the tracker is bounced back off of the eye. When participants look away from the screen, tracking is lost. This often occurs when the participant looks up and to the left while thinking, or when the participant looks down at the desk while working problems on paper. To accommodate for the precision of the tracker, the pages were coded with a font size large enough to allow for judgment as to what the

participant was looking at. This was the set-up for the RED tracker running at 60Hz. The remainder of the data collection occurs via manual work or student entered information.

Participants enter in their answers via the computer system through the method discussed in the interview design of the tracker. Meanwhile, the interviewer (myself) recorded notes on comments made by participants. The task solution, ME, and MC selection were also recorded manually. All of this information is available via the computer database that is created for each participant; however, due to errors encountered during initial trial testing of the system, information was kept manually via a record sheet in case the data needed to be accounted for.

### 2.7.3 Data Evaluation

Upon completion of the interview the data is available for analysis through the BeGaze software (Sensomotoric Instruments, 2011a)that accompanies the system. The BeGaze software allows for analysis of responses through looking at time-on-task, or the amount of time a participant spent on a given area of the task or the task as a whole; heat maps, areas of focus by a participant generated via the amount of time spent looking at a given area; scan paths, or the path the participants eyes created when solving the task; fixations, which are places a participants eyes lingered and the time for that fixation was recorded; areas of interest (AOI) in which a specific area of a task or its component may be specified for extra information (this information includes the number of times the piece of information was referenced and the total time spent on the area); and pupil diameter, where the tracker takes measurements of the pupil size throughout tracking. All of this data may be viewed individually for participants or for all participants as a whole in the BeGaze program through videos and screenshots of the information as it was recorded. The information may also be exported into excel as raw numerical data for analysis.

## All of the above mentioned methods of data analysis were utilized during evaluation of the tracker data collected for this project. 2.7.4 Statistical Methods and Data Collection

### *2.7.4.1 Likert Scale*

A Likert scale is a scale which measures attitude or opinion based on a range of provided responses (Likert, 1932). Since its development in 1932 the Likert scale has been used in research (many examples of which are given in chapter 1.2.2.2 ). The scale ranges based on the number of response options provided. Typically Likert scales range between five and seven response options. In the case of the Rapid Knowledge Assessment, a five point Likert scale is used to determine mental effort (Chapter 2, section 4.2.2).

### *2.7.4.2 Pearson Correlations versus Spearman Correlations*

Pearson product-moment correlations coefficient, r, is used to assess the direction and strength of the relationship between two continuous variables (Gravetter and Wallnau, 2007). The closer the value is to 1 (or -1, depending on the direction of the relationship expected), the stronger the relationship between the two variables. The significance of the value (α) is then compared to the probability of obtaining the values calculated (p-value). If the α -value is 0.01 (1%) or 0.05 (5%), it means that there is less than a 1% or 5% chance of the result occurring by chance. Therefore, if the p-value comes back less than the α-value it is statistically significant. This means that under the most stringent of circumstances it is highly unlikely that the result occurred by chance, i.e. it is 99% or 95% certain that that is the true correlation. In most social science research, an alpha of 0.05 is the minimum acceptable value, which is the standard followed in this research.

Spearman correlations are correlations run between two variables where the variables have no implied numerical value. Data for comparison is therefore assigned numerical values. Since

there is no set numerical value, or integers between the assigned values of the data, the Spearman correlation accounts for this fact during analysis. Spearman correlations are measured using Spearman's rank correlation coefficient, rho. The Spearman correlation then measures the relationship between the two values, as is done in the Pearson correlation. The significance of this value is also measured in the same manner as it is for the Pearson correlation.

#### *2.7.4.3 Cronbach's Alpha*

Cronbach's alpha is a measurement of reliability. The analysis compares the number of useable cases, cases that include all pertinent information, against one another. It then looks at the repeatability of the information analyzed for consistency amongst the cases. In social science statistical analysis, a value of 0.70 is considered an acceptable value for consistency. This same value will be used to check for acceptable consistency, and therefore reliability, in this work.

# Chapter 3: Assessing viability of the Instrument

## 3. Introduction to proof of concept

The concept of rapidly assessing student knowledge in the classroom holds great potential and use in the lecture setting. Kayluga and Sweller (2004, 2005) found this to be true for the linear domain of math. In order to develop such an instrument in a complex domain like chemistry, one first needs to determine if students approach problem solving in similar measurable manners, and, if so, whether those processes relate enough to one another to be categorized in recognizable units by novices. This means that one must first determine the multiple ways students approach solving the same task.

To test the concept that an in-classroom instrument may be able to measure knowledge in real time, based on the theory of previous work, the collection and validity of the first steps needed to be established for continuation of the project. To accomplish this, first steps of the participants needed to be evaluated in comparison to other measures of assessment used within the classroom. The traditional model of assessment is performance, and therefore is widely accepted for use in assessment studies. Therefore, the aggregate scores for the areas of performance and first-step performance were used for each participant in comparison to his or her third exam score, two final exam scores, and percent grade in the course. Aggregate scores were used to allow for a comparison of the class as a whole over the course of the semester. Each first-step performance was coded as correct or incorrect based on the submitted open responses application to the task. This allowed for generation of a multiple choice list of first steps (discussed later). If the solution provided for the step was completed properly, data was generated based on the type and number (correct or incorrect) of responses submitted by each

lecture section. Through comparison of students' performance on the task response and first step a better relationship between the initial thought and final solution is visible. All three lecture sections of the course were compared using Pearson correlations, and then combined together to test for overall validity of the instrument. This is an important comparison as the three lecture sections were taught using either didactic or active learning. This final piece allows one to look at the possibility of the validity to be disrupted based on the type of learning environment.

## 3.1 Determination of step viability

To determine if students commonly agreed on how to start solving tasks, a range of fifteen tasks were developed for testing and comparison. Of the fifteen tasks, twelve were valid for comparison. Tasks one and ten were not included in the analysis, as only one lecture section was used to collect data for the tasks, and task thirteen was excluded based on missing task performance data. The areas within the tasks compared included performance on the task, performance on the first step reported (i.e. if the first step was performed correctly or not), score on the ACS standardized final exam, the second final exam (written by the course instructors), and the third hourly exam (which covered aspects of each of the twelve task categories listed in Appendix B, and the final percent grade in the class. These areas were compared using Pearson product-moment correlations (Appendix D) to determine the degree to which students' reported first steps are valid in the creation of a rapid knowledge assessment instrument.

Two of the six areas used for the correlations were coded dichotomously (task performance and first step performance). These two aspects were coded this way as they fit the profile for categorical data, or data that only fits directly into a specific description (i.e. correct or

incorrect). The data was coded in this way for both lecture types. Table 3 shows the data for the correlations comparing the traditional didactic lecture (lecture 401, treatment 3) and the two active learning lecture sections (lectures 402, treatment 1, and 403, treatment 2). In the didactic lecture, significant correlations were present at the 95% and 99% confidence intervals across all six areas of comparison for the tasks. In the first active learning lecture significant correlations were present for all areas except first step performance. Pearson correlations for first step performance to Final 1, Final 2, Exam 3, and class percent were found to not be significant. Upon examining the first steps submitted, it was found that a higher number of students in treatment one attempted to solve the problem completely. Often the students were incorrect in this attempt. It is possible that this does not correlate to the external measures because the complete problem as a first step was incorrect due to calculation error, or that the students became clear in their error post task. However, in the second active learning lecture section these correlations were found to be significant at the 0.01 level. In this section the only discrepancy was found between Task performance and the Exam 3 score. Here no significance was noted. In the addition, if all three lecture sections are combined for comparison each area correlated within the tasks comes back as significant at the 0.01 level.

**Table 3. Correlation Values Spring 2007**

| | Lecture 401 (n=51) | Lecture 402 (n=69) | Lecture 403 (n=43) | All Lectures (n=163) |
|---|---|---|---|---|
| Task Performance to: | | | | |
| First Step Performance (TP:FS) | 0.642** | 0.274* | 0.511** | 0.458** |
| Final 1 (TP:F1) | 0.415** | 0.450** | 0.539** | 0.388** |
| Final 2 (TP:F2) | 0.546** | 0.398** | 0.470** | 0.406** |
| Exam 3 (TP:E3) | 0.486** | 0.563** | 0.278 | 0.426** |
| Percent in Class (TP:PC) | 0.650** | 0.527** | 0.379* | 0.520** |
| First Step Performance to: | | | | |
| Final 1 (FS:F1) | 0.547** | 0.170 | 0.433** | 0.353** |
| Final 2 (FS:F2) | 0.459** | 0.203** | 0.584** | 0.384** |
| Exam 3 (FS:E3) | 0.291* | 0.117 | 0.408** | 0.263** |
| Percent in Class (FS:PC) | 0.511** | 0.136 | 0.446** | 0.310** |
| Final 1 to: | | | | |
| Final 2 (F1:F2) | 0.664** | 0.843** | 0.708** | 0.777** |
| Exam 3 (F1:E3) | 0.621** | 0.709** | 0.426** | 0.583** |
| Percent in Class (F1:PC) | 0.714** | 0.897** | 0.641** | 0.742** |
| Final 2 to: | | | | |
| Exam 3 (F2:E3) | 0.626** | 0.624** | 0.583** | 0.599** |
| Percent in Class (F2:PC) | 0.792** | 0.881** | 0.797** | 0.803** |
| Exam 3 to: | | | | |
| Percent in Class (E3:PC) | 0.811** | 0.791** | 0.664** | 0.735** |

*p<0.05 (2-tailed) **p<0.01 (2-tailed)

### 3.1.1 Didactic versus Active learning

Since the three lecture sections were not taught by the same instructor, but by two different instructors, any differences needed to be examined for a relationship to the instructors' methods. One of the active learning lectures was taught by one instructor (treatment 1), while the other active learning lecture (treatment 2) and the didactic lecture (treatment 3) were taught by another instructor. This allowed for a comparison of differences between didactic and active-learning, as being taught by the same instructor allowed for a direct comparison. The fact

that a second active learning section was taught by another instructor allowed for comparison of instructor effect through examination of the two active learning sections to one another. When looking at the didactic lecture section significant correlations appeared between all six areas of comparison. Significance between these areas shows that there is internal (between the areas of first step performance and performance on task) and external validity (first step performance and performance to final exam scores exam scores, grade in class, and third class exam) for use of the correctness of first steps to measure a student's ability to problem solve within a task. This is not yet looking at the efficiency of the step, so much as it is looking at if the student performed their first step correctly. Since there is significance at the 0.01 level between the first step performance and the task performance, one can see that there is a relationship between correctly identifying and carrying out a first step and the performance to any of the exam scores or the percent grade in the course. In terms of treatment 3, it proved worthwhile to investigate the use of first steps to rapidly measure one's understanding in the problem solving tasks.

In comparison to treatment 1, however, one sees that significance was not always found when looking at external measures of validity. Here, first step performance to exam three and percent course grades was not found to be significant, but the final exam scores were found to be significant. External validity here appears to be contradictory, but when compared to treatment 2 one sees significance for each area compared to the first step performance. It is not clear why there are contradictory measures of significance between the two lecture sections; one possible reason is the difference in class size. However, enough evidence is present to support using first performance measures as a means to determine a students' understanding of a concept within an active learning environment.

When comparing the two types of lecture style, didactic versus active learning, one sees that overall independent of the type of lecture there is a relationship between the first step submitted by students and the students' performance and understanding of the concept of the task. This is seen in the combination of the three lecture sections, when significance is shown across all six areas of comparison at a 0.01 level.

## 3.2 First step viability

When determining the usefulness of first steps in rapidly assessing chemistry it helps to look at whether the steps were carried out correctly. This is something that lends towards the reliability of using first steps in assessment measurements. Kalyuga and Sweller (2004, 2005) did not include this in their study on algebra. The correlations described in this section of the study will show that success on the task is directly related to the success on the first step, which is why it is not necessary to measure each separately. The performance on the task alone will give a measurement of the success of the first step.

In addition to examining the teaching methods employed in the lectures, the fifteen tasks were analyzed and coded for first steps (described in the methods section). The number of steps submitted by the students per lecture varied depending on the task. In general each lecture section provided similar numbers of first steps (Table 4), except on tasks 9, 12, and 13 where a discrepancy was noted. Task 9 showed that treatment 3 submitted a higher number of responses than treatments 1 and 2. On task 12 treatment 3 submitted four more steps than treatment 1, which submitted four more steps than treatment 2. On task 13 treatment 2 submitted a greater number of steps than the other two lecture sections. Figure 11 shows the comparison of the number of responses by step per lecture section to the steps provided in the open response. This figure demonstrates that students gravitate towards certain first steps

when solving tasks. The tasks varied in the number of steps submitted, and, in general, it can be seen that the number of responses remained similar across the sections for the majority of the tasks. However, this does not mean that the first steps submitted were the same or similar for each lecture section. By comparing the open-response submissions to one another, and the number of responses of each type of open-response, it is possible to determine if the students largely agree on a series of first steps for certain concepts.

**Table 4. Spring 2007 Rapid Knowledge Assessment**

| Task Number | Number of Steps Per Lecture Section | | | Total Steps/Task | # Excluded as Steps |
|---|---|---|---|---|---|
| | 401 (treatment 2) | 402 (treatment 1) | 403 (treatment 3) | | |
| 1 | 6 | - | - | 6 | 1 |
| 2 | 4 | 4 | 4 | 4 | 1 |
| 3 | 6 | 6 | 6 | 6 | 1 |
| 4 | 7 | 7 | 7 | 7 | 1 |
| 5 | 7 | 8 | 8 | 9 | 2 |
| 6 | 5 | 5 | 5 | 5 | 1 |
| 7 | 9 | 9 | 9 | 9 | 2 |
| 8 | 10 | 10 | 11 | 13 | 1 |
| 9 | 17 | 12 | 11 | 19 | 1 |
| 10 | - | 6 | - | 6 | 1 |
| 11 | 13 | 12 | 13 | 14 | 1 |
| 12 | 16 | 12 | 8 | 21 | 1 |
| 13 | 11 | 10 | 24 | 28 | 1 |
| 14 | 13 | 15 | 14 | 22 | 2 |
| 15 | 6 | 9 | 7 | 10 | 1 |

“Others”/”No Clue”/Miscellaneous Steps excluded from total steps per task. This count is given in the last column.

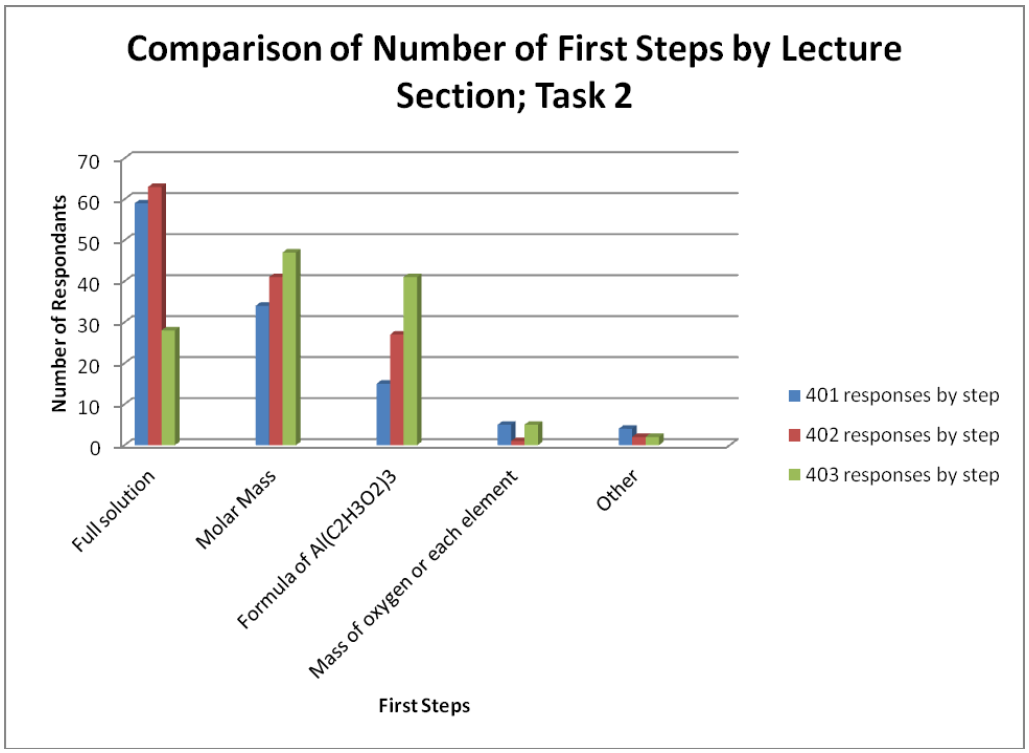



**Figure 11. Comparison of first steps via OR by lecture. Where 402 is treatment 1, 401 is treatment 2, and 403 is treatment 3.**

### 3.2.1Step analysis

Because the number of steps varied based on the task provided to the students, and some tasks differed by lecture sections (tasks 9, 12, and 13), one must look at the variation in the number of steps across the lecture sections. In task 9, treatment 3 submitted a higher number of responses than treatments 1 and 2, which might lead one to believe that the type of lecture plays a role in a student's method to solving a task. When looking at task 12, one sees that there is a difference in four submitted steps between each treatment group, and in task 13 treatment 2 submitted a greater number of steps than the other two treatments. There is no explanation why this occurred, and because the variations were across both the didactic and the active learning lectures, one cannot make the assumption that the differences demonstrate a relationship to the type of learning (active versus didactic) done in the lecture sections. This is due to the fact

that the number of responses varied within the active learning sections as well, and not just between the didactic and active learning sections. Also, one cannot claim to know the type of learning taking place by the students, just the instructional method presented. This is why the instrument is designed to evaluate student process, which lends toward student efficiency in problem solving. The instructor's role is to facilitate a student's own construction of knowledge.

## 3.3 Step Processing

In task 2 for Figure 11 there are 5 steps provided by the students across the three lecture sections. It can be seen that of the five first steps provided only four steps are actual processes that would lead to the solution of the task. For this reason, the final step labeled "other" is discarded. "Other" is considered to be anything not related to the solution of the task at hand. Figure 12 gives the responses by step for task 12 across all three lecture sections. The figure shows that although many steps are present – 21 first steps were provided in the open response section, with the step "other" once again discarded – not all of them are present for each lecture section, nor are they represented in similar proportions when present in multiple lecture sections.

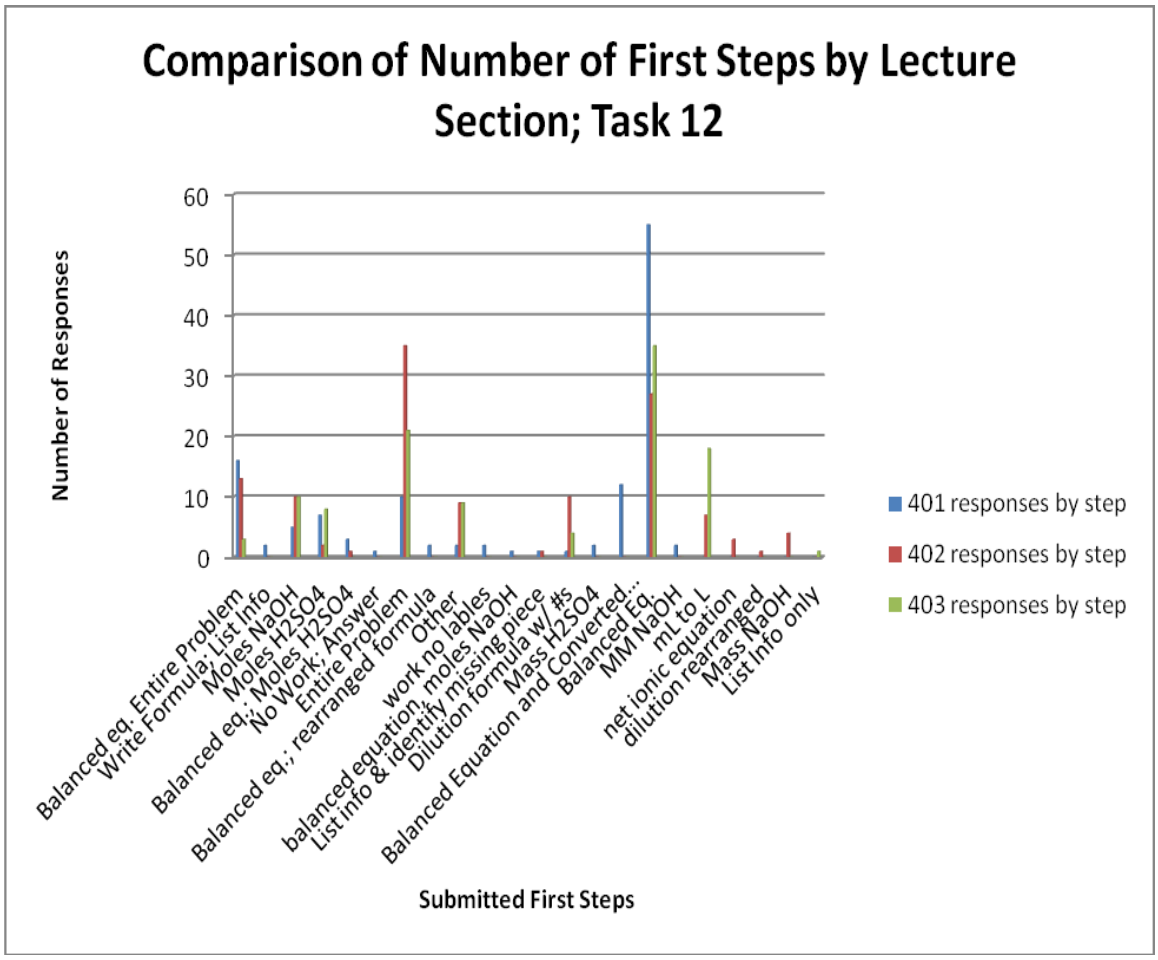



**Figure 12. Comparison of first steps via OR by lecture section, where many first steps were submitted.**

In comparison, Figure 13 and Figure 14 show the collection of first steps in comparison by total across lecture sections. It is seen that in Figure 13, with only four first steps submitted, there is a high percentage of students who agree on the number of overall applicable steps. In Figure 14 even though there are 21 valid steps provided by the students, only nine of the steps show a high agreement amongst the steps submitted. However, many of the steps did not represent enough of the students to justify inclusion in the instrument. This information distracts from outcome of the matching student population. Because the goal of the instrument is to include as many students as possible while coding open-ended responses into potential first steps, by setting a threshold of 5% for inclusion in coded first step responses those open-ended responses

not in agreement with the class are sorted out (Figure 15). This allows for a clearer view of where the class lies as a whole, and shows that there is still agreement among first steps.

With the possible diversity in the number and type of steps submitted by students, one might have cause to argue that the steps cannot be paired down into a useable list for an electronic instrument. However, if one examines the first step by frequency as in Figure 14 versus Figure 16, where the average percent response to a given first step is shown, one sees that the nine first steps that were submitted and agreed upon the most still cover a large range of the class (from 32.87% to 3.37%, which covers 92.42% of the class as a whole). This same agreement among over 90% of the respondents is found within the majority of the other tasks.

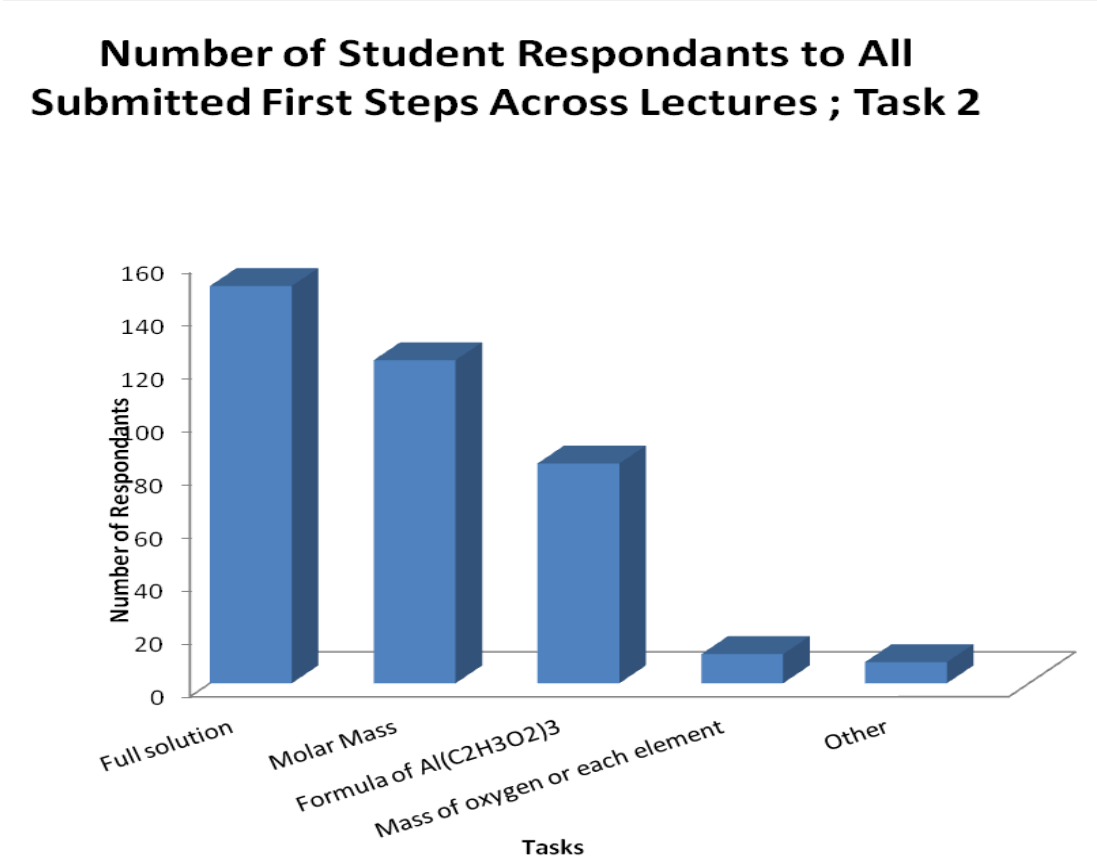



**Figure 13. Open-ended response submissions and the number of responses for Gen Chem I.**

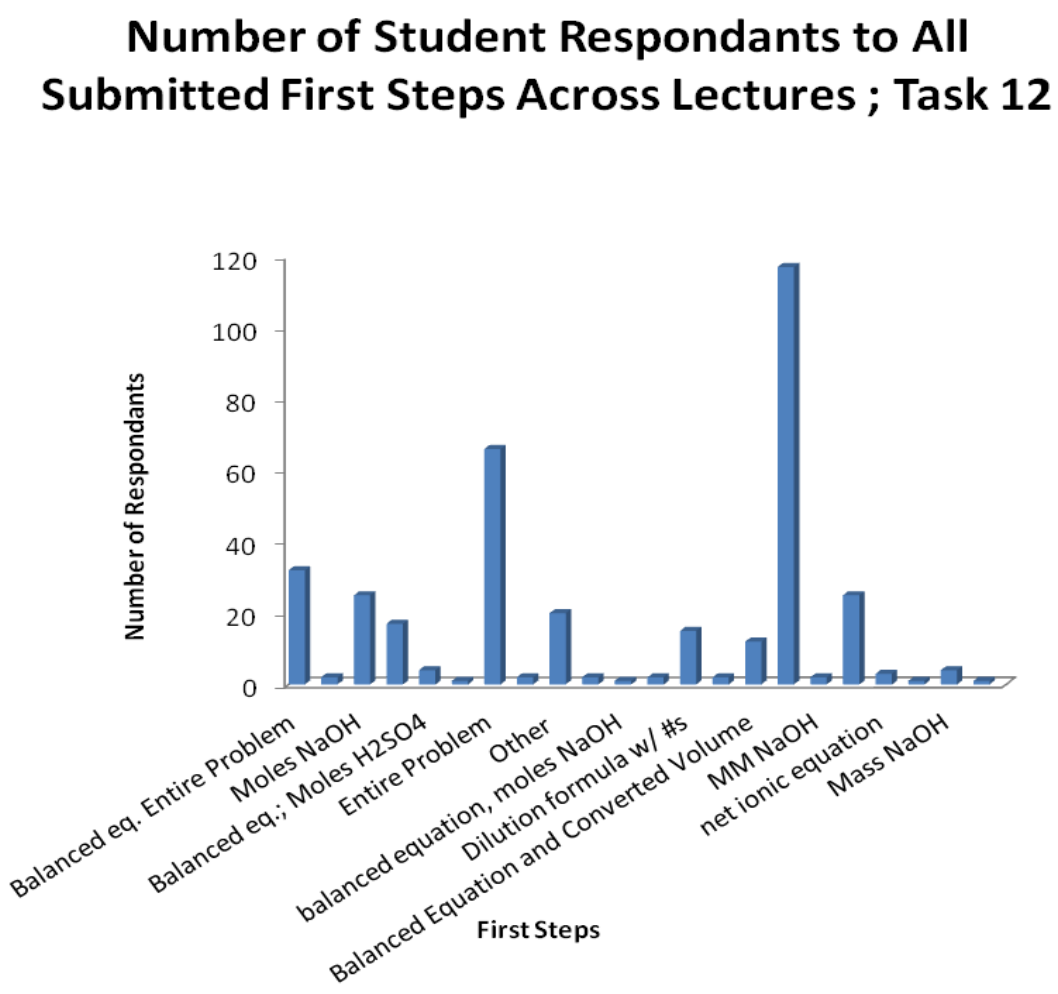



**Figure 14. All open-ended response submissions for task 12 from Gen Chem I.**

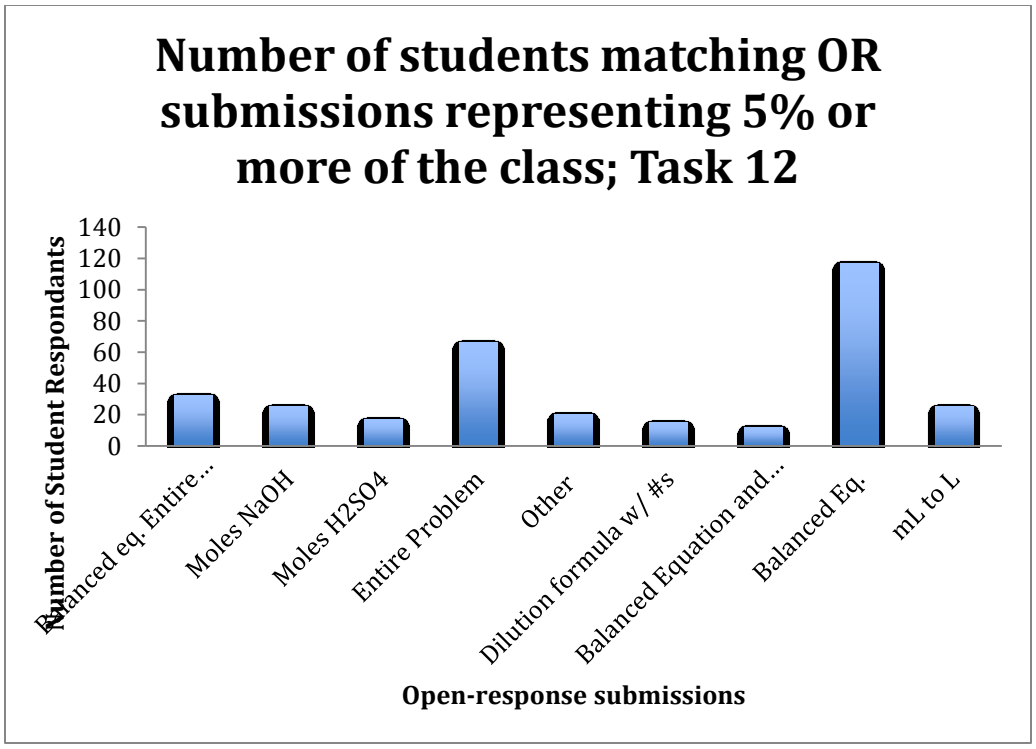



**Figure 15. Viable open-ended response submissions for task 12 after inclusion threashold of 5% was set.**

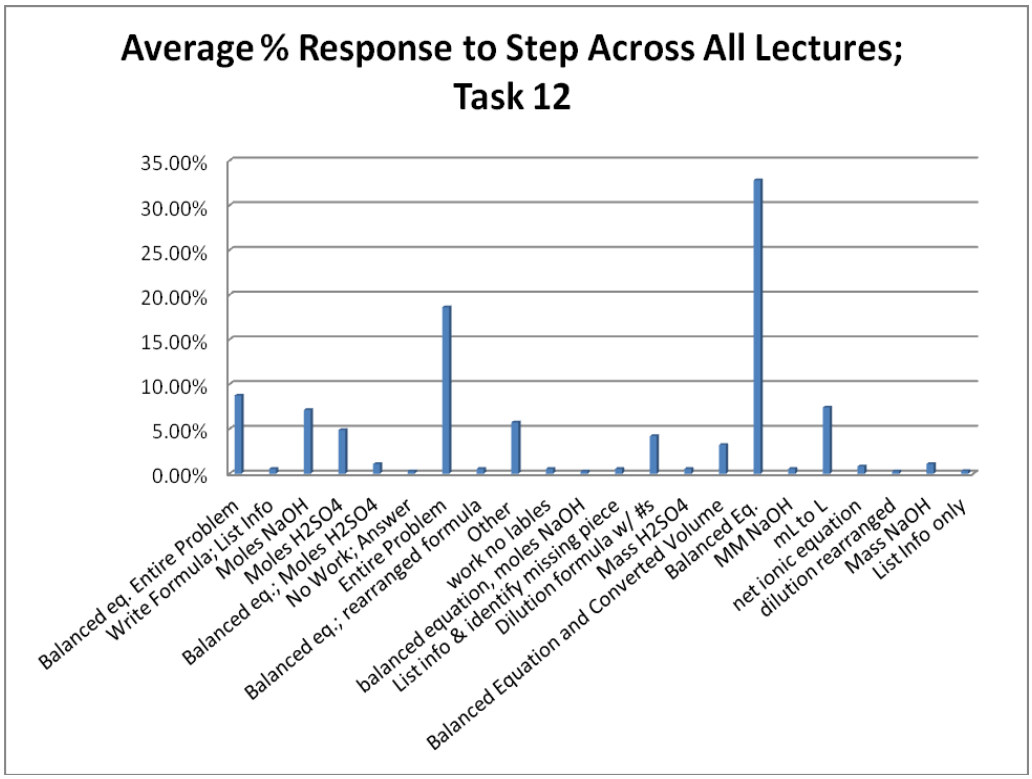



**Figure 16. Percent response of students to each open-ended response before 5% threshold set for task 12.**

### 3.3.1 Process analysis

Upon analyzing the first steps submitted by the students, it is seen that the number of first-steps responses varied based on the task assigned to the students. In tasks 2 and 12, shown in Figures 11 and 12, the number of steps varied greatly. Although the number of steps varied, it is seen that there is agreement across the sections on the steps submitted, even if it is not complete agreement. Figure 11 and Figure 12 demonstrate that the number of steps present allows for providing as little as five first steps for a student to consider when solving a task. In comparison to Figures 13 and 14, where tasks 2 and 12 are viewed by total respondents to a step across all three lecture sections, one can see that not every step submitted is necessary for student to have an option when it came to solving the task. That is, more than one step can be combined into a single option presented to the students. In fact, if one considers what was discussed in

section III above, and what is seen in Figures 13,14, and 16, the great amount of agreement upon the top most first steps shows that the majority of the students are in agreement upon the range of first steps to solve the tasks. This demonstrates that student generation of steps may be combined to provide a set of first steps for the student to choose from, and to accommodate the majority of the students utilizing the instrument. As discussed in section III, not all tasks found high levels of agreement for a few steps within each task. However, when only a small subset of students selected these options it was not an issue, but when a high percentage selected them it showed actual difference between the treatment groups. This demonstrated that there was a difference between the types of learning (active versus didactic) taking place. This does not mean that an instrument designed to rapidly assess knowledge would not work for a task such as this, but that the tasks require refinement for further testing and analysis.

## 3.4 What this all means

When developing a multiple-choice list from open responses the number of first steps given in open response format by students to a complex task is limited for even a large sample. This is beneficial for turning open-ended responses into multiple-choice responses. The performance on the task in comparison to these first step responses correlate well with other measures, indicating that performance alone is a valid way of assessing performance on the first step submissions. Therefore, it is not necessary to develop a measure to assess the performance on the step itself. Through analysis of the type and number of first steps provided by the students, in comparison to various forms of external and internal sources to test validity, first steps are found to a compelling way to assess students' knowledge. From the number of first steps provided by students, to the categorization of the these steps while eliminating "other" responses, and determination that performance on task will correctly evaluate performance on

first step, one can use external measures of proficiency to understand how a student starts to solve a task, while providing valuable insight into how well a student has learned a concept. First steps provide a rich set of information not see in performance alone, demonstrating that more research on the development of the instrument will be beneficial towards understanding how to determine appropriate assessment of students' knowledge. Therefore, through development and implementation of a rapid knowledge assessment instrument in the classroom, one can gain a better understanding of how a student's processes are related to his or her knowledge base in chemistry. The way a student approaches solving a problem is directly related to his or her schema.

# Chapter 4: Confirmation of student and expert agreement for use of electronic first step implementation

## 4.1 Establishing Reliability

It was explained earlier that the multiple-choice first steps were determined through open-ended response. This process stemmed from the work done by Kalyuga and Sweller (2004 and 2005), where students listed a first step in the task solution. The math tasks used in their work followed a linear process, and therefore only had a set range of first steps. Chemistry tasks, not being linear processes, therefore, need to be vetted for first step options. Then, it needs to be determined if open-responses are able to be grouped for a reasonable number of electronic options. A larger question which needs to be addressed in this phase is, "How does the presentation of first steps in a multiple-choice format affect the reported first step of students?" Another way to ask this is, "Can a student's first step be accurately captured through the use of multiple choice responses compared to open responses?" The following chapter reflects upon the analysis of the percent response by students as a means to establish reliable first steps for the instrument. It will also discuss how these steps were then analyzed for use in the instrument to determine student efficiency using Cronbach's alpha (a reliability analysis) to determine agreement amongst experts. The agreement amongst experts, in terms of efficiency, is used to determine the efficiency rating applied to each of the first-steps provided by the students. To ensure the reliability of the efficiency ratings applied to the first-steps, percent agreement amongst the experts on the ratings will also be examined.

### 4.1.1 Collection of First Steps

As discussed in the methods section all first steps were collected on paper starting in the fall of 2008. During this time the coded steps were counted per grouped answers (i.e. if one person gave the formula "$CO_2$" and someone else wrote out "carbon dioxide" that would be grouped together). These answers were then taken as a percentage of the total responses (Figure 17). The purpose was to determine if the students as a whole agreed on any number of steps. It was determined from the collected data that in general at least 90% of students could agree on between four to eight steps. This was very important for reliability, as the instrument needed to have a set of reliable student based answers to move forward with the multiple choice steps. It also showed validity in ability to capture a majority of possible student responses. A retesting of the pilot phase (from spring 2007) was carried out for one semester in fall 2009, in both general and preparatory chemistry before compilation of the first multiple choice selection list in spring 2009. This retesting was important for multiple reasons, including establishing preliminary validity and reliability. The second testing also included a new course (General Chemistry I). Here, retesting and finding the same or similar responses to the first step request supports the conclusion that the tasks themselves, and the first step reported for each task, was valid for introductory chemistry in general. Secondly, the reliability of the first step reporting, in an open-response format, was established for both preparatory and general chemistry students. The multiple-choice selection list was then used as a way to the test the degree to which the students' responses during the open-response phase accurately were reflected in the multiple-choice selections.

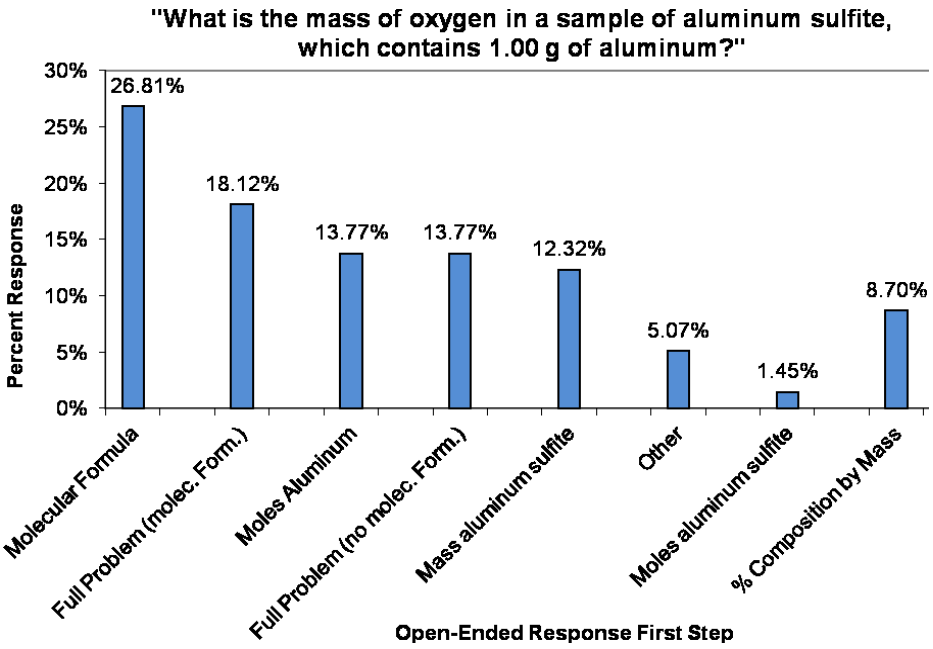



**Figure 17. Responses from general chemistry students to a task regarding properties of matter.**

### 4.1.1.1 Generation of electronic response options

As discussed earlier, multiple choice options for the electronic portion of the system were generated from open-response items that were coded into groups of same responses (just different variations), and then scored as “1” or “0” for matching and non-matching (Figure 18). It is important to note that students were provided with an “out” response “reading the exercise, however I am not sure how to start the exercise”, and that full solutions of the problem were excluded, as they did not signify the use of a specific first step but rather a misunderstanding of how to report their first step, which was addressed with more refined training on the use of the instrument. The coded and grouped steps were then organized into a list for testing, in which the tasks were ordered from shortest to longest length (with the exception of the final step

always being the “out” step). The ordering of the tasks was done to reduce cuing of a particular step being more accurate than another.

**Exercise: What is the percent composition by mass of oxygen in aluminum acetate?**

| Steps- “For this exercise, my first step is…” | | |
|---|---|---|
| A | 1 | writing the chemical formula |
| B | 0 | determining the mass of oxygen |
| C | 0 | determining the number of moles of oxygen |
| D | 0 | determining the molar mass of aluminum acetate |
| E | 0 | Reading the exercise, however I am not sure how to start the exercise |

Example of OR answers that were used to determine Step A:

“$Al(CH_3COO)_3$”

“$Al(C_2H_3O_2)_3$”

“writing the formula for aluminum acetate”

“writing the chemical formula”

**Figure 18. Example of a task where OR options were matched to MC responses in the CPS system.**

Through use of the discussed coding measures, class OR types were compared. Figure 19 shows a comparison of open-responses between the two courses, and how these were related to the generation of the first step items in the electronic instrument. Once MC lists were generated, it was possible to compare OR submissions to MC selections of students utilizing phase 2 of the instrument. Figure 20 demonstrates the ability of students to match their first step during OR to the MC list provided.

Task: What is the mass of oxygen in a sample of aluminum sulfite which contains 1.00 g of aluminum?

1. Writing the chemical formula
2. Finding moles of oxygen
3. Finding moles of aluminum
4. Finding the mass of aluminum sulfite
5. Finding the moles of aluminum sulfite
6. Identifying the mole ratio of oxygen to aluminum
7. Reading the exercise, however I am not sure how to start the exercise

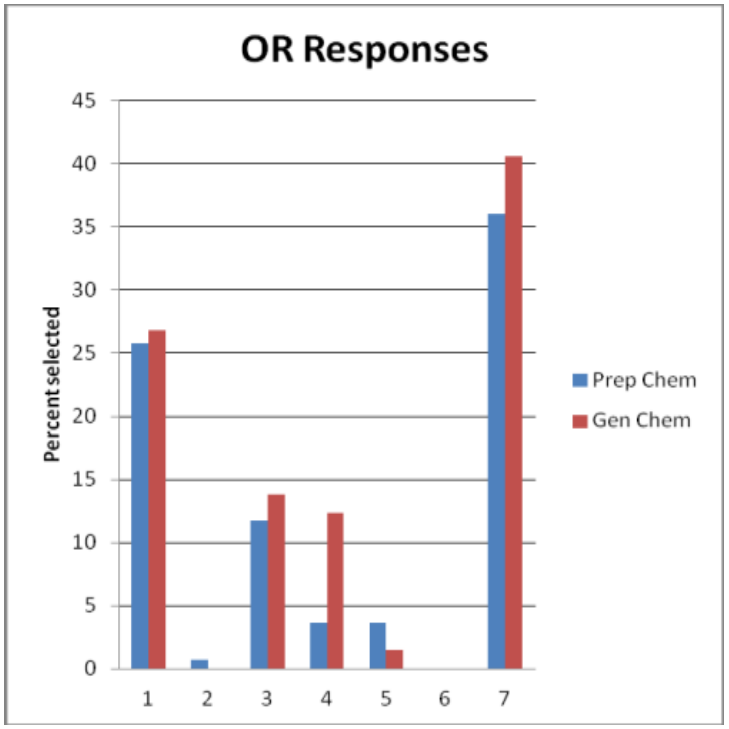



**Figure 19. Comparison of a formula calculation tasks multiple-choice first steps generated through OR submissions by preparatory and general chemistry students. Numbers on the x-axis correspond to the numbers of the multiple choice selections listed for the electronic instrument.**

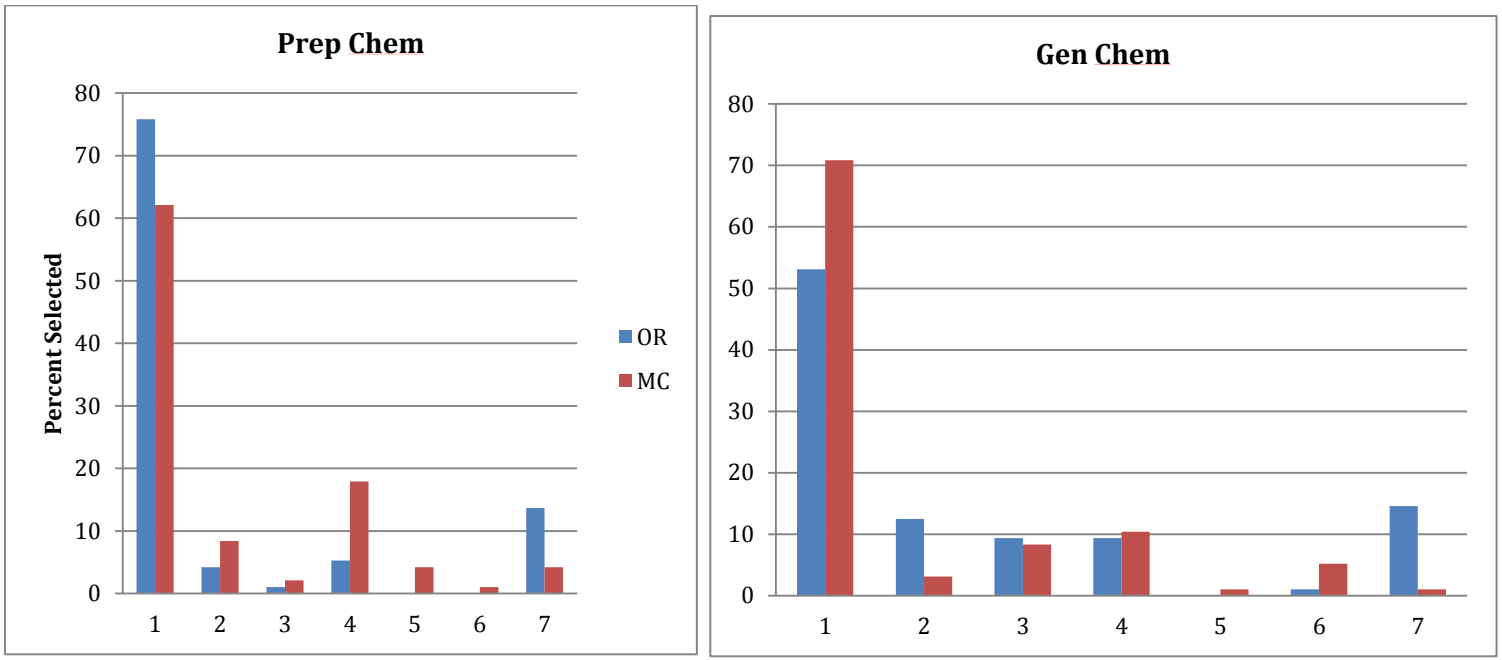



**Figure 20. Distribution of open-response and multiple choice responses for general chemistry and preparatory chemistry students for the same task and steps represented in figure 19. The numbers on the x-axis here correspond to the step numbers listed in figure 19.**

From comparison of the OR alone one can see that students' open-responses matched within seven similar steps at least 85% of the time for both general and preparatory chemistry students (Figure 20). While this number is not the 90% mentioned earlier, it does show that a set of steps are validly determined for a majority of students through OR collection. From Figure 19, one can see that when comparing open-response answers to multiple-choice answers by step type for both classes (n=228), 74.6% of students match their open-ended response to their multiple-choice response. However, this also demonstrates that not every student matches. While students not matching their open-response to their multiple choice response presents a concern, at the point in time this data was collected it was not yet able to be determined if the students were changing responses due to cueing (cueing is when a stimulus alerts the subject to another piece of knowledge not originally accessed or connected to his or her original thought process), pressure to present the what the student perceived to be the best or most accurate step, or because they could not remember what they had originally submitted on the open-response form. Because previous studies on the capacity of working memory with access to long-term memory showed the capacity of memory to be 7 ± 2 pieces of information in working memory (Miller, 1956), and the fact that this phase of the experiment was designed to minimize the amount of time between submission of the open-response and the selection of the multiple-choice first step, it is reasonable that students do not forget the first-step that was submitted on paper. For this reason our attention turned to the factors of cueing and the pressure to provide what the student perceived to be the best possible step. In order to rule out the perception that there was a "correct" response for a first step, the following semester instructions were given to the class on what the first-step means and that there was no such thing as a correct or incorrect first step. By providing this instruction a comparison can be made between open-ended and

multiple choice responses including or excluding external responses (Figure 21, Table5). By removing students whose open-ended responses did not relate to the task, or the solution of the task, there is improved reliability of the students selecting the open-response submitted during the first portion of the task response. However, it is also noted that after instruction the reliability of the students to select the same multiple-choice step as was submitted during the open-response time also improved for those students who gave an external type response (a response that was not coded into a multiple choice option due to it being reported by less than 5% of respondents for the class). This can be seen in the changes that occur for content areas that had multiple tasks tested. Since each task had steps that did not fit into the steps coded into the instrument, it was expected that some reliability percentages would be low. In the oxygen example (Figure 18) it was shown how tasks were coded (1=match, 0=no match), and how there was no room given for partial matches. The strictness of this coding ensures there is no interpretation as to what the student may have been thinking. Therefore, steps not coded into the multiple-choice list may have an aspect to them that the student called out when choosing a first step from the multiple-choice list. This would cause the reliability to seem lower than it actually stands. This shows that the electronic first steps provided to the students are a reliable way to collect first steps needed for the instrument.

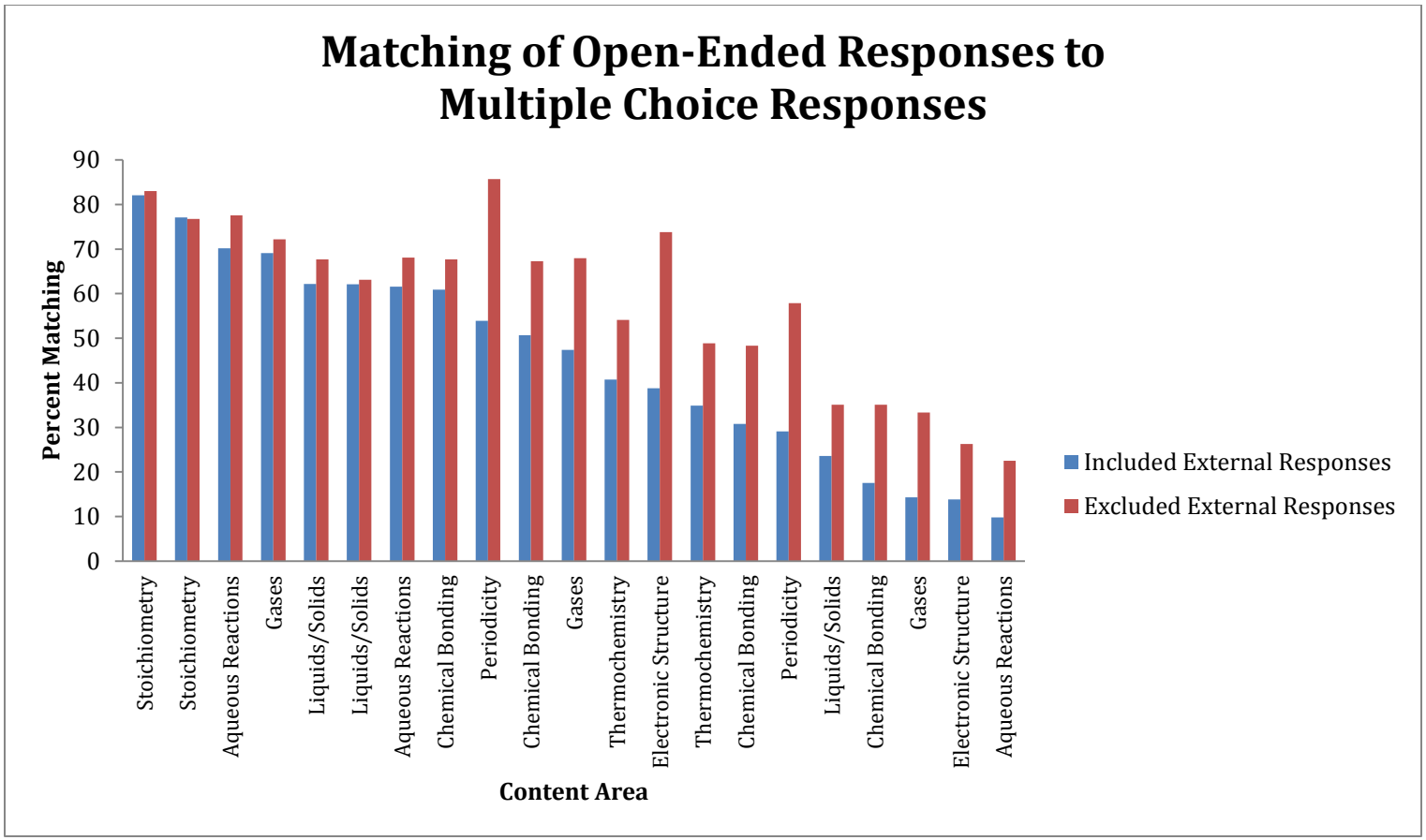



**Figure 21. Comparison of open-ended responses to multiple choice responses across task areas. Areas that are listed multiple times represent clones of the original task or another task presented in the category.**

**Table5. Percent Matching for Open-Response to Multiple Choice Response**

| Task Number | Content Area | Included External Responses | Excluded External Responses |
|---|---|---|---|
| 35 | Stoichiometry | 82.1 | 83.0 |
| 33 | Stoichiometry | 77.1 | 76.8 |
| 42 | Aqueous Reactions | 70.2 | 77.6 |
| 47 | Gases | 69.1 | 72.2 |
| 18 | Liquids/Solids | 62.2 | 67.7 |
| 19 | Liquids/Solids | 62.1 | 63.1 |
| 43 | Aqueous Reactions | 61.6 | 68.1 |
| 16a | Chemical Bonding | 60.9 | 67.7 |
| 13 | Periodicity | 53.9 | 85.7 |
| 17 | Chemical Bonding | 50.7 | 67.3 |
| 48 | Gases | 47.4 | 68.0 |
| 50 | Thermochemistry | 40.7 | 54.1 |
| 11 | Electronic Structure | 38.8 | 73.8 |
| 49 | Thermochemistry | 34.9 | 48.9 |
| 16b | Chemical Bonding | 30.8 | 48.3 |
| 12 | Periodicity | 29.1 | 57.9 |
| 21 | Liquids/Solids | 23.6 | 35.1 |
| 15 | Chemical Bonding | 17.5 | 48.3 |
| 44 | Gases | 14.3 | 33.3 |
| 8 | Electronic Structure | 13.8 | 26.3 |
| 40 | Aqueous Reactions | 9.8 | 22.5 |

Comparison of percent matching for open-response to multiple choice options when external responses are included or excluded. External responses are those responses that did not have a multiple choice option. In most cases these responses were incorrect paths to solving the task, or were more basic than the most basic step provided based on formation of steps from other semesters

With a list of compiled first steps completed (Appendix B) and reliability of first steps confirmed for those specific tasks, it needed to be determined if the same steps would be reliable if the task was cloned. Clones of tasks are tasks that should test exactly the same as the original task. However, while performance would be expected to remain constant assuming no change in schema, the problem-solving process may change. For this reason, the first step list must also be tested for clones. Because slight wording and content changes should not affect the outcome of the task there is no difference expected in reliability between clones and original tasks in terms of reliability for electronic first steps. The clones tested for open-ended response showed that the electronic first steps held reliable for the cloned tasks (Figure 22). What we see in figure 6 is the number of students that OR related to one of the multiple-choice options listed for the original task. The cloned task given in figure 22 yielded a 43% match between OR and MC with external responses removed.

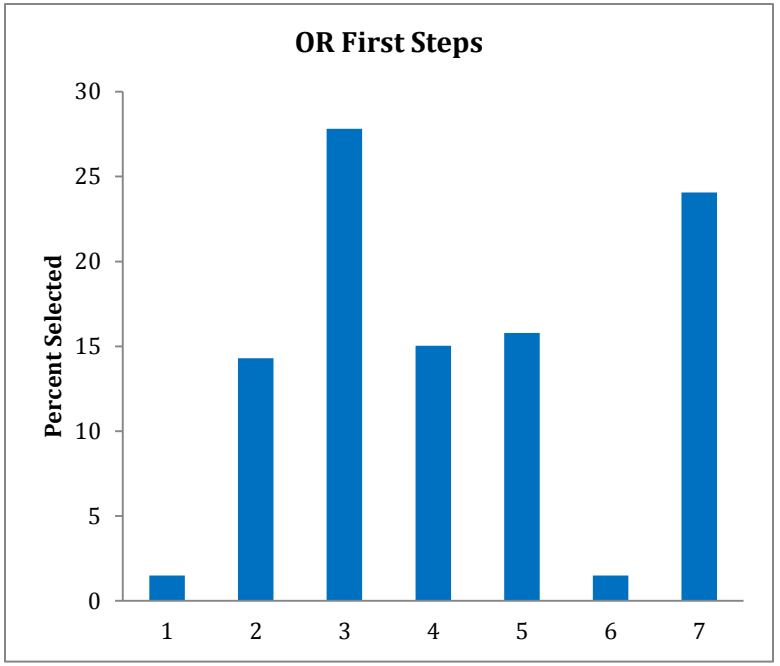
**Original item**

How many atoms of hydrogen are in the empirical formula for a hydrocarbon containing 83.6% carbon by mass?

**Cloned item**

How many atoms of hydrogen are in the empirical formula for a hydrocarbon containing 93.70% carbon by mass?

1. 83.6% carbon = 83.6g carbon
2. Calculating the number of moles of carbon
3. Calculating the number of moles of hydrogen
4. 100g total – 83.6g carbon = 16.4g hydrogen
5. 100% total – 83.6% carbon = 16.4% hydrogen
6. Knowing that hydrocarbons contain only carbon and hydrogen
7. Reading the exercise, however I am not sure how to start the exercise




The large number of students in the "other" category reflects both complete problem solving and steps not selected for electronic responses (highly inefficient).

**Figure 22. Comparison of original to cloned item with open-ended responses compared to electronic first steps provided. Numbers on the x-axis represent the electronic first step given to the left of the graph.**

The training on “first step in your problem-solving process” to attempt to address the problem of disagreement between MC and OR, discussed earlier, was determined though implementation of “Think-a-loud” interviews. The original design of the interviews was meant to collect further information from students on their thoughts during problem solving. During these interviews little data was collected on first steps, as students did not provide a common set of methods on how their first steps were determined. Instead, students provided their first steps as how they viewed the organization of the material to be used. This approach to the entire task solution, versus one step in the task solution, even when prompted for the first step in the problem solving process, led to the decision to implement the instruction in class on what is meant by the “first step” in problem solving.

### 4.1.2 Generation of Efficiency Measurement

With the first-step-generated response list proving to be reliable from the student generated steps, it needed to be determined if the list of steps provided could be ranked for efficiency as discussed in the methodology, and if the number of appropriate steps could be agreed upon for inclusion in the instrument. Expert responses to the first steps provided electronically were assessed through a reliability analysis using Cronbach’s alpha. The steps were ranked with 1 being the lowest efficiency (“Reading the exercise, however I am not sure how to start the exercise”), and the highest efficiency being determined by the maximum number of steps available. A threshold of 0.8 was set to ensure significant agreement on efficiency across all tasks and content areas. Table 6 shows that all α-cronbach measurements surpassed the determined threshold, thereby suggesting that the the experts can agree upon what step is more or less efficient than another. The reliability of efficiency ratings was determined to be 0.9303 using the α-cronbach measurement.

When determining the reliability by content area, the method used (Cronbach's-alpha) makes the assumption that the data is set in intervals. The ratings given by the raters have no definite interval, but are subjective. Therefore, while there is high agreement amongst the raters, to confirm that the reliability is being accurately measured it is important to look at the efficiency ratings amongst the experts in terms of agreement by percentage (Table 7). Here the percent agreement amongst the raters surpasses 50%, for the highest number of steps available, demonstrating that efficiency of problem solving is measurable by a developed rating scale. While the agreement may not appear high, it does demonstrate that the reliability values calculated for the agreement on efficiency for the content areas is valid. Two options were available for determining the efficiency ratings for steps based on these findings. These options were to meet with all of the raters and discuss the ratings as a group until a unanimous agreement could be reached amongst the group, or to use the mode of the raters by developing a logical progression of the provided ratings. Example A below shows the submitted efficiency ratings for each of the four experts for a single task, and the final ratings based on the logical progression determined by the mode of the supplied ratings.

| First Step Response: | A | B | C | D | E |
|---|---|---|---|---|---|
| Rater 1 | 4 | 2 | 5 | 3 | 1 |
| Rater 2 | 4 | 2 | 5 | 3 | 1 |
| Rater 3 | 3 | 2 | 5 | 4 | 1 |
| Rater 4 | 5 | 2 | 4 | 3 | 1 |
| Average Rating | 4 | 2 | 4.75 | 3.25 | 1 |
| Final Rating | 4 | 2 | 5 | 3 | 1 |

**Figure 22b. Rater efficiency responses for task 23, and their development into a final efficiency rating.**

**Table 6. Reliability analysis of steps**

| Cronbach's alpha | Content Area |
|---|---|
| **0.950** | General Chemistry |
| **0.921** | Atomic Structure |
| **0.885** | Electronic Structure |
| **0.986** | Periodicity |
| **0.960** | Chemical Bonding |
| **0.924** | Liquids and Solids |
| **0.924** | Formula Calculations |
| **0.943** | Reactions |
| **0.932** | Stoichiometry |
| **0.866** | Aqueous Reactions |
| **0.840** | Gases |
| **0.878** | Thermochemistry |

**Table 7. Percentage agreement by number of steps**

| Number of Steps | Percent Agreement (including out response) | Percent Agreement (excluding out response) |
|---|---|---|
| Five | 75 | 68.75 |
| Six | 74 | 68 |
| Seven | 64 | 58 |
| Eight | 61 | 56 |

With the reliability established for the tasks and steps it was also important to consider the complexity of the tasks provided. Complexity ratings were obtained via the previously discussed method. The overall complexity, difficulty, of the different content areas (Table 8) when examined using alpha-Cronbach's found a reliability of 0.8694. This too surpasses the threshold for supported reliability, and therefore the determined complexities are accurately applied to the content areas within the instrument. This shows that when assigning complexity the experts had high agreement on the overall difficulty of the tasks. The complexity ratings therefore allow for proper organization of the task within the content areas. By starting with the lowest complexity within each content area, when utilizing the instrument, the goal is to decrease the load on the working memory to allow for better integration of information into the students schema. If the information is integrated properly, then as the tasks build

in complexity, the chance to overload the working memory decreases. The goal of this is to help make sure student efficiency is being properly assessed when being measured by the instrument.

**Table 8. Expert complexity ratings by content area**

| Content Area | Complexity |
|---|---|
| **General** | 4.3 |
| **Atomic Structure** | 4.8 |
| **Electronic Structure** | 5.7 |
| **Periodicity** | 5.3 |
| **Chemical Bonding** | 6.0 |
| **Liquids and Solids** | 6.6 |
| **Formula Calculations** | 5.7 |
| **Reactions** | 6.3 |
| **Stoichiometry** | 7.4 |
| **Aqueous Reactions** | 6.9 |
| **Gases** | 6.3 |
| **Thermochemistry** | 8.6 |

## 4.2 Discussion

Through examining student open-ended responses it was determined that students were able to use the instrument in an open-ended form, and that there was agreement amongst students on a general set of "first steps" when it came to task solutions. These generated first steps from open-responses were successfully coded and grouped into electronic options for open-ended responses and implemented in the second phase of the instrument. The electronic options given in multiple choice format were found to match student open-responses and therefore were reliable, but students did not always pick the matching multiple choice option for their given open-response. Training on instrument use increased this agreement between open-response and multiple-choice, thereby increasing the reliability of the electronic options. Finally, analysis using Cronbach's alpha to determine reliability between experts was performed to determine complexity of the tasks, and to determine efficiency of electronic first steps for

the tasks, for the different content areas within the instrument. Experts were able to agree on the complexity of the tasks and a general efficiency for each step within the tasks. This agreement allows for implementing the instrument in the most effective manner by providing an efficiency rating scale in which the majority of students fit, and by confirming organization of the items within the instrument to most accurately measure the load on working memory. These ratings also provided the ranking of the first steps by efficiency, allowing the comparison of the problem-solving efficiency by first step, task performance, mental effort and the combination of these measures into an overall assessment of the student's efficiency in problem solving as a measure of their schema development.

# Chapter 5: Confirmation of successful instrument development

In the previous chapter the concept of creating a usable list of electronic multiple-choice first steps was established through analyzing the matching and non-matching responses of students' open-ended responses to their electronic first step selection. The reliability of efficiency ratings for the same electronic first steps was established using experts. This information is the support for the data found here. This chapter will examine the use of efficiency of an individual's first step in conjunction with his or her task performance and mental effort for the generation of an overall assessment score using Pearson and Spearman correlations. T-test data examining the differences between preparatory chemistry and general chemistry will also be shown to give support as to why the two courses data must be analyzed separately.

## 5.1 Determination of Data Evaluation

The method of data collection for phase 2 of the instrument was applied in the spring semester of 2009 for both preparatory and general chemistry courses involved in the study. This was the first semester in which both course levels were provided tasks with electronic multiple-choice first steps. As discussed in the methods section of this work, efficiency ratings for an individual's first step were converted into percent efficiency of the step. It was important to determine if the instrument would accurately measure information for multiple course levels, and to gain an understanding in the differences between the knowledge levels of the learners. To accomplish this only tasks that were asked in both courses were evaluated. An aggregate score for each of the constructs was generated for each participant, and descriptive statistics (Table 9; see Appendix E for graphs) and independent t-tests (Table 10) were performed for both groups on the four major measurements of the instrument (task performance, mental effort, percent efficiency of the first step, and generated assessment scores). The two types of assessment scores listed show the order of importance in which the three areas of data

collection were analyzed, where the first area is given the highest priority (see Appendix B for list of tasks included). It was determined from comparison of the t-value, -8.498 (p=0.000), that there was a significant difference between the two groups for performance on the tasks. The t-values for percent efficiency on selection of first step and the generation of assessment scores showed similar results. However, no significant difference was found between the two groups on self-reporting of mental effort, t=-1.075(p=0.284).

**Table 9. T-test Descriptive Statistics for Preparatory Chemistry and General Chemistry**

| | Chem 100 (n=101) | Chem 102 (n=80) |
|---|---|---|
| | Mean (SD) | Mean (SD) |
| Task Performance | 0.290317 (0.1581) | 0.539049 (0.2345) |
| Mental Effort | 0.529096 (0.1299) | 0.550527 (0.1372) |
| Percent Efficiency of First Step | 0.569062 (0.0948) | 0.53748 (0.0931) |
| Assessment Score 1 (Task Performance: Mental Effort: Percent Efficiency of First Step) | 1.708725 (0.4909) | 2.249869 (0.7597) |
| Assessment Score 2 (Task Performance: Percent Efficiency of First Step: Mental Effort) | 1.69565 (0.4683) | 2.214881 (0.7389) |

**Table 10. T-test Table for Preparatory Chemistry and General Chemistry**

| **Construct** | **t** | **df** | **Sig. (2-tailed)** |
|---|---|---|---|
| TP | -8.498 | 179 | 0.000 |
| ME | -1.075 | 179 | 0.284 |
| Percent Efficiency of First Step | 2.243 | 179 | 0.026 |
| Assessment Score 1 (Task Performance: Mental Effort: Percent Efficiency of First Step) | -5.794 | 179 | 0.000 |
| Assessment Score 2 (Task Performance: Percent Efficiency of First Step: Mental Effort) | -5.754 | 179 | 0.000 |

The use of the t-test allowed for the direct comparison of the two courses on each of the analyzed constructs to determine significance. It was expected that the general chemistry I students would outperform the students in preparatory chemistry. Interestingly, the lack of a significant difference in the means of the means of the reported mental effort suggests that although there is increased performance in general chemistry I over preparatory chemistry students on these items, there is not a significant increase in mental efficiency. This is further corroborated in the efficiency measure where the difference in the means of these values is significantly different with the preparatory chemistry students solving these problems more efficiently. A possible explanation of this unexpected result could be that the instruction of the preparatory chemistry students includes much discussion on problem solving of these basic problems. In general chemistry I, less attention is given to the problem solving of these basic tasks and more time is focused on the understanding and problem solving of more complex concepts. It is also possible that this reflects the many types of backgrounds of students in general chemistry I, including high school preparation that may give a specific process for solving problems without further elaboration or generalization of the process. To further analyze the two courses it was necessary to run correlations to determine any specific differences that occurred.

When comparing the preparatory and general chemistry data both Pearson and Spearman correlations were performed. Pearson correlations were used for data that was nominal or scale (task performance and mental effort), while Spearman correlations were used for data that was ordinal (percent efficiency of the first step and assessment scores 1 and 2). The task performance, mental effort, and percent efficiency of the first step were used to generate the two assessment values analyzed (Appendix I). It was expected that task performance would be negatively correlated to mental effort, as performance should increase as load on working memory decreases due to an improved schema. The correlation of mental effort to efficiency should be negative as well. If the first step is becoming more efficient, then schema should have improved causing load on working memory to decrease. What was found is that

these trends were supported by the collected data. Table 11 (see Appendix F for graph) shows the correlations between the two courses for each of the constructs to one another and to external measures (the fourth class exam (cumulative), the two standardized final exam scores, and the percent score earned in the class) where "*" (p=0.01) and "**" (p=0.05) indicate the threshold for the level of significance. Pearson correlations showed significance for the two courses in a majority of the constructs. For the general chemistry sample and the preparatory chemistry sample, both of the Assessment Score variables were moderately to strongly correlated with the final exam scores and the students' percent grade in class scores. Correlations ranged from 0.328 to 0.573, indicating that as Assessment Score1 and 2 increased, the scores on Final 1, Final 2, and the percent grade in class also increased ($p<0.001$). There was also a significant correlation in both samples between the students' task performance scores and the percent efficiency of their first step. While the correlation for the general chemistry sample ($r=0.245$, $p<0.05$) would be considered a weak correlation, it is still significant and in the direction expected. The correlation coefficient for the preparatory chemistry sample was somewhat higher ($r=0.337$, $p<0.05$) and would be considered a moderate correlation. Both correlation coefficients suggest that as task performance scores are higher, the percent efficiency of the first step is also higher, or more efficient. This finding is also in the expected direction.

There were a few variables that were not significantly correlated in either sample. Task performance was not significantly correlated with mental effort, and the correlation coefficient for both samples was negative. In addition, the mental effort score was not significantly correlated with the percent efficiency of the first step. In this case, the correlation coefficient for both samples was also negative. Although the correlation was not significant in either case, it was in the expected direction for both. It was expected that higher task performance scores would be associated with lower mental effort, and that higher efficiency of the first step would also be associated with lower mental effort. Figure 23 illustrates these trends.

Despite the lack of significant correlation between mental effort with task performance scores and percent efficiency of first step scores, it is apparent that when examining the combination of mental effort, performance, and efficiency to create an assessment score that there is a significant correlation to all of the external measures (p=0.000). This supports the use of the instrument in measuring the efficiency as a gauge for complementary classroom instruction, as well as for the use of first steps to predict student performance.

**Table 11. Pearson and Spearman Correlations for Preparatory and General Chemistry**

| Correlation Area | Preparatory Chemistry (n=100) | General Chemistry (n=79) |
|---|---|---|
| | Correlation Value (sig.) | Correlation Value (sig.) |
| Task Performance to: | | |
| Mental Effort | -0.170 (0.091) | -0.145 (0.202) |
| Percent Efficiency First Step | .337** (0.001) | .245* (0.030) |
| Mental Effort to: | | |
| Percent Efficiency First Step | -0.148 (0.142) | -0.117 (0.306) |
| Assessment Score 1 to: | | |
| Final 1 | .545** (0.000) | .392** (0.000) |
| Final 2 | .492** (0.000) | .328** (0.003) |
| Percent Grade in Class | .513** (0.000) | .551** (0.000) |
| Assessment Score 2 to: | | |
| Final 1 | .553** (0.000) | .392** (0.000) |
| Final 2 | .518** (0.000) | .348** (0.002) |
| Percent Grade in Class | .573** (0.000) | .555** (0.000) |

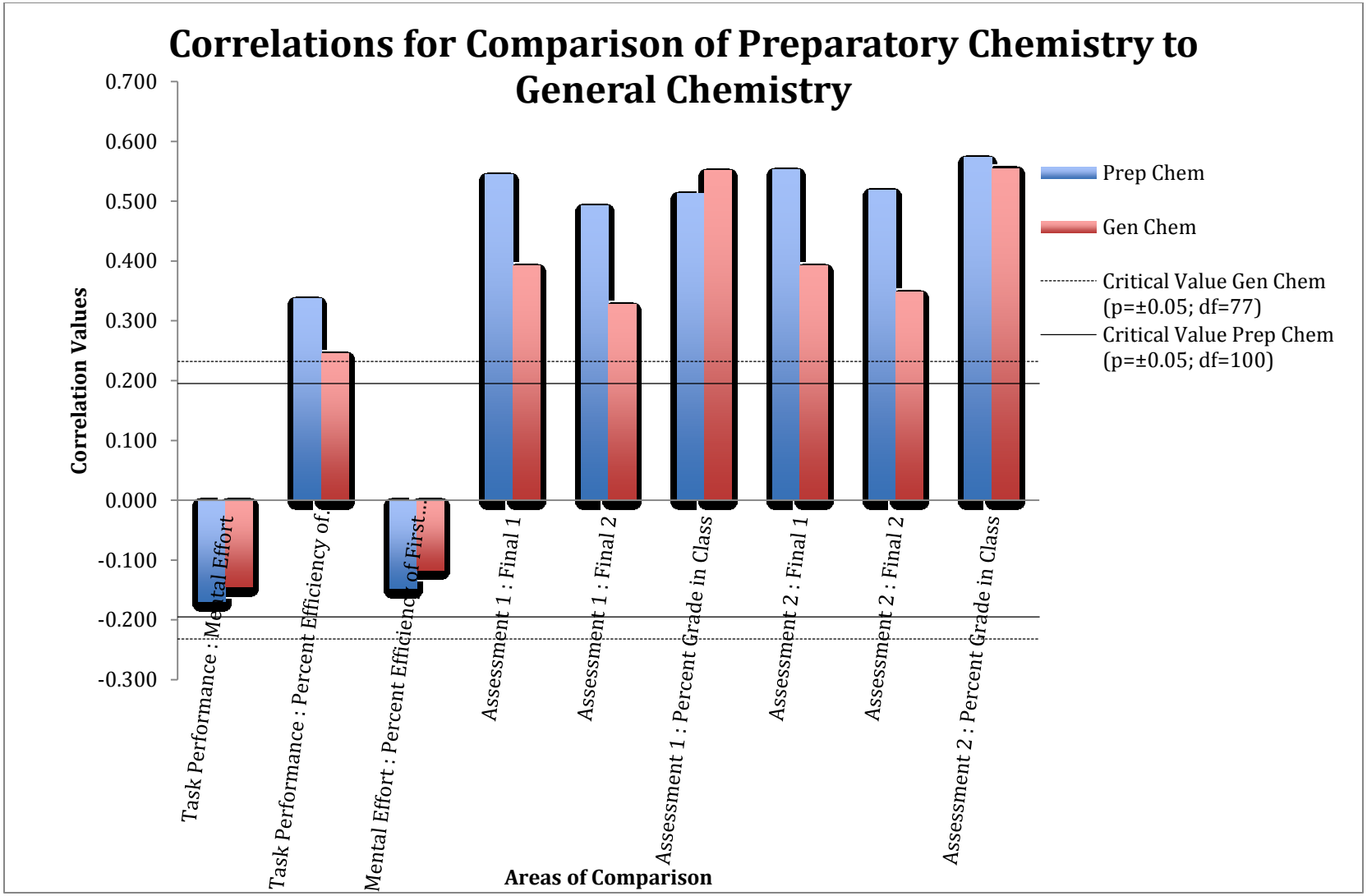



**Figure 23. Correlation graph of preparatory and general chemistry constructs.**

### 5.1.1 Discussion

The t-test scores for comparison of the preparatory and general chemistry data indicate that there is a significant difference between the two courses for task performance (p=0.000), percent efficiency of the first step (p=0.026), and assessments 1 and 2 (p=0.000), but not for mental effort (p=0.284). The t-scores for each of these areas indicates that the difference appears as expected showing greater expertise for students taking general chemistry, except for in percent efficiency of the first step t=2.243 (p=0.026) where preparatory chemistry shows to be more efficient in the selection of their first steps. This difference between the two groups in terms of first step efficiency may be the result of differences in sample numbers, or have to do with the amount of extra information general chemistry I students are being taught causing extra load on working memory. This shows that the two groups of students cannot be directly compared due to the differences in their levels of understanding. The correlation data supports the use of the instrument in both preparatory and general chemistry courses. Both courses showed correlations for areas in the direction expected, and showed significant at the 0.01 and 0.05 level.

## 5.2 Semester Comparison Using Correlations

Phase two of the study was carried out through the fall of 2009. In the spring of 2010 phase three was implemented, and data began to be collected solely electronically. However, the tasks are still comparable across semesters from the spring of 2009 through the fall of 2010, as the task list remained consistent with only addition of task clones. The data collected in these semesters consisted of task performance, mental effort, and first step efficiency. The first step selected was coded into a percent efficiency rating, and then compared to the task performance and mental effort to generate an assessment score as described in the methods. The aggregate scores of task performance, mental effort, percent efficiency of the first step, and the assessment values (calculations described in methods) were

correlated to one another using Pearson and Spearman correlations as described above in the spring 2009 analysis. Running the correlations in the same manner allowed for an overall comparative analysis of each semester. Table 12 shows the correlation data for all of the tasks collected each semester, including the clones, for each internal and external construct. Correlation values marked with asterisks indicate the threshold for the level of significance met for the value ["*" (p=0.01); "**" (p=0.05)]. Table 13 shows the correlations for each semester with the clones excluded from the analysis. By analyzing the data with and without the correlations it is possible to determine if tasks generate a significant result due to the nature of the task, or based on how the task was presented. Of interest to note, is that in comparison to the spring 2009 semester data, the fall of 2009 through the fall of 2010 showed significant correlations between task performance and mental effort (p=0.000 to p=0.002) with the clones included or excluded. This data again trended as expected (see Appendix G for figures).

**Table 12. Pearson and Spearman Correlations by Semester: Including Clones of Tasks**

| | Preparatory Chemistry | General Chemistry | | | |
|---|---|---|---|---|---|
| | S09 | S09 | F09 | S10 | F10 |
| | (n=100) | (n=79) | (n=195) | (n=184) | (n=114) |
| | Correlation Value (Sig.) | Correlation Value (Sig.) | Correlation Value (Sig.) | Correlation Value (Sig.) | Correlation Value (Sig.) |
| Task Performance to: | | | | | |
| Mental Effort | -0.170 (0.091) | -0.145 (0.202) | -.372** (0.000) | -.305** (0.000) | -.290** (0.002) |
| Percent Efficiency of First Step | .337** (0.001) | .245* (0.030) | .270** (0.000) | .163* (0.027) | 0.148 (0.117) |
| Mental Effort to: | | | | | |
| Percent Efficiency of First Step | -0.148 (0.142) | -0.117 (0.306) | -.331** (0.000) | -.232** (0.002) | -.250** (0.007) |
| Assessment 1 to: | | | | | |
| Final 1 | .545** (0.000) | .392** (0.000) | .533** (0.000) | .387** (0.000) | .391** (0.000) |
| Final 2 | .492** (0.000) | .328** (0.003) | .527** (0.000) | .365** (0.000) | .376** (0.000) |
| Percent Grade in Class | .513** (0.000) | .551** (0.000) | .608** (0.000) | .470** (0.000) | .480** (0.000) |
| Assessment 2 to: | | | | | |
| Final 1 | .553** (0.000) | .392** (0.000) | .535** (0.000) | .394** (0.000) | .435** (0.000) |
| Final 2 | .518** (0.000) | .348** (0.002) | .525** (0.000) | .367** (0.000) | .395** (0.000) |
| Percent Grade in Class | .573** (0.000) | .555** (0.000) | .604** (0.000) | .479** (0.000) | .507** (0.000) |

**Table 13. Pearson and Spearman Correlations by Semester: Excluding Clones of Tasks**

| | Preparatory Chemistry | General Chemistry | | | |
|---|---|---|---|---|---|
| | S09 | S09 | F09 | S10 | F10 |
| | (n=100) | (n=79) | (n=195) | (n=184) | (n=114) |
| | Correlation Value (Sig.) | Correlation Value (Sig.) | Correlation Value (Sig.) | Correlation Value (Sig.) | Correlation Value (Sig.) |
| Task Performance to: | | | | | |
| Mental Effort | -0.139 (0.168) | -0.145 (0.202) | -.373** (0.000) | -.305** (0.000) | -.286** (0.002) |
| Percent Efficiency of First Step | .252* (0.011) | .245* (0.030) | .256** (0.000) | .163* (0.027) | 0.137 (0.147) |
| Mental Effort to: | | | | | |
| Percent Efficiency of First Step | -0.145 (0.150) | -0.117 (0.306) | -.334** (0.000) | -.232** (0.002) | -.248** (0.008) |
| Assessment 1 to: | | | | | |
| Final 1 | .542** (0.000) | .392** (0.000) | .530** (0.000) | .387** (0.000) | .376** (0.000) |
| Final 2 | .462** (0.000) | .328** (0.003) | .529** (0.000) | .365** (0.000) | .362** (0.000) |
| Percent Grade in Class | .494** (0.000) | .551** (0.000) | .603** (0.000) | .470** (0.000) | .466** (0.000) |
| Assessment 2 to: | | | | | |
| Final 1 | .560** (0.000) | .392** (0.000) | .532** (0.000) | .394** (0.000) | .412** (0.000) |
| Final 2 | .487** (0.000) | .348** (0.002) | .520** (0.000) | .367** (0.000) | .370** (0.000) |
| Percent Grade in Class | .556** (0.000) | .555** (0.000) | .598** (0.000) | .479** (0.000) | .490** (0.000) |

Significant correlation values at the 0.01 and 0.05 level for both internal and external constructs across the semesters supports the use for analyzing the data as a whole. Because it was previously determined that preparatory chemistry and general chemistry show a significant difference in task performance and assessment scores in the independent t-test (p=0.000), only data from general chemistry courses may be successfully combined for correlations. As in the previous correlations in this section aggregate numbers were calculated for task performance, mental effort, percent efficiency of first step, and the assessment scores. The data was generated for the combined general chemistry sections, and then

correlated using Pearson correlation coefficients and Spearman rho correlation coefficients including and excluding clones (Table 14). All correlations including and excluding clones proved to be significant ($p=0.000$, $n=572$) (exception: excluded clones for percent efficiency of first step to final 2 $p=0.013$) for both internal and external constructs. Figure 24 shows a comparison of the correlations for included and excluded clones of tasks.

**Table 14. Pearson and Spearman Correlations: All General Chemistry Semesters Combined**

| | Including Clones (n=572) | Excluding Clones (n=572) |
|---|---|---|
| | Correlation Value (Sig.) | Correlation Value (Sig.) |
| Task Performance to: | | |
| Mental Effort | -.250** (0.000) | -.247** (0.000) |
| Percent Efficiency of First Step | .178** (0.000) | .171** (0.000) |
| Mental Effort to: | | |
| Percent Efficiency of First Step | -.227** (0.000) | -.235** (0.000) |
| Assessment 1 to: | | |
| Final 1 | .444** (0.000) | .440** (0.000) |
| Final 2 | .411** (0.000) | .408** (0.000) |
| Percent Grade in Class | .513** (0.000) | .511** (0.000) |
| Assessment 2 to: | | |
| Final 1 | .459** (0.000) | .456** (0.000) |
| Final 2 | .406** (0.000) | .402** (0.000) |
| Percent Grade in Class | .520** (0.000) | .518** (0.000) |

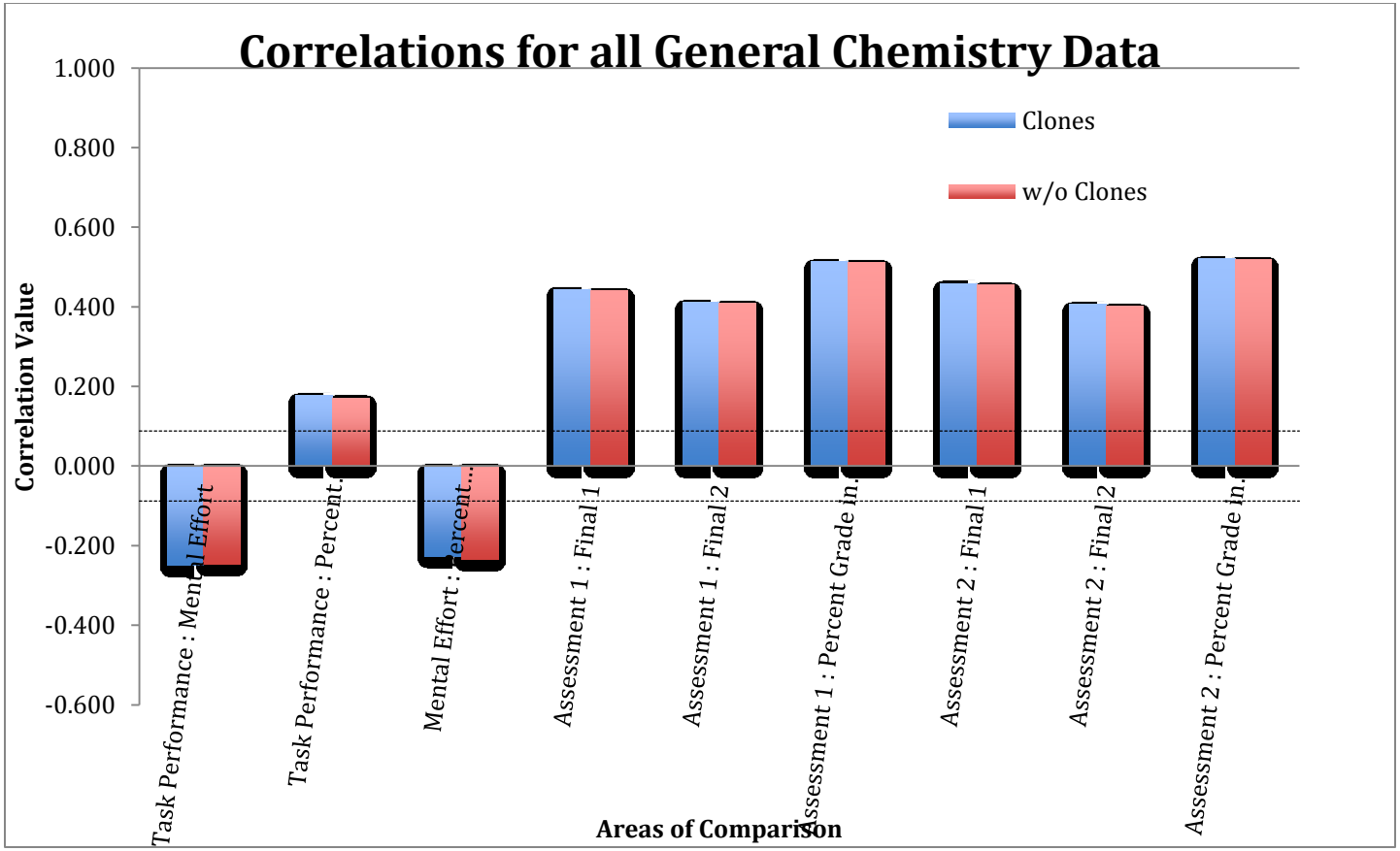



**Figure 24. Correlations for all combined general chemistry courses.**

### 5.2.1 Discussion

Examination of independent t-test descriptive statistics and t-scores showed preparatory chemistry and general chemistry students cannot be combined when analyzing data in regards to student efficiency. This was found in the significant difference between performance and assessment scores of the two groups. The results found for mental effort in this test showed that students' ability to report load on working memory does not vary by course, and therefore is still accurately utilized when applied to preparatory chemistry. Results from the Pearson and Spearman correlation coefficients suggest that the instrument accurately measures task performance, mental effort, and efficiency of first steps for both general and preparatory chemistry. This is seen the internal validity of the trends between the constructs. External validity is also seen in the significant correlations between the internal and external measures. High significance and accurate trends of the internal measures to external measures of final exams along with an internal measure to the fourth class exam indicates that the values for the instrument measurements are consistent. Finally the combination of the general chemistry data for an overall correlation analysis shows that correlations between internal and external measures are highly significant for the course as a whole, and that therefore the measures taken with the instrument are in fact valid.

# Chapter 6: Use of Eye-tracking to objectively evaluate student subjective reporting of mental effort, and confirm reliability of electronic first step options

## 6.1 Eye-tracking technology in Rapid-Knowledge Assessment

In order to establish continual reliability of the multiple choice steps provided to the students, an eye tracker was used to study the reliability of students' self-reported mental effort and the relationship between time spent on tasks, mental effort, performance, and efficiency of first steps. An eye-tracker was used specifically for its ability to measure pupil size and movement in response to a stimulus, known as task evoked pupillary response (TEPR). The use of pupil size allows for the measurement of mental effort (Beatty & Wagoner, 1978; Kahneman & Jacson, 1966; Stone et al., 2004), and together with the student's self- reported mental effort, validity of using student self-reported effort is confirmed. The use of the tracker also allows for further collection of data for checking the use of a multiple choice for first steps in the instrument. As in the classroom testing, open-ended responses to multiple choice responses were scored on a matching to non-matching basis and taken as a percentage for overall matching. The tracker also records the time on task, allowing for comparison of time spent on the task to pupil size from task evoked pupillary response (TEPR) and mental effort, using Pearson's correlation coefficient, *r*, in relation to the logic of numerical scale associated with these types of measures, while Spearman correlations for normally non-scaled data are applied when a scale is imposed on a piece of information to allow for comparative analysis. Because individual responses to multiple levels of measurement vary, a direct value cannot be reported for these correlations,

but will lend towards establishment of a pattern to lend for future use in the next phase of the project.

### 6.1.1 Reliability of Electronic First Steps

Just as in the classroom experiment, the eye-tracking experiment looked to confirm the reliability of the multiple choice options for first steps provided to the students by comparing to the open response of their first step. In addition, we investigated the pathways used by the students to select their first step. To accomplish this, tracking experiments incorporated an open-ended response section as described in the methodology. The data from this section was coded in the same fashion as for the lecture portion of the experiment, and scored using "1" for matching and "0" for non-matching responses between OR and MC. Reliability was determined as a matching percentage for each content area. Figure 25 shows the reliability analysis in comparison to the content areas tested. Reliable data is defined as any category that meets or exceeds an alpha value of 0.7 (this is considered an acceptable mark in social sciences). The higher the alpha value, the more reliable the construct is and the more internal consistency that exists between the open-ended response and the electronic multiple choice option chosen. When developing an instrument based on open-ended response coded into groupings a goal of 0.6 is reasonable, as the stringency of the coding often alters the actual reliability of the information to rule out any type of "guessing" as to what the person writing the response meant their remark. Of the seven content areas tested during the first part of the interviews, 5 areas showed an alpha score at or above 0.7. The atomic structure content area fell in the 0.6-0.7 range showing low reliability, but not complete dismissal of the electronic steps provided. The periodicity content showed reliability in the 0.5 -0.6 range, showing poorer reliability of provided electronic first steps. However, the number of participants in this part of the study (n=73) was smaller, and because not every participant was usable for each task in every content

area the actual number of participants varied for each area. Because the majority of the content areas surpass the 0.6 to 0.7 threshold for reliability even at such a small n value, the use of multiple choice first steps formed from open-ended responses is a reliable way to collect first steps in individual students' problem solving processes. Therefore, it can also be concluded that the use of multiple choice first steps to assign problem solving process efficiency is valid.

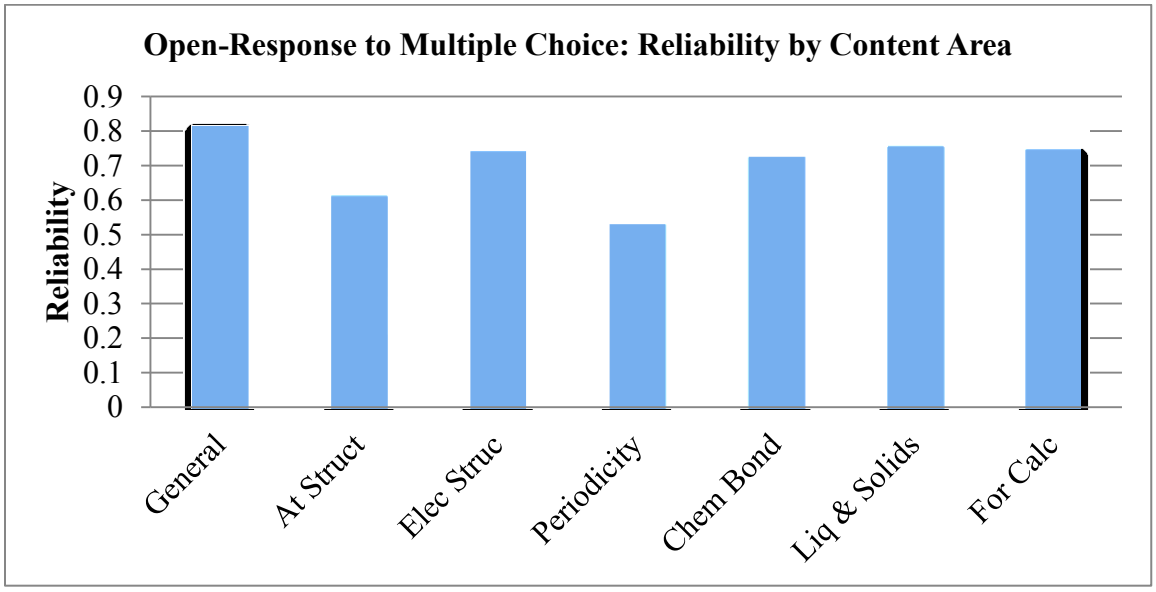



**Figure 25 Reliability of open-ended responses to electronic first steps during eye-tracking interviews.**

### 6.1.2 Use of Self-Reported Mental Effort

Another area of study in the eye-tracking interviews was used to confirm that students accurately reported mental effort during the classroom data collection. The methodology was applied as described in the methods section, and was based on previous research that demonstrates a relationship between mental effort and pupil diameter (Beatty & Wagoner, 1978; Kahneman & Jacson, 1966; Stone et al., 2004). While these studies looked at the relationship between mental effort and pupil diameter, none of the studies examined the

measure as a way to confirm self-reporting of mental effort as a measure of load on working memory and the subjective complexity of the material. The studies also did not examine the use of TEPR for measurement of maximum pupil dilation in conjunction with self-reported mental effort. If TEPR and self-reported mental effort are measuring the same load on working memory, then appropriate correlations between the two areas and objective measures (such as time on task, performance, and efficiency) should be found.

To test for accurate reporting of mental effort by students, time on task (measured in seconds) was compared to maximum pupil diameter (measured in millimeters) (Figure 26). The correlations suggest that there is a relationship between time on task and the maximum pupil diameter, but that there is not enough information to conclude that both measures are measure the same load on working memory. Time on task was also compared to students self-reported mental effort ratings (Likert scale rating of 1 through 5) (Figure 27). Results are reported as individual student examples for 13 participants across a set of 23 tasks. Here it can be seen that the two are positively correlated, suggesting that there is a relationship between the amount of time spent on a task and student reported mental effort. A direct comparison of subjective mental effort ratings by students was correlated with the average maximum pupil diameter (Figure 28). In all three cases Pearson correlation coefficients were generated. The values for participants are expected to be positively correlated for time on task (TOT) to maximum pupil diameter due to the relationship each has with cognitive load or subjective complexity. The results shown do not match the expected outcome, nor do they contradict the expected outcome. With the TOT to mental effort ratings a positive correlation is expected based on the cognitive load concepts discussed earlier. The expected trend is demonstrated for all 13 participants. For the comparison of mental effort to maximum pupil diameter the same results

are predicted as with the TOT to maximum pupil diameter based on the relationship both share with mental effort.

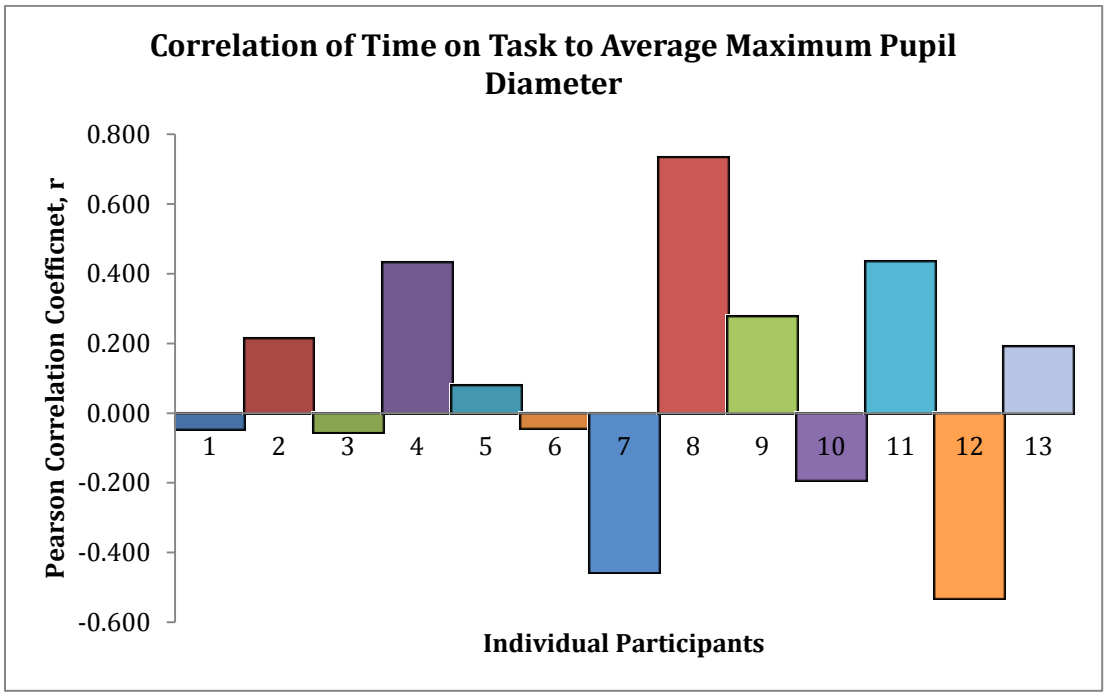



**Figure 26. Pearson correlation coefficients for 13 participants on 23 tasks.**

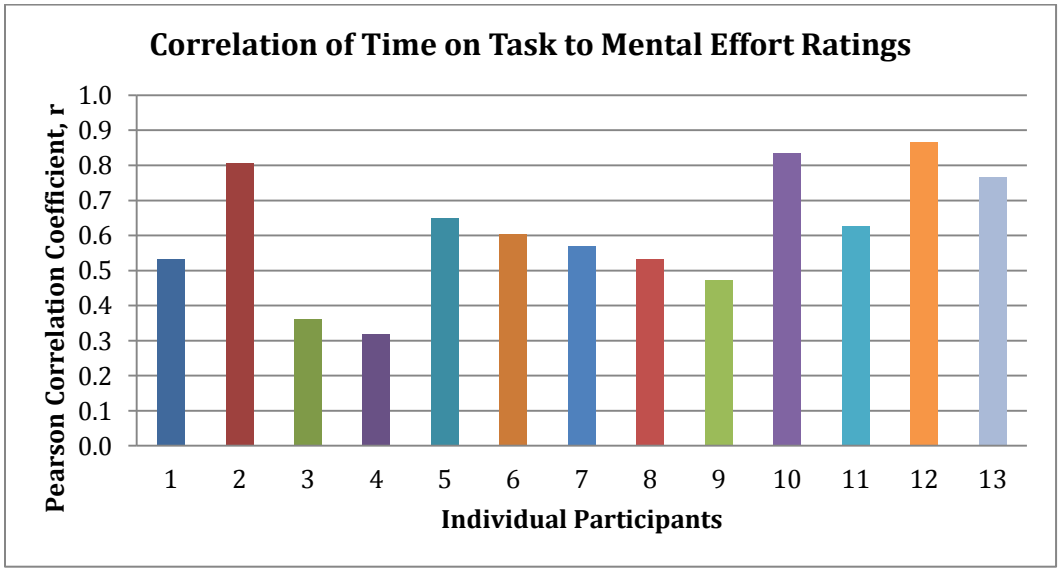



**Figure 27. Individual participant Pearson correlations for TOT and mental effort based on 23 eye-tracking tasks.**

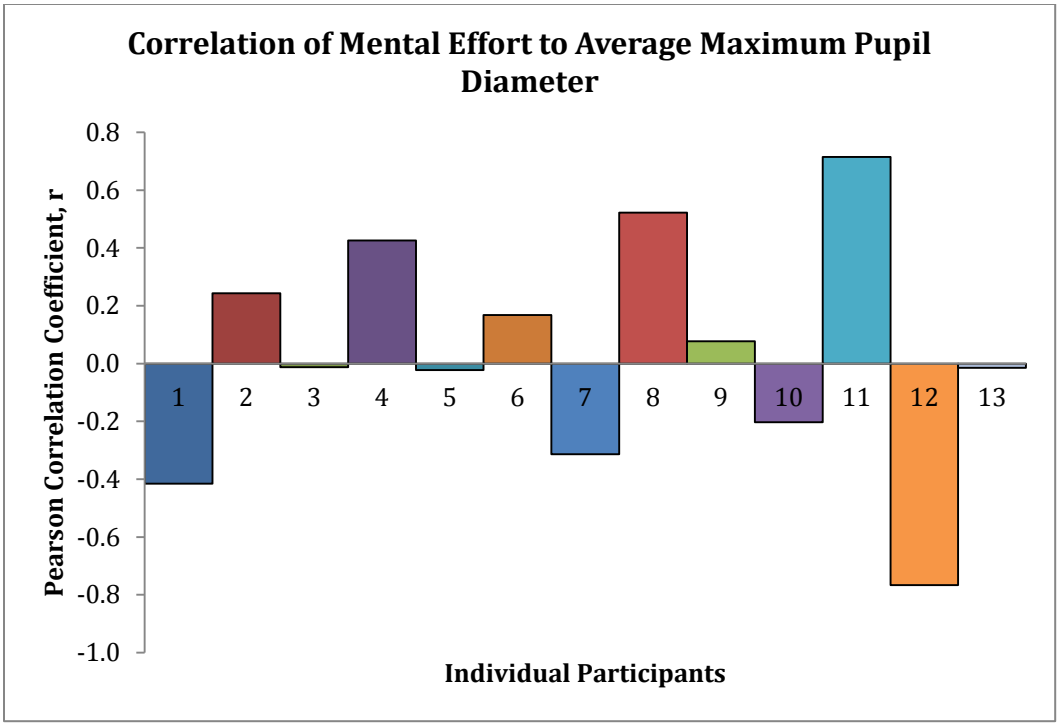



**Figure 28. Pearson correlations for mental effort and maximum pupil diameter for 23 eye-tracking tasks.**

### 6.1.3 Validity of performance measures

The data from eye-tracking is rich in many areas outside of understanding mental effort. Interviews with students allowed for review of validity of the classroom instrument via testing RKA tasks in a monitored setting, where all information was recorded. Through this method task performance was analyzed in the same method as the classroom data, with additional information on average maximum pupil size available for comparative analysis. Figure 29 shows the relationship of task performance to time spent on the task. Individual participants were compared for variation in time and pattern in time spent in task solution to outcome in performance. Measurements taken were those of Pearson correlation coefficients, r, and were found to be negatively correlated in respect to one another. This was predicted to be the case. Time on task has been a proven predictor of performance (Paas et al., 2003). With TOT

correlating as predicted, the next logical step was to examine the maximum pupil diameter in relation to performance. Pearson correlations were run to determine the relationship (Figure 30). Because data here was inconclusive, the performance was then compared to mental effort to determine if the two areas would change inversely as student efficiency increased lessening cognitive load (Figure 31). This final analysis involved running Spearman correlations of the efficiency measurements of the participants' first steps to the performance outcome (Figure 32). Here Spearman correlations were used in place of Pearson correlations accommodate for the shift in data type. While the steps for the tasks were ranked numerically for efficiency, the efficiency rating is an applied value and not an interval or scaled value. Therefore, the use of Spearman correlations acts as a more accurate predictor of the relationship of these two values.

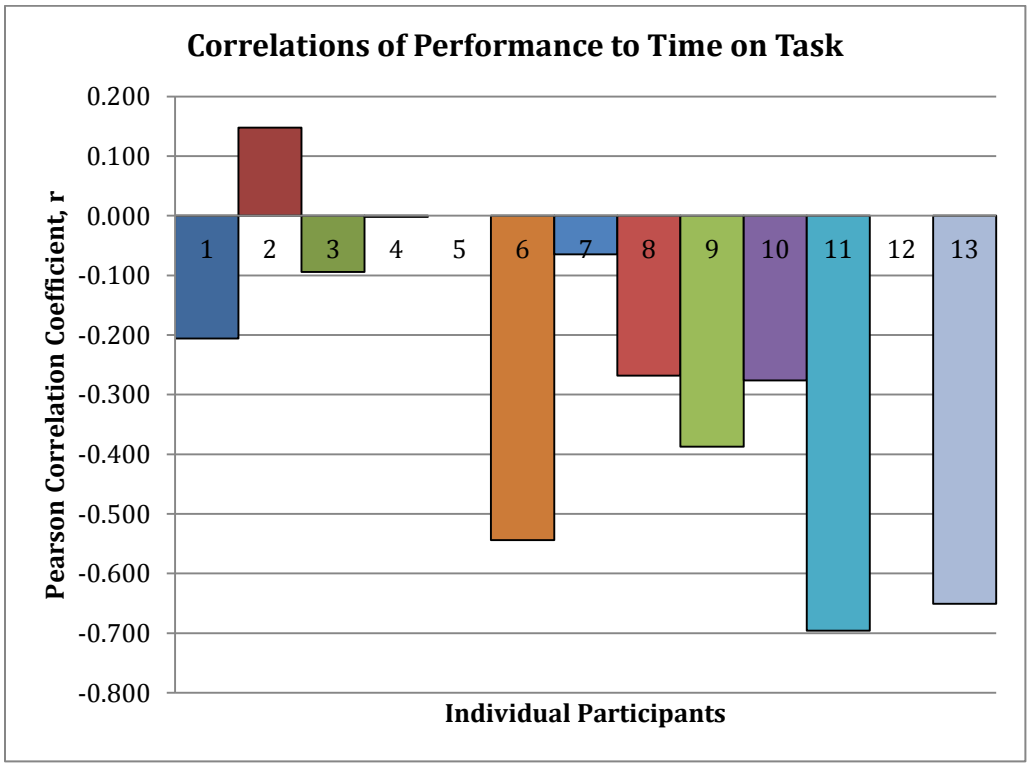



**Figure 29. Comparison of 13 participants performance to their time on task.**

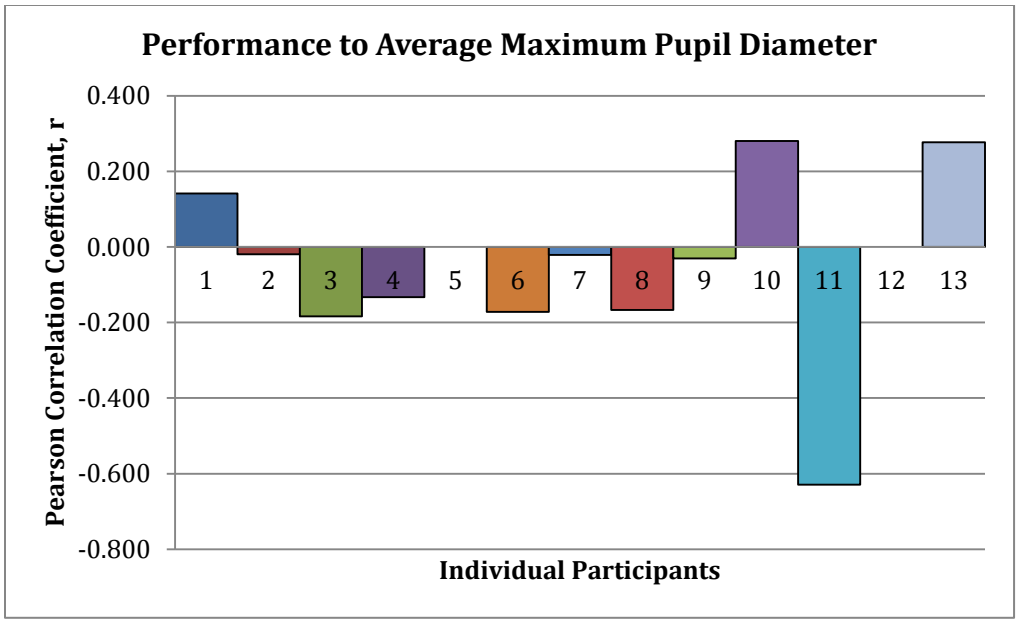



**Figure 30. Comparison of performance with task evoked pupillary response data.**

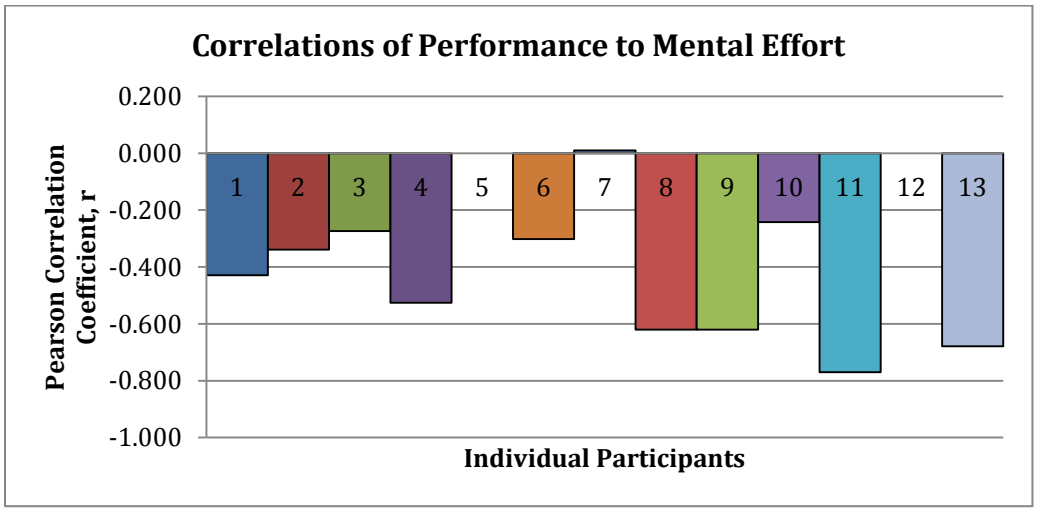



**Figure 31. Comparison of performance and mental effort for 13 participants.**

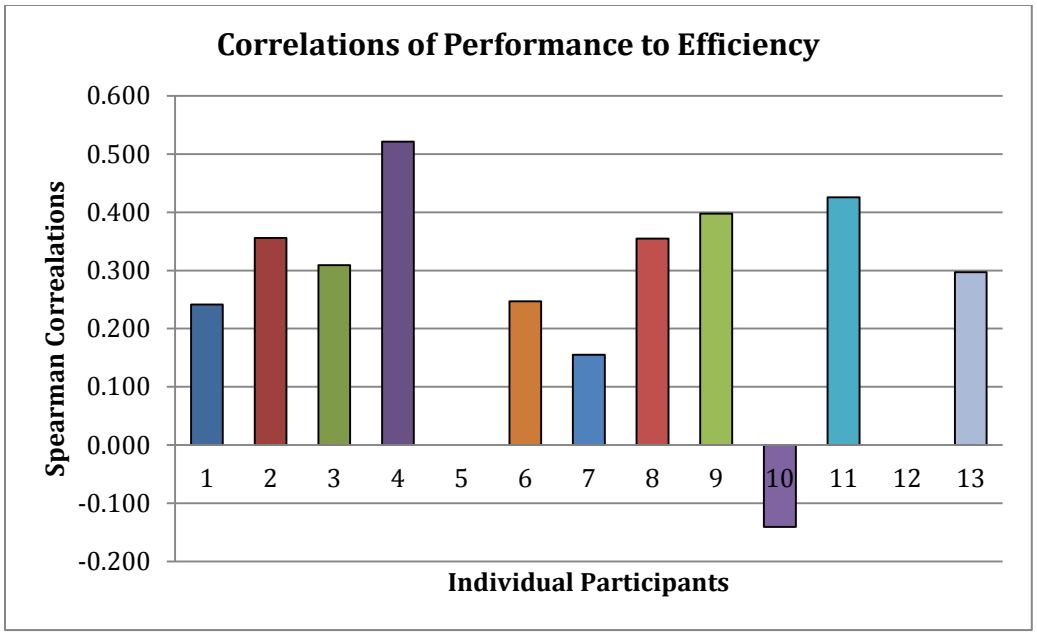



**Figure 32. Comparison of performance to efficiency for 13 participants, where two participants did not contain values that allowed for Pearson correlations to be performed.**

### 6.1.4 Reviewing efficiency of first steps in relation to alternate mental effort measurements

Part of establishing the use of the RKA instrument in the classroom was to determine the validity of the efficiency measurement. To establish validity of measurements taken for step efficiency, correlations were done with TOT and maximum pupil diameter. Correlations of TOT to efficiency showed a general trend for a decrease in time with an increase in efficiency (n=13). Spearman correlations are used for efficiency measurements here as in the previous section on performance. Efficiency measurements are scaled as described in the methods section on electronic first steps. With the number of steps varying by task, the efficiency scores used in the correlations are based on the percent efficiency out of the number of available steps (if the efficiency score for the step is 6 and there are 8 total options, the score is 6/8 or .75) just as in the classroom calculations of efficiency. The least efficient score always received a rating of 1,

and the most efficient score always received the highest number. Because individual efficiency and time on task varies by task and by person, the information does not yield a single correlation value, but rather gives an indication of correlation trend between the efficiency and the time on task. Rather than normalize all time on task values between the participants (necessitating more extensive question calibration with the participants), the analysis was only conducted across individual participants. Variation in time is expected to fluctuate with efficiency of the step, as with mental effort. The correlations for 13 individual participants show a trend in relationship between the two areas (Figure 33).

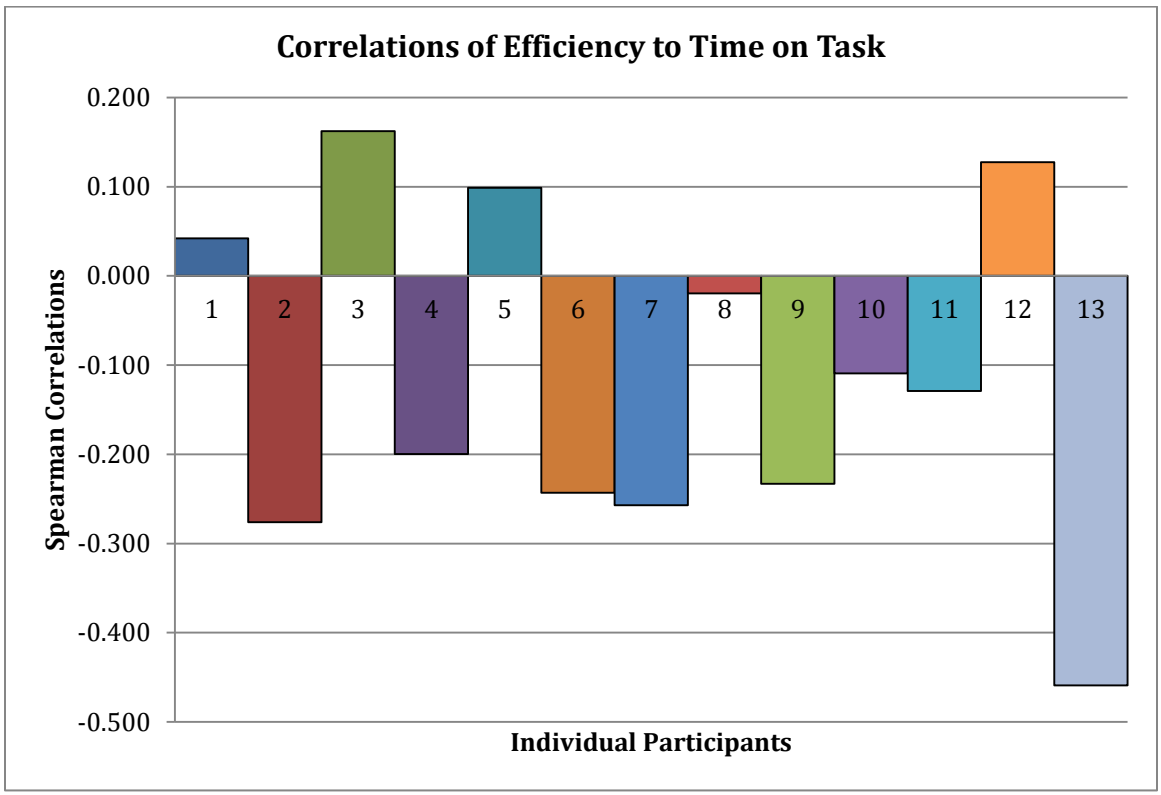



**Figure 33. Comparison of efficiency of electronic multiple-choice response with the time on task for 13 participants.**

Efficiency measures are also compared to maximum pupil diameter for the same 13 participants. As before the correlations were expected to be negative due to the expected

inverse relationship of the two pieces of cognitive information. While 9 of the 13 participants show this trend to be accurate, the four participants that do not follow the trend show maximum pupil diameter to be inconclusive in determination of a pattern (Figure 34).

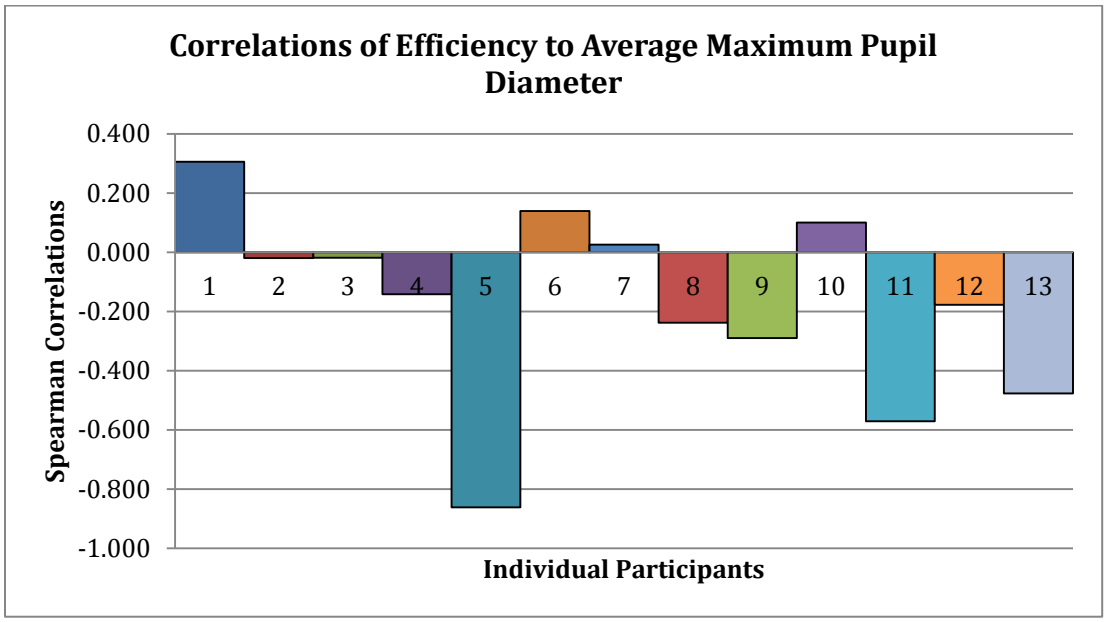



**Figure 34. Comparison of efficiency with TEPR data.**

### 6.1.5 Discussion

The high reliability measures of open-ended response to electronic multiple choice options shows that it is possible to use electronic options as a measure of efficiency. This then allows further advancement of information in the instrument design. The alternate areas associated with measurement in the instrument measured using the eye tracker showed support of each measure for use in the classroom. The construct of mental effort showed positive correlations with time on task, lending towards time being a predictor that the load on working memory is related to the amount of mental effort applied during task solving. However, when examining the measure of maximum pupil diameter in comparison to mental effort, the mix of positive and negative correlations are inconclusive. Further investigation into this shows that the issue may

not be in measuring average or single pupil diameters themselves, but in use of the tracking technology and the type of task used in the instrument. The methodology applied provides participants with use of paper to perform task solutions. In order for participants to utilize the paper focus must be taken away from the screen. This causes loss of tracking during task solution, and leads to missing pupil data during overall data evaluation (Figure 35).

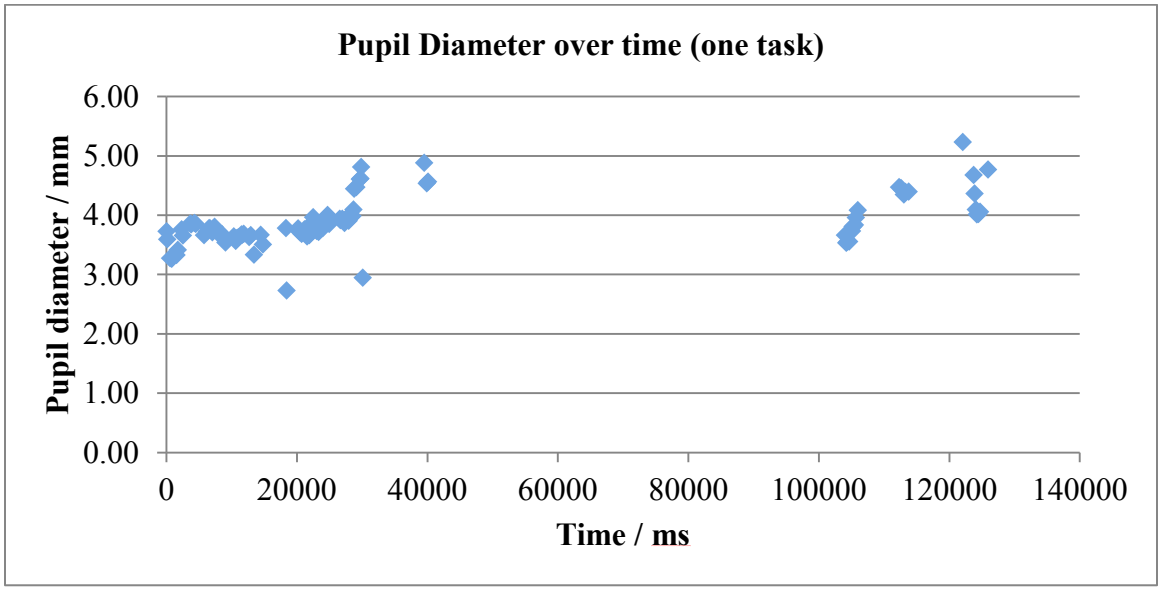



**Figure 35. Pupil Diameter versus time map as an explanation for why TEPR is inconclusive, but may still be valid.**

This could also explain why the results of mental effort to maximum pupil diameter are inconclusive. The subjective mental effort measure cannot be compared to maximum pupil diameter measures if data on pupil diameter is lost. However, the general trend shown in figure 35 supports the concept of change in pupil diameter during task solution, showing a link between pupil measurements and mental effort. Additionally, the trend of pupil diameter and loading of working memory while reading and executing a task prior to the loss of tracking also supports this conclusion. The pupil diameter profile shown in Figure 35 exemplifies the correlation of the pupil diameter to the loading on working memory, with the loss of tracking

taking place while further loading is presumably occurring. The tracking resumes with the delivery or reporting of the response. While the data here may be inconclusive for a relationship between task-evoked pupillary response (as measured through the maximum pupil diameter) and mental effort, the use of subjective mental effort ratings is still supported through the use of time on task. Support is also seen in the performance and efficiency measurements. The expectation of a negative correlation for performance is expected because performance should increase as information in schema becomes more automated. This then leads to a decrease in demand on working memory, lowering mental effort. Negative r-values for the performance to mental effort comparison demonstrates this relationship. If students were not able to accurately gauge their own load on working memory, a more positive correlation would be expected. This was also demonstrated in the class wide data collected during the semester of these interviews (Figure 36).

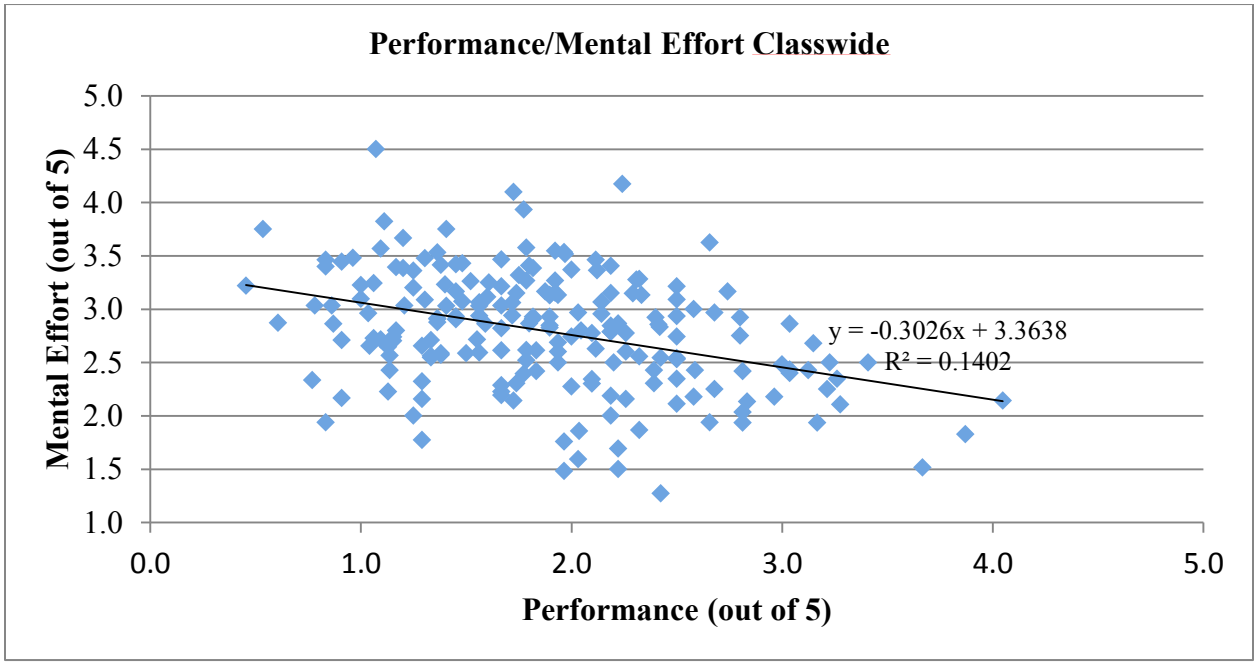



**Figure 36. Trend line demonstrating the relationship of mental effort to performance on a single task during classroom application of the instrument.**

In comparison using performance and efficiency in comparison of one another a positive r-value would be expected. Here if efficiency increases so should performance, and vice versa. Therefore efficiency acts as a predictor of performance. If a comparison of the three areas of mental effort, performance, and efficiency are taken; each pair of constructs serves as a predictor of the third construct, showing validity of the instrument in turn. Efficiency measurement comparisons of TOT and maximum pupil diameter are, however, not conclusive. While the negative r-values are seen as expected, it is not a prevalent outcome for the 13 participants. In both of these measurements approximate 30% of the participants correlated positively when compared to the efficiency construct. While the use of maximum pupil diameter in this study has already been discussed, the use of TOT up to this point has proved to be an accurate measure. The information shown suggests that as with the TOT/maximum pupil diameter correlation, maximum pupil diameter does not follow the expected trend. Reasons for the inconclusive results for TOT to efficiency and maximum pupil diameter then will require further investigation, as will the use of maximum pupil diameter as a single construct in future work on this project. In the meantime other issues involving technology glitches that occurred during tracking, such as screen freezes, may also account for some of the time issues seen when comparing efficiency and TOT and TOT to maximum pupil diameter.

# Chapter 7: Future work

## 7.1 Eye Tracking

In continuation of this study in the future, inconsistencies in maximum pupil diameter and TOT can be addressed through developing a method allowing students to perform problem-solving while working on the screen. Integration of a tablet or SMART board system and a hat mounted versus desk mounted eye-tracker is one possible solution. Research has been done that shows coding may be done to integrate tools such as a calculator into the screen(Tang, Topczewski, Topchewski, & Pienta, 2012). Further information is also available through a more in-depth analysis of individual student information using scan path and heat maps (Figure 37). These give more detailed information of the exact information from a task that is accessed, how often it is accessed, and the amount of time it is utilized.

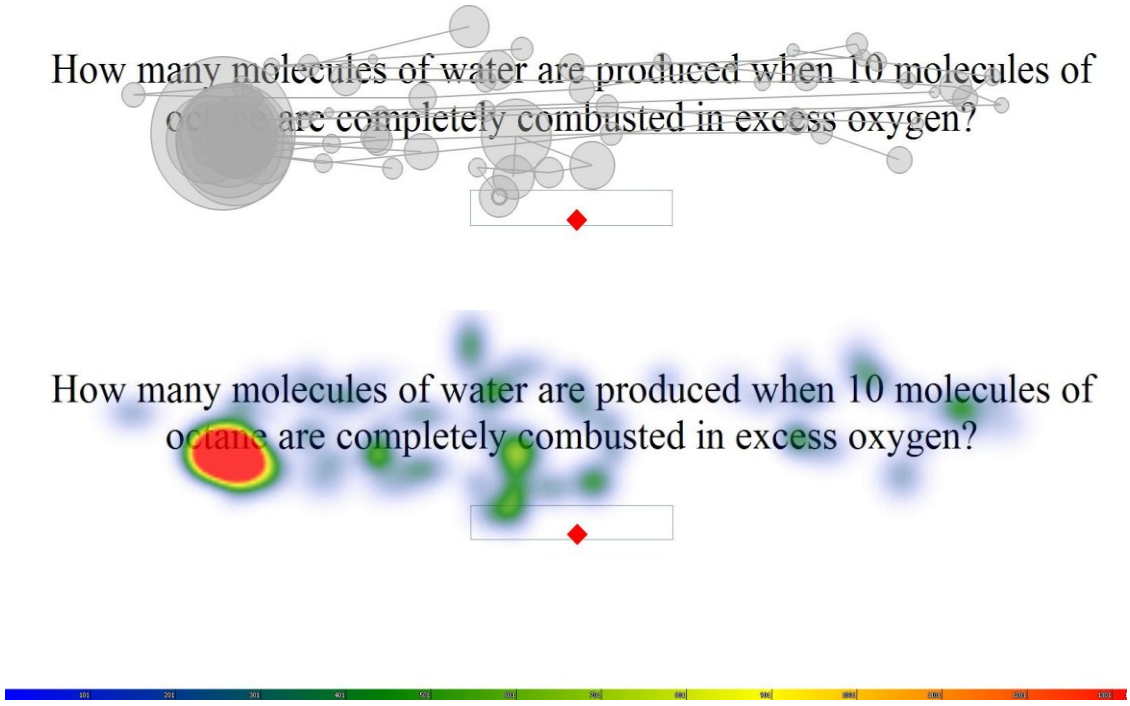



**Figure 37. Data is shown for one participant for one task. The top version of the task shows scan path data with fixations represented as overlapping circles. The darker the color of the overlapping circles the more times the item was accessed. The bottom figure represents the amount of time an item was accessed. The scale on the bottom demonstrates the color intensity associated with the time. The cooler the color the less time was spent looking at the area, while the warmer the color the more time that was spent on the area.**

## Chapter 8: Conclusions

The use of cognitive load theory in conjunction with schema learning theory and adaptive control of thought-rational gives a solid foundation for the use of an instrument that measures load on working memory as part of an efficiency assessment. Through mapping the beginning of student's problem solving strategies in first steps, and comparing this to the load on working memory, a higher predictability of performance on current and future tasks is possible. This ability to measure the students' problem-solving skills and relate it back to an expected outcome based on the mapped compared efficiency of problem processes, allows for the establishment of appropriate interventions when cognitive mapping demonstrates inefficient processes. Through intervening earlier in the schema formation of new information, the likelihood of redirecting schema pathways to more efficient methods is possible.

By developing an instrument that is completely electronic it is in fact possible to have real-time, formative assessment in the classroom based on task performance, mental effort, and efficiency of first step. The use of student generated first step lists for a large scale electronic instrument is supported by the 90% agreement of students across open-response items for a task, and the continual agreement of students at the 90% level in multiple iterations of testing. Combining this with efficiency ratings developed by experts, and a valid method of testing load on working memory through a five-point Likert scale, allowed for the creation of student individual and class wide assessment scores when combined with performance on a task. The eye-tracking data reported demonstrates that load on working memory is in fact reliable when self-reported by students, through use of measurements involving time-on-task and task evoked pupillary response and is accurately measured when posed on a five point quasi-interval scale.

With support from classroom data measurements and eye-tracking data collected in interviews, the development of the RKA for use in the chemistry classroom is a valid tool for assessing student knowledge on individual concepts within the chemistry curriculum. The data also supports the need for an instrument to give a rapid measure of knowledge to allow for timely interventions to aid in the development of more efficient schema and ultimately better learning of the domain.

## Sources Cited:

Anderson, J. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. (1996). A Simple Theory of Comblex Cognition. *American Psychologist, 51*(4), 355-365.

Anderson, J., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Lawrance Erlbaum Associates, Inc., Publishers.

Baddeley, A. (1986). *Working Memory*. Oxford: Oxford Science Publications.

Beatty, J., & Wagoner, B. (1978). Pupillometric Signs of Brain Activation Vary with Level of Cognitive Processing. *Science, 199*, 1216-1218. doi: 10.1126/science.628837

Brunning, R., Schraw, G., Norby, M., & Ronning, R. (2004). *Cognitive Psychology And Instruction* (4 ed.). Upper Saddle River: Parson Education, Inc.

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

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10*, 1-20.

Chase, W., & Simon, H. (1973). Perception in Chess. *Cognitive Psychology, 4*, 55-81.

Chi, M., Glaser, R., & Rees, E. (1982). Expertise in Problem Solving. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence* (Vol. 1, pp. 7-75). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

einstruction. (2007). Classroom performance system (Version 3.2): einstruction.

Ericsson, K. A., & Kintsch, W. (1995). LONG-TERM WORKING-MEMORY. *Psychological Review, 102*(2), 211-245. doi: 10.1037//0033-295x.102.2.211

Gravetter, Frederick J., Wallnau, Larry B. (2007). *Statistics for the Behavioral Sciences*. Belmont, CA: Thomson Wadsworth.

Hendy, K., Hamilton, K., & Landry, L. (1993). Measuring Subjective Workload: When Is One Scale Better Than Many? *Human Factors, 35*(4), 579-601.

Jacob, R., & Karn, K. (2003). Eye Tracking in Human–Computer Interaction and Usability Research: Ready to Deliver the Promises. In R. D. Hyona (Ed.), *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research* (Vol. 2, pp. 573-605). Oxford, England: Elsevier Science BV.

Kahneman, D., & Jacson, B. (1966). Pupil Diameter and Load on Memory. *Science, 154*, 1583-1585.

Kalyuga, S. (2006). Rapid cognitive assessment of learners' knowledge structures. *Learning and Instruction, 16*, 1-11. doi: 10.1016/j.learninstruc.2005.12.002

Kalyuga, S. (2007). Expertise Reversal Effect and Its Implications for Learner-Tailored Instruction. *Educational Psychology Review, 19*, 509-539. doi: 10.1007/s10648-007-9054-3

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The Expertise Reversal Effect. *Educational Psychologist, 38*(1), 23-31.

Kalyuga, S., & Sweller, J. (2004). Measuring Knowledge to Optimize Cognitive Load Factors During Instruction. *Journal of Education Psychology, 96*(3), 558-568. doi: 10.1037/0022-0663.96.3.558

Kalyuga, S., & Sweller, J. (2005). Rapid Dynamic Assessment of Expertise to Improve the Efficiency of Adaptive E-learning. *ETR&D, 53*(3), 83-93.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the Task-Evoked Pupillary Response with a Remote Eye Tracker. *P. 2008 on Sym. on Eye Tracking Res. & Appl.*, 69-73.

Knaus, K.J.; Murphy, K.L.; Blecking, A.; Holme, T.A.; A Valid and Reliable Instrument for Cognitive Complexity Rating Assignment of Chemistry Exam Items, *Journal of Chemical Education*, 2011, *88*, 554-560.

Larkin, J., McDermott, J., Simon, D., & Simon, H. (1980). Models of Competence in Solving Physics Problems. *Cognitive Science, 4*, 317-345.

Likert, R. (1932). A technique for the measurment of attitudes. *Archives of PsychologyI,* 22, 1-54.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding Instruction. *Journal of Education Psychology, 88*(1), 49-63.

Marshall, S. (1995). Schemas in Problem Solving. New York: Cambridge University Press.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43-52. doi: 10.1207/s15326985ep3801_6

Miller, G. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits On Our Capacity For Processing Information. *The Psychological Review, 63*(2), 81-97.

Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. (2003). Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Psychologist, 38*(1), 63-71.

Paas, F., & van Merrienboer, J. (1994a). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*, 51-71.

Paas, F., van Merrienboer, J. J. G., & Adam, J. J. (1994). MEASUREMENT OF COGNITIVE LOAD IN INSTRUCTIONAL-RESEARCH. *Perceptual and Motor Skills, 79*(1), 419-430.

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin, 124*(3), 372-422.

Sensomotoric Instruments, SMI. (2011a). BeGaze (Version 3.2).

Sensomotoric Instruments, SMI. (2011b). Experiment Center (Version 3).

Sensomotoric Instruments, SMI. (2011c). RED 250.

Stone, B., Lee, M., Dennis, S., & Nettelbeck, T. (2004). *Pupil Size and Mental Load*. Paper presented at the 1st Adelaide Mental Life Conference, Adelaide, AU. http://www.psychology.adelaide.edu.au/cognition/aml/S.

Sweller, J. (1994). Cognitive Load Theory, Learning Difficulty, and Instructional Design. *Learning and Instruction, 4*, 295-312. doi: 0959-4752(94)00010-7

Sweller, J., van Merrienboer, J., & Paas, F. (1998). Cognitive Architecture and Instructional Design. *Educational Psychology Review, 10*(3), 251-294.

Tang, H., Topczewski, J., Topchewski, A., & Pienta, N. (2012). *Permutation test for groups of scanpaths using normalized Levenshtein distances and application in NMR questions*. Paper presented at the Proceedings of the Symposium on Eye Tracking Research and Applications, Santa Barbara, California.

van Merrienboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147-177. doi: 10.1007/s10648-005-3951-0

## Appendix A: Example of Spot Review Sheet

**Spot Review Problems (Lec 401)** **December 1, 2008**

Problem: How many types of intermolecular forces does ammonia have?

Your work: Your answer: _____

**Please tear off and hand in:**

Name:_____ DS Section: _____

For today's problem, your **first step** is:

# Appendix B: Master Task Document

**Rapid Knowledge Assessment Problems**

**General**

---

1 What is 225 Kelvin in degrees Fahrenheit? **$-54^{o}F$**

A solving for the temperature in degrees Celsius

B writing the equation to convert Kelvin to Celsius

C writing the equation to convert Celsius to Kelvin

D writing the equation to convert Kelvin to Fahrenheit

E combining the equation to convert Kelvin to Celsius and the equation to convert Celsius to Fahrenheit

F reading the exercise, however I am not sure how to start the exercise

2 How many 1.0 $cm^3$ cubes of gold have a collective mass of 1.0 kg? **52**

A calculating the volume

B looking up the density of gold

C converting kilograms to grams

D writing the equation for density

E solving the density equation for volume

F reading the exercise, however I am not sure how to start the exercise

3 A nanoparticle is 85 nm in size. What magnification must be used to view this particle to a size of 3 cm? **350k X**

A converting nanometers to meters

B convert all of the units into meters

C converting nanometers to centimeters

D finding the difference in the order of magnitudes

E writing an equation and filling in the appropriate numbers

F reading the exercise, however I am not sure how to start the exercise

**Atomic Structure**

---

4 What is the atomic number of an alkali metal with 20 neutrons? **19**

F09 steps

A listing the alkali metals

B A - (number of neutrons) = Z

C Z + (number of neutrons) = A

D finding the element with A = 40

E finding group 1 on the periodic table

F finding the group 1 element with A – Z = 20

G reading the exercise, however I am not sure how to start the exercise

5 What is Z for an ion with a charge of 3– and 36 electrons? **33**

A writing Z = 36 – 3

B finding Z on the periodic table

C # protons = # electrons in a (neutral) atom

D solving for the # electrons in a (neutral) atom

E solving for # protons, given the number of electrons

F reading the exercise, however I am not sure how to start the exercise

6 What is the mass of nitrogen-15, if nitrogen has two stable isotopes and nitrogen-14 has a mass of 14.003074 amu and a relative abundance of 99.632? Use 14.007 amu for the average atomic mass of nitrogen. **15.1**

A solving $\%^{15}\mathrm{N} = 0.378\%$

B writing $\%^{15}\mathrm{N} = 100 - \%^{14}\mathrm{N}$

C making a table of the given data

D writing the equation $\mathrm{N} = \left(\frac{\%^{15}\mathrm{N}}{100}\right)\left({}^{15}\mathrm{N}\right) + \left(\frac{\%^{14}\mathrm{N}}{100}\right)\left({}^{14}\mathrm{N}\right)$

E writing the equation $14.007\ \text{amu} = \left(\frac{99.632\%}{100}\right)(14.003074\ \text{amu}) - \left(\frac{\%^{15}\mathrm{N}}{100}\right)\left({}^{15}\mathrm{N}\right)$

F reading the exercise, however I am not sure how to start the exercise

**Electronic Structure**

---

7 Of the four visible lines in the hydrogen atomic emission spectrum, what is the value of $n_{\text{hi}}$ for the line of the shortest wavelength? **6**

A writing $\Delta E = h\nu = R_{\mathrm{H}}\left(\frac{1}{n_{\mathrm{i}}^2} - \frac{1}{n_{\mathrm{f}}^2}\right)$

B drawing a picture of the Bohr model

C looking at the hydrogen emission spectrum

D knowing that the Balmer series is for nlow = 2

E listing the series of transitions for the Balmer series

F knowing the relationship between wavelength and energy

G reading the exercise, however I am not sure how to start the exercise

8 What is the lowest value of $m_l$ for any *d* atomic orbital? **–2**

A listing the values of ml

B knowing that ml = -l…0…+l

C finding the values for a d orbital

D calculating the number of orbitals

E knowing that the highest value of l is n - 1

F finding the value of l which corresponds to a d sublevel

G reading the exercise, however I am not sure how to start the exercise

9 There are two elements in period 3 which have 2 unpaired electrons in the ground state. What is the sum of Z for these two elements? **30**

A writing an orbital diagram

B finding period 3 on the periodic table

C writing the electron configurations for period 3 elements

D knowing that all of these elements contain 1s2 2s2 2p6

E writing the orbital diagram for only the nonmetals in period 3

F writing the orbital diagram for only silicon, phosphorus and sulfur

G knowing that only the 3p sublevel will have more than one unpaired electron

H reading the exercise, however I am not sure how to start the exercise

10 How many electrons are in the 3rd energy level for a ground state atom of manganese? **13**

A finding Mn on the periodic table

B writing the full electron configuration

C writing the noble gas electron configuration

D finding the number of electrons in 3s, 3p, and 3d

E knowing that there are only s, p, and d sublevels in the 3rd energy level

F reading the exercise, however I am not sure how to start the exercise

11 How many elements in period 3 have atoms which are paramagnetic in the ground state? **6**

A writing an orbital diagram

B finding period 3 on the periodic table

C finding the number of paramagnetic elements

D finding the number of paired and unpaired spins

E knowing how many electrons are in s and p orbitals

F writing the electron configurations for period 3 elements

G reading the exercise, however I am not sure how to start the exercise

**Periodicity**

12 What is the atomic number for a period 3 element with 578 kJ·mol$^{-1}$, 1820 kJ·mol$^{-1}$, 2750 kJ·mol$^{-1}$, 11,600 kJ·mol$^{-1}$ as the first four ionization energies? **13**

A listing the elements in period 3.

B locating the largest difference in IE.

C following the trend for IE in period 3.

D finding period 3 on the periodic table.

E knowing that the largest IE increases indicates core electrons.

F finding the number of valence electrons for the elements in period 3.

G reading the exercise, however I am not sure how to start the exercise

13 What is the atomic number for the element in group 2 most likely to form covalent bonds with hydrogen? **4**

A defining a covalent bond

B listing the group 2 elements

C locating group 2 on the periodic table

D knowing the periodic trend of electronegativity

E using trend for electronegativity to find the highest and lowest electronegative element.

F reading the exercise, however I am not sure how to start the exercise

**Chemical Bonding and Molecular Structure**

14 What is the subscript for lead when lead with 78 electrons combines to form an ionic compound with sulfate? **1**

A writing the atomic number of lead

B writing the charge on the sulfate ion

C writing the formula of the sulfate ion

D determining the charge on the lead ion

E determining the number of electrons on lead

F reading the exercise, however I am not sure how to start the exercise

15 How many lone pairs of electrons on one molecule of nitrogen trichloride? **10**

A identifying the central atom

B writing the skeletal structure

C writing the Lewis dot structure

D determining the number of valence electrons

E determining the number of bonds in the molecule.

F determining the number of electrons pairs in the molecule

G reading the exercise, however I am not sure how to start the exercise

15b What is the total number of valence electrons in one molecule of $PCl_3$? **26**

A finding P and Cl on the periodic table.

B finding the group number for P and Cl.

C reading the formula for which elements are present.

D determining the number of valence electrons for each atom.

E determining the number of valence electrons for each element.

F reading the formula for how many atoms of each element are present.

G reading the exercise, however I am not sure how to start the exercise

16 What is the bond angle in the carbonate ion? **120**

A finding the central atom

B writing the Lewis dot structure

C placing the atoms in the structure

D finding the number of electron groups

E writing the formula for the carbonate ion

F calculating the total number of valence electrons

G finding the number of valence electrons for each element

H reading the exercise, however I am not sure how to start the exercise

16b How many sigma bonds are in one molecule of allene, $H_2CCCH_2$? **6**

A defining a sigma bond.

B drawing the structural formula

C drawing the Lewis dot structure

D placing the atoms in the structure

E identifying the number of single bonds

F determining the total number of valence electrons

G reading the exercise, however I am not sure how to start the exercise

17 What is the bond order is $N_2^+$? **2.5**

A drawing a Lewis dot structure

B writing the equation for bond order

C writing the molecular orbital diagram

D finding nitrogen on the periodic table

E determining the number of electrons for the ion

F finding the number of valence electrons in nitrogen

G reading the exercise, however I am not sure how to start the exercise

**Liquids and Solids**

18 How many types of intermolecular forces does carbon tetrachloride have? **1**

A writing the chemical formula

B drawing the Lewis dot structure

C determining the polarity of the molecule

D determining the number of valence electrons

E determining the types of bonds in the molecule

F reading the exercise, however I am not sure how to start the exercise

19 How many types of intermolecular forces does ammonia have? **3**

A determining the shape

B determining the polarity

C writing the chemical formula

D writing the Lewis dot structure

E listing the possible intermolecular forces

F listing the intermolecular forces for ammonia

G reading the exercise, however I am not sure how to start the exercise

20 What is Z for the noble gas with the lowest boiling point? **2**

A knowing that Z = # of electrons for atoms

B finding the noble gases on the periodic table

C knowing the noble gases only have dispersion forces

D knowing the correlation between strength of IMF and boiling point

E knowing the correlation between strength of dispersion forces and total number of electrons

F reading the exercise, however I am not sure how to start the exercise

21 Polonium is the only element which packs in a simple cubic structure in its crystalline form. Polonium has an atomic radius of 164 pm. What is the density of polonium?

**9.83 g·cm$^{-3}$**

A converting pm to cm

B defining simple cubic

C writing the equation for density

D drawing a picture of the unit cell

E finding the molar mass of polonium

F determining the volume of one atom

G writing the equation to solve for one side of the unit cell

H reading the exercise, however I am not sure how to start the exercise

22 The molar heat of vaporization of benzene is 31.0 $kJ \cdot mol^{-1}$ and the normal boiling point is 80.1°C. What is the vapor pressure of benzene at room temperature (25°C)? **108 mmHg**

A writing the value for R

B writing the ideal gas law

C converting the temperature into kelvin

D writing the Clausius-Clayperon equation

E knowing the vapor pressure at the normal boiling point

F reading the exercise, however I am not sure how to start the exercise

**Properties of Matter – Formula Calculations**

23 What is the mass of $1.0 \times 10^{23}$ molecules of oxygen molecules? **5.32 g**

A determining the molar mass of oxygen

B writing the formula for molecular oxygen

C calculating the number of moles of oxygen

D understanding that $6.022 \times 10^{23}$ molecules = 1 mole

E reading the exercise, however I am not sure how to start the exercise

24 What is the percent composition by mass of oxygen in aluminum acetate? **47.03%**

A writing the chemical formula

B determining the mass of oxygen

C determining the number of moles of oxygen

D determining the molar mass of aluminum acetate

E reading the exercise, however I am not sure how to start the exercise

25 What is the mass (in g) of sodium in a 10.0 g sample of sodium phosphite? **4.66 g**

A writing the chemical formula

B writing the molar mass of sodium

C determining the molar mass of sodium phosphite

D determining the number of moles of sodium phosphite

E reading the exercise, however I am not sure how to start the exercise

27 What is the mass of oxygen in a sample of aluminum sulfite which contains 1.00 g of aluminum? **2.67 g**

A writing the formula

B finding the moles of oxygen.

C finding the moles of aluminum.

D finding the mass of aluminum sulfate.

E finding the moles of aluminum sulfate.

F identifying the mole ratio of oxygen to aluminum.

G reading the exercise, however I am not sure how to start the exercise

27_cl What is the mass of oxygen in a sample of sodium nitrate which contains 1.00 g of sodium? **2.09 g**

Use same steps as 27 above and swap substance

28 How many atoms of hydrogen are in the empirical formula for a hydrocarbon containing 83.6% carbon by mass? **7**

A 83.6% carbon = 83.6 g carbon

B calculating the number of moles of carbon.

C calculating the number of moles of hydrogen.

D 100 g total – 83.6 g carbon = 16.4 g hydrogen.

E 100% total – 83.6% carbon = 16.4% hydrogen.

F knowing that hydrocarbons contain only carbon and hydrogen.

G reading the exercise, however I am not sure how to start the exercise

28_cl How many atoms of hydrogen are in the empirical formula for a hydrocarbon containing 93.70% carbon by mass? **4**

Use same steps as 28 above and swap values

**Reactions**

29 What is the sum of the stoichiometric coefficients (balanced to the lowest common denominator) for the balanced reaction of benzene ($C_6H_6$) reacting with oxygen to form carbon dioxide and water? **35**

A writing the formula for water

B writing the formula for oxygen

C identifying products and reactants

D writing the formula for carbon dioxide

E writing the skeletal equation for the reaction

F reading the exercise, however I am not sure how to start the exercise

29_cl What is the sum of the stoichiometric coefficients (balanced to the lowest common denominator) for the balanced reaction of methanol ($CH_3OH$) reacting with oxygen to form carbon dioxide and water?**11**

A writing the formula for water

B writing the formula for oxygen

C identifying products and reactants

D writing the formula for carbon dioxide

E writing the skeletal equation for the reaction

F reading the exercise, however I am not sure how to start the exercise

30 What is the sum of the stoichiometric coefficients (balanced to the lowest common denominator) for the balanced reaction of zinc with hydrochloric acid (HCl)? **5**

A identifying the type of reaction

B identifying the reactants and products

C writing the skeletal chemical equation

D writing the chemical formula for the products

E writing the chemical formula for the reactants

F reading the exercise, however I am not sure how to start the exercise

31 What is the sum of the stoichiometric coefficients (balanced to the lowest common denominator) for the balanced net ionic reaction of iron(III) chloride and potassium hydroxide? **5**

A writing the skeletal equation

B writing the total balanced equation

C writing the formula for the products

D writing the formula for the reactants

E writing the total ionic balanced equation

F reading the exercise, however I am not sure how to start the exercise

**Stoichiometry**

32 How many molecules of water are produced when 10 molecules of octane are completely combusted in excess oxygen? **90**

A calculating mass of water

B calculating moles of water

C calculating mass of octane

D calculating moles of octane

E writing the balanced equation

F use a mole ratio for comparing moles of octane to water

G use the mole ratio for comparing molecules of octane to water

H reading the exercise, however I am not sure how to start the exercise

33 What mass of oxygen is needed to completely combust 1.00 g of ethanol to produce carbon dioxide and water vapor? **2.08 g**

A using a mole ratio

B writing the balanced equation

C calculating the moles of oxygen

D calculating the moles of ethanol

E calculating molar mass of ethanol

F writing the chemical formula for ethanol.

G reading the exercise, however I am not sure how to start the exercise

35 What is the percent yield for a reaction of lithium hydroxide with carbon dioxide to yield lithium bicarbonate in which 50.0 g of lithium hydroxide is reacted and 72.8 g of lithium bicarbonate is experimentally obtained? **51.3%**

A writing the balanced equation

B writing the formula of all compounds

C writing the equation for percent yield

D calculating moles of lithium hydroxide

E finding the molar mass of lithium hydroxide

F finding the molar mass of lithium bicarbonate

G calculating the theoretical yield of lithium bicarbonate

H reading the exercise, however I am not sure how to start the exercise

**Aqueous Reactions**

37 How many grams of sodium nitrate must be used in order to prepare $5.00 \times 10^2$ mL of a 0.100 M solution? **4.25 g**

A knowing that M = mol / L

B converting the volume from mL to L

C determining the molar mass of sodium nitrate

D writing the chemical formula for sodium nitrate

E calculating the number of moles of sodium nitrate.

F reading the exercise, however I am not sure how to start the exercise

38 It is desired to add $H_2O$ to 50.0 mL of a 9.00 M aq solution of sodium phosphate in order to decrease the concentration to 0.245 M. What should the final volume be (in L)?

40 37.2 mL of 0.142 M NaOH is needed for complete neutralization of 25.0 mL of a solution of $H_2SO_4$. What is the molar concentration of the acid? **0.106 M**

A writing the balanced equation

B converting the volume into liters

C determining the molar mass of NaOH

D determining the molar mass of H2SO4

E determining the number of moles of NaOH

F determining the number of moles of H2SO4

G knowing that the dilution equation is M1V1 = M2V2

H reading the exercise, however I am not sure how to start the exercise

42 If 20.25 mL of calcium nitrate and an excess of sodium phosphate are mixed, and 0.250 g of the precipitate forms at a 95.5% yield, what is the molar concentration of calcium nitrate?(310.18) **0.125 M**

A converting mL to L

B writing the balanced chemical equation.

C determining the formula of the precipitate.

D determining the molar mass of calcium nitrate

E determining the theoretical yield of the precipitate.

F determining the number of moles of calcium nitrate

G reading the exercise, however I am not sure how to start the exercise

43 What is the oxidation state (entered without a sign) for chromium in iron(III) dichromate? **6**

A writing the charge on each ion

B writing the charge on dichromate ion

C determining the oxidation number on iron

D writing the chemical formula of dichromate

E determining the oxidation number on oxygen

F writing the chemical formula for iron(III) dichromate.

G reading the exercise, however I am not sure how to start the exercise

43_c What is the oxidation state (entered without a sign) for chlorine in iron(III) perchlorate? **7**

A writing the charge on each ion

B writing the charge on perchlorate ion

C determining the oxidation number on iron

D writing the chemical formula of perchlorate

E determining the oxidation number on oxygen

F writing the chemical formula for iron(III) perchlorate.

G reading the exercise, however I am not sure how to start the exercise

**Gases**

---

44 What is molar mass of a gas with a density of 1.96 $g \cdot L^{-1}$ at STP? **44.0**

A writing the ideal gas law

B listing standard conditions

C writing the equation for density

D writing the formula, density = MMP / RT

E calculating molar mass from the ideal gas law

F combining the equations for density and the ideal gas law

G reading the exercise, however I am not sure how to start the exercise

46 Ammonia, $NH_3$, is made commercially by reacting $N_2$ and $H_2$. How many liters of $NH_3$ can be made from 4.62 L of $H_2$ if both gases are measured at the same temperature and pressure?

47 What is the root mean square speed (in m/s) of hydrogen at $0^oC$? **1836**

A writing the equation for speed.

B writing the formula for hydrogen.

C writing the value for R in J/(mol K).

D converting the temperature to Kelvin.

E determining the molar mass for hydrogen.

F reading the exercise, however I am not sure how to start the exercise

48 What mass (in g) of nitrogen must be added to 15.0 g of neon in a 50.0 L container at 50.0$^{o}$C to yield a final pressure of 1050 mmHg? **52.2 g**

A writing the ideal gas law.

B converting mmHg to atm.

C converting the temperature into Kelvin.

D writing Dalton's Law of Partial Pressures.

E determining the partial pressure of neon.

F determining the number of moles of neon.

G reading the exercise, however I am not sure how to start the exercise

48_c What mass (in g) of oxygen must be added to 15.0 g of neon in a 50.0 L container at 50.0$^{o}$C to yield a final pressure of 950 mmHg? **51.6 g**

**Thermochemistry**

49 What is the final temperature (in $^{o}C$) when 1 gallon of water evolves 118.8 kJ of heat when it cools from 32.5$^{o}C$? 1 gallon = 3790 mL; $s_{H2O}$ = 4.18 $J \cdot g^{-1} \cdot {}^{o}C^{-1}$ **25.0$^{o}C$**

A converting kJ to J.

B writing the equation for heat.

C converting the volume into metric.

D converting the mass of water to grams.

E solving the heat equation for change in temperature.

F writing the equation for heat with both initial and final temperature.

G reading the exercise, however I am not sure how to start the exercise

50 How much heat (in kJ) is generated when 100.0 g of iron reacts with excess oxygen under standard conditions (and constant pressure) to form iron(III) oxide if the standard enthalpy of formation for iron(III) oxide is –822.2 $kJ \cdot mol^{-1}$ **–736 kJ**

A writing the formula for iron(III) oxide

B determining the number of moles of iron

C determining the molar mass of iron(III) oxide

D determining the number of moles of iron(III) oxide

E writing the chemical equation for the formation of iron(III) oxide

F reading the exercise, however I am not sure how to start the exercise

50_c1 How much heat (in kJ) is generated when 100.0 g of copper reacts with excess oxygen under standard conditions (and constant pressure) to form copper(I) oxide if the standard enthalpy of formation for copper(I) oxide is –166.69 kJ·mol$^{-1}$ **–131 kJ**

A writing the formula for copper(I) oxide

B determining the number of moles of copper

C determining the molar mass of copper(I) oxide

D determining the number of moles of copper(I) oxide

E writing the chemical equation for the formation of copper(I) oxide

F reading the exercise, however I am not sure how to start the exercise

50_c2 How much heat (in kJ) is generated when 100.0 g of aluminum reacts with excess oxygen under standard conditions (and constant pressure) to form aluminum oxide if the standard enthalpy of formation for aluminum oxide is –1669.8 kJ·mol$^{-1}$ **–3095 kJ**

A writing the formula for aluminum oxide

B determining the number of moles of aluminum

C determining the molar mass of aluminum oxide

D determining the number of moles of aluminum oxide

E writing the chemical equation for the formation of aluminum oxide

F reading the exercise, however I am not sure how to start the exercise

51 What mass (in g) of ethane ($C_2H_6$) is needed to boil to completion 1 gallon (3790 g) of water initially at 20°C if the standard molar enthalpy of combustion for ethane is –1558.8 kJ·mol$^{-1}$? Assume the process is only 50% efficient. $s_{H2O}$ = 4.18 J·g$^{-1}$·°C$^{-1}$ and $\Delta H_{vap}$ = 40.656 kJ·mol$^{-1}$ **379 g**

A writing the formula for heat

B determining the molar mass of water

C determining the molar mass of ethane

D determining the number of moles of water

E calculating the change in temperature of the water

F writing the balanced chemical equation for the combustion of ethane

G reading the exercise, however I am not sure how to start the exercise

51_c What mass (in g) of methane is needed to boil to completion 1 liter of water initially at 40°C if the standard molar enthalpy of combustion for methane is –890.25 kJ·mol$^{-1}$? Assume the process is only 50% efficient. $s_{H2O}$ = 4.18 J·g$^{-1}$·°C$^{-1}$ and $\Delta H_{vap}$ = 40.656 kJ·mol$^{-1}$ **90.3 g**

A writing the formula for heat

B determining the molar mass of water

C determining the molar mass of methane

D determining the number of moles of water

E calculating the change in temperature of the water

F writing the balanced chemical equation for the combustion of methane

G reading the exercise, however I am not sure how to start the exercise

## Appendix C: Complexity Rubric

Rubric for assigning task complexity and an example of the worksheet used by raters to assign task complexity.

To assign complexity

(1) Read the task.

(2) Count the number of pieces of knowledge

(3) Estimate from the perspective of a student a relative difficulty rating to each

(4) Use the rubric to add up the component complexities to determine a numerical complexity rating

(5) Increase the overall complexity rating of an item by estimating how interrelated or interactive the chemistry knowledge must be for the task

***Number & relative difficulty/complexity of component concepts or skills needed to master item***

| Rating | easy | medium | difficult |
|---|---|---|---|
| 1 | 1 | | |
| 2 | 2 | 1 | |
| 3 | 3 – 4 | 2 | |
| 4 | 5 – 6 | 3 – 4 | 1 |
| 5 | | 5 – 6 | 2 |
| 6 | | | 3 – 4 |
| 7 | | | 5 – 6 |

***Concept/skill interactivity***

| | |
|---|---|
| Non-significant | 0 |
| Basic | +1 |
| complex | +2 |

Complexity Analysis of Item #1

| **Elements** | **Difficulty of each element** | **Total number of each category** | **Ranking of elements** | **Sum of ranking** | **Complexity rating** |
|---|---|---|---|---|---|
| | | Easy: | Easy: | | |
| | | Medium: | Medium: | | |
| | | Hard: | Hard: | | |
| **Interactivity rating:** | | | | | |

## Appendix D: Comparison of Preparatory Chemistry Lectures in 2007

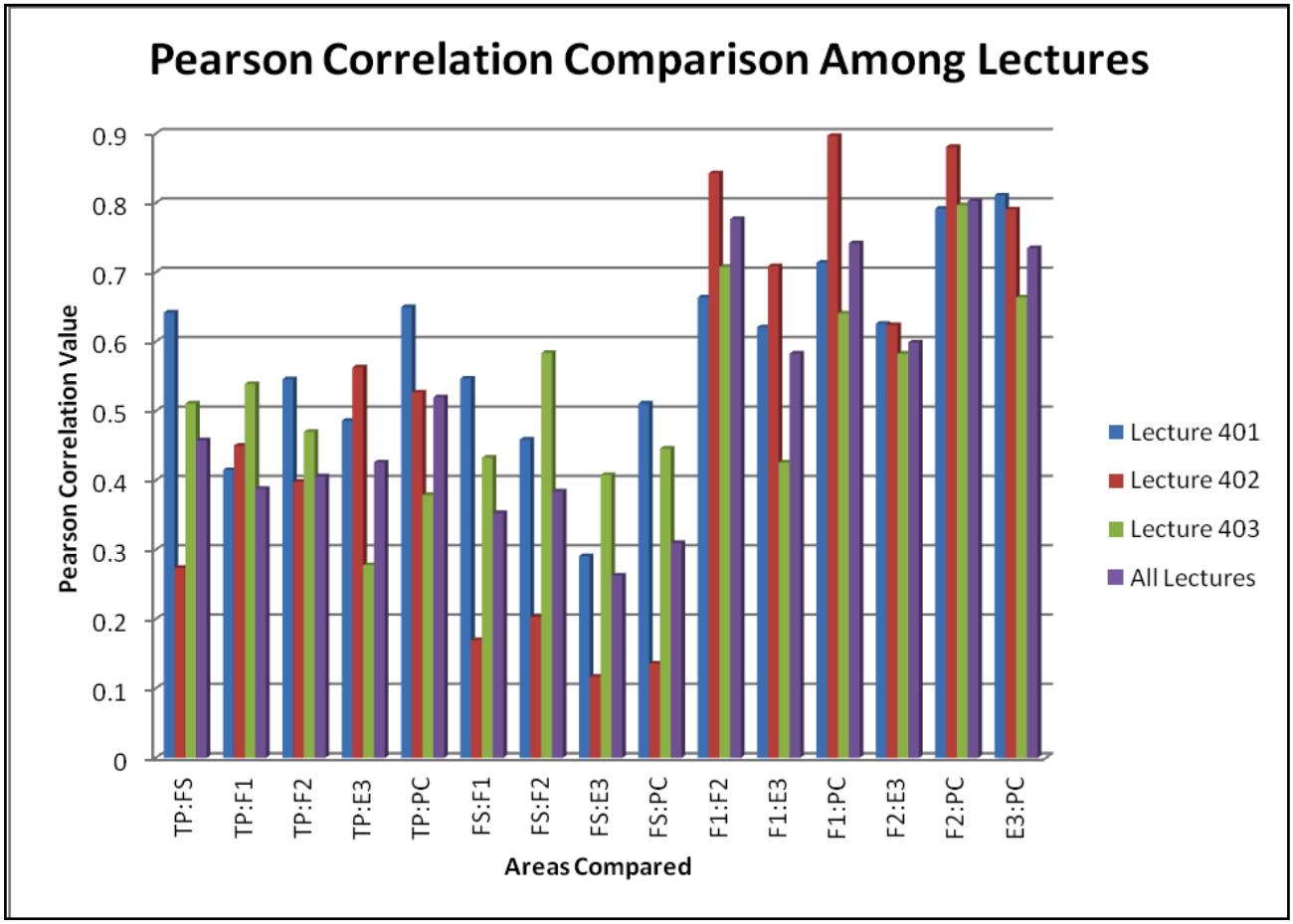
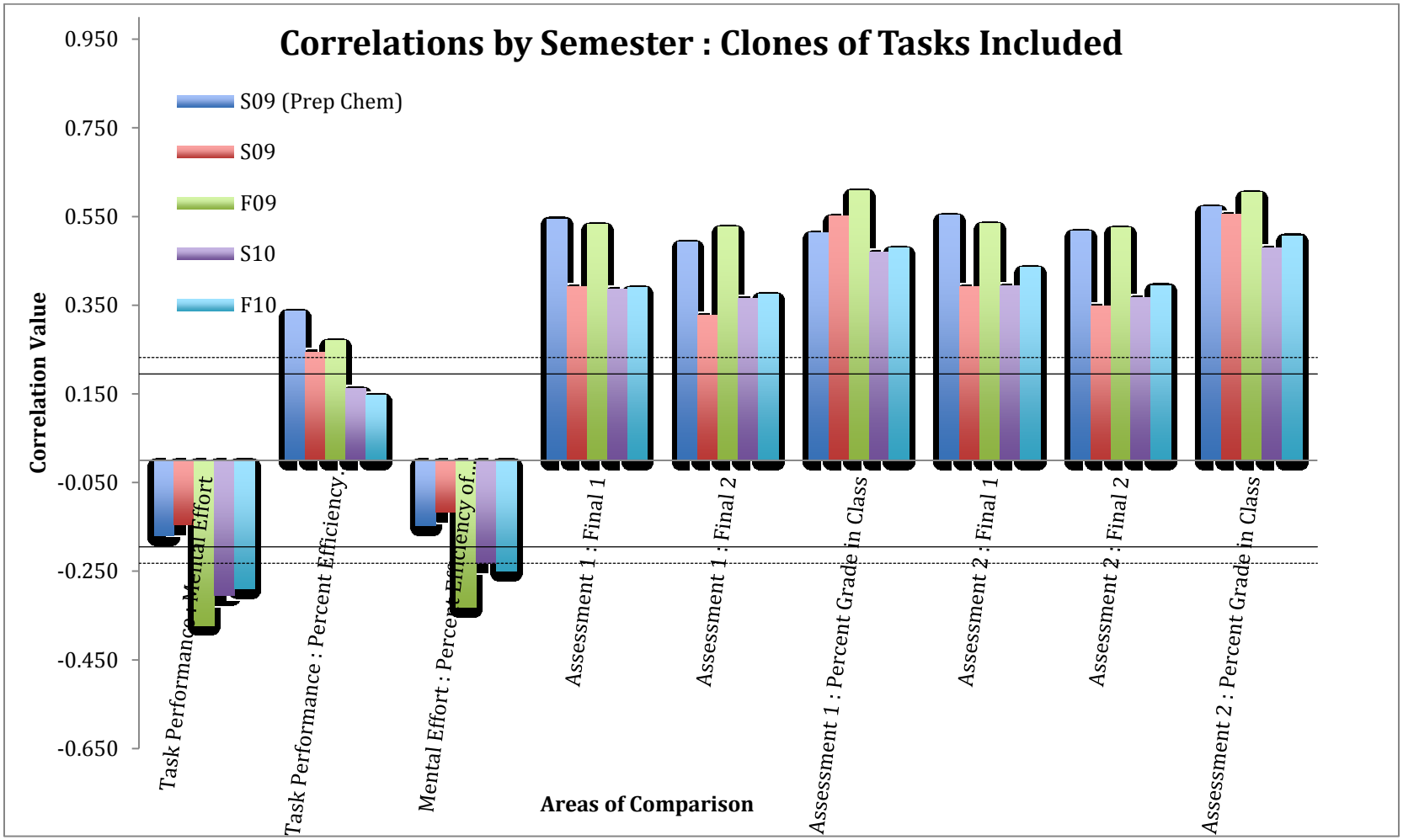
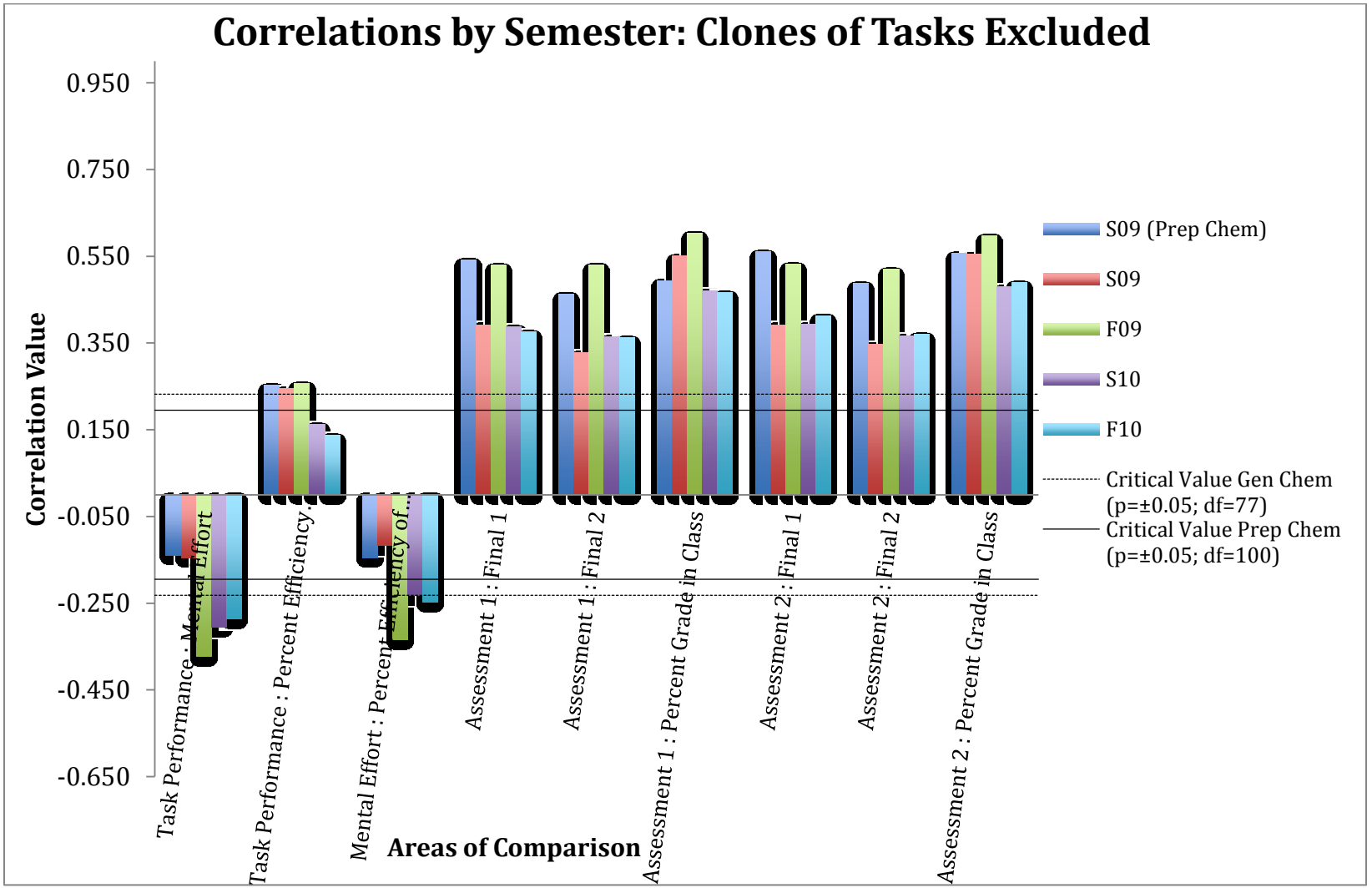



Figure : Comparison of Pearson Correlation values for Task Performance (TP), First Step Performance (FS), Final 1 (F1), Final 2 (F2), Third Exam (E3), and Percent Grade in Class (PC).

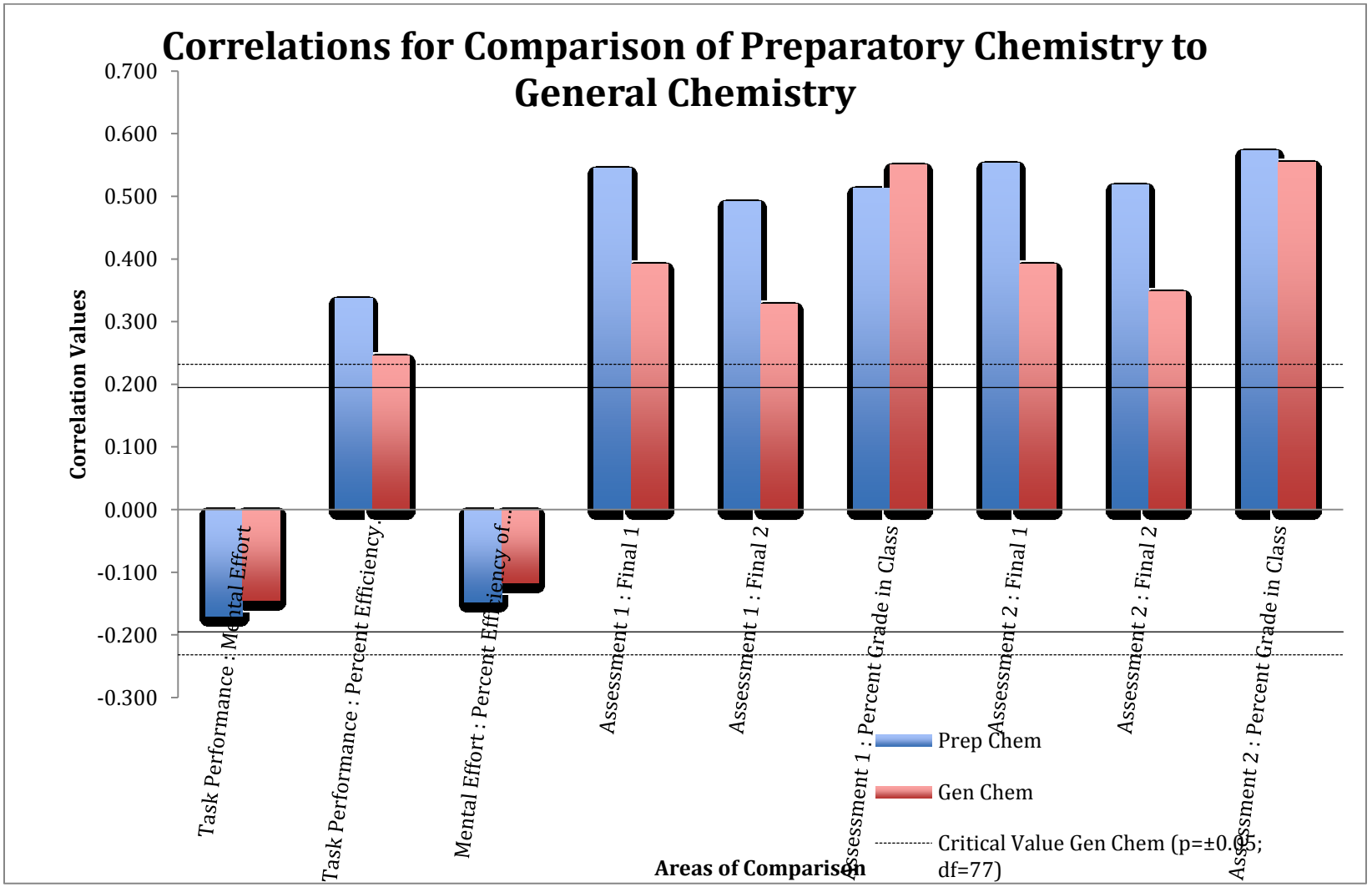
## Appendix E: Comparison of Data for 2009 and 2010 Including Cloned Tasks



**Appendix G: Comparison of Data for 2009 and 2010 Excluding Cloned Tasks**



**Appendix F: Comparison of Preparatory and General Chemistry for 2009**

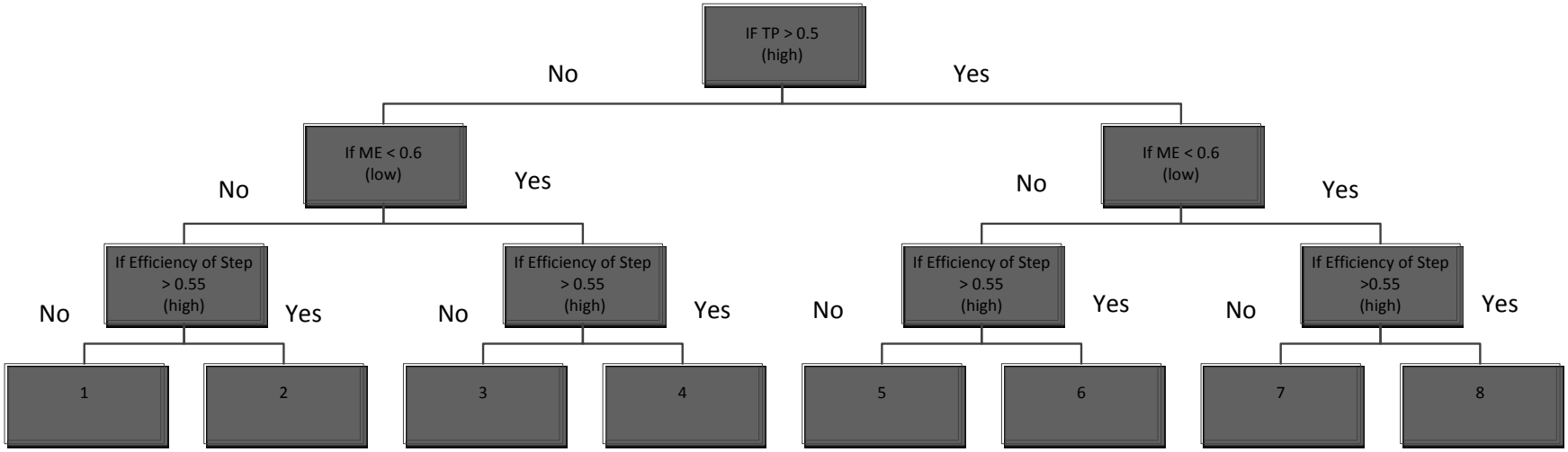


## Appendix H: Spring 2007 Rapid Knowledge Assessment Problems

1. What is the mass of $1.0 \times 10^{23}$ molecules of oxygen molecules? (5.32 g)
2. What is the percent composition by mass of oxygen in aluminum acetate? (47.03%)
3. What is the molar mass for an ionic compound which contains 75.9% nickel by mass and nitrogen? (232.11 $g \cdot mol^{-1}$)
4. What is the mass of oxygen in a sample of aluminum sulfite which contains 1.00 g of aluminum? (2.67 g)
5. How many molecules of carbon dioxide are produced when 10 molecules of octane are completely combusted in excess oxygen? (80)
6. What mass of oxygen is needed to completely combust 1.00 g of ethanol to produce carbon dioxide and water vapor? (2.08 g)
7. 56.6 g of calcium combines with excess nitrogen to form calcium nitride. If 32.4 g of calcium nitride is recovered, what is the percent yield for the reaction? (46.5%)
8. What is the percent yield for a reaction of lithium hydroxide with carbon dioxide to yield lithium bicarbonate in which 50.0g of lithium hydroxide is reacted and 72.8 g of lithium bicarbonate is experimentally obtained? (51.3%)
9. What mass of octane is combusted in excess oxygen to produce 355 L of carbon dioxide (density = 1.96 $g \cdot L^{-1}$) at an 80.0% yield for the reaction? (282 g)
10. How many grams of sodium nitrate must be used in order to prepare $5.00 \times 10^2$ mL of a 0.100 M solution? (4.25 g)
11. It is desired to add water to 50.0 mL of a 0.900 M aqueous solution of sodium phosphate in order to decrease the concentration to 0.245 M. What should the final volume be (in L)? (0.184 L)
12. 32.7 mL of 0.142 M NaOH is needed for complete neutralization of 25.0 mL of a solution of $H_2HO_4$. What is the molar concentration of the acid? (0.106 M)
13. If 25.0 mL of each 0.200 M $Ca(NO_3)_2$ and 0.100 M $Na_3PO_4$ are mixed, how many grams of solid $Ca_3(PO_4)_2$ are mixed, how many grams of solid $Ca_3(PO_4)_2$ (molar mass = 310 $g \cdot mol^{-1}$) are formed? (0.388 g)
14. The density of phosphine gas has been found to be 1.26 $g \cdot L^{-1}$ at 50°C and 747 mmHg. Calculate the molar mass of phosphine. (34.0 g/mol)

15. Ammonia, $NH_3$, is made commercially by reacting $N_2$ and $H_2$. How many liters of ammonia can be made from 4.62 L of $H_2$ if both gases are measured at the same temperature and pressure? (3.08 L)

## Appendix I: Organization to Determine Assessment Scores I and II




Assessment Score I