2013

Development of Critical Thinking in Secondary Students

Blair Marlin Lee
Lehigh University

Development of Critical Thinking Skills in Secondary Students

by

Blair Lee

A Dissertation

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Education

in

Educational Leadership

Lehigh University

May 20, 2013

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the Degree of Doctor of Education.

_____
Date

_____
Dissertation Advisor

_____
Accepted Date

Committee Members:

_____
Jill Sperandio

_____
Peggy Kong

_____
Louise Donohue

_____
Yu-Hao Kate Lin

Dedication

This dissertation is dedicated to my wife and children. I cannot adequately express my appreciation for all of the encouragement you gave and the sacrifices you made during the entire doctoral program. Your support gave me the perseverance and determination to finish this chapter of my life.

To my best friend and wife, Kim, I thank you for the sacrifices you made through all of the summers, holidays, weekends, and long nights. Your selfless sacrifice allowed me to focus on accomplishing this goal while you took on the role of father and mother during my absence. You always knew when to give me words of praise and when I needed a kick in the pants to keep me moving forward. The belief you had in me kept me going through my moments of doubt. Without your blessing, sacrifice, love, and support I would not have been able to finish. I have accomplished much more than I ever thought possible because you are my friend and partner.

To my children, Hudson and Maguire, thank you for your understanding and patience during this whole process. I could not have accomplished this without the undeserved belief that you had in me. I am certain that your thousands of little prayers helped me get through every step of this journey. I look forward to witnessing the exciting plans that God has for you. I could not be more proud to be your father.

Acknowledgements

I want to acknowledge and thank my dissertation chair, Jill Sperandio, for the expertise and guidance she provided through this entire process. I am grateful for her encouragement and support as I struggled through the writing and revision process.

I am also very appreciative of the time, effort and support from the members of my committee, Peggy Kong, Louise Donohue, and Kate Lin. I am honored to have worked with such knowledgeable, discerning, and gracious individuals.

I want to thank Daphne Hobson and Roger Douglas for their leadership of the Office of International Programs at Lehigh University. Without their support and guidance during the early years of this journey I am sure I would have lost my way and not reached this goal.

I also want to thank Laura Roberts for her mentorship throughout this process. Her assistance with my statistical analysis was critical; however, her support, guidance, and patience were essential to the success of this study.

I want to acknowledge the efforts of the American International School of Kuwait to quantify the development of critical thinking. Many schools state in their mission and beliefs the desire to develop critical thinking, but you have taken the brave step of measuring it. Thank you for sharing your efforts and data with me.

Finally, I want to thank my dear friend, Russ McLean, for igniting the spark in me to understand the development of critical thinking. His passion for education and belief that critical thinking must be developed in our students if we are to ever have a better more peaceful world is palpable. You inspire me. Thank you.

Abstract

The purpose of this non-experimental, causal comparative study was to examine the development of critical thinking in eleventh-and-twelfth grade students registered to study the International Baccalaureate-Diploma Programme (IB-DP) and its mandatory Theory of Knowledge (TOK) course. *Ex-post facto* data, collected by the American International School of Kuwait from the fall of 2008 to the spring of 2011, was analyzed to determine whether enrollment in the TOK course significantly impacted critical thinking skill development, as measured by the Test of Everyday Reasoning (TER).

The analysis of the data showed that students enrolled in the IB-DP developed critical thinking skills to a greater degree than those students in the comparison group. Further analysis, utilizing Analysis of Covariance (ANCOVA), revealed that prior academic ability, as demonstrated by PSAT scores, eliminated the effect of treatment on student TER performance. Therefore, it was determined that students' performance on the PSAT was a greater indicator of critical thinking skill improvement than participation in the TOK course.

However, the rudimentary supplemental analysis of high and low performers on the PSAT-Critical Reading may suggest that the IB-DP and the TOK course are beneficial to those students that self-select to enter a rigorous program despite previous poor performance on standardized tests. It is suggested that further research be done with these students to determine if significant differences do indeed exist for this student.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

Democratic societies are destined for failure unless their citizens purposefully participate in the development of their own thinking processes (ten Dam & Volman, 2004; Glaser, 1985; Paul, 1984). Successful democratic societies depend on an educated and knowledgeable citizenry forming public policy, a public policy that is generated by informed and robust debate, and evaluated with a spirit of understanding and tolerance (Facione, 2011; Walters, 1986). For societies to develop rationally informed and tolerant policies, they must provide arenas in which the critical thinking abilities of each individual citizen can become involved in the democratic process (Daly, 1998; ten Dam & Volman, 2004). Democratic citizens must develop the ability to think independently and critically (Sternberg & Baron, 1986) and it is incumbent upon the educational institutions to develop these skills (Bailin & Segel, 2003; Dewey 1909/2008; Giancarlo, Blohm & Urdan, 2004; Norris, 1985; Scheffler, 1973).

Educational leaders in the United States realized the importance of instilling strong critical thinking skills in students (Glaser, 1985; Norris, 1985; Paul, 1992). National studies conducted in the 1980s and early 1990s suggested that both secondary and post-secondary students lacked critical thinking skills (National Education Goals Panel, 1992; Association of American Colleges and Universities [AAC&U], 1985; National Institute of Education Study Group, 1984). These findings resulted in a call for the mandatory incorporation of objectives, designed to improve critical thinking skills, in the curricula for American students (Facione, P., Facione, N., Tiwari, A., & Yuen, F.,

2009; Pithers & Soden, 2000).

Education curricula reform aimed at improving critical thinking skills faltered, however, as a failure to form consensus on the definition of critical thinking hindered practitioners and researchers alike from incorporating critical thinking objectives into curricula (Abrami, Bernard, Borokhovski, Wade, Surkes, Tamim & Zhang, 2008; Bissell & Lemons, 2006; Facione, 1990; Geersten, 2003; Greenlaw & DeLoach, 2003). Various definitions of critical thinking skills exist but most focus on two categories, cognitive skills and affective dispositions (Facione, 1990; Halpern, 1998; Chambers, Bartle, Carter-Wells, Bagwell, Greenbaum, Padget, & Thomson, C., 2000; ten Dam & Volman, 2004). Critical thinking requires the ability to use cognitive powers to make purposeful, self-regulatory judgment requiring the use of a core set of cognitive skills (Facione, Sanchez, Facione, & Gainen, 1995). The cognitive skills most often associated with critical thinking include: 1) interpretation, 2) analysis, 3) evaluation, 4) inference, 5) explanation, and 6) self-regulation (Facione, 1990). The affective disposition toward critical thinking consists of an intellectual persistence to consistently use rational thought to solve problems and make decisions (Facione, Facione, & Giancarlo, 1997). Definitions of affective dispositions most commonly include: 1) open-mindedness, 2) flexibility, 3) inquisitiveness, 4) judiciousness, and 5) persistence (Facione, Facione, & Giancarlo, 2000; Chambers, et al., 2000; Tishman & Andrade, 1996).

While educational leaders agree on the importance of critical thinking and its two primary components, cognitive skills and affective dispositions, little consensus exists about how best to integrate and foster critical thinking in school curricula (Ben-Chaim, Ron, & Zoller, 2000; Grauerholz & Bouma-Holtrop, 2003). Some curriculum developers

have sought to incorporate skills throughout the curriculum (Geersten, 2003; McPeck 1990a). Geertsen (2003) and McPeck (1990a) argued that certain critical thinking skills were specific for different disciplines and are not transferable. They argued that a simple course or two of "critical thinking" would have little effect on the development of critical thinking skills (McPeck, 1990b).

Other researchers contended that critical thinking involved abilities and skills that were transferable to any problem or subject matter (Ennis 1990; Vaughn 2008). They supported the development of single courses that introduced students to critical thinking skills and provided them with practice in using them within a liberal arts education (Paul, 1992; Halpern, 1998, Vaugh, 2008). Little evidence exists to suggest that one approach was better than another, or that the direct teaching of cognitive skills and affective dispositions enhances critical thinking in students (Geertsen, 2003).

The International Baccalaureate (IB) – Diploma Programme (DP) represents one example of a curriculum consisting of a course designed to teach critical thinking skills within a liberal arts framework. The development of critical thinking skills is one of the IB's primary goals (IBO, 2006a). Accordingly, the IB (2012a) requires that all students enrolled in the DP take the Theory of Knowledge (TOK) course. This requirement is designed to ensure that "students become critical thinkers who understand the interdisciplinary nature of learning" (IBO, 2006a, p.7). The design of the TOK course encourages teachers and students to "reflect critically on diverse ways of knowing and on areas of knowledge" (IBO, 2009a, p.4). Zemplen (2007) believed that the approach of the IB was superior to many other curricula as a result of its purposeful inclusion of critical thinking development. However, the generic cognitive skills learned in the TOK course

may not be adequate for students to successfully navigate the specific critical thinking that is required in the IB sciences (Zemplen, 2007).

Study

This study attempted to address whether critical thinking can be developed by what Ennis (1989) described as a general course designed to enhance the critical thinking skills of its students. This causal comparative study used previously collected data to measure gains in the development of critical thinking in students at the American International School of Kuwait (AISK) from the fall of 2008 to the spring of 2011. At AISK all eleventh-and twelfth-grade classes have been modeled on the IB-DP curriculum. Therefore, all students received similar curriculum objectives whether they were enrolled in the IB-DP or not. However, students that were registered to study the IB-DP must have taken the TOK course along with other requirements, while the non-IB-DP students were not required to take the TOK course.

Research Questions

My research seeks to examine the relationship of the development of critical thinking in students participating in and completing the TOK course in comparison with students who have not completed the TOK course. The research questions guiding the study include:

1. Is there a significant difference between the growth of critical thinking skills, as

demonstrated by Test of Everyday Reasoning (TER) – Total Score, between students who participated in a course designed to improve critical thinking skills (the International Baccalaureate – Diploma Programme Theory of Knowledge course) and those who did not participate in the course?

a. Is there a significant difference between the growth of targeted cognitive skills, as demonstrated in TER – Scale Scores (analysis and interpretation; evaluation and explanation; and inference), between students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the course?
b. Is there a significant difference between the growth of targeted reasoning skills, as demonstrated in TER – Deductive and Inductive Reasoning Scale Scores, between students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the course?

Research Design and Method

I examined these questions using a quantitative *ex post facto* research method, employing a pretest, treatment, and posttest design. This causal comparative study analyzed data collected in the fall of 2008 to the spring of 2011 by AISK from two consecutive cohorts of students progressing through grades eleven and twelve. Participants were drawn from eleventh-and twelfth-grade students at AISK, some of whom were enrolled in the IB-DP TOK course, while others were not. Students' critical thinking skills were tested prior to the treatment, the IB-DP Theory of Knowledge course,

and a posttest was administered after the completion of the course.

At the beginning of the eleventh-grade and then again at the end of the twelfth-grade all AISK students were administered the Test of Everyday Reasoning (TER). The TER is a critical thinking test designed by California Critical Thinking Skills Test (CCTST) for high school students in grades ten through twelve.

The CCTST based its family of tests on the Delphi expert consensus definition of critical thinking. The Delphi panel of experts identified six core critical thinking cognitive skills: interpretation, analysis, evaluation, inference, explanation, and self-regulation (Facione, 1990). The TER reports an overall score on critical thinking skills and five scale scores (analysis and interpretation, inference, evaluation and explanation, inductive reasoning, and deductive reasoning) for each test taker. The five TER scale scores are considered essential elements for critical thinking (Facione, Facione, & Winterhalter, 2011).

This study compared two naturally occurring groups of students, one which has self-selected to enroll as Diploma Programme students and one which has not. Students entering eleventh-grade at AISK must choose between enrollment in the International Baccalaureate-Diploma Programme (IB-DP), with the required Theory of Knowledge (TOK) course, and enrollment as a regular AISK student following the IB-DP curriculum, but with no TOK course requirement.

Two cohorts were examined for this study. Cohort 1 students were in eleventh-grade during the 2008 – 2009 school year; and Cohort 2 students were in eleventh-grade during the 2009 – 2010 school year. The growth of critical thinking skills was examined

for cohorts 1 and 2. Students fully enrolled in the IB-DP were compared with students enrolled as regular students by analyzing TER data collected by AISK.

Significance of the Study

Despite the increasing importance of developing critical thinking skills in education (Atkinson, 1997; Daly, 1998; Grauerholz & Bouma-Holtrop, 2003), a dearth of empirical research remains a problem (Baker, 1981; Geertsen, 2003). Leading theorists disagree on the most effective methods of teaching critical thinking skills (Grauerholz & Bouma-Holtrop, 2003). Debate persists as to whether or not a single course can influence critical thinking. However, few studies have indicated whether courses such as TOK increase the ability of students to think critically (Geertsen, 2003). This study sought to establish whether the TOK course influences the development of critical thinking skills in students.

This study adds to the body of empirical research seeking to understand the development of critical thinking skills. The understanding gained from this research should aid curriculum designers in determining if courses designed to develop critical thinking are effective.

Limitations

The limitations of this study arise mainly from the *ex post facto* design that has several inherent threats to internal validity in this research design (Fraenkel & Wallen,

2006). Because this type of design examines data that was previously collected, it does not allow the researcher to control, manipulate, or isolate any variables within the study or allow for random assignment to treatment (Fraenkel & Wallen, 2006). The treatment and comparison groups may be significantly different based on their self-selection of groups. For example, if it is determined that IB-DP students develop critical thinking at a greater rate than non-IB-DP students, then the explanation may be that students that are predisposed to critical thinking development may be more readily attracted to enrolling in the IB-DP and thus the results would not be related to the treatment. However, specific analytical tools were incorporated to increase the strength of various correlational inferences.

Another limitation affecting this study is that the TOK course is not the only difference in curriculum between the two cohorts. While TOK is the most significant difference specifically targeting the development of critical thinking skills, other differences exist between the treatment group and the comparison group that may affect the development of these skills. These differences are the result of whether students meet the requirements to be an IB-DP student. IB-DP students are required to take three Higher Level (HL) courses, write an Extended Essay (EE), and complete a Creativity, Action, Service (CAS) experience. While all students in the comparison group do receive a liberal arts education following the IB-DP curriculum, they do not need to meet the other requirements.

Although students at AISK receive instruction following the IB-DP curriculum, it is not possible to determine whether significant differences in scores can be solely attributed to the TOK course. Students in the comparison group differ in course selection,

teachers, and enrollment in IB-DP certificate courses (students receive subject specific credit from the IB but the students are not enrolled as IB-DP students). Students in the treatment group also differ in their course selection, teachers for classes, and teachers for TOK instruction. However, the IB central office expends considerable energy in an attempt to ensure similar instruction, not only at the same school, but also all across the world (IBO, 2006b; IBO, 2007). Although it is impossible to achieve complete standardization among all schools offering the IB-DP, the IB central office has put practices in place in an attempt to provide a level of quality and standardization of curriculum and practices in the schools that offer it (IBO, 2001; IBO, 2009c; IBO, 2010).

Definition of Terms

The following definitions are provided to clarify the terminology used for this study:

International Baccalaureate (IB)-Diploma Programme (DP): The IB Diploma Programme is a two-year curriculum for students in grades eleven and twelve. Courses of study are divided into six different groups: group 1 - language A1, group 2 - second language, group 3 - individuals and societies, group 4 - experimental sciences, group 5 - mathematics and computer, and group 6 - the arts. Subjects may be taken at either the standard or higher level. The programme also has three core requirements designed to broaden the educational experience: Creativity, Action, Service; Extended Essay; and Theory of Knowledge (IBO, 2007, 2008, 2009c).

Standard Level (SL): Courses studied at the standard level require 150 hours to

adequately complete (IBO, 2007, 2008).

Higher Level (HL): Courses studied at the higher level require students to explore these subjects at a greater depth. Two hundred and forty teaching hours are recommended to adequately complete this course (IBO, 2007, 2008).

Theory of Knowledge (TOK): The interdisciplinary TOK course is designed to unify the six academic groups and develop an appreciation of differing cultural perspectives (IBO, 2006a). The course encourages critical thinking to help students evaluate different approaches to knowing. TOK is designed to ask questions. The most fundamental question is "How do we know [a particular knowledge claim] to be true?" (IBO, 2009a, p. 5). The role that language and thought have upon the development of critical thinking is studied throughout the course (IBO, 2002; 2012b).

Extended Essay (EE): The extended essay requires students to independently research, under the guidance of a supervisor, a question relating to one of the subjects they are studying (IBO, 2006a). This research culminates in an essay that may not exceed 4000 words (IBO, 2002).

Creativity, Action, Service (CAS): Creativity, action, service requires students to actively engage in real life tasks beyond their academic endeavors (IBO, 2006a). Students are required to do meaningful activities involving these three areas. They may choose an activity that combines all three components or target each one separately (IBO, 2002). Creativity often involves the arts but can be any experience that requires creative thinking; Action requires physical activity leading to a healthy lifestyle; Service requires voluntary contributions to the local or global community (IBO, 2012b).

International Baccalaureate-Diploma Programme (IB-DP) student: IB-DP students meet all the requirements to potentially obtain a diploma from the IB. They participate in six subjects at higher level or standard level; complete the Extended Essay; meet the Creativity, Action, Service requirements, and complete the Theory of Knowledge course (IBO, 2007).

Certificate Students: Certificate students do not meet the requirements to be awarded the IB diploma but are taking IB-DP courses. Certificate students that successfully pass an examination for an IB-DP course are awarded a certificate for that course (IBO, 2007).

Delphi Method: The Delphi Method is a qualitative research method utilizing a panel of experts willing to utilize their expertise to work toward an agreement (Facione, Giancarlo, Facione, & Gainen, 1995). The Delphi Method allows experts to work collaboratively about a particular topic of interest. The central investigator organizes the group and initiates discussion regarding an initial question. All responses are organized, summarized, and transmitted back to all panelists for reactions, replies, and additional questions. When consensus is near, the central investigator proposes a final statement and asks for agreement. If a large majority of the panel agrees, a consensus statement will emerge with points of disagreement registered (Facione, 2011).

Test of Everyday Reasoning (TER): The TER is a 35 item multiple-choice test designed for secondary students which measures the reasoning skills of the test-taker. This progressive test measures the test-taker's ability to analyze and evaluate various forms of information. The TER reports "provide an overall score on critical thinking

skills (Total Score) and 5 scale scores (Analysis, Inference, Evaluation, Inductive Reasoning and Deductive Reasoning) are provided for each test taker" (Facione et. al., 2012).

Total Score: The Total Score indicates the overall critical thinking skill level of an individual. This score is useful for the purpose of comparing the critical thinking skills of one individual to another (Facione, et. al., 2012).

Analysis: For the purposes of this study, analysis has a dual meaning. First of all, it means "to comprehend and express the meaning or significance of a wide variety of experiences, situations, data, events, judgments, conventions, beliefs, rules, procedures or criteria" (Facione, et. al., 2012, p. 11). Analysis also means, "to identify the intended and actual inferential relationships among statements, questions, concepts, descriptions or other forms of representation intended to express beliefs, judgments, experiences, reasons, information or opinions" (Facione, et. al., 2012, p. 11).

Inference: Inference means to "identify and secure elements needed to draw reasonable conclusions; to form conjectures and hypotheses; to consider relevant information; and to educe the consequences flowing from data, statements, principles, evidence, judgments, beliefs, opinions, concepts, descriptions, questions, or other forms of representation" (Facione, et. al., 2012, p. 11).

Evaluation: Two aspects of evaluation exist for the TER. First of all it means, "to assess the credibility of statements or other representations which are accounts or descriptions of a person's perception, experience, situation, judgment, belief or opinion; and to assess the logical strength of the actual or intended inferential relationships among

statements, descriptions, questions, or other forms of representations" (Facione, et. al., 2012, p. 11). On the TER, evaluation also means "to state the results of one's reasoning; to justify that reasoning in terms of the evidential, conceptual, methodological, criteriological and contextual considerations upon which one's results were based; and to present one's reasoning in the form of cogent arguments" (Facione, et. al., 2012, p. 11).

Inductive Reasoning: Inductive reasoning occurs when a conclusion to an argument is justified by the assumed facts stated in the proposition. The conclusion, however, is not absolute. Drawing conclusions about something less familiar based on the characteristics of something familiar is reasoning by analogy or also known as inductive reasoning. Inductive reasoning is also used when a prediction of what will happen in the future is based on past experiences. Quality inductive reasoning requires the conclusion to be probabilistic based on an accurate inference of the premise (Facione, et. al., 2012). Inductive reasoning arrives at a general conclusion, presumed to be true, based on specific information known to be true.

Deductive Reasoning: In deductive reasoning, the conclusion of an argument is presumed true if the premise that it is based on is assumed to be true (Facione, et. al., 2012). Deductive reasoning moves from generalizations, presumed to be true, to a specific conclusion that is assumed to be true.

Interpretation: Interpretation is being able to "comprehend and express the meaning or significance of a wide variety of experiences, situations, data, events, judgments, conventions, beliefs, rules, procedures or criteria" (Facione, 1990, p. 16).

Explanation: Explanation is "to state the results of one's reasoning; to justify the

reasoning in terms of the evidential, conceptual, methodological, criteriological and contextual considerations upon which one's results were based; and to present one's reasoning in the form of cogent arguments" (Facione, 1990, p. 21).

Self-Regulation: Self-regulation requires an individual to consciously "monitor one's cognitive activities, the elements used in those activities, and the results educed, particularly by applying skills in analysis and evaluation to one's own inferential judgments with a view toward questioning, confirming, validating, or correcting either one's reasoning or one's results" (Facione, 1990, p. 22).

Comparison Group: The comparison group is composed of students attending AISK who are not IB-DP students. These students are in the regular liberal arts program and are enrolled in courses that follow the IB-DP curriculum. These students do not take the Theory of Knowledge course, write the Extended Essay, or participate in Creativity, Action, Service. They are also not required to take three Higher Level courses. A high percentage of the comparison group are certificate students and are enrolled in Standard Level courses with the IB-DP students.

Treatment Group: The treatment group is composed of students attending AISK who are IB-DP students. These students take the Theory of Knowledge course, write the Extended Essay, and participate in Creativity, Action, Service. They are also required to take three Higher Level courses.

Chapter 2

Literature Review

This review of the literature examines the broader issue of whether a course designed to develop generic critical thinking skills can be successful in achieving this objective. It presents the theoretical and empirical approaches that have been used to understand critical thinking development. The review focuses in particular on four domains of research surrounding critical thinking education theory. The first domain discusses the history of critical thinking theory and its application to modern education models. The second domain investigates the development of a definition of critical thinking. The third domain focuses on instructional techniques used to develop critical thinking skills. The fourth domain examines the IB-Diploma Programme (IB-DP) and its application to research on critical thinking in the development of curriculum and instructional techniques.

History of Critical Thinking

The Ancient Greek philosophers are celebrated for their propensity for questioning authority and basing their opinions on observations and rational thought. At the time, the concept of asking questions was considered radical and dangerous because it implied that traditional authority figures, such as the gods or the government, were not infallible or in possession of specialized knowledge (Paul, 1992). They encouraged the

young people of their time to demand explanations from their elders and to recognize that some explanations were more worthy of acceptance than others (Thayer-Bacon, 2000). Socrates, the most famous and influential of the Greek philosophers, recognized the need to question authority and traditional beliefs and in so doing laid the foundation for critical thinking (Hare, 1998).

Although Greek philosophy laid the foundation for critical thinking, Immanuel Kant's philosophy marked the high point of the Enlightenment, which embraced humanism and the benefits of human rationality, thus introducing the modern era of democracies (Mingers, 1992; Toulmin, 1990). Kant's seminal work *The Critique of Pure Reason*, published in 1781, questioned the basic assumptions and judgments used to establish fundamental truths. By challenging humanity to not accept knowledge as given, Kant questioned what previous thinkers and societies had ordained as absolutes.

Kant played a pivotal role in the development of *critical theory*, but Hegel, Marx, and the Frankfurt School also heavily influenced the examination and critique of society (Magee 1990; Rorty, 1991). Hegel (1991), in his 1822 political work *Elements of the Philosophy of Right*, criticized the contemporary subjugated state. He believed this subjugation was a historical condition and not a natural state, and saw the need for liberation. He sought to reveal the social and political structures that dominate and shape our world.

Marx and Engels (1848/2010) developed a theory intended to influence the alienated, underprivileged masses to revolt against the capitalist system of exploitation in their 1848 pamphlet the *Communist Manifesto*. The Frankfurt School of critical theorists

saw the Enlightenment as having faltered in bringing about the emancipation of humankind. They argued that modern advancements had created technological and political structures that removed the scarcity of resources repressing the poor. They believed that critical theorists must work for universal freedom from toil and domination and bring about human emancipation (Brocklesby & Cummings, 1996).

In 1962, Jurgen Habermas (1989) extended the work of the Frankfurt School by publishing his first major work, *The Structural Transformation of the Public Sphere.* In this immensely controversial text he conceptualized critical theory as knowledge that empowered human beings to liberate themselves through self-reflection. The project of human emancipation seeks to achieve the maximum development for all individuals by improving the quality of work and life in the societies in which they participate.

Collective emancipation requires individual understanding and knowledge about one's values and beliefs. Habermas (1989) contended that critical theory assists people to understand what needs to be done to support self-reflection. Postmodern thinkers, such as Habermas, wanted to expose the traditionally imposed structures and categories that govern our lives as historically situated and prone to deconstruction. They realized the importance of being able to articulate one's thoughts, questions, and feelings as a necessary attribute to initiate social change. Communication through dialogue was seen as a means to bring about mutual understanding and rational consensus (Heydebrand, 2004).

The concept of developing critical thinking to improve social conditions evolved from the originators of critical theory. Theorists encouraged the development of critical

thinking skills to challenge the oppressive constructs that limited the freedom of societies (Brocklesby & Cummings, 1996). Questioning, arguing, re-thinking, and reflecting were seen as necessary skills for an individual to contribute to society and provide a platform for change (Brookfield, 1987). The changes and challenges made by members of postmodern society resulted from the willingness to question, to argue, and to think at a higher level of cognitive function such as application, analysis, and evaluation. Critical thinking skills were seen as necessary in the postmodern era.

While Habermas was developing the theory of human emancipation educational scholars began to apply critical theory to the practical purposes of education. John Dewey, in *How We Think: A Restatement of the Relation of Reflective Thinking to the Educative Process*, presents a concept of education that incorporates the characteristics of critical thinking as "reflective thinking… active, persistent, and careful consideration of a belief or supposed form of knowledge in light of the grounds which support it and the further conclusions to which it trends" (1933, p. 9). Dewey, who had studied such famous American pragmatists as Charles Saunders Pierce and William James, recognized the necessity for individuals to actively engage their own thinking through reflection and evaluation, giving reasons and explanations to defend their assumptions. He encouraged individuals to participate in the process of reflective thinking in order to improve their understanding and learning. Dewey began to develop "a concept of education which stressed the importance of thought, inquiry, and intellectual independence" (Hare, 1998, p. 39). Dewey exercised a great influence on the intellectual traditions that concerned the next generation of scholars and educational theorists (Hare, 1998).

Edward Glaser (1941), in *An Experiment in the Development of Critical Thinking*,

is credited with leading what is often referred as the "First Wave of the Critical Thinking Movement" (Walters, 1994). Glaser argued that critical thinking must become a primary goal of educational curricula. His work generated a discourse on the need for the development of critical thinking within educational circles. Glaser worked with Watson to identify intellectual skills, abilities, and dispositions necessary to make decisions unaffected by emotions or bias (Watson & Glaser, 1942).

Benjamin Bloom (1956) entered into the critical thinking discourse with his attempt to further define these intellectual skills and abilities. He articulated six levels of cognition: knowledge, comprehension, application, analysis, synthesis, and evaluation. These levels proceed from the most basic levels of thinking to the most complex. The three highest levels (analysis, synthesis, and evaluation) represent critical thinking (Clark, 2010; Kennedy, Fisher, & Ennis, 1991). Bloom's influence on the understanding of critical thinking is considerable but interest and discussion continued to exist primarily in the field of philosophy (Hare, 1998).

However, several key incidents in the early 1980s engendered a second wave of critical thinking enthusiasm and this time its applications were seen to be multi-disciplinary (Walters, 1994). National studies conducted in the 1980s and early 1990s suggested that both secondary and post secondary students lacked critical thinking skills (National Education Goals Panel, 1992; Association of American Colleges and Universities [AAC&U], 1985; National Institute of Education Study Group, 1984). These findings resulted in a call for mandatory incorporation of critical thinking skills in the curricula for American students (Facione, Facione, Tiwari, & Yuen, 2009; Pithers & Soden, 2000). With critical thinking again receiving considerable attention, scholars from

various academic disciplines were forced to adapt the concept of critical thinking to their respective disciplines (Norris & Ennis, 1989; Walters, 1994). Suddenly, critical thinking had become an interdisciplinary concern (Paul, 1995).

Definition of Critical Thinking

In 1987, the American Philosophical Association (APA) asked Peter Facione, a leading critical thinking researcher from California State University Fullerton, to conduct a systematic inquiry into the state of critical thinking. Facione utilized the Delphi Method and convened an interdisciplinary interactive panel of experts willing to share their expertise and work toward a consensus on the state of critical thinking (Facione, Sanchez, Facione, & Gainen, 1995). Forty-six persons, widely recognized by their professional colleagues to be leaders in critical thinking instruction, participated in the APA project coordinated by Facione (Facione, 1990).

The panel of experts understood that divergent conceptualizations of critical thinking had hindered curricular efforts. The panel determined that its "most worthwhile contribution would be an articulation of a clear and correct conceptualization of CT" (Facione, 1990, p. 6). After two years of discussion and reflection, their project culminated in this consensus statement regarding critical thinking:

> We understand critical thinking to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based. As such, CT is a liberating

> force in education and a powerful resource in one's personal and civic life. While not synonymous with good thinking, CT is a pervasive and self-rectifying human phenomenon… It combines developing CT skills with nurturing those dispositions which consistently yield useful insights and which are the basis of a rational and democratic society (Facione, 1990, p.3).

The panel went on to identify six cognitive skills that were included in critical thinking a) interpretation, b) analysis, c) evaluation, d) inference, e) explanation, and f) self-regulation (Facione, 1990, p.8). The panel acknowledged that it is not necessary to be proficient at every skill in order to have critical thinking ability (Facione, 1990).

The panel also recognized that not every useful cognitive process should be thought of as critical thinking. Problem-solving, decision-making, and creative thinking are other forms of higher-order thinking that are not identified as critical thinking. The expert panel was adamant that the existence of other forms of higher-order thinking does not preclude the development of a careful and accurate conceptualization of critical thinking (Facione, 1990).

Ennis (1987), a member of Facione's expert panel, identified logic as the key component of critical thinking, which he defined as the correct assessing of statements. Ennis (1979) made a connection between critical thinking and those who were capable of engaging in rational thought. He espoused good judgment as a necessary component in the process of arriving at the correct assessment. Ennis' description helps to understand critical thinking in terms of the explicit underlying processes involved.

Paul (1992) agrees with Ennis that arriving at a correct assessment, or as he puts it

"truth", is the ultimate goal of critical thinking. He believes that this process requires reasoning, the application of standards, and the use of logic. Critical thinking, in his view, involves "figuring out" something which cannot simply be a matter of arbitrary determination: "If what we figure out can be anything we want it to be, anything we fantasize it as being, then there is no logic to the expression 'figure it out'" (Paul, 1992, p. 18). In the process of using reason and logic, he suggests that standards should be judiciously applied. The application of standards and logic requires that views be evaluated judiciously with the intent of determining truth.

Hemming (2000) believes that good critical thinking is not simply arriving at the correct or true assessment. She believes answers and responses should occur as part of a rational process in which well-formulated and substantiated views are adopted with the understanding that these views may continue to come under critical review. This is not to say that the critical thinking process is merely stating diverse opinions, or understanding and appreciating others' perspectives. Lipman (1991) theorizes that all views are not equally valid. Instead, positions are to be well thought-out, plausible, and defensible. It is only when relevant, new information is brought to bear on the matter that one would engage in revision to one's views. Ultimately, critical thinking involves the use of cognitive skills (interpretation, analysis, evaluation, inference, explanation, and self-regulation as articulated by Facione in 1990) to form correct assessments.

While almost all scholars agree that cognitive skills are required to participate in critical thinking, many scholars also suggest that an individual must have the disposition to do so (Ben-Chaim et al., 2000; Facione, 1990; Halpern, 1998; Chambers, et al., 2000; ten Dam & Volman, 2004). It has been argued that critical thinking involves two related,

but conceptually distinct, aspects: the ability to use cognitive skills to reason well but, at the same time, to also possess the disposition to do so (Ben-Chaim et al., 2000; Bailin & Seigel, 2003). The affective disposition required for critical thinking is the internal motivation to consistently use thinking to engage problems and make decisions (Facione, Facione, & Giancarlo, 1997). Affective dispositions are defined as certain tendencies of action toward particular patterns of intellectual behavior (Ennis, 1987, Tishman & Andrade, 1996). Definitions of affective dispositions most commonly include a) open-mindedness, b) flexibility, c) inquisitiveness, d) judiciousness, and e) persistence (Facione, P., Facione, N., & Giancarlo, C., 2000; Chambers, et al., 2000).

This study focused on the development of the cognitive skills required for effective critical thinking. Cognitive skills have been defined as the ability to use cognitive powers to make purposeful, self-regulatory judgments (Facione, Sanchez, Facione, & Gainen, 1995). The cognitive skills most commonly associated with critical thinking include a) interpretation, b) analysis, c) evaluation, d) inference, e) explanation, and f) self-regulation (Facione, 1990).

Teaching Critical Thinking

Enhancing students' critical thinking abilities has become one of the most important goals in education (Bailin & Segel, 2003; Bitter & Legacy, 2007; Daly, 1998; Grauerholz & Bouma-Holtrop, 2003; Scheffler, 1973). Despite this widespread focus on critical thinking, it has remained a complex task for curriculum developers to determine the best method of developing this important skill in students (Pithers & Soden, 2000).

Many theorists have proposed strategies to encourage students' development of critical thinking skills, but few have empirically-tested these strategies to determine if critical thinking skills are being developed (Geersten, 2003; Grauerholz & Bouma-Holtrop, 2003; Pithers & Soden, 2000). Educators and curriculum developers continue to speculate as to the best way to teach these important skills.

Is Critical Thinking a Generic Skill?: Most scholars agree that critical thinking is necessary in all disciplinary areas, but debate persists about whether it consists of a set of generic skills that apply across various disciplines or whether it depends on the specific subject domain in which it is being applied (Abrami et al., 2008). If critical thinking is dependent on knowledge and expertise in a specific discipline, then it should be learned by solving problems within that specific subject area (Smith, 2002). If it is generic, then specialized courses designed to develop critical thinking skills should be taught (Vaughn, 2008).

The subject-specific position is represented in the views expressed by McPeck (1990a). He believes that in the zeal to articulate the need for critical thinking in classrooms, curricula were developed incorrectly. McPeck (1990a) suggests that critical thinking is field dependent. He argues that teaching critical thinking outside of a specific subject area is impossible. For McPeck, critical thinking involves a specific knowledge component for a specified discipline, and a critical component involving reflection of that specific knowledge. Barrow (1991) agrees and voices concern over attempts to implement a critical thinking curriculum which does not relate critical thinking to content and argues that such programs typically avoid critical thought within complex, sophisticated, and important areas of inquiry. Therefore, critical thinking in one subject

does not necessarily translate into critical thinking within another subject. That is why McPeck (1990b) and Barrow (1991) argue that the great bulk of critical thinking programs stressing techniques and strategies without regard to a solid knowledge base are misguided.

Conversely, many scholars argue that critical thinking involves abilities and skills that *are* applicable to any problem or subject matter (Paul, 1992; Seigel, 1990; Vaughn, 2008; Wade & Tavris, 1993). Vaughn describes critical thinking as a systematic evaluation or formulation of beliefs that can be used across all disciplines. Critical thinking is comprised of a set of techniques, attitudes, and principles that we can use in all disciplines (Vaughn, 2008). These scholars would argue that critical thinking is a set of skills that can be systematically applied to all situations regardless of the subject matter.

Halpern (1998) agrees, stressing the importance of students learning skills that can be used in multiple contexts. Students need to develop a repertoire of skills that can be called upon when necessary. Learning should be designed to enable students to retrieve necessary skills regardless of content area. Critical thinking should be taught in such a way as to facilitate the appropriate and spontaneous transfer of skills. When this is achieved students are able to focus their attention on the construct of the arguments or problems presented (Halpern, 1998).

Do Ennis' Four Instructional Techniques Develop Critical Thinking?: Regardless of the theory embraced by the practitioner in the classroom most instructional techniques can be classified within Ennis' (1989) typology. The four instructional

interventions described by Ennis (1989) are: general, infusion, immersion, and mixed. In the general course, specific subject content is not addressed but rather critical thinking skills and dispositions are the targeted learning objectives. The infusion method teaches specific critical thinking learning objectives within the subject matter. Critical thinking and content knowledge are both explicit learning objectives. Immersion, in contrast, does not explicitly identify the critical thinking learning objectives, which exist alongside the content outcomes but are not expressed within the instructional environment. In the mixed approach, critical thinking is taught as a separate track within a specific subject. In recent years there has been a revival to determine which of Ennis' types of instruction best develop critical thinking learning objectives.

General: According to Ennis (1989), the general approach attempts to teach critical thinking skills and dispositions in a separate course. The presentation of these objectives is separate from the content of any specific subject. The general approach usually involves some content but it is not required or specific to any particular discipline (Abrami et al., 2008). This approach suggests a universal set of critical thinking skills or dispositions that can be applied to all situations regardless of the discipline (Ennis, 1989).

A study of 38 fourth-and fifth-grade students in St. Louis public schools examined the effects of Ennis' *general* approach on the development of critical thinking skills (Riesenmy, Mitchell, Hudgins, & Ebel, 1991). This quasi-experiment examined whether students receiving training and practice on Hudgins and Edelman's (1988) four self-directed critical thinking roles could retain and transfer these skills on a problem-solving posttest better than students who were not given instruction or practice. The students in the treatment group had higher retention scores on the three variables that

were tested. The treatment group also demonstrated superior skills when answering the problems requiring the transfer of the critical thinking skills. This study supported the general approach demonstrating that schoolchildren given training in reasoning methods perform better than untrained children.

Infusion: The infusion approach to teaching critical thinking requires purposeful and well-understood subject matter instruction in which students are explicitly encouraged to think critically. It is important to note that general critical thinking principles are made explicit within the instruction (Ennis, 1989).

An example of critical thinking infusion instruction is provided by Zohar, Weinberger, and Tamir (1994) in the Biology Critical Thinking Project (BCT). The BCT project was designed to support seventh-grade biology students in Israel develop critical thinking skills. This study incorporated specific critical thinking activities into the biology curriculum. The objectives of the BCT project were to determine if critical thinking development would occur with an infusion instructional approach. The 678 seventh-grade students were randomly distributed into two groups. Both groups studied the same seventh-grade biology textbook but the treatment group experienced the BCT activities and the comparison group did not. The results indicated that the students in the treatment group improved their critical thinking skills compared to the students in the control group. Critical thinking skills improved in new areas of applied biological knowledge as well as other everyday topics, suggesting transfer of skills across domains. Another interesting component of the study revealed that students in the treatment group also scored significantly higher on a biological knowledge test. This study indicates that knowledge and critical thinking can be developed simultaneously using the infusion

method.

In another study, Barak, Ben-Chaim, and Zoller (2007) examined whether critical thinking skill development occurs when instruction of these skills is embedded within the framework of science education. This longitudinal pretest, posttest, post-posttest experimental study divided students in an Israeli high school into three groups. The experimental group consisting of 57 science majors was exposed to teaching strategies designed to enhance critical thinking skills. Two other control groups, one with 41 science majors and the other with 79 non-science majors, were taught traditionally and received no critical thinking instruction in their lessons. The treatment group significantly improved critical thinking skills and dispositions in comparison to the control groups on posttests of *The California Critical Thinking Disposition Inventory* (CCTDI) (Facione & Facione, 1992) test and *The California Critical Thinking Skills Test* (CCTST) (Facione & Facione, 1994). This study demonstrated that critical thinking skills transfer across domains as the CCTST tests generic critical thinking skills and is not specific to science. Barak, et al. (2007) also determined through qualitative interviews that teacher training on critical thinking instructional strategies must improve to meaningfully enhance this learning objective.

An experimental study in two high schools in Latvia examined the effectiveness of the infusion approach to critical thinking instruction within the context of foreign language education (Sokol, Oget, & Khomenko, 2008). A pretest/posttest design was employed to determine if 26 students receiving critical thinking instruction as part of the foreign language curriculum demonstrated significant increase in critical thinking skills in comparison to the 27 students in the control group. A number of factors including

difference in instructional time, previous exposure to critical thinking instruction, and teachers varied approaches, could not be controlled for in the study and thus brings into question the reliability of the study. The researchers are confident, however, that the results support critical thinking instruction imbedded into the curriculum as an instructional method to enhance the development of critical thinking and the transfer of these skills across multiple disciplines (Sokol, et al. 2008).

Immersion: In the immersion approach, teachers design subject matter to be thought-provoking and force students to think critically by immersing them in the subject matter. However, critical thinking learning objectives and principles are not explicit within the instruction of the course (Ennis, 1989).

Kamin, O'Sullivan, and Deterding (2002) tested the immersion method to develop critical thinking in 128 third-year medical students using digital video case simulations followed by group discussions. Their study sought to determine if critical thinking as demonstrated by discussion among students during group sessions differed by presentation format. One group of students viewed the cases on video and proceeded to discuss the cases online, the second group saw the videos and discussed the cases face-to-face, and the third group had face-to-face discussions after reading a text account of the case. Discussions were later transcribed and coded according to a predetermined critical thinking metric. The results showed critical thinking development was greatest with the groups that received the video presentation of the content, with the online discussion group scoring the highest. The authors speculated that the online format forced the students to formulate and articulate their ideas to a greater degree than the face-to-face discussions.

Mixed: The mixed approach combines either infusion or immersion with the general approach. This approach has a separate course aimed at teaching generic principles of critical thinking, but students are still exposed to critical thinking instruction in a subject specific setting (Ennis, 1989).

McCarthy-Tucker (1998) used the mixed critical thinking instructional strategy in a study of 190 students enrolled in a multicultural urban school. Pretests were given to all students prior to the treatment. The treatment group received instruction in formal logic as a supplement to their curricular instruction. Posttests revealed a significant increase in critical thinking for students who received supplemental instruction in logic. McCarthy-Tucker (1998) recommended formal logic instructional training for all teacher-training programs.

A study of the effect of a mixed instructional approach on fifth- and seventh-grade students in Newark, New Jersey examined the development of critical thinking skills in the subjects of language, math, and reading (Hartman-Haas, 1984). The experimental group was taught a supplemental thinking curriculum along with integrated skill development. A pretest/posttest design was used and demonstrated that the experimental group showed significant increases in critical thinking, language, and math ability in comparison to the control group.

Which Instructional Approach is the Most Effective?: A meta-analysis of empirical studies conducted from 1960 to 2005 examined the impact of instructional interventions affecting the development of critical thinking skills and dispositions (Abrami, et al., 2008). The researchers searched the term “critical thinking” in

commonly-used academic databases. This search produced 3,720 studies for consideration. These studies were evaluated against the following criteria to determine if they would be included in the study: "(a) accessibility – the study must be publicly available or archived; (b) relevancy – the study addresses the issue of CT development, improvement, and/or active use; (c) presence of intervention – the study presents some kind of instructional intervention; (d) comparison – the study compares outcomes that resulted from different types or levels of treatment (e.g., control group and experimental group, pretest and posttest, etc.); (e) quantitative data sufficiency – measures of relevant dependent variables are reported in a way that enables effect size extraction or estimation: (f) duration – the treatment in total lasted at least 3 hrs; and (g) age – participants were no younger than 6 years old" (p.1108). A total of 117 studies met all of the criteria (Abrami, et al., 2008).

Ennis's (1989) critical thinking typology of instructional approaches (general, infusion, immersion, and mixed) was used to classify the various instructional interventions in these 117 studies. The mixed method, where critical thinking is taught as an independent or separate component within the course of study, had the greatest positive effect on critical thinking development. The immersion method, where critical thinking is an implicit expectation of the course, had the smallest effect on critical thinking development. The general approach, where critical thinking skills are taught as the explicit objectives of the course, and the infusion approach, where critical thinking is explicitly stated as an objective but is embedded into the course content, had moderate effects on critical thinking development (Abrami, et al., 2008). Based on these studies, whether critical thinking is taught as a separate course or embedded within content

appears to have no significant difference. The authors concluded that the mixed method of instruction, where critical thinking objectives are taught separately and then applied as part of a specific course design, appears to have the greatest impact on critical thinking development.

The Abrami et al. (2008) study provides valuable insight into the most effective instructional methods for the development of critical thinking. However, of the 117 studies analyzed, only eight targeted high-school (16-18) age students (Abrami et al., 2008). The vast majority (80 of 117) of the studies examine undergraduate postsecondary students (Abrami, et al., 2008). A dearth of research on the development of critical thinking in high-school age children continues to be a problem.

Critical Thinking Development in the International Baccalaureate: Diploma Programme

Alec Peterson (2003), the first director general of the IB, stated that "the aim of general education is not the acquisition of general knowledge, but the development of the general powers of the mind to operate in a variety of ways of thinking" (p. 41). The Diploma Programme continues to be guided by this principle in the planning of curriculum and the methods of assessment (IBO, 2009a; IBO, 2009b).

The International Baccalaureate Diploma Programme is considered to be a demanding educational program designed for students aged 16 to 19 (IBO, 2009a; Mathews & Mathews, 2012; DiGiorgio, 2010). Diplomats and others involved in the reconstruction effort of post-World War II Europe required a common high-school

curriculum for their children. They needed a curriculum that would fulfill the entrance requirements of universities worldwide. As a result, in 1968 the IB Diploma Programme was established to provide international students with a well-rounded education, to facilitate mobility, and to promote multicultural understanding (IBO, 2009a). Seven schools participated in the original pilot IB examinations (IBO, 2012a). The IB has undergone noteworthy growth since those modest beginnings, and as of May 2011, 2,283 schools worldwide offered the Diploma Programme (IBO, 2011).

The IB Diploma Programme is a two-year programme that culminates in students sitting for internationally standardized exams. The programme requires students to select one course from each of six different subject areas: language A1 or the students' best language; a second language, focusing on spoken and written communication; individuals and societies; experimental sciences; mathematics and computer science; and the arts. This requirement is designed to have students gain a better understanding and knowledge base in all of the subject areas. However, the particular entrance requirements of many postsecondary institutions have influenced the IB Diploma Programme to allow students some flexibility. Students are able to forego a course in the arts and substitute another course from one of the other five subjects. In each of these six courses, students are instructed for 150 hours of core material. Students are also required to study three courses at a higher level, demanding an additional 90 hours of instructional time (IBO, 2009a).

The TOK course, another requirement of the DP, challenges students to question knowledge claims, and to appreciate differences in cultural experiences and beliefs. Diploma Programme candidates must also write a 4,000 word essay based on independent research. Finally, candidates are required to develop a portfolio during the

two years of the program through the Creativity, Action, and Service (CAS) requirement. CAS requires involvement by the students in creative pursuits, physical activities, and service projects. Students must meet all these requirements concurrently during two years of study (IBO, 2009c; IBO 2012a).

Theory of Knowledge (TOK). The TOK requirement was the focus of this study. In 1999, the IB Diploma Programme launched a new course, TOK, to unify the academic areas, develop critical thinking, and encourage appreciation of other cultural perspectives. The TOK requirement became central to the educational philosophy of the Diploma Programme. It was designed to challenge students and teachers to consider the role of knowledge in their own culture and belief system. It also encouraged participants to reflect critically on diverse ways of knowing and on others' areas of knowledge. It acts as a catalyst to students' awareness of themselves as thinkers, encouraging them to grapple with the complexity of knowledge and to ultimately act responsibly in an increasingly interconnected world (IBO, 2006).

The TOK course is promoted by the IB as the flagship of the Diploma Programme (IBO, 2006). The course encourages critical thinking in an attempt to help students make sense of what they encounter. In its attempt to promote considered inquiry into different ways of knowing the TOK course is composed almost entirely of questions. The most central of these questions is "How do we know?". Critical reflection is considered to be the foundation for developing awareness of international and multicultural issues. Reflection represents a necessary component for examination of our moral, political, and aesthetic judgments. Additionally, the role of language and thought and the development of students' critical thinking skills is a key component of the class (IBO, 2009b).

It is clear through an examination of the Aims and Objectives of the TOK course that critical thinking is a significant component (Appendix B). One of the aims is to "develop an awareness of how knowledge is constructed, critically examined, evaluated and renewed, by communities and individuals" (IBO, 2006, p. 5). The first stated objective of the course is that the student should be able to: "analyse critically knowledge claims, their underlying assumptions and their implications" (IBO, 2006, p. 5). In fact, almost all of the Aims and Objectives of the TOK course can be aligned with at least one of the six cognitive skills (interpretation, analysis, evaluation, inference, explanation, and self-regulation) identified by Facione (1990) and his panel of experts (see Table 1). This strongly links the TOK course to the identified cognitive skills of critical thinking used for this research.

Table 1

TOK Aims and Objectives: Cognitive Skills

The Aims of the TOK course are to:	Cognitive Skill
develop a fascination with the richness of knowledge as a human endeavor, and an understanding of the empowerment that follows from reflecting upon it	Self-Regulation
develop an awareness of how knowledge is constructed, critically examined, evaluated and renewed, by communities and individuals	Evaluation
encourage students to reflect on their experiences as learners, in everyday life and in the Diploma Programme, and to make connections between academic disciplines and between thoughts, feelings and actions	Self-Regulation
encourage an interest in the diversity of ways of thinking and ways of living of individuals and communities, and an awareness of personal ideological assumptions, including participants' own	Inference
encourage consideration of the responsibilities originating from the relationship between knowledge, the community and the individual as citizen of the world.	Self-Regulation
Having followed the TOK course, students should be able to:	
1. analyse critically knowledge claims, their underlying assumptions and their implications	Analysis Inference
2. generate questions, explanations, conjectures, hypotheses, alternative ideas and possible solutions in response to knowledge issues concerning areas of knowledge, ways of knowing and students' own experience as learners	Explanation Interpretation
3. demonstrate an understanding of different perspectives on knowledge issues	Interpretation Explanation
4. draw links and make effective comparisons between different approaches to knowledge issues that derive from areas of knowledge, ways of knowing, theoretical positions and cultural values	Inference
5. demonstrate an ability to give personal, self-aware response to a knowledge issue	Explanation Self-Regulation
6. formulate and communicate ideas clearly with due regard for accuracy and academic honesty.	Explanation Self-Regulation

A significant change to the 2006 edition of the TOK Guide recognizes that students must make links between TOK and their Diploma Programme subjects. It is necessary for DP subject teachers to dialogue with TOK teachers and purposefully plan to make TOK links within their classroom instruction. TOK teachers are also required to look for connections with their course and the other courses the students are studying. Diploma Programme subject curricula are evaluated on a seven-year cycle. As the new subject guides are reviewed, a significant component will be the recognition of the relationship between the subjects and the TOK course. Subject teachers are directed to make connections with TOK questions as they plan the delivery of their courses. Suggestions of theoretical concerns that arise in the subject classroom can be further explored in the TOK classroom (IBO, 2006).

The fact the TOK course is in existence would suggest the IB curricula designers agree with scholars such as Paul (1992), Vaughn (2008), and Wade and Tavris (1993) in the belief that critical thinking involves abilities and skills that are applicable to any problem or subject matter. However, the recent changes to the TOK guide suggest they realize the need for subject teachers to teach for transfer of specific critical thinking skills. A definite shift in design has occurred within the IB to ensure the subject teachers and the TOK teachers are working in conjunction to deliver the flagship course of the Diploma Programme, Theory of Knowledge. This design would be categorized in Ennis's (1989) instructional typology as the mixed approach.

The mixed approach has been identified by Abrami and his colleagues (Abrami et

al., 2008) as the most effective instructional technique to develop critical thinking in students. It is clear by the structure of the Diploma Programme and the requirements of the TOK course that the IB subscribes to the philosophy that a balanced, mixed approach to critical thinking instruction is best. However, even Abrami's study (Abrami et al., 2008) was only able to examine eight studies applicable to high school age children. There is a dearth of applicable empirical research surrounding this age of student. This study adds to the paucity of research on the development of critical thinking in high school age students.

Chapter 3

Methodology

Introduction

This study sought to establish if there is a difference in critical thinking skill development between students who have participated in a course designed to improve critical thinking, specifically the International Baccalaureate – Diploma Programme (IB-DP) Theory of Knowledge (TOK) course, and students who have no specific critical thinking education. This chapter will explain the study's purpose, research questions, choice of research design, instrumentation, population and sample, setting, procedure for data collection, and the data analysis utilized to answer the research questions.

Purpose of the Study

Despite the recognized importance of developing critical thinking skills, a level of uncertainty exists among curricula designers and practitioners as to how best accomplish this task (Geersten, 2003; Pithers & Soden, 2000). Debate persists among leading theorists as to the most effective instructional methods for developing critical thinking skills in students (Grauerholz & Bouma-Holtrop, 2003). A lack of empirical research on the proposed strategies remains a significant problem (Geersten, 2003; Grauerholz & Bouma-Holtrop, 2003; Pithers & Soden, 2000). This lack of research has caused curriculum developers to speculate as to the best way to teach these important skills

(Atkinson, 1997; Daly, 1998; Grauerholz & Bouma-Holtrop, 2003). The purpose of this study is to evaluate the impact of a specific course designed to teach generic critical thinking skills and add to the body of empirical research on the development of critical thinking skills in high school age students.

Research Questions

This study sought to answer the following questions:

1. Is there a significant difference between the growth of critical thinking skills, as demonstrated by the Test of Everyday Reasoning (TER)– Total Score, between students who participated in a course designed to improve critical thinking skills (the International Baccalaureate –Diploma Programme Theory of Knowledge course) and those who did not participate in the course?
    a. Is there a significant difference between the growth of targeted cognitive skills, as demonstrated in TER – Scale Scores (analysis and interpretation; evaluation and explanation; and inference), in students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the course?
    b. Is there a significant difference between the growth of targeted reasoning skills, as demonstrated in TER – Deductive and Inductive Reasoning Scale Scores, in students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the course?

Research Design

The design chosen for this study is a quantitative, causal comparative, non-experimental design using historical data. Causal comparative research analyzes for possible cause-and-effect relationships between dependent and independent variables (Fraenkel & Wallen, 2006). *Ex post facto*, from the Latin for "after the fact" describes situations that do not permit the randomization and manipulation of variables (Ary, Jacobs, Razavieh, & Sorenson, 2009). *Ex post facto* research is used "after variation in the variable of interest has already been determined in the natural course of events" (Ary, et al., 2009, p. 332). It is useful when the events or treatments studied already have occurred and therefore cannot be manipulated for the research (Cohen, Manion, & Morrison, 2007).

Ex post facto, causal comparative designs are often used in educational research to study retrospective cause-and-effect relationships (Cohen, et al., 2007). Individuals cannot be randomly assigned to treatment or control groups because they are already established prior to the beginning of the research (Gay, Mills, & Airasian, 2009). The two groups of students compared in the proposed study self-selected to enroll in the IB-DP or not.

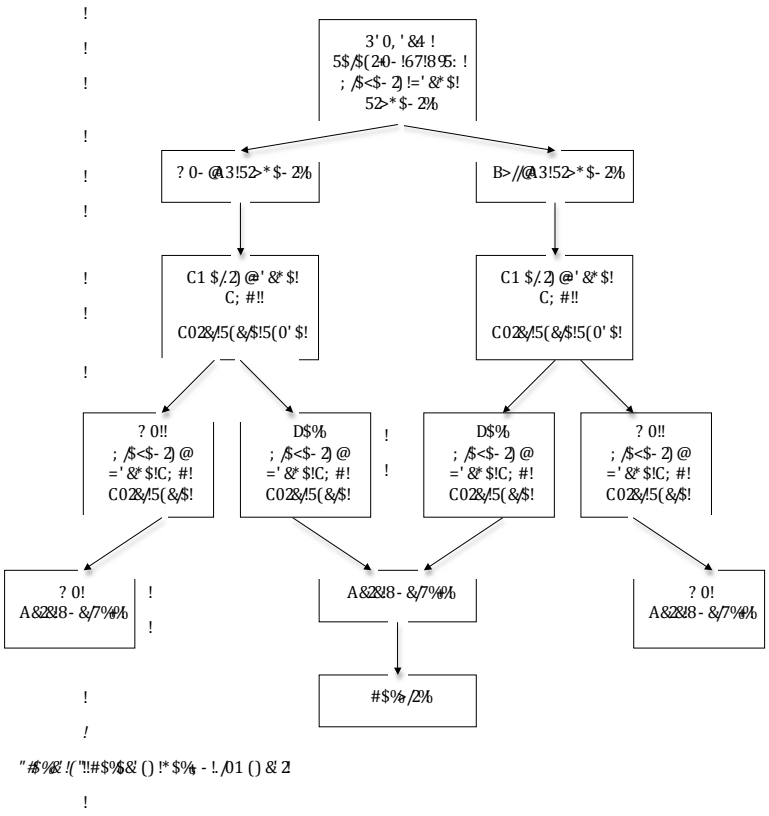
The data for this study are drawn from research collected by the American International School Kuwait (AISK) as part of its on-going assessment and evaluation process of the development of critical thinking. The larger school study collected data from two separate cohorts: Cohort 1 graduated from AISK in 2010 and Cohort 2 graduated in 2011. When these students were in eleventh-grade, they self-registered for

either the IB Diploma sequence of courses or to take IB courses but not be IB Diploma students.

One factor that influences students' selection into the IB-DP is the perception that is held by various nationalities if it will be of benefit to them. Most nationalities see the DP as an opportunity to be identified as a diligent, hard-working, and capable student by college admissions officers; however, this same perception does not exist for Kuwaiti students. The government of Kuwait has extremely lucrative scholarship programs available to national scholars. The Kuwait Investment Authority's (2012), a branch of the Kuwait government, Masters in Business Administration scholarship of approximately $270,000 USD for two years is an example of monies that are available to Kuwaiti citizens (see Appendix D). There are generally three types of scholarships available to high school graduates and a myriad of others available to post-secondary graduates. All scholarships consider GPA as one of the leading requirements for eligibility. The difficulty of the program or curriculum is not taken into consideration. Thus Kuwaiti students are indirectly encouraged to register for the easiest courses or program possible while still maintaining other eligibility requirements (Ray, personal communication October 3, 2012).

The TOK course is mandatory for IB-DP students and their participation forms the intervention or treatment portion of this study. Students not enrolled as IB-DP students cannot participate in the TOK class. As part of the assessment of critical thinking being conducted at AISK, the students in 11th-and 12th-grade were asked to complete the Test of Everyday Reasoning (TER). Both cohorts completed the TER pretest in the fall of 11th-grade and the same TER posttest in the spring of 12th-grade.

Figure 1 outlines the research design schema. The class of 2010 was the first class for which both a pretest and posttest were available for analyses.



Instrumentation

Test of Everyday Reasoning. AISK used the Test of Everyday Reasoning (TER), a version of the California Critical Thinking Skills Test (CCTST), for all students in grade eleven (pretest) and grade twelve (posttest). The TER is used by AISK to examine the development of critical thinking, an identified learning objective in need of development and measurement by AISK as part of its accreditation process with the Middle States Association (MSA) in 2008. Facione (1990) developed this instrument in response to a request from the American Philosophical Association (APA) to make a

systematic inquiry into the state of critical thinking. Facione used the Delphi method, involving forty-six "experts" in critical thinking instruction, to articulate a "clear and correct conceptualization of critical thinking" (Facione, 1990, p. 6). The CCTST is based on the expert panels consensus definition of critical thinking as the process of "purposeful, self-regulatory, reflective judgment" (Facione, 1990, p. 6).

The CCTST used the expert panel consensus definition to evaluate an item pool that had been developed in college-level testing projects for ten years. The items had been continually evaluated and refined for their ability to capture the reasoning process and to reveal common human reasoning errors in test takers (Facione, Facione, & Winterhalter, 2012). The validation studies of the first forms of the CCTST were conducted in college-level institutions in California. They were conducted using the case control methodology and lead to the first version of the CCTST. Since that time the item pool has been expanded significantly and has been used to support critical thinking tests for students ranging from grade three, to the doctoral level, to high-level professionals (Facione, et al, 2012). The TER, based on the CCTST, is intended for students in secondary school or in the first two years of post-secondary education (Facione, et al., 2012).

The TER is designed to be an intellectually challenging test, which addresses the cognitive and reasoning skills of students. The 35 item multiple-choice test targets the core cognitive and reasoning skills believed to be essential elements in the development of critical thinking. Each of the 35 questions is assigned to the cognitive skill categories: analysis and interpretation, evaluation and explanation, or inference. Additionally the items are reassigned to the areas of inductive and deductive reasoning (Facione, et al.,

2012).

The reliability of the TER was computed using Kuder-Richardson internal reliability coefficients. The Kuder-Richardson -20 (KR-20) function as the comparable measurement to Cronbach's alpha used for dichotomously scored instruments (Facione, et al., 2012). A KR-20 above .70 denotes a high level of internal consistency for an instrument with multi-dimensional scales (Facione, et al., 2012). The current aggregated KR-20 estimates of the TER, ranging from .71 - .86, demonstrate sufficient evidence of the internal consistency of these instrument's ability to measure the overall conceptual construct: critical thinking (Facione, et al., 2012).

The TER is designed to involve the test-taker's reasoning skills using recognizable issues and contexts. Test-takers must use cognitive skills to evaluate, analyze, and interpret information presented in a variety of forms. They must also analyze inferences and determine if they represent strong or weak reasoning and then explain why a given evaluation of an inference is strong or weak. No specialized content knowledge is required as information needed to answer questions correctly is provided within the questions themselves. The paper-and-pencil test requires 50 minutes to complete and the Flesch-Kincaid Readability Level of the TER is equivalent to 6th-grade (Facione, et al., 2012).

Population and Sample

The participants in this study were chosen from students enrolled as 11th-graders in 2008 and 2009 at the American International School Kuwait (AISK). AISK is

accredited by the Middle States Association of Colleges and Schools (MSA) and authorized by the International Baccalaureate Organization to offer the Primary Years Programme (PYP) for pre-kindergarten to grade five, the Middle Years Programme (MYP) for grades six to ten, and the Diploma Programme (DP) for grades 11 and 12.

Two cohorts were examined for this study. Cohort 1 students were in 11th-grade during the 2008 – 2009 school year and graduated in 2010. Cohort 2 students were in 11th-grade during the 2009 – 2010 school year and graduated in 2011. Cohort 1 had 71 students in the fall of 2008. Of the 71 students enrolled in 11th-grade, 67 completed the Test of Everyday Reasoning (TER) in October 2008 and again in April 2010, at the end of their senior year. Of the 67 students that completed both tests of the TER, 24 were IB-DP students. These students were grouped according to full enrollment in the IB-Diploma Programme. There are 24 students in the treatment group of IB-DP students and 43 students in the comparison group of non-IB-DP students.

Cohort 2 had 106 students in the fall of 2009. Of these106 students, 104 completed the Test of Everyday Reasoning (TER) in October 2009 and again in April 2011, at the end of their senior year. Of the 104 students that completed both tests of the TER, 38 were IB-DP students. These students were grouped according to enrollment in the IB-Diploma Programme. There are 38 students in the treatment group of IB-DP students and 66 students in the comparison group of non-IB-DP students.

Generalizability. AISK is typical of schools authorized to offer the IB programmes. The International Baccalaureate has authorization standards that are strictly adhered to and monitored. An essential feature of the IB is that standards for schools are

similar world-wide. The IB employs three primary components to ensure standardization among schools delivering the IB-DP. First of all, any school that offers the IB-DP must be authorized to do so (IBO, 2007; IBO, 2010). The process of becoming authorized to offer the DP is the same for all schools. This procedure is designed to ensure all schools are well prepared and able to successfully implement the programme (IBO, 2007). Schools wishing to offer the DP must work through two key stages, the feasibility study and an authorization visit from an IB visiting team. Furthermore, the IB requires all authorized schools to complete a self-study process. This process is designed to reinforce appropriate support and commitment from the community for the programme.

Secondly, the IB offers extensive professional development for teachers working in authorized schools. In 2006, over 30,000 teachers and administrators were trained at IB sponsored workshops and events around the world (IBO, 2007). These workshops encourage IB teachers to reflect upon and improve their practice. Schools seeking authorization must meet mandatory levels of IB authorized training to continue the process.

The final essential element of standardization for the DP is student assessment. The IB assessment standards are the same for all students world-wide. Student performance is measured according to standards and criteria that are consistent from school-to-school and year-to-year (IBO, 2007). Final examinations occur two times a year, in May (for schools in the northern hemisphere) and November (for those in the southern hemisphere). International teams of examiners, who are trained and monitored by the IB, assess the students' work (IBO, 2007).

As a result, in terms of curriculum and pedagogy, AISK is similar to other schools authorized to offer the IB-Diploma Programme. Therefore, the results of this study should be generalizable to the 2,283 schools worldwide offering the Diploma Programme as of May, 2011 (IBO, 2011).

Procedure

The superintendent of AISK granted permission to proceed with the study. This permission allowed specific data to be accessed in order to analyze the graduating classes of 2010 and 2011. School personnel searched school records to retrieve specific data regarding sex, nationality, IB-DP status, grade-point average, PSAT results, and TER results. These data were given to the researcher using an unidentifiable unique identification number.

AISK began testing the critical thinking skills of its 11th-and 12th-grade students in October of 2008 using the Test of Everyday Reasoning (TER), a version of the California Critical Thinking Skills Test (CCTST). All 11th-and 12th-grade students were given the opportunity to take the TER in the fall of their 11th-grade year and the spring of their senior year.

The paper-and-pencil Test of Everyday Reasoning was administered according to the instructions contained in the *Test Manual: Test of Everyday Reasoning* (Facione, et al., 2012) by the Curriculum Coordinator at AISK. The paper-and-pencil test was chosen due to internet instability in Kuwait. This test enabled the majority of the students to take the test during the same testing period. Students were given 50 minutes to complete the

test. Students used computer-generated Unique Identification Numbers as identifiers on the CapScore answer sheet to ensure no two duplicate numbers were used on the answer sheet.

Data Analysis

The data were organized categorically using *Microsoft Excel 2007.* After reviewing for data entry accuracy, all data were imported into the Statistical Package of Social Sciences (SPSS) Base 19.0 computer program for statistical analysis. SPSS is a comprehensive data analysis system that can take data and complete complex statistical analyses as well as descriptive statistics (Weinberg & Abramowitz, 2008).

Students in both graduating classes were divided into two groups for the purposes of analysis: IB Diploma candidate students and non-IB Diploma students. TER results were collected in October of Grade 11 and April of Grade 12 for all students. The TER results used in the study include: Total score; Cognitive scale scores: Analysis and Interpretation scale score, Inference scale score, Evaluation and Explanation scale score; and Reasoning scale scores: Inductive Reasoning scale score, and Deductive Reasoning scale score.

Individual student TER gain scores were calculated by comparing the initial TER scores obtained in the fall of the students' junior year to those obtained in the spring of their senior year. The gain scores for each pair of data/testing points were aggregated for the IB and non-IB groups, and simple descriptive statistics (N, mean gain score, standard deviation, median gain score) were calculated and reported.

Mean gain scores for each of the two groups (IB and non-IB) underwent further analysis to test the following hypotheses:

$H_0$1. There is no significant difference between the growth of critical thinking skills, as demonstrated by Test of Everyday Reasoning (TER)–Total Score, of students who participated in the International Baccalaureate–Diploma Programme (IB-DP) Theory of Knowledge (TOK) course and those who did not participate in the course.

Dependent Variable: The dependent variable identified to test $H_0$1 is an estimate of the systematic gain on TER – Total score from pre-treatment to post-treatment.

Independent Variable: The independent variable identified to test $H_0$1 is group and it has two attributes: treatment and comparison.

$H_0$1a. There is no significant difference between the growth of targeted cognitive skills, as demonstrated in TER - Scale Scores (analysis and interpretation; evaluation and explanation; and inference), of students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the course.

Dependent Variable: The dependent variable identified to test $H_0$1a is an estimate of the systematic gain on TER – Scale Scores (analysis and interpretation; evaluation and explanation; and inference) from pre-treatment to post-treatment.

Independent Variable: The independent variable identified to test $H_0$1a is group and it has two attributes: treatment and comparison.

$H_0$1b. There is no significant difference between the growth of targeted reasoning skills, as demonstrated in TER – Deductive and Inductive Reasoning Scale Scores, of students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the course.

Dependent Variable: The dependent variable identified to test $H_0$1b is an estimate of the systematic gain on TER – Deductive and Inductive Reasoning Scale Scores from pre-treatment to post-treatment.

Independent Variable: The independent variable identified to test $H_0$1b is group and it has two attributes: treatment and comparison.

This analysis used univariate analysis of covariance (ANCOVA) to determine whether obtained differences of the mean gain of the TER – Total score are statistically significant. An analysis of covariance, or ANCOVA, statistically adjusts or equalizes the initial differences between groups when participants are not randomly assigned to treatment groups (Weinberg & Abramowitz, 2008). ANCOVA adjusts the dependent variable (estimate of the systematic gain on TER) for differences on an independent variable (group: treatment and comparison) to be controlled (Gay, Mills, & Airasian, 2009).

ANCOVA is most effective in predicting the value of a dependent variable when the independent variable being studied has a strong correlation to the dependent variable but does not interact with other independent variables (Gay et al., 2009). ANCOVA neutralizes the effect of the more powerful, non-interacting variable. The effects of the interacting independent variables can become ambiguous if this measure is not utilized.

Accordingly, when the data include independent variables that are designed to be independent of each other, it is better to use ANOVA or ANCOVA than multiple regression. A researcher is able to analyze more complex relationships of the data using ANOVA or ANCOVA as an analytical tool. Most importantly ANCOVA allows the researcher to address questions about interactions. ANCOVA is differentiated from other techniques in that it is more effective in neutralizing the effect of a continuous independent variable on the experiment (Gay et al., 2009). ANCOVA also allows for the equating of non-equivalent (intact) groups.

In addition to the data analysis described above, the research project involved a multivariate analysis of covariance (MANCOVA) to determine whether obtained differences of the mean gain of cognitive scale scores as demonstrated by the TER – Analysis and Interpretation, Evaluation and Explanation, and Inference scale scores exhibit significant differences between the treatment group and the comparison group. The MANCOVA analyzed the differences of the mean gain of the reasoning scale scores as demonstrated by the TER – Deductive and Inductive Reasoning scale scores to determine if there was a significant difference between students who participated in the treatment and those who did not.

Limitations

Ex post facto designs possess inherent threats to research validity. They do not allow the researcher to control, manipulate, or isolate any variables or allow for random assignment to treatment (Fraenkel & Wallen, 2006). Analytical tools were used to limit

the effects of intervening variables; however, the complexity of the variables limit this study to establishing whether or not significant differences exist. If significant differences do indeed exist, it will be beyond the scope of this study to assign causality to participation in the Theory of Knowledge course.

Summary

This study utilized data collected from two successive cohorts of students entering their 11th-grade at American International School Kuwait from 2009 until 2010. This non-experimental study employs a pretest/posttest design to establish the effects of a two-year course of study designed to develop critical thinking as measured by the Test of Everyday Reasoning. Data was analyzed using Statistical Packages for the Social Sciences (SPSS) software, version 19.0, using ANCOVA and MANCOVA. These analyses determined if there is a significant difference in the development of critical thinking skills between students who participated in a course designed to improve critical thinking skills (the International Baccalaureate –Diploma Programme Theory of Knowledge course) and those who did not participate in the course.

CHAPTER 4

Results and Analysis

The purpose of this study is to determine if any statistically significant difference exists when comparing the growth of critical thinking between students enrolled in a specific course designed to teach generic critical thinking skills and students not enrolled in the course. The specific questions the study sought to answer were:

1. Is there a significant difference between the growth of critical thinking skills, as demonstrated by Test of Everyday Reasoning (TER) – Total Score, between students who participated in a course designed to improve critical thinking skills (the International Baccalaureate – Diploma Programme Theory of Knowledge course) and those who did not participate in the course?
    a. Is there a significant difference between the growth of targeted cognitive skills, as demonstrated in TER – Scale Scores (analysis and interpretation; evaluation and explanation; and inference), between students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the course?
    b. Is there a significant difference between the growth of targeted reasoning skills, as demonstrated in TER – Deductive and Inductive Reasoning Scale Scores, between students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the course?

The results of the tests used to examine the data are detailed here. The purpose of the statistical analysis was to determine if any statistically significant differences emerged. Finally, a summary of their contribution to providing answers to the research questions is provided.

This study examined 171 high school students, from two cohorts, who were administered the TER at AISK in the fall of eleventh-grade and the spring of twelfth-grade. The Cohort 1 subsample, was comprised of 67 AISK students who graduated in 2010. This group represented 39.2 percent of the total sample of 171 students. The Cohort 2 subsample was comprised of 104 AISK students who graduated one year later than Cohort 1, in 2011, and represented 60.8 percent of the total sample. Table 2 represents the frequency and percentage of students for each cohort.

Table 2

Frequency and Percent of Students in Each Cohort

Cohort	Frequency	Percent	Valid Percent	Cumulative Percent
1	67	39.2	39.2	39.2
2	104	60.8	60.8	100.0
Total	171	100.0	100.0	

Descriptive Analysis of Comparison and Treatment Groups

The analysis of these data began with an evaluation of the descriptive statistics of the treatment and comparison groups. The comparison group was comprised of all AISK

students in the graduation classes of 2010 and 2011, with a recorded score for the TER scale scores in 11th-and 12th-grade, but who were *not* enrolled in the IB-DP. The treatment group was comprised of all AISK students in the classes of 2010 and 2011, with a recorded score for the TER scale scores in 11th-and 12th-grade, and who *were* enrolled in the IB-DP.

The comparison group had a cumulative total sample over the two cohorts of 109 students, representing 43 students in Cohort 1 and 66 students in Cohort 2. The treatment group had a cumulative total sample over the two cohorts of 62 students, representing 24 students in Cohort 1 and 38 students in Cohort 2. Table 3 shows the frequency and percent of students in the treatment and comparison groups.

Table 3

Frequency and Percent of Students in the Treatment and Comparison Groups

	Comparison		Treatment		Total	
	n	*%*	*n*	*%*	*n*	*%*
Cohort 1	43	39.4	24	38.7	67	39.2
Cohort 2	66	60.6	38	61.3	104	60.8
Total	109	63.7	62	36.3	171	100.0

It is necessary to determine if certain demographic variables in the dataset need to be examined as alternate explanations of the effect that the treatment may have on the outcome. These alternate explanations are referred to as third variable explanations of the

hypothetical treatment effect. The demographic variables that may effect the outcome of the treatment are examined further below.

Comparison of Sex

The number of male and female students in each group, comparison and treatment, and the percentage of each group are noted in Table 4. The table reflects the number and percentage of each sex of students organized by cohort and the number and percentage of each sex in the combined total. The demographic data showed that there were 13.4% more males than females in the study; however, there is little difference (3.2% more males) between each sex in the treatment group.

Table 4

Proportion of Students in the Treatment and Comparison Groups by Sex

	Comparison		Treatment		Total	
	n	*%*	*n*	*%*	*n*	*%*
Male	65	59.6	32	51.6	97	56.7
Female	44	40.4	30	48.4	74	43.3
Total	109	100.0	62	100.0	171	100.0

Comparison of Ethnicity

The ethnic designations chosen for this study were Kuwaiti, Other Middle Eastern, American/Canadian, European, Other. The number and percentage of students in each of

these ethnic designations are given in Table 5. The number and percentage of each group, comparison and treatment, are given in Table 6. There are 93 Kuwaiti students involved in the study representing 54.4% of the total population. An interesting finding is that only 17.2% of Kuwaiti students chose to enter the IB-DP or treatment. The other ethnic designations all have over 57% of their students selecting the IB-DP.

Table 5

Number and Percentages by Nationality

	Kuwaiti		Other Middle Eastern		American/Canadian		European		Other		Total	
	n	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Total	93	54.4	35	20.5	29	16.9	3	1.8	11	6.4	171	100.0

Table 6

Group Proportions by Nationality

	Kuwaiti		Other Middle Eastern		American/Canadian		European		Other		Total	
	n	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Treatment	16	17.2	20	57.1	17	58.6	2	66.7	7	63.6	62	36.3
Comparison	77	82.8	15	42.9	12	41.4	1	33.3	4	36.3	109	63.7
Total	93	100.0	35	100.0	29	100.0	3	100.0	11	100.0	171	100.0

Comparison of Sex in Each Ethnic Grouping

The number of female and male students in each ethnic designation and the percentage of each group, comparison and treatment, are given in Table 7. There are a higher percentage of males to females throughout all of the ethnic groups with the exception of European; however, the limited sample number in the European ethnic grouping limits the generalizability of those particular findings. One other important finding is the large proportion of males in the Other ethnic grouping (approximately, 72%) as contrasted with approximately 57% of males in the sample as a whole.

Another important ethnic/sex difference among the groups is the lower proportion of Kuwaiti males, or 37.5%, in the treatment group. In comparison, the American/Canadian ethnic grouping has the next lowest proportion of males in the treatment group at 52.5%.

The proportion of males to females distinguished by nationality needs to be explored as a possible third variable explanation as the difference between the two groups is noteworthy. The percentage of males in the Kuwaiti group is considerably less than that of the other nationality groupings and needs to be examined.

Table 7

Group and Sex Proportions by Nationality

	Kuwaiti		Other Middle Eastern		American/Canadian		European		Other		Total	
	n	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Treatment												
Male	6	37.5	11	55	9	52.5	0	0	6	85.7	32	51.6
Female	10	62.5	9	45	8	47.1	2	100	1	14.3	30	48.8
Comparison												
Male	46	59.7	9	60	7	58.3	1	100	2	50	65	59.6
Female	31	40.3	6	40	5	41.7	0	0	2	50	44	40.4
Total												
Male	52	55.9	20	57.1	16	55.2	1	33.3	8	72.37	97	56.7
Female	41	44.1	15	42.9	13	44.8	2	66.7	3	27.3	74	43.3

The difference in academic ability between the two groups needs to be studied to determine if it could be a possible third variable explanation. Preliminary SAT results, Grade Point Averages, pretest TER scores, and posttest TER scores were identified as measures of academic ability and were subsequently examined.

Comparison of Preliminary SAT (PSAT) Scores

One measure of prior academic ability for this study was the PSAT. This standardized test of academic ability was chosen as all students in this study wrote the PSAT in the fall of the eleventh-grade prior to the start of the TOK class, or treatment. The corresponding means, medians, minimum scores, maximum scores, and standard deviations for the comparison and treatment groups are displayed in Table 8.

On average, the treatment group scored higher in PSAT Critical Reading (+9.24), Math (+8.74), Writing (+8.39), and Total score (+26.37) than the comparison group.

Table 8

Means, Medians, Minimums, Maximums and Standard Deviations on PSAT Scores for Each Group

Group		PSAT Critical Reading	PSAT Math	PSAT Writing	PSAT Total
Comparison	Mean	39.26	40.63	40.21	120.10
	Median	40.00	41.00	39.00	119.00
	Min	20	20	20	68
	Max	67	57	68	181
	SD	8.35	7.31	9.17	21.03
	N	94	94	94	94
Treatment	Mean	48.50	49.37	48.60	146.47
	Median	49.00	49.00	49.00	149.00
	Min	27	37	35	109
	Max	70	67	63	196
	SD	8.30	7.83	6.75	19.29
	N	60	60	60	60
Total	Mean	42.86	44.03	43.48	130.37
	Median	43.00	44.00	43.00	129.00
	Min	20	20	20	68
	Max	70	67	68	196
	SD	9.45	8.63	9.24	24.06
	N	154	154	154	154

Comparison of Grade Point Average (GPA) Scores

GPAs were another academic ability data point used to evaluate students' academic ability. The corresponding means, medians, minimum scores, maximum scores, and standard deviations for the comparison and treatment groups are displayed in Table 9. It should be noted that there is a relatively small sample of ninth-graders for whom GPAs are available.

The treatment group exhibited higher GPAs, on average, than the comparison group every year from grade 9 to grade 12 (+.78, +.52, +.49, +.24, respectively). It is important to note that the difference in GPA was reduced every year from grade 9 to grade 12.

Table 9

Means, Medians, Minimums, Maximums and Standard Deviations for GPA for Grades 9 through 12 for each Group

		GPA Grade 9	GPA Grade 10	GPA Grade 11	GPA Grade 12
Comparison	Mean	2.60	2.92	2.79	2.97
	Median	2.63	2.89	2.78	3.00
	Min	1.58	1.53	1.43	1.02
	Max	3.86	4.00	4.00	4.00
	SD	.54	.52	.64	.59
	N	33	97	105	109
Treatment	Mean	3.38	3.44	3.28	3.21
	Median	3.39	3.53	3.39	3.30
	Min	2.38	2.50	1.93	2.10
	Max	3.89	3.98	3.96	3.96
	SD	.36	.36	.51	.49
	N	17	50	57	62
Total	Mean	2.87	3.09	2.96	3.06
	Median	2.85	3.16	2.99	3.09
	Min	1.58	1.53	1.43	1.02
	Max	3.89	4.00	4.00	4.00
	SD	.61	.53	.64	.56
	N	50	147	162	171

Comparison of TER Pretest Scores

Students' TER pretest scores were compared in order to evaluate whether the students in the comparison and treatment groups were equivalent prior to treatment. The TER pretest is administered to all students in the fall of the eleventh-grade prior to the start of the treatment, the TOK class. The corresponding means, medians, minimum scores, maximum scores, and standard deviations for the comparison and treatment groups are displayed in Table 10.

The treatment group outscored the comparison group on the TER Total score on average by +4.16. The treatment group also scored higher on average on the TER cognitive skills of Analysis (+1.03), Inference (+1.53), and Evaluation (+1.41). Similarly, the treatment group scored higher on average in the areas of Deductive and Inductive reasoning on the TER by +2.28 and +1.77 respectively.

Table 10

Means, Medians, Minimum Scores, Maximum Scores and Standard Deviations for TER Pretest Scores for Each Group

		Total pretest	Analysis pretest	Inference pretest	Evaluation pretest	Deduction pretest	Induction pretest
Comparison	Mean	18.14	5.05	8.24	5.01	9.82	8.43
	Median	18.00	5.00	8.00	5.00	10.00	8.00
	Min	6	1	3	1	1	2
	Max	32	8	15	11	21	15
	SD	5.03	1.64	2.60	2.23	3.28	2.68
	N	100	99	99	99	100	100
Treatment	Mean	22.30	6.08	9.77	6.45	12.10	10.20
	Median	22.50	6.00	10.00	6.00	12.00	10.50
	Min	12	3	5	2	5	6
	Max	32	9	14	11	17	16
	SD	4.53	1.49	2.32	2.11	2.89	2.58
	N	60	60	60	60	60	60
Total	Mean	19.70	5.44	8.82	5.55	10.68	9.09
	Median	19.00	5.00	9.00	5.00	11.00	9.00
	Min	6	1	3	1	1	2
	Max	32	9	15	11	21	16
	SD	5.24	1.66	2.59	2.29	3.32	2.77
	N	160	159	159	159	160	160

Comparison of TER Posttest Scores

In order to evaluate whether the students in the comparison and treatment groups were equivalent, students' TER posttest scores were compared. The TER posttest is administered to all students in the spring of the twelfth-grade after the completion of the treatment, the TOK class. The corresponding means, medians, minimum scores, maximum scores, and standard deviations for the comparison and treatment groups are displayed in Table 11.

The treatment group outscored the comparison group on the posttest TER Total score on average by +6.76. The treatment group also scored higher on average on the TER cognitive skills of Analysis (+1.82), Inference (+2.91), and Evaluation (+1.94). Similarly, the treatment group scored higher on average in the areas of Deductive and Inductive reasoning on the TER by +3.82 and +2.92 respectively.

Comparison of Change from TER Pretest to Posttest

The comparison group increased in the TER Total Score as well as the Cognitive Skills and Reasoning scale scores (Total = +.29, Analysis = +.02, Inference = +.05, Evaluation = +.66, Deduction = +.20, Induction = +.50); however, the treatment group improved more than the comparison group on all TER scores (Total = +.3.39, Analysis = +.81, Inference = +1.43, Evaluation = +1.16, Deduction = +1.74, Induction = +1.65). These scores represent an increase by the treatment group over the comparison group in the average score from pretest to posttest by +2.60 on the TER-Total score, +.79 on the TER Analysis scale score, +1.38 on the TER Inference scale score, +.53 on the TER

Evaluation scale score, +1.54 on the TER-Deductive Reasoning scale score, and +1.15 on the TER Inductive Reasoning scale score.

The academic ability of the treatment group was noticeably higher than the comparison group in all four of the measures examined. The treatment group outperformed the comparison group in PSAT scores, GPAs, TER pretest scores, and TER posttest scores.

Three different teachers taught the TOK class to the IB-DP students causing a varying treatment experience between the students. Therefore, the teacher giving the treatment to the students needs to be explored to determine if it is a possible third variable explanation of the effect of treatment.

Table 11

Means, Medians, Minimums, Maximums and Standard Deviations on TER Posttest Scores for Each Group

		Total posttest	Analysis posttest	Inference posttest	Evaluation posttest	Deduction posttest	Induction posttest
Comparison	Mean	18.93	5.07	8.29	5.67	10.02	8.93
	Median	19.50	5.00	8.00	6.00	10.50	9.00
	Min	5	1	2	1	1	2
	Max	34	9	15	11	18	16
	SD	7.10	2.20	3.07	2.64	4.31	3.30
	N	104	102	104	104	104	104
Treatment	Mean	25.69	6.89	11.20	7.61	13.84	11.85
	Median	26.00	7.00	11.00	8.00	15.00	12.00
	Min	7	1	3	1	2	5
	Max	35	9	15	11	19	16
	SD	5.80	1.80	2.77	2.33	3.67	2.71
	N	61	61	61	61	61	61
Total	Mean	21.43	5.75	9.36	6.39	11.43	10.01
	Median	22.00	6.00	10.00	7.00	12.00	11.00
	Min	5	1	2	1	1	2
	Max	35	9	15	11	19	16
	SD	7.39	2.24	3.27	2.69	4.48	3.39
	N	165	163	165	165	165	165

Comparison of Treatment Group By Teacher

The number and percent of treatment group students in each of three teachers' classrooms is represented in Table 12. There were a total of 62 students receiving treatment. Teacher A taught 26, or 41.9 percent, of students in the treatment group; Teacher B taught 22, or 35.5 percent of students in the treatment group; and Teacher C taught 14 or 22.6 percent of the students in the treatment group.

Table 12

Frequency and Percent of Students in the Treatment Group by Teacher and in the Comparison Group

	Frequency	Percent
Teacher A	26	41.9
Teacher B	22	35.5
Teacher C	14	22.6
Treatment Group Total	62	100

Correlational Tests of Demographic Variables

In order to determine if particular demographic variables in the dataset would need to be examined as potential third variable explanations of a hypothetical treatment effect, a series of correlational tests were performed.

Cohort: Table 13 shows that none of the TER mean pretest or posttest scores significantly differed between the two cohorts. Thus cohort is not a candidate for a third variable explanation of a hypothetical treatment effect and does not need further analysis.

Table 13

ANOVA Test for the Effects of Cohort on the TER Pretest and Posttest Scores

		Sum of Squares	df	Mean Square	F	Sig.
Analysis pretest	Between Groups	6.361	1	6.361	2.329	.129
	Within Groups	428.822	157	2.731		
	Total	435.182	158			
Inference pretest	Between Groups	.946	1	.946	.140	.709
	Within Groups	1062.765	157	6.769		
	Total	1063.711	158			
Evaluation pretest	Between Groups	10.410	1	10.410	1.996	.160
	Within Groups	818.886	157	5.216		
	Total	829.296	158			
Deduction pretest	Between Groups	1.927	1	1.927	.174	.677
	Within Groups	1749.173	158	11.071		
	Total	1751.100	159			
Induction pretest	Between Groups	14.884	1	14.884	1.955	.164
	Within Groups	1202.710	158	7.612		
	Total	1217.594	159			
Analysis posttest	Between Groups	.687	1	.687	.137	.712
	Within Groups	810.000	161	5.031		
	Total	810.687	162			
Inference posttest	Between Groups	.897	1	.897	.083	.773
	Within Groups	1757.285	163	10.781		
	Total	1758.182	164			

Evaluation posttest	Between Groups	.063	1	.063	.009	.926
	Within Groups	1185.113	163	7.271		
	Total	1185.176	164			
Deduction posttest	Between Groups	.614	1	.614	.030	.862
	Within Groups	3283.834	163	20.146		
	Total	3284.448	164			
Induction Posttest	Between Groups	1.972	1	1.972	.171	.680
	Within Groups	1884.004	163	11.558		
	Total	1885.976	164			

Sex: The ANOVA test for difference of sex shows that one of the TER scale scores, pretest inferential reasoning, differs for males and females. The results are shown in Table 14.

Further analysis, shown in Table 15, revealed that males scored higher on the pretest TER inference scale score. Table 15 shows the further analysis on the inference pretest scores for males and females. This study is not specifically designed to study the effects of sex; however, the fact that it is correlated with the critical thinking cognitive skill - inference suggests that one model of analysis should include sex as a covariate. This analysis eliminates the effect of sex and allows the independent variable of group, treatment or comparison, to be more accurately analyzed.

Table 14

ANOVA Test of Mean Sex Differences on TER Mean Scores

		Sum of Squares	df	Mean Square	F	p
Analysis pretest	Between Groups	4.70	1	4.70	1.72	.19
	Within Groups	430.48	157	2.74		
	Total	435.18	158			
Inference pretest	**Between Groups**	**66.51**	**1**	**66.51**	**10.47**	**.001**
	Within Groups	997.20	157	6.35		
	Total	1063.71	158			
Evaluation pretest	Between Groups	.22	1	.22	.04	.84
	Within Groups	829.08	157	5.28		
	Total	829.30	158			
Deduction pretest	Between Groups	11.57	1	11.57	1.05	.31
	Within Groups	1739.53	158	11.01		
	Total	1751.10	159			
Induction pretest	Between Groups	.04	1	.04	.01	.94
	Within Groups	1217.55	158	7.71		
	Total	1217.59	159			
Analysis posttest	Between Groups	3.36	1	3.36	.67	.41
	Within Groups	807.33	161	5.01		
	Total	810.69	162			
Inference posttest	Between Groups	.51	1	.51	.05	.83
	Within Groups	1757.67	163	10.78		
	Total	1758.18	164			

Evaluation posttest	Between Groups	.69	1	.69	.10	.76
	Within Groups	1184.48	163	7.27		
	Total	1185.18	164			
Deduction posttest	Between Groups	21.25	1	21.25	1.06	.30
	Within Groups	3263.20	163	20.02		
	Total	3284.45	164			
Induction Posttest	Between Groups	4.89	1	4.89	.42	.52
	Within Groups	1881.08	163	11.54	1.72	
	Total	1885.98	164			

Table 15

Mean Scores on Inference Pretest for Males and Females

	N	Mean	SD
Female	67	8.06	2.58
Male	92	9.37	2.48
Total	159	8.82	2.59

GPA: Only the 9th-and 10th-grade GPAs were considered to be legitimate possible covariates as these were assessed prior to the treatment, the TOK course. The 11th-and 12th-grade GPAs were assessed after the start of the TOK course; therefore, eliminating these scores as legitimate covariates, third variable explanations, for any treatment effects that may be discovered. The sample size for 9th-grade GPA is relatively small and may reduce the usefulness of this variable.

Table 16 shows that both 9th-and 10th-grade GPAs were significantly correlated with all five TER scale scores. Thus, it will be necessary to consider a series of models that include the grade 9 and 10 GPA scores to control for the effects of these possible third variable explanations of the hypothetical treatment effect. It is necessary to control for initial dependent variable, or group, differences on GPA scores by examining these scores as covariates in the analyses. If it is determined that the effect for group is a nonfactor when we control for the effects of GPA, the data will become consistent with a spurious effect. Any significant growth or difference between the two groups could be rationalized as one group being more capable than the other as measured by GPA performance.

Table 16

Correlations Between GPAs and TER Scales at Posttest

		Analysis	Inference	Evaluation	Deduction	Induction
GPA Grade 9	Pearson Correlation	.336*	.478**	.518**	.437**	.500**
	Sig. (2-tailed)	.021	.001	.000	.002	.000
	N	47	47	47	47	47
GPA Grade 10	Pearson Correlation	.323**	.327**	.204*	.284**	.299**
	Sig. (2-tailed)	.000	.000	.015	.001	.000
	N	140	142	142	142	142

** Correlation is significant at $p < 0.01$ level (2-tailed).

* Correlation is significant at $p < 0.05$ level (2-tailed).

PSAT: Table 17 reveals that all five of the scale scores on the TER test are significantly correlated with the three PSAT scores. Therefore, it will be necessary to run a series of models that include each of the PSAT scores. These models will control for initial group differences on PSAT scores and analyze these scores as covariates. If the effect for group remains constant when the effects of PSAT are controlled, PSAT scores can be disregarded as a third variable explanation of the hypothesized treatment effect.

Table 17

Correlations Between PSAT Scales and TER Scales at Posttest

		Analysis	Inference	Evaluation	Deduction	Induction
PSAT	*r*	.656**	.587**	.605**	.631**	.642**
Critical	*p*	.000	.000	.000	.000	.000
Reading	*N*	148	150	150	150	150
PSAT	*r*	.532**	.595**	.488**	.586**	.543**
Math	*p*	.000	.000	.000	.000	.000
	N	148	150	150	150	150
PSAT	*r*	.579**	.571**	.508**	.579**	.566**
Writing	*p*	.000	.000	.000	.000	.000
	N	148	150	150	150	150

**$p < .0001$ (2-tailed).

Treatment Group Teacher: Students in the treatment group received instruction for TOK, or treatment, from one of three different teachers creating the possibility of correlated error. One of the assumptions, when utilizing ANOVA for the analysis, is that error terms are uncorrelated (Garson, 2012). Thus, error terms should be independent, random, and normally distributed around a zero mean. Garson (2012) suggests that the best way to ensure this is through randomization of samples. This was not done, therefore,

we must test this assumption. If there are no mean differences on the TER scales between teachers, it can be concluded that the error terms are uncorrelated across classrooms and the assumption of uncorrelated error is met. Table 18 shows that there are no mean differences on the TER scales across classrooms. It is concluded that the assumption of uncorrelated error is met.

Table 18

ANOVA Test of Mean Differences on TER Scores by Teacher

		Sum of Squares	df	Mean Square	F	p
Analysis pretest	Between Groups	2.02	2	1.01	.45	.64
	Within Groups	128.57	57	2.26		
	Total	130.58	59			
Inference pretest	Between Groups	15.60	2	7.80	1.48	.24
	Within Groups	301.13	57	5.28		
	Total	316.73	59			
Evaluation pretest	Between Groups	7.50	2	3.75	.84	.44
	Within Groups	255.35	57	4.48		
	Total	262.85	59			
Deduction pretest	Between Groups	6.07	2	3.04	.36	.70
	Within Groups	487.33	57	8.55		
	Total	493.40	59			
Induction pretest	Between Groups	2.66	2	1.33	.19	.82
	Within Groups	388.94	57	6.82		
	Total	391.60	59			
Analysis posttest	Between Groups	8.80	2	4.40	1.38	.26
	Within Groups	185.39	58	3.20		
	Total	194.20	60			
Inference posttest	Between Groups	19.18	2	9.59	1.26	.29
	Within Groups	442.46	58	7.63		
	Total	461.64	60			
Evaluation posttest	Between Groups	25.50	2	12.75	2.47	.09
	Within Groups	299.06	58	5.16		
	Total	324.56	60			
Deduction posttest	Between Groups	42.97	2	21.48	1.63	.21
	Within Groups	765.39	58	13.20		
	Total	808.36	60			
Induction Posttest	Between Groups	31.73	2	15.87	2.26	.11
	Within Groups	407.94	58	7.03		
	Total	439.67	60			

Summary of the Description of the Treatment and Comparison Groups

The statistical descriptions of the treatment and comparison groups in this study show that the two groups are comparable in terms of ethnicity, teacher, and cohort and can thus be eliminated as third variable explanations of a hypothetical treatment effect.

I examined correlations between TER scores and several other variables. There was a relationship between sex and the pretest of the TER scale score – Inference. In addition, the statistical analysis demonstrated TER scores were significantly linked to 9th-and 10th-grade GPA and PSAT scores. Therefore, it will be necessary to compute a series of additional models that include each of the PSAT scores and GPA for 9th-and 10th-grade for all five of the TER subscales as well as another model that includes sex. These additional models will include covariates so that I can determine whether there are third variable explanations of any hypothetical effects of treatment that emerge.

Analysis of Research Questions

Prior to examining the research questions it is necessary to determine if conclusion validity exists for this study. Conclusion validity informs the researcher if the suspected connections or assumptions among the variables are significant. If they are significant it means the assumptions are so strong that they are not likely to occur by chance coincidence, but rather due to some causal link among the variables being studied (Trochim, 2006). All of the assumptions of the inferential tests were examined and the results of those are presented in Appendix D. In most cases the assumptions were met and in those few cases where the assumptions were not met, the sample sizes were

deemed large enough to render the tests robust to violations of the assumptions. Thus conclusion validity was strong for all tests.

Research Question 1

Is there a significant difference between the growth of critical thinking skills, as demonstrated by the Test of Everyday Reasoning (TER)– Total Score, between students who participated in a course designed to improve critical thinking skills (the International Baccalaureate –Diploma Programme Theory of Knowledge course) and those who did not participate in the course?

TER Total Score pretest and posttest scores were analyzed to determine if there was a difference between the growth of critical thinking skills in the treatment group when compared to the comparison group. A series of models for analysis were used to examine a set of possible 'third variable explanations' (sex, 9th-and 10th-grade GPAs, and PSAT scores) of the significant effect linked to treatment. This analysis used univariate analysis of covariance (ANCOVA) to determine whether obtained differences of the mean gain of the TER – Total score are statistically significant. An analysis of covariance, or ANCOVA, statistically adjusts or equalizes the initial differences between groups when participants are not randomly assigned to treatment groups (Weinberg & Abramowitz, 2008). ANCOVA adjusts the dependent variable (estimate of the systematic gain on TER) for differences on an independent variable (group: treatment and comparison) to be controlled (Gay, Mills, & Airasian, 2009).

The results for the models used to analyze the first research question are presented in Table 19. The outcome studied was change on the TER Total Score from pretest to posttest. The question predictor in Model A1 was treatment. A question predictor is a variable on the independent side of the equation in which the researcher has a strong interest. In this case, the question predictor was treatment. The control predictors are the variables on the independent side of the equation in which the researcher does not have a strong interest. These variables are included in the equation to control for correlated error or other similar reasons. In this case sex, GPA, and PSAT are control predictors (Trochim, 2006). A categorical variable was entered into the model: it was coded 0 for students in the comparison group and 1 for students in the treatment group. The analysis demonstrated that students in the treatment group exhibited greater gains on TER Total Score. As shown in Table 19, treatment was a significant predictor of change on the TER Total Score from pretest to posttest (b = -3.06, $p < .005$). A standard alpha criterion of .05 was used for this analysis. This dictates that any variable with a probability level (p) less than .05 was deemed to have a significant effect.

The comparison group is the reference group in the b coefficient analysis. These coefficients are interpreted as the amount of change in the dependent variable (Y) that is associated with a change in one unit of the independent variable (X). All b coefficients are unstandardized, which means that the slopes can be interpreted directly in terms of the raw values of X and Y (Janda, 2001).

In this analysis, the b coefficient has a negative valence (b = -3.06), signifying that the comparison group demonstrated less improvement than the treatment group from pretest to posttest. The magnitude of the b coefficient indicates the scope of the

difference in growth between the treatment group and the comparison group. More specifically, the average gain on the Total Score for the treatment group was 3.06 points greater than the comparison group.

It is necessary to determine the strength of the effect size to evaluate whether or not 3.06 points is a significant difference between the treatment group and the comparison group. To determine the significance of the 'effect' you divide the *b* coefficient by the Standard Deviation (SD) of the full sample (treatment group combined with the control group) which is 5.24 (Cohen, 1988). Therefore, we divide 3.06 by 5.24 to find the effect size statistic. Thus, the effect size is 3.06/5.24 = .58. Cohen's (1988) guidelines for analysis of power have been embraced as the standard gauge for effect sizes. He established that effect sizes smaller than .2 were small; effect sizes between .2 and .8 were moderate in size; and statistics larger than .8 were large effect sizes. Thus, the effect size of .58 is considered moderate in size.

Despite the moderate effect size, I could not simply conclude that the treatment caused the greater gains on TER-Total Score as the *ex-post facto* study design did not allow for random assignment to groups. Further analysis was required to understand the growth patterns for the treatment group and the comparison group in this study. Possible alternative explanations had to be explored to better understand the superior performance in TER Total Score of students in the treatment group. Possible third variable explanations were tested in an attempt to explain the greater success of the treatment group. A series of models utilizing variable explanations' (sex, 9th-and 10th-grade GPAs, and PSAT scores) that were uncovered in the preliminary analyses previously conducted.

Table 19

Regression Coefficients for Predictors of TER Total Scores

Predictors	Model A	Model B	Model C	Model D	Model E	Model F
	TER Total	TER Total	TER Total	TER Total	TER Total	TER Total
Total Pre	.86***	.87***	1.04***	.89***	.52***	.52***
Sex		-.21				
GPA (9th)			.17			
GPA (10^{th})				-.65		
PSAT_CRT					.22**	.22**
PSAT_MT					.04	.04
PSAT_WRT					.10	.10
Treatment	-3.06**	-3.04**	-3.03+	-3.51	-1.24	-1.24

[a]The difference between Model E and Model F is that Model E uses type 3 sums of squares which assumes equal cell sizes; Model F uses type 1 sums of squares does not assume equal cell sizes.

+$p < .10$. *$p < .05$. **$p < .005$. ***$p < .0005$.

Model B (Table 19) incorporated the control predictor sex (n = 150). I considered this model because a preliminary examination of pretest TER scale scores (see Table 14 and 15) showed males had an advantage on the ability to draw Inferences. Model B takes sex into account while analyzing the change from pretest scores on the TER to the posttest scores on the TER. The variable "sex" was not a significant predictor of growth of any of the TER scales; therefore, we can conclude that males and females grew at equal rates on the TER. After controlling for the effects of sex, the effect of treatment persisted as a significant predictor of growth on TER Total Score (b = -3.04) Thus, we can conclude that the effect of treatment observed in Model A was not a spurious effect of sex differences in ability.

Model C (Table 19) depicted growth from pretest to posttest on TER Total Score while taking 9th-grade GPA into account (n = 42). This analysis was completed because 9th-grade GPA was found to be a predictor of test performance on a series of preliminary correlations (Table 16). An argument could be made that the superior growth experienced by the treatment group was due to the fact that brighter, more academically gifted students enrolled in the IB-DP and was not due to the treatment (TOK course) at all.

Model C (Table 19) demonstrates that the effect of treatment decreased to a trend level of significance (b = -3.03, p <.10). This suggests that the ability differences between the treatment group and the comparison group, as measured by 9th-grade GPA, may explain the effect of treatment observed in Model A. However, it should be noted that the number of observations included in Model C was relatively low (n = 42) as 9th-grade GPA was unavailable for many participants. Also, this variable studied in isolation is of minor importance as a covariate because students participated in this study at least

sixteen months after 9th-grade GPA was calculated. However, the hypothesis that ability differences explain the effect of treatment is reinforced by the results for Model D (Table 19).

Model D depicts growth on TER Total Score for treatment group students versus comparison group students while accounting for the control predictor 10th-grade GPA. The sample size was 130 for this test. Although there were missing data on 20 cases for 10th-grade GPA, this analysis was more useful than the analysis with 9th-grade GPA as a covariate for two reasons: (a) the sample size was larger for Model D and (b) the data from Model D was collected just three months prior to the students entering the DP or treatment.

The central finding from Model D was that the significant effect of treatment growth in TER Total Score that was evident for Model A, disappeared once the covariate 10th-grade GPA was considered. Two possible explanations for this finding are: (a) The effect of treatment observed in Model A was a spurious effect attributable to the self-selection of students into the treatment group. The fact that more academically capable students selected to enter the treatment group may account for the greater growth experienced by this group. Once those ability differences were controlled by entering 10th-grade GPA into the model, the differential growth patterns for the treatment group and the comparison group disappeared. Or (b) perhaps the smaller size of the D sample (n = 150 for Model A versus 130 for Model D) and the concomitant reduction in power explains the non-significant effect of treatment in Analysis in Model D.

When I examined Models E and F, I found support for explanation (a): that the effect of treatment is a spurious effect due to ability differences between the students in the treatment group versus the control group (Table 19). Model E portrays growth from the pretest to the posttest on the TER Total Score while accounting for the predictor variable of scores on the PSAT. It demonstrates that PSAT Critical Reading is a significant predictor of performance on TER Total Score ($b = .22$, $p < .005$). PSAT Math and Writing scores were not significant predictors of performance on TER Total Score. However, when Model A is contrasted with Model E the significant effect of treatment on growth of TER Total Score disappears once PSAT scores are accounted for. Thus, the alternative (competing) hypothesis that the treatment was not the cause of the superior growth of the students in the treatment group is gaining legitimacy. The data support the hypothesis that the significantly larger growth from pretest to posttest TER Total Scores among the treatment group of students was due to the fact that more academically gifted students self-selected into the IB-DP.

Another model was constructed to account for the differences in size between the treatment group ($n = 60$) and the comparison group ($n = 94$). Model F includes the PSAT scores just as Model E did, but it also takes into account the different cell sizes. The only change from Model E to Model F is the change from a type 3 sum of squares to a type 1 sum of squares algorithm. A type 3 sum of squares analysis assumes that the cell sizes are the same. A type 1 sum of squares analysis adjusts for the differences in cell sizes. The results from Model F were identical to Model E.

In conclusion, the analyses of growth from the pretest to the posttest TER Total Score demonstrates that the mean ability differences between students in the two groups

explain the significant effect of treatment observed in Model A. Unfortunately, these analyses do not support the hypothesis that the TOK course leads to increased achievement in critical thinking as measured by TER Total Scores.

Analysis Comparing Effect of Treatment to PSAT-CR Score

On average, students in the treatment group gained 3.39 points on the TER Total Score from pretest to posttest. In comparison, the students in the comparison group gained an average of .79 points from pretest to posttest. Thus the mean gain for the IB-DP students, or treatment group, is approximately 4.3 times greater than the mean gain for the non-IB-DP students, or comparison group. This is an impressive difference; however, the *ex-post facto* study design requires further analysis of possible third variable explanations prior to reaching the conclusion that treatment caused the greater gains on the TER-Total Score.

In order to contrast the effect of treatment to the PSAT-CR skills effect, I ran another analysis in which I bisected the sample into two groups. Group 1 consisted of students that scored in the top 50% on the PSAT-CR test. Group 2 consisted of students that scored in the bottom 50% on the PSAT-CR test. Students in Group 1 gained, on average, 3.14 points on the TER from pretest to posttest. Those in Group 2 gained, on average, .70 points on the TER. Thus, the students who scored in the top 50% on PSAT-CR gained, on average, 4.5 times more than those who scored in the bottom 50% on the PSAT-CR.

This rudimentary analysis concludes that the advantage of scoring high on the PSAT-CR (TER average gain scores that are 4.5 times higher) is greater than the

advantage of being in the treatment group (TER gain scores that are only 4.3 times higher). This analysis demonstrates in a more simplistic method what the more refined ANCOVA analysis determined.

Research Question 1a

Is there a significant difference between the growth of targeted cognitive skills, as demonstrated in TER – Scale Scores (analysis and interpretation; evaluation and explanation; and inference), in students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the course?

TER – Cognitive Scale scores (Analysis and Interpretation, Evaluation and Explanation, and Inference) from pretest and posttest were analyzed to determine if there was a difference between the growth of critical thinking skills in the treatment group when compared to the comparison group. A series of models (sex, 9th-and 10th-grade GPAs, and PSAT scores) were analyzed with different covariates entered into the analyses to determine if they were legitimate third variables acting as predictors of growth of specific cognitive skills. This analysis involved a multivariate analysis of covariance (MANCOVA) to determine whether obtained differences of the mean gain of cognitive scale scores exhibit significant differences between the treatment group and the comparison group.

The first baseline is represented in Model A1 in Table 20 (n=150). The TER pretest scale scores on Analysis, Inference, and Evaluation are the first three predictors or covariates in this model. These variables were entered into the model to control for

performance in these areas prior to the beginning of treatment. This allowed for the measurement of growth that occurred on each cognitive skill from pretest to posttest. It should be noted that the pretest scores on each skill are significant predictors of the posttest scores in the majority of the models; however, this is to be expected, as reason would dictate that performance on these cognitive skills in 11th-grade would be correlated with performance in 12th-grade.

The question predictor in Model A1 was treatment. The analysis demonstrated that students in the treatment group exhibited greater gains on Analysis and Inference scores relative to students in the comparison group (b = -.71, $p < .05$ and b = -1.16, $p < .005$, respectively). A standard alpha criterion of .05 was used for this analysis.

The comparison group is the reference group in the b coefficient analysis. In this analysis, the b coefficient have a negative valence (b = -.71 and b = -1.16), signifying that the comparison group demonstrated less improvement than the treatment group from pretest to posttest. More specifically, the average gain on the Analysis scores for the treatment group was .71 points greater than the comparison group. Likewise, the average differential growth for the treatment group in Inference scores was 1.16 points greater than the comparison group. The b coefficient associated with the effect of treatment on Evaluation gain scores (b = -.50, p = n.s) was non-significant. The interpretation of a non-significant b coefficient is that there were no significant differences in the skill growth in Evaluation, on average, between the treatment group and the comparison group.

To determine the significance of the effect of treatment the b coefficient is divided by the Standard Deviation (SD) of the full sample (treatment group combined

with the control group). Therefore, to determine the effect size of the Analysis average gain score we divide the *b* coefficient of .71 by the SD of 1.66. Thus, the Analysis effect size is .71/1.66 = .42. According to Cohen's guidelines this effect size would be considered moderate. The effect size of the Inference average gain score (*b* coefficient = 1.16) is calculated by dividing 1.16 by 2.59. Thus, the effect size of .44 is considered moderate according to Cohen's guidelines.

I cannot simply conclude that the treatment caused the greater gains on Inference. Alternative explanations will need to be explored to better understand the superior performance in Inference of students in the treatment group. Possible third variable explanations must be tested in an attempt to explain the greater success of the treatment group. A series of models utilizing ANCOVA for analysis were used to examine a set of predetermined hypothetical 'third variable explanations' (sex, 9th-and 10th-grade GPAs, and PSAT scores) of the treatment effect uncovered in Model A1.

Table 20

Regression Coefficients for Predictors of TER Scores for Analysis, Inference, and Evaluation

Predictors	Model A1 Analysis Post	Model B1 Analysis Post	Model C1 Analysis Post	Model D1 Analysis Post	Model E1 Analysis Post	Model F1[a] Analysis Post
Analysis (pre)	.41***	.43***	.56*	.38**	.28*	.28*
Inference (pre)	.23***	.21**	.30	.23**	.18*	.18*
Evaluation (pre)	.19*	.19**	.12	.22*	.08	.08
Sex		-.23				
GPA (9th)			-.77			
GPA (10th)				.30		
PSAT CR					.08**	.08**
PSAT Math					-.02	-.02
PSAT Writing					.03	.03
Treatment	-.71*	-.74*	-1.13	-.57	-.24	-.24
	Inference Post	Inference Post	Inference Post	Inference Post	Inference Post	Inference Post
Analysis (pre)	.54***	.53***	.68*	.52**	.39*	.39*
Inference (pre)	.41***	.42***	.51**	.38***	.31**	.31**
Evaluation (pre)	.31**	.31**	.18	.38***	.20*	.20*
Sex		.074				
GPA (9th)			-.13			
GPA (10th)				.26		
PSAT CR					.04	.04
PSAT Math					.03	.03
PSAT Writing					.06	.06
Treatment	-1.16**	-1.16*	-1.45+	-1.25*	-.50	-.50

Table 20 (continued)

Predictors	Model A1	Model B1	Model C1	Model D1	Model E1	Model F1[a]
	Evaluation Post	Evaluation Post	Evaluation Post	Evaluation Post	Evaluation Post	Evaluation Post
Analysis (pre)	.30*	.31*	.23	.27*	.16	.16
Inference (pre)	.12	.11	.21	.15	.01	.01
Evaluation (pre)	.53***	.53***	.45**	.56***	.39***	.39***
Sex		-.14				
GPA (9th)			.70			
GPA (10th)				-.20		
PSAT CR					.08*	.08*
PSAT Math					.00	.00
PSAT Writing					.03	.03
Treatment	-.50	-.52	-.57	-.56	-.05	-.05

[a]The difference between Model E1 and Model F1 is that Model E1 uses type 3 sums of squares which assumes equal cell sizes; Model F1 uses type 1 sums of squares does not assume equal cell sizes.

+$p < .10$. *$p < .05$. **$p < .005$. ***$p < .0005$

Model B1 (Table 20) incorporated the covariate sex (n = 150). This model was required as a preliminary examination of pretest TER scale scores (see Table 14 and 15) presented males had an advantage on the ability to draw Inferences. Model B1 takes sex into account while analyzing the change from pretest scores on the TER to the posttest scores on the TER. The variable "sex" was not a significant predictor of growth of any of the TER scales; therefore, we can conclude that males and females perform equally on the TER. After controlling for the effects of sex, the effect of treatment was a significant predictor of growth on the Analysis and Inference scales (b = -.74, p <.05 and b = -1.16, p < .05). Treatment was not a significant predictor of change on the cognitive skill of Evaluation on the TER (b = -.52, p = n.s.). Thus, we can conclude that the effect of treatment observed in Model A1 was not a spurious effect of sex differences in ability.

Model C1 (Table 20) depicted growth from pretest to posttest on Analysis, Inference, and Evaluation while taking 9th-grade GPA into account (n = 42). This analysis was completed because 9th-grade GPA was found to be a predictor of test performance on a series of preliminary correlations (Table 16). Model C1 demonstrates that the effect of treatment was not significantly linked to growth on the TER scale scores of Analysis or Evaluation after controlling for 9th-grade GPA. However, the effect of treatment was linked to growth at the trend level for the outcome variable: Inference (b = -1.45, p < .08). Therefore, a trend exists for students in the treatment group to experience greater growth on the cognitive skill of Inference, even when the effect of 9th-grade GPA is controlled.

It should be noted that 9th-grade GPA was insignificant in each case, indicating no relationship between 9th-grade GPA and the change students experience in the TER

cognitive skill scale scores. This non-significant effect may have been a result of the fact that the number of observations included in model C1 was relatively low (n = 42) as 9th-grade GPA was unavailable for many participants. It is likely that the non-significant findings for most of the variables in the model were due to type 2 errors and the concomitant low power. Also, this variable is less relevant as a covariate as students participated in this study at least sixteen months after 9th-grade GPA was calculated.

Model D1 depicts growth on Analysis, Inference, and Evaluation for treatment group students versus comparison group students while accounting for 10th-grade GPA. The sample size was 130 for this test due to missing data on 20 cases for 10th-grade GPA.

The central finding from the Model D1 analysis is that the significant effect of treatment growth in Analysis that was evident for Model A1, disappeared once the covariate 10th-grade GPA was considered. In contrast to growth in Analysis, the treatment effect persisted with regard to growth in the ability to make Inferences (b = -1.25, $p < .02$) for Model D1. This result appears to be in direct conflict with the earlier examination of Analysis. One important piece of data to notice is the effect size is larger for Inference than for Analysis. Thus, the data are consistent with the hypothesis that the disappearance of the significant effect for Analysis in Model D1 was a result of the reduction of power. This, however, does not rule out the possibility that the alleged treatment effect is really due to the academic ability differences between the students in the treatment group and in the comparison group. It will be necessary to gather further evidence to sufficiently defend this hypothesis.

Treatment was not a significant predictor for growth in the cognitive skill of Evaluation (b = -.56, p = n.s.) in Model D1. This analysis continues with the same conclusion set in Models A1 through C1.

Model E1 portrays growth from the pretest to the posttest on the TER cognitive skill scale scores while accounting for the covariate of scores on the PSAT. It demonstrates that PSAT Critical Reading is a significant predictor of performance on the Analysis and Evaluation posttests (b = .08, p < .005 and b = .08, p < .05, respectively). PSAT Math and Writing scores were not significant predictors of performance on any of the TER cognitive skills posttest scores. However, when Model A1 is contrasted with Model E1 the significant effect of treatment on growth Analysis and Inference cognitive skills disappears once PSAT scores are accounted for.

Another model was constructed to account for the differences in size between the treatment group (n = 60) and the comparison group (n = 94). Model F1 was the same as Model E1 except that the analysis uses type 1 sums of squares instead of the default method of type 3 sums of squares. Type 1 sums of squares is valuable because it corrects for any differences that may exist due to differing subsample size. The results from Model F1 were not significantly different from Model E1.

Research Question 1b

b. *Is there a significant difference between the growth of targeted reasoning skills, as demonstrated in TER – Deductive and Inductive Reasoning Scale Scores, between students who participated in the IB-DP Theory of Knowledge*

course and those who did not participate in the course?

TER – Reasoning Scale scores (Deductive and Inductive) from pretest and posttest were analyzed to determine if there was a difference between the growth of critical thinking skills in the treatment group when compared to the comparison group. A series of models (sex, 9th-and 10th-grade GPAs, and PSAT scores) were analyzed with different covariates entered into the analyses to determine if they were legitimate third variables acting as predictors of growth of specific reasoning skills. This analysis involved a multivariate analysis of covariance (MANCOVA) to determine whether obtained differences of the mean gain of cognitive scale scores exhibit significant differences between the treatment group and the comparison group.

Table 21 presents a series of analyses on models examining the effect of treatment and various covariates on the gains on the posttest TER – Deductive and Inductive Reasoning scale scores for the treatment group. The baseline model A2 (n = 150) tests the covariates (predictor variables) Inductive and Deductive Reasoning. The pretest scores on each of these reasoning skills are significant predictors of the posttest scores in all of the subsequent models. Once again this is to be expected, as it is logical that performance on these skills at the beginning of 11th-grade would be correlated with performance in 12th-grade.

Treatment (enrollment in TOK course) was the question predictor entered in Model A2. I created a categorical variable coded 0 for students in the comparison group and 1 for students in the treatment group. The data revealed students in the treatment group demonstrated greater gains on Deductive and Inductive Reasoning scores relative

to the comparison group (b = -1.91, $p < .005$ and b = -1.33, $p < .005$, respectively). The standard alpha criterion of .05 was used for this model. Any variable with a probability level (p) less than .05 was deemed to have a significant effect. The b coefficient indicated the treatment group scored 1.91 points greater than the comparison group for Deductive Reasoning on average. The b coefficient for Inductive Reasoning indicated that the treatment group scored 1.33 points higher than the comparison group on average.

To determine the significance of the effect of treatment the b coefficient is divided by the Standard Deviation (SD) of the full sample (treatment group combined with the control group). Therefore, to determine the effect size of the Deductive average gain score we divide the b coefficient of 1.91 by the SD of 3.32. Thus the Deductive effect size is 1.91/3.32 = .57. According to Cohen's guidelines this effect size would be considered moderate. The effect size of the Inductive average gain score (b coefficient = 1.33) is calculated by dividing 1.33 by 2.77. Thus the effect size of .48 is considered moderate according to Cohen's guidelines.

The use of historical data for this study only allowed for a causal-comparative design for this study. Such a design does not allow for the random assignment to groups. Therefore, we cannot simply conclude that the treatment caused the greater gains on Deductive and Inductive Reasoning. The growth patterns for the treatment and comparison group will need to be further analyzed to adequately understand the greater performance of the treatment group in reasoning skills. Correlational tests revealed a set of hypothetical 'third variable explanations' (sex, 9th-and 10th-grade GPAs, and PSAT scores) that would need to be further analyzed to eliminate them as spurious causes for

the superior performance of the treatment group. A series of models, utilizing ANCOVA for analysis, were used to examine the treatment effect uncovered in Model A2.

Table 21

Regression Coefficients for Predictors of TER Scores for Deductive and Inductive Reasoning

Predictors	Model A2 Deduction Post	Model B2 Deduction Post	Model C2 Deduction Post	Model D2 Deduction Post	Model E2 Deduction Post	Model F2[a] Deduction Post
Deduction(pre)	.59***	.58**	.69**	.59***	.37**	.37**
Induction(pre)	.31*	.31*	.48	.33*	.09	.09
Sex		-.40				
GPA (9^{th})			-.22			
GPA (10^{th})				-.40		
PSAT CR					.13**	.13**
PSAT Math					.05	.05
PSAT Writing					.06	.06
Treatment	-1.91**	-1.95**	-1.73	-.21**	-.70	-.70
	Induction Post	Induction Post	Induction Post	Induction Post	Induction Post	Induction Post
Deduction (pre)	.29***	.30***	.39*	.30***	.18*	.18*
Induction (pre)	.48***	.48***	.49*	.50***	.31**	.31**
Sex		.53				
GPA (9^{th})			.43			
GPA (10^{th})				-.21		
PSAT CR					.10**	.10**
PSAT Math					.00	.00
PSAT Writing					.04	.04
Treatment	-1.33**	-1.27**	-1.20	-.53**	-.68	-.68

[a]The difference between Model E2 and Model F2 is that Model E2 applied type 3 sums of squares, which assume equal cell sizes, and ModelF2 applied type 1 sums of squares, which does not assume equal cell sizes.

+$p < .10$. *$p < .05$. **$p < .005$. ***$p < .0005$

Model B2 examined the covariate sex (n = 150). This model accounted for sex while analyzing the change on the TER – Deductive and Inductive Reasoning scale scores from pretest to posttest. After controlling for the effects of sex, the effect of treatment was a significant predictor of growth on the Deductive and Inductive scales (b = -1.95, $p < .005$ and b = -1.27, $p < .005$). The third variable sex was not a significant predictor of change on any TER scale scores. Subsequently, we can conclude that males and females demonstrate similar development changes. The most important conclusion from these results is that the effect of the treatment observed in model A2 was not attributable to ability differences in sex.

Model C2 depicts growth in TER – Reasoning skills while accounting for 9th-grade GPA. Grade 9 GPA was found to be a predictor (Table 16) on a series of preliminary correlation analyses forcing a closer examination of its influence on the treatment. In Model C2 the effect of treatment on Deductive and Inductive Reasoning was nullified when 9th-grade GPA was taken into account. It should be noted, however, that 9th-grade GPA was insignificant when examined as a predictor of performance on these skills. The fact that Model C2 had a relatively low number of participants (n = 42) may have been the cause of this finding. These low numbers would likely cause type 2 errors and concomitant low power, rendering grade 9 GPA an unreliable predictor variable.

Model D2 examines the effect of treatment on the growth of the TER – Deductive and Inductive Reasoning posttest while taking 10th-grade GPA into account. After controlling for the effects of 10th-grade GPA, the effect of treatment was a significant predictor of growth on the Deductive and Inductive scales (b = -.21, $p < .005$ and b = -53, $p < .005$). This indicates that students in the treatment group would score, on average, .21 and .53 points greater in Deductive and Inductive Reasoning respectively than their peers in the comparison group. The third

variable 10th-grade GPA was not a significant predictor of change on any TER scale scores. Subsequently, we can conclude that regardless of GPA in grade 10, treatment group students will demonstrate greater development change than those in the comparison group on the TER – Reasoning skills. The most important conclusion from these results is that the effect of the treatment observed in model A2 was not attributable to ability differences as measured by10th-grade GPA.

Model E2 accounted for PSAT scores while examining the students' growth on the TER – Reasoning skills from pretests to posttests. PSAT Critical Reading scores proved to be a significant predictor on both Inductive and Deductive Reasoning skills posttests ($b = .13$, $p < .005$ and $b = .10$, $p < .005$, respectively); however, PSAT Math and Writing scores were not significant predictors of performance on Reasoning skills posttest scores. Most importantly, when Model A2 is contrasted with Model E2 the effect of treatment on growth on the TER – Reasoning skills from pretests to posttests disappears when PSAT scores are accounted for. This finding dictates further examination of the alternative (competing) hypothesis that the treatment did not cause the superior growth demonstrated by the students in the treatment group. It is highly possible that the greater growth exhibited by the treatment group was due to the fact that students with a superior ability in Reasoning skills self-selected to enter the IB-DP (treatment).

Model F2 was constructed to account for the differences in size between the treatment group ($n = 60$) and the comparison group ($n = 94$). Model F2 includes the PSAT scores just as Model E2 did, but it also takes into account the different group sizes. The only change from Model E2 to Model F2 is the change from a type 3 sum of squares to a type 1 sum of squares algorithm. A type 1 sum of squares analysis adjusts for the differences in the number of students in each group, while a type 3 sum of squares analysis assumes that the groups were the same size.

The results from Model F2 were found to be similar to the results from Model E2. These results once again support an alternative hypothesis to explain the superior growth in Deductive and Inductive Reasoning skills. The evidence is consistent with the alternative hypothesis that students with superior critical thinking skills self-selected for entrance into the IB-DP (treatment).

Chapter 5

Conclusions and Recommendations

Introduction

Educational leaders agree that critical thinking skills are needed to survive in the complex, digital and global economy of the 21st century (Jacobs, 2010). Thus, it is necessary to find ways to effectively develop these skills in today's students. However, little consensus exists on how to best teach these important skills (Ben-Chaim, Ron, & Zoller, 2000; Grauerholz & Bouma-Holtrop, 2003). The purpose of this study was to evaluate the impact of a single course on critical thinking skills as measured by the TER. The goal was to provide a quantitative analysis to determine if any statistically significant differences exist in the development of critical thinking skills as related to the TOK class offered to students who were enrolled in the IB-DP at the American International School Kuwait from the fall of 2008 to the spring of 2011.

Each research question will be addressed in the following summary and discussion of findings section within this chapter. Additionally, this chapter will present limitations of the study, implications of the study, conclusions of the study, recommendations for further study, and the researcher's conclusions.

Summary and Discussion of Findings

I began this study by gathering data from research collected by AISK as part of its on-going assessment and evaluation process. The larger school study collected data from two separate cohorts: Cohort 1 graduated from AISK in 2010 and Cohort 2 graduated in 2011. These

students could choose whether or not to take the IB Diploma sequence of courses. As part of the assessment of critical thinking at AISK, the students in 11th-and 12th-grade were required to complete the TER. The class of 2010 was the first class for which both a pretest and posttest were available.

Data on the student graduation year, sex, ethnicity, IB-DP status, GPA, PSAT scores, and TER pretest and posttest scores were collected on 171 students: 67 students who graduated in 2010 (Cohort 1) and 104 students who graduated in 2011 (Cohort 2). Cohort 1 had 24 students in the treatment group of IB-DP students and 43 students in the comparison group of non-IB-DP students. Cohort 2 was comprised of 38 students in the treatment group of IB-DP students and 66 students in the comparison group of non-IB-DP students.

Discussion of Descriptive Statistics

Sex, ethnicity, cohort, teacher, previous academic ability (GPA, PSAT), and prior critical thinking ability (pretest TER scores) were compared between treatment and comparison groups. The population was made up of 97 males (56.7%) and 74 females (43.3%). The smaller number of females in the population could be attributed to a number of all-girls schools that are present in Kuwait. Kuwait is a conservative Islamic-State and many of the parents are hesitant to have females educated in a co-educational institution (Ray, personal communication October 3, 2012). The proportion of males in the comparison group was eight percent higher than in the treatment group.

The total population was made up of 93 (54.4%) Kuwaitis. A noticeably smaller percentage of Kuwaiti students (17.2%) chose to enter the IB-DP in comparison to the other

nationalities (on average 61.5%). Perhaps the reason for the relatively low percentage of Kuwaiti students in the IB-DP is because Kuwaiti students need to maintain a high GPA in order to be eligible for the lucrative government scholarships that are readily available. GPA is a significant variable when determining eligibility and the difficulty of the program is not taken into account; therefore, the perception exists that attempting to meet all of the IB-DP Higher Level course requirements is unnecessary and may in fact limit a student's chances of achieving a high GPA and consequently limit their chances of being awarded a scholarship. An additional analysis showed that ethnicity was not correlated to treatment. Therefore, ethnicity could be eliminated as a third variable explanation of the treatment effect.

The treatment group exhibited higher previous academic ability as measured by GPAs and PSAT scores. The GPAs for the treatment group were moderately higher, on average, than the comparison group in 9th-and 10th-grade. On average, the treatment group scored moderately higher in PSAT Critical Reading, Math, Writing, and Total score than the comparison group. The treatment group demonstrated higher prior critical thinking ability in the pretest TER. On average, the treatment group outscored the comparison group on the TER Total score (+4.16) and all five of the scale scores: Analysis (+1.03), Inference (+1.53), Evaluation (+1.41), Deductive Reasoning (+2.28), and Inductive Reasoning (+1.77).

The statistical analysis demonstrated TER scores were significantly linked to 9th-and 10th-grade GPA and PSAT scores. Therefore, a series of additional models were developed that included each of the PSAT scores and GPA for 9th-and 10th-grade for all five of the TER scale scores as well as another model that included sex. These covariates were used to determine whether they could be considered as third variable explanations of the effects that were linked to treatment.

The treatment group exhibited higher critical thinking ability than the comparison group on average in the posttest TER (Total score = +6.76; Analysis = +1.82; Inference = +2.91; Evaluation = +1.94; Deductive Reasoning = +3.82; and Inductive Reasoning = +2.92). This is logical as the treatment group demonstrated higher prior critical thinking ability in the pretest TER.

A statistical analysis demonstrated that the treatment and comparison groups were comparable when accounting for cohort and thus cohort could be eliminated as a third variable explanation. An analysis of treatment teacher also demonstrated that the three treatment groups were comparable and teacher was not a feasible third variable explanation of the effect linked to treatment.

Discussion of Research Questions

This non-experimental, causal-comparative research study was designed using *ex post facto* data to determine whether there was a significant difference in the development of critical thinking skills between students who were enrolled in a course designed specifically to teach critical thinking skills and students not exposed to this course. I will present a summary of this study organized in the order of the research questions.

Research Question 1

Is there a significant difference between the growth of critical thinking skills, as demonstrated by Test of Everyday Reasoning (TER) – Total Score, between students who

participated in a course designed to improve critical thinking skills (the International Baccalaureate – Diploma Programme Theory of Knowledge course) and those who did not participate in the course?

TER Total Score pretest and posttest scores were analyzed to determine if there was a significant difference between the growth of critical thinking skills in the treatment group when compared to the comparison group. The outcome studied was change on TER Total Score from pretest to posttest. The analysis demonstrated that students in the treatment group exhibited greater gains on TER Total Score. In this analysis, the average gain on the Total Score for the treatment group was moderately greater than the comparison group using Cohen's (1998) guidelines to determine the strength of the effect size.

The study design requires further analysis of possible third variable explanations prior to reaching the conclusion that treatment caused the greater gains on the TER-Total Score. A series of models were examined to determine if previously identified covariates (sex, 9th-and 10th-grade GPA, and PSAT scores) could explain the effect linked to treatment. An analysis of 9th-grade GPA revealed that academic ability was a predictor of success on the TER. This variable studied in isolation is of minor importance as a covariate because of the relatively low number of available participants (n = 42) and the fact that this GPA was calculated 16 months prior to the students participating in the study.

Further analysis revealed that the significant effect of treatment disappeared when mean ability differences (10th-grade GPAs and PSAT scores) were added as covariates. Unfortunately, these analyses do not support the hypothesis that the treatment (TOK course) leads to increased achievement in critical thinking as measured by change on TER Total Scores. In fact, the data

suggest that Treatment is a spurious effect due to prior academic ability differences between the students in the treatment group versus the comparison group.

The results for research question 1 were the most critical to this study because the designers of the TER have indicated that Total Score is the most reliable measure of critical thinking ability (Facione, et. al., 2012). Further analyses of the cognitive and reasoning scale scores (Analysis, Inference, Evaluation, Induction, Deduction) increases our understanding of critical thinking development but are not as reliable as measures of critical thinking when studied in isolation.

Research Question 1a

Is there a significant difference between the growth of targeted cognitive skills, as demonstrated in TER – Scale Scores (analysis and interpretation; evaluation and explanation; and inference), in students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the course?

The analysis demonstrated that students in the treatment group exhibited greater gains on Analysis and Inference scores relative to students in the comparison group. The effect size of the Analysis average gain score is considered moderate according to Cohen's (1988) guidelines. The effect size of the Inference average gain is also considered moderate according to Cohen's guidelines.

The *b* coefficient associated with the effect of treatment on Evaluation gain scores was non-significant. The interpretation of a non-significant *b* coefficient is that there were no

significant differences in the skill growth in Evaluation, on average, between the treatment group and the comparison group.

A series of models utilizing MANCOVA for analysis were used to examine a set of possible 'third variable explanations' (sex, 9th-and 10th-grade GPAs, and PSAT scores) of the significant effect linked to treatment. This analysis revealed that 9th-and 10th-grade GPAs and PSAT Critical Reading scores are predictors of growth on TER scale scores of Analysis and Evaluation. PSAT Math and Writing scores were not significant predictors of growth on any of the TER cognitive skills posttest scores.

The significant effect of treatment linked to the ability to make Inferences remained when examining the effect of 9th-and10th-grade GPA. However, when examining the effect of treatment while accounting for PSAT scores, the significant effect of treatment on cognitive skill growth in Analysis and Inference disappeared.

Prior academic ability is a plausible explanation for the significant advantage that the treatment group exhibited in growth of critical thinking skills. The analysis of PSAT scores indicated that performance on the PSAT was a good predictor of growth on the TER cognitive skills, particularly Analysis and Inference.

Research Question 1b

Is there a significant difference between the growth of targeted reasoning skills, as demonstrated in TER – Deductive and Inductive Reasoning Scale Scores, between students who participated in the IB-DP Theory of Knowledge course and those who did not participate in the

course?

Treatment was the question predictor; the data revealed students receiving treatment demonstrated greater gains on Deductive and Inductive Reasoning scores relative to the comparison group. The effect size of average gain score for both Deductive and Inductive Reasoning, according to Cohen's guidelines, would be considered moderate.

The use of historical data only allowed for a causal-comparative design for this study. Such a design does not allow for the random assignment to groups. Therefore, we cannot simply conclude that treatment caused the greater gains on Deductive and Inductive Reasoning. A series of MANCOVA models (including sex, 9th-and 10th-grade GPAs, and PSAT scores), was used to test the hypothesis that the effect of treatment was spurious.

Upon further study the effect of treatment on Deductive and Inductive Reasoning was nullified when 9th-grade GPA was taken into account. It should be noted, however, that 9th-grade GPA was insignificant when examined as a predictor of growth on these skills. The fact that this model had a relatively low number of participants (n = 42) may have been the reason for this finding. These low numbers would likely cause type 2 errors and concomitant low power, rendering grade 9 GPA an unreliable predictor variable.

After controlling for the effects of 10th-grade GPA, the effect of treatment remained a significant predictor of growth on the Deductive and Inductive scales. Students in the treatment group scored higher in Deductive and Inductive Reasoning respectively than their peers in the comparison group when accounting for 10th-grade GPA. In addition, 10th-grade GPA was not a significant predictor of change on any TER scale scores. Subsequently, we can conclude that regardless of GPA in grade 10, treatment group students will demonstrate greater growth than

those in the comparison group on the TER – Reasoning skills.

PSAT Critical Reading scores proved to be a significant predictor on both Inductive and Deductive Reasoning skills posttests. PSAT Math and Writing scores were not significant predictors of growth on Reasoning skills. Most importantly, when both treatment and PSAT scores were included in the model the effect of treatment on growth on the TER – Reasoning skills disappeared. This further supports the finding from Research Question 1 that treatment may not be the cause for the superior growth demonstrated by the students in the treatment group. It is highly possible that the greater growth exhibited by the treatment group was due to the fact that students with superior ability self-selected to enter the IB-DP (treatment).

Further Analysis of Performance on PSAT-Critical Reading and TER Gain Scores

Although the examination of PSAT-CR and its correlation to the development of critical thinking skills was not a part of this study the results of the analyses of the research questions warranted further, albeit rudimentary, analysis (see Appendix F). This analysis grouped students according to high and low scorers on the PSAT-CR. Further study revealed little difference between students in the IB-DP (treatment) and those students not enrolled in the IB-DP (comparison) when comparing the average gain score from pretest to posttest on the TER-Total Score for students that were high scorers on the PSAT-CR. Therefore, if a student was a high scorer on the PSAT-CR it did not matter whether a student was enrolled in the IB-DP or not, he or she demonstrated impressive gains on TER-Total Score from pretest to posttest.

Interestingly, significant differences existed between the comparison and treatment groups when comparing average gain scores of the low scorers on the PSAT-CR. Students that

were *not* enrolled in the IB-DP showed unimpressive gains on the TER-Total Score from pretest to posttest. However, students in the IB-DP (treatment) demonstrated impressive gains on the TER-Total Score from pretest to posttest. In fact, this group had the highest average gain score of all four groups. The difficulty is that this group has a small sample size (n = 15) and the effect was not significant due to low power.

These findings appear to indicate that the IB-DP has the greatest effect on developing critical thinking skills in students that have not previously demonstrated high academic achievement as measured by the PSAT-CR. When considering the development of critical thinking skills this group of students appears to benefit the most from the IB-DP. These students are not traditionally considered to be the best candidates for the IB-DP; however, these findings may challenge previously held beliefs about potential IB-DP students.

Limitations of Study

This study is limited by several factors. These include issues related to the sample available, course of study for both the treatment and comparison group, and challenges associated with equating the groups.

One limitation of this study is that the research design did not allow for random assignment of groups; therefore, the two groups, treatment and comparison, were not randomly separated into two statistically equal groups. This did not allow me to conclude that the treatment was the cause for differential growth of TER scores for the two groups. Further analysis of various covariates needed to be conducted in order to eliminate the possible third variable explanations for the effects linked to treatment.

Although the results did not demonstrate that the TOK course influenced critical thinking development this may have been caused by several factors. One of these factors may have been that the population sample self-registered for the IB-DP. This may have caused students that had demonstrated a greater affinity to schooling or enhanced critical thinking skills to choose to enter a more difficult and rigorous program. Students who have experienced previous success in school and do well in this environment may have chosen the IB-DP, which may be seen as a third variable explanation for the superior growth in critical thinking.

Another limitation may be the similar testing format of the two primary instruments used in the analysis. The PSAT and the TER are multiple-choice tests utilizing a testing booklet and bubble sheet to record responses. These tests are both timed forcing students to move through the questions quickly and efficiently in order to complete the test. It is plausible that students that inherently find these types of tests manageable would do better on the PSAT and subsequently would also perform well on the TER.

Once again the instrument used to measure critical thinking (TER) could be a limiting factor in assessing the effectiveness of the TOK course in developing this objective. Although the TER appears to align well with the Aims and Objectives of the TOK course (see Table 1) it was never designed to measure these objectives. There may be other measures of critical thinking that align more closely with the goals of the TOK course. Other researchers have encountered this same challenge. Walther (2009), assessing the critical thinking skills of high school students in the Advanced Placement and the IB-DP, used the Ennis-Weir Critical Thinking Essay Test (EW). She believed the open-ended essay test format encouraging application of critical thinking skills in the context of solving real life problems would be best suited to measure the critical thinking that these programs hope to develop. However, Walther

also acknowledged the limitation of this test in regards to achieving the objectives of her study for a variety of reasons. She felt the scenarios used in the test were dated and may not be engaging for the students; thus, limiting their motivation to generate critical thought. Another limitation of this test that she identified was that examiners were not provided with more objective criteria when evaluating responses. This highlights the difficulty in finding a test that matches exactly with the goals of a study or a course but ultimately may affect the results. There may not be any one instrument that will adequately measure the development of critical thinking that the IB-DP is attempting to accomplish with the TOK course.

It is recognized that although critical thinking is one of the primary goals of the TOK course it is not the only objective. The TER obviously does not measure these other goals. It is important to note that although this study found that the TOK course did not develop critical thinking, as measured by the TER, it did not evaluate the overall effectiveness of the TOK course. In this aspect this study is limited in its evaluation of the TOK course.

Implications

Educational leaders agree that critical thinking is a necessary skill that students need in the 21st century (Jacobs, 2010). The literature review found that limited research exists on critical thinking development in secondary education (Abrami et al., 2008). Despite this lack of research, Abrami and his colleagues found that what Ennis' (1989) describes as 'mixed method' was the most effective instructional strategy to enhance critical thinking development. The IB-DP requires all students to complete a course entitled Theory of Knowledge. The aims and objectives of this course are, in part, to develop the critical thinking skills of its students (IBO,

2006). A study of the curriculum indicates that the IB has chosen the mixed-method for its delivery model for this course in its attempt to accomplish these goals.

Students in the treatment group, or IB-DP students, experienced significantly greater growth than students in the comparison group; however, further analysis revealed that treatment could not be isolated as the cause of critical thinking development. Students with greater academic achievement, as measured by performance on PSAT, were seen as the likely cause for the superior critical thinking development exhibited by the treatment group. Even though the treatment may have played a role in the development of critical thinking in the IB-DP students, it could not be isolated as the primary cause. The data support the hypothesis that the significantly larger growth from pretest to posttest TER Total Scores among the treatment group of students was due to the fact that students that have previously demonstrated greater academic achievement self-selected into the IB-DP. This does not mean that the TOK course does not increase students' critical thinking skills but rather further study is required to determine its effectiveness.

The further, rudimentary analysis comparing students' performance on the PSAT-CR and average gain scores on the TER generated interesting results. Despite the small sample size of students in Quadrant 4 (low scores on the PSAT-CR and IB-DP students) limiting the significance of the results it is still worth mentioning the findings. It appears that the IB-DP and the TOK course has the greatest impact on the development of critical thinking on students that performed poorly on the PSAT-CR. This challenges some of the traditionally held beliefs that the IB-DP is designed for high academically achieving students. Although, the IB takes great pains to promote the belief that hard work and motivation are the key components to success in

the IB-DP it is commonly thought that the IB-DP is designed for the academically gifted student. These results challenge this long held belief.

Recommendations for Future Research

Several recommendations for research can be made. Valuable information could be gained from studying two credit-based transitional programs that purport to develop critical thinking skills but utilize a different method of delivery. Credit-based transitional programs are designed to offer secondary students opportunities to take first year college-level coursework and earn college credit prior to entering college (Bailey & Karp, 2003). Two common credit-based transitional programs are the International Baccalaureate – Diploma Programme and the Advanced Placement (AP) program. The IB-DP utilizes Ennis' (1989) mixed method instructional approach while the AP appears to use the infusion instructional method. The AP does not offer a course similar to the required TOK course for IB-DP students. Valuable information could be gained from studying two comparable groups of students who self-selected to enter into rigorous and challenging programs that approach critical thinking instruction differently.

A study researching the affective domain of critical thinking in conjunction with the cognitive domain would add critical insight into the development of critical thinking. CCTST, the designers of the TER, have a series of tests that measure both the affective and cognitive domains. The affective domain targeted by the CCTST is a student's motivation to learn and apply critical thinking skills. This is another area with limited empirical research. A study combining the two domains of critical thinking would help educational leaders determine

necessary learning objectives to target in a critical thinking curriculum.

An additional area of study would be a longitudinal study that linked high school student success on a critical thinking test, similar to the TER, and success in post secondary education. This would provide information as to the value of critical thinking in secondary education. It would also provide college admissions personnel with important information as to skills that are beneficial for success at college.

A qualitative study would provide further information to enhance the quality of this research. It would be valuable to include a review of curricula; interviews with students, teachers, administrators, parents, and counselors; observing classes; and reviewing assessment practices developed to evaluate critical thinking development. This research would provide valuable information about the effectiveness of critical thinking development in a school. This would assist school leaders who are attempting to effectively implement a critical thinking program in their school.

A longitudinal study examining the critical thinking development throughout students' academic careers would provide valuable information for educational leaders. This study highlighted that students' previous academic ability directly correlated with growth of critical thinking. Little empirical research has been completed on school-aged children. Most critical thinking research has targeted post-secondary students. A longitudinal study could help identify the best age for developing the capacity for critical thinking. Further research in this area could enhance our understanding of critical thinking development and the most appropriate time to target its development.

Lastly, further study should target the development of critical thinking in students that

have not performed well on traditional standardized tests yet still self-select to enter an academically challenging program. A study combining the affective dispositions of critical thinking (open-mindedness, flexibility, inquisitiveness, judiciousness, and persistence) combined with critical thinking skill development in students identified in Quadrant 4 (low scores on standardized tests but self-selecting into a rigorous program) would provide interesting and important data. The understanding of how and when these important skills are developed would provide educators with valuable information. The pocket of students that exhibit the characteristics of intellectual perseverance, motivation, and belief in one-self would provide valuable and interesting information for educators. Can these skills be developed in a traditional school setting?

Conclusion

This study measured the development of critical thinking skills of students enrolled in the IB-DP and students who were not. It examined research on effective critical thinking strategies and programs designed to teach critical thinking. The review of the literature highlighted the need for further research in this area.

The analysis of the data gathered in this research study showed that students enrolled in the IB-DP with its mandatory TOK class developed critical thinking skills to a greater degree than those students in the comparison group. Further analysis, however, revealed that prior academic ability, as demonstrated by PSAT scores, eliminated the effect of treatment on student TER performance. This was a disappointing finding for the researcher as it was hoped that this study would show the TOK course caused growth on critical thinking.

Further research needs to be conducted to build upon the results found in this study. Researching the impact of critical thinking development on student learning is a complex enterprise that is difficult to conduct for a myriad of reasons. These reasons include gaining access to schools, documenting the extent to which teachers follow the curriculum, collecting data over a sustained period of time, identifying appropriate comparison groups, isolating variables, and accessing valid measures of student critical thinking achievement. I conclude that the superior growth for the treatment group on the TER Total Score is likely due to mean academic ability differences between students in the two groups rather than due to the treatment. Unfortunately, these analyses do not support the hypothesis that the TOK course leads to increased achievement in critical thinking as measured by TER Total Scores. However, the rudimentary, supplemental analysis of high and low performers on the PSAT-CR may suggest that the IB-DP and the TOK course is extremely beneficial to those students that self-select to enter a rigorous program despite previous poor performance on standardized tests.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., et al. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78*, 1102-1134.

Ary, D., Jacobs, L., Razavieh, A., & Sorenson, C. (2009). *Introduction to research in education* (8th ed.). Belmont, CA: Wadsworth Publishing.

Association of American Colleges and Universities (1985). *Integrity in the college curriculum.* Washington, DC: Association of American Colleges and Universities.

Atkinson, D. (1997). A critical approach to critical thinking in TESOL. *TESOL Quarterly, 31*, 71-94.

Bailey, T. & Karp, M. (2003). *Promoting college access and success: a review of credit-based transitional programs.* New York: Teachers College Columbia University, Community College Research Center.

Bailin, S., & Siegel, H. (2007). Critical thinking, in *The Blackwell guide to the philosophy of education* (eds N. Blake, P. Smeyers, R. Smith and P. Standis) (pp. 181-193), Blackwell Publishing Ltd., Oxford, UK. doi: 10.1002/9780470996294.ch11

Baker, P. J. (1981). Learning sociology and assessing critical thinking. *Teaching Sociology, 8*, 325-363.

Barak, M., Ben-Chaim, D. & Zoller, U. (2007). Purposefully teaching for the promotion of higher-order thinking skills: a case for critical thinking. *Research in Science Education, 37*, 353-369.

Barrow, R. (1991). The generic fallacy. *Educational Philosophy and Theory, 23*, 7-17.

Ben-Chaim, D., Ron, S., & Zoller, U. (2000). The disposition of eleventh-grade science students toward critical thinking. *Journal of Science Education and Technology, 9*, 149-159.

Bissell, A. N., & Lemons, P. P. (2006). A new method for assessing critical thinking in the classroom. *Bioscience, 56*, 66-72.

Bitter, G. G., & Legacy, J. M. (2007). *Using technology in the classroom* (7th ed.). Boston: Allyn & Bacon.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals (1st ed.).* Harlow, Essex, England: Longman Group.

Brocklesby, J., & Cummings, S. (1996). Foucault plays Habermas: An alternative philosophical underpinning for critical systems thinking. *The Journal of the Operational Research Society, 47*, 741-754.

Brookfield, S. D. (1987). *Developing critical thinkers: Challenging adults to explore alternative ways of thinking and acting.* San Francisco: Jossey-Bass.

Chambers, A., Bartle, K., Carter-Wells, A. J., Bagwell, J., Greenbaum, J., Padget, D., et al. (2000). Creative and active strategies to promote critical thinking. *Yearbook (Claremont Reading Conference),* 58-69.

Clark, D. R. (2010). *Bloom's taxonomy of learning domains: The three types of learning.* Retrieved 1/28, 2012, from http://www.nwlink.com/~donclark/hrd/bloom.html

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hilldale, NJ: Lawrence Erlbaum.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). London: Routledge Falmer.

Daly, W. M. (1998). Critical thinking as an outcome of nursing education. what is it? Why is it important to nursing practice? *Journal of Advanced Nursing, 28,* 323-331.

Dewey, J. (1909/2008). *Moral principles in education* (EBook #25172 ed.) Project Gutenberg EBook. Retrieved from http://www.pgdp.net

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. Boston: Heath and Company.

DiGiorgio, C. (2010). Choices of Students, Parents, and Teachers and Their Effects on Schools and Communities: A Case Study of a New Enriched High School Program. *Journal of School Choice*, *4*(3), 278-292.

Ennis, R. H. (1979). Presidential address: A conception of rational thinking. In Coombs, J. (Ed.), *Philosophy of education*. Normal, Il: Philosophy of Education Society.

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In Baron, J. B.; Sternberg, R. J. (Ed.), *Teaching thinking skills: Theory and practice* (pp. 9-26). New York: W. H. Freeman/Times Books/Henry Holt & Co.

Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher, 18,* 4-10.

Ennis, R. H. (1990). The extent to which critical thinking is subject-specific: Further clarification. *Educational Researcher, 19,* 13-16.

Ennis, R. H. (1994). Critical thinking dispositions: Their nature and assessability. *Informal Logic, 18* (2 & 3), 165-182.

Facione, P. (1990). *Critical thinking: A statement of expert consensus for the purpose of educational assessment and introduction. The Delphi Report: Research findings and recommendations.* Milbrae, CA: American Philosophical Association. (ERIC Document Reproduction Service No. 315 423)

Facione, P. (2011). *Critical thinking: What it is and why it counts.* Retrieved February 9, 2012, from http://www.insightassessment.com/CT-Resources/Independent-Critical-Thinking-Research/pdf-file/Critical-Thinking-What-It-Is-and-Why-It-Counts-PDF/(language)/eng-US

Facione, P., & Facione, N. (1992). *The California Critical Thinking Disposition Inventory (CCTDI).* Millbrae, CA: California Academic.

Facione, P., & Facione, N. (1994). *The California Critical Thinking Skills Test - Test Manual.* Millbrae, CA: California Academic.

Facione, P. A., Facione, N. C., & Giancarlo, C. A. (1997). Professional judgment and the disposition toward critical thinking. *The California Academic Press,* 1-17.

Facione, P., Facione, N., & Giancarlo, C. (2000). The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal Logic, 20*, 61-84.

Facione, P. A., Facione, N. C., Tiwari, A., & Yuen, F. (2009). Chinese and American perspectives on the pervasive human phenomenon of critical thinking. *Journal of Peking University (Philosophy and Social Sciences), 46*, 55-62.

Facione, P., Facione, N., & Winterhalter, K. (2012). *The test of everyday reasoning – (TER): Test Manual*. Millbrae, CA: California Academic Press.

Facione, P. A., Sanchez, C. A., Facione, N. C., & Gainen, J. (1995). The disposition toward critical thinking. *Journal of General Education, 44*(1), 1-25.

Fraenkel, J., & Wallen, N. (2006) *How to design and evaluate research in education* (6th ed.). New York: McGraw-Hill.

Garson, G. D. (2012). General Linear Models: Univariate GLM, Anova/Ancova, Repeated Measures. Asheboro, NC: Statistical Publishing Associates.

Gay, L. R., Mills, G. E., & Airasian, P. (2009). *Educational research: Competencies for analysis and application* (9th ed.). Upper Saddle River, NJ: Pearson.

Geertsen, H. R. (2003). Rethinking thinking about higher-level thinking. *Teaching Sociology, 31*, 1-19.

Giancarlo, C. A., Blohm, S. W., & Urdan, T. (2004). Assessing secondary students' disposition toward critical thinking: Development of the California measure of mental motivation. *Education and Psychological Measurement. 64*, 347-364.

Glaser, E. M. (1941). *An experiment in the development of critical thinking.* New York: Teachers College Columbia University.

Glaser, E. M. (1985). Critical thinking: Educating for responsible citizenship in a democracy. *National Forum: Phi Kappa Phi Journal, 65*, 24-27.

Grauerholz, L., & Bouma-Holtrop, S. (2003). Exploring critical sociological thinking. *Teaching Sociology, 31*, 485-496.

Greenlaw, S. A., & DeLoach, S. B. (2003). Teaching critical thinking with electronic discussion. *Journal of Economic Education, 34*, 36-52.

Habermas, J. (1989). *The structural transformation of the public sphere* (T. Burger Trans.). Cambridge, MA: MIT Press. (Original work published 1962).

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American Psychologist, 53*, 449-455.

Hare, W. (1998). Critical thinking as an aim of education. *Critical Thinking Across the Disciplines, 18*(2), 38-51.

Hartman-Haas, H. J. (1984). An evaluation of the holistic approach to improving thinking. (ERIC Document Reproduction Service No. 242 759).

Hegel, G. W. F. (1991). *Elements of the philosophy of right* (H. B. Nisbett Trans.). New York: Cambridge University Press. (Originally published 1822).

Hemming, H. (2000). Encouraging critical thinking: "but… what does that mean?" *McGill Journal of Education, 35*, 173-186.

Heydebrand, W. (2004). Critical theory. *Contemporary Sociology, 33*, 732-735.

Hudgins, B. B., & Edelman, S. (1988). Children's self-directed critical thinking. *The Journal of Educational Research, 81*, 262-273.

International Baccalaureate Organization. (2001) (*updated 2005). Application procedure for candidate schools.* Retrieved March 28, 2012, from http://www.ibo.org/school/candidate/documents/applicationprodecure.pdf

International Baccalaureate Organization. (2002). *Diploma programme school's guide.* Retrieved January 4, 2012, from http://www.ibo.org/diploma/documents/schools_guide_diploma.pdf

International Baccalaureate Organization. (2006a) (*updated November, 2008*). *Diploma programme theory of knowledge - guide*. Clippenham, Wiltshire: Antony Rowe Ltd.

International Baccalaureate Organization. (2006b). *Rules for IB world schools: Diploma programme.* Retrieved March 28, 2012, from http://www.ibo.org/documentlibrary/rules_ibworldschools/documents/DP_rules_en.pdf

International Baccalaureate Organization. (2007). *IB diploma programme: A guide for universities and colleges.* Retrieved January 7, 2012, from https://www.ibo.org/communications/publications/documents/e_uni_recognition_brochure_07.PDF

International Baccalaureate Organization. (2008). *The IB diploma programme – flyer* Retrieved January 7, 2012, from https://www.ibo.org/communications/publications/documents/DPflyerEng.pdf

International Baccalaureate Organization. (2009a). *Diploma programme A basis for practice.* Retrieved January 4, 2012, from http://www.ibo.org/diploma/documents/basis_diploma.pdf

International Baccalaureate Organization. (2009b). *Mission and Strategy.* Retrieved May 13, 2009, from http://www.ibo.org/mission/index.cfm

International Baccalaureate Organization. (2009c). *The IB diploma programme: From principles to practice.* Cardiff, Wales: Peterson House.

International Baccalaureate Organization. (2010). *Guide to school authorization: Diploma Programme.* Retrieved March 28, 2012, from *http://www.ibo.org/become/guidance/documents/DP_Guidetoschoolauthorization.pdf*

International Baccalaureate Organization. (2011). *The IB diploma programme statistical bulletin.* Retrieved March 28, 2012, from http://www.ibo.org/facts/statbulletin/dpstats/documents/statistical_bulletin_may_2011.pdf

International Baccalaureate Organization. (2012a). *Diploma programme curriculum core requirements: Theory of knowledge.* Retrieved January 3, 2012, from www.ibo.org/diploma/curriculum/core/knowledge/

International Baccalaureate Organization. (2012b). *Diploma programme curriculum-core requirements.* Retrieved January 16, 2012, from http://ibo.org/diploma/curriculum/core/

Jacobs, H. H. (2010). *Curriculum 21: Essential education for a changing world.* Alexandria, VA: ASCD.

Kamin, C., O'Sullivan, P., & Deterding, R. (2002). *Does project L.I.V.E. case modality impact critical thinking in PBL groups?* (April ed.). New Orleans, LA: Paper presented at the annual meeting of the American Educational Research Association.

Kant, I. (1998). In Guyer P., Wood A. (Eds.), *Critique of pure reason* [Kritik der reinen Vernunft] (P. Guyer, A. Wood Trans.). New York: Cambridge University Press. (Original published in 1781).

Kennedy, M., Fisher, M.B., & Ennis, R.H. (1991). Critical thinking: Literature review and needed research. In L. Idol and B.F. Jones (Eds.), *Educational values and cognitive instruction: Implications for reform* (pp. 11-40). Hillsdale, NJ: Erlbaum.

Kuwait Investment Authority, (2012). The MBA Scholarship Program. Retrieved October 5, 2012, from http://www.kia.gov.kw/En/MD_Office/Training_Dep/MBA/Pages/default.aspx

Lipman, M. (1991). *Thinking in education.* New York: Cambridge University Press.

Magee, B. (1990). Hegel and Marx. In Magee, E (Ed.), *The Great Philosophers – An Introduction to Western Philosophy* (pp. 188-208). Oxford: Oxford University Press.

Marx, K., & Engels, F. (2010). The communist manifesto. *The two narratives of political economy* (Capaldi, N & Lloyd, G. Trans) (pp. 389-408). New York: John Wiley & Sons, Inc. (Originally published in 1822).

Mathews, J., & Mathews, L. (2012). Going Global, for Rich and Poor. *School Administrator*, *69*(8), 44-51.

McCarthy-Tucker, S. N. (1998). Teaching logic to adolescents to improve thinking skills. *The International Journal of Creativity & Problem Solving*, *8*(1), 45-66.

McPeck, J. E. (1990). Critical thinking and the "trivial pursuit" theory of knowledge. *Teaching Philosophy, 8*, 295-308.

McPeck, J. E. (1990). Critical thinking and subject specificity: A reply to Ennis. *Educational Researcher, 19*, 10-12.

Mingers, J. (1992). Recent developments in critical management science. *The Journal of the Operational Research Society, 43*, 1-10.

National Education Goals Panel. (1992). *Executive summary: The national education goals report-building a nation of learners.* Washington, DC: Author.

National Institute of Education Study Group. (1984). *Involvement in learning: Realizing the potential of American higher education.* Washington, DC: National Institute of Education, US Department of Education.

Norris, S. P. (1985). Synthesis of research on critical thinking. *Educational Leadership, 42*, 40-45.

Norris, S. P., & Ennis, R. H. (1989). *The practitioners' guide to teaching thinking series: Evaluating critical thinking*. Pacific Grove, CA: Midwest Publications.

Paul, R. W. (1984). Critical thinking: Fundamental to education for a free society. *Educational Leadership, 42*(1), 4-14.

Paul, R. W. (1992). *Critical thinking: What every person needs to survive in a rapidly changing world* (2nd ed.). Santa Rosa, CA: The Foundation for Critical Thinking.

Paul, R. W. (1995). *Critical thinking: How to prepare students for a rapidly changing world* (4th ed.). Santa Rosa, CA: Foundation for Critical Thinking.

Peterson, A. D. C. (2003). *Schools across frontiers: The story of International Baccalaureate and the United World Colleges* (2nd ed.). Chicago: Open Court.

Pithers, R. T., & Soden, R. (2000). Critical thinking in education: A review. *Educational Research, 42*, 237-249.

Riesenmy, M. R., Mitchell, S., Hudgins, B. B., & Ebel, D. (1991). Retention and transfer of children's self-directed critical thinking skills. *The Journal of Educational Research, 85*, 14-25.

Rorty, R. (1991). Habermas and Lyotard on postmodernity. In Hostelry, I (Ed.), *Zeitgeist in Babel* (pp. 84-97). Indianapolis: University of Indiana Press.

Scheffler, I. (1973). *Reason and teaching*. Indianapolis: Bobbs-Merrill.

Siegel, H. (1990). *Educating reason: Rationality, critical thinking and education*. New York: Routledge.

Smith, G. (2002). Are there domain-specific thinking skills? *Journal of Philosophy of Education, 36*, 207-227.

Sokol, A., Oget, D., Sonntag, M., & Khomenko, N. (2008). The development of inventive thinking skills in the upper secondary language classroom. *Thinking Skills and Creativity, 3*, 34-46.

Sternberg, R. J. (1986). *Critical thinking: Its nature, measurement, and improvement*. (ERIC Document Reproduction Service No. ED 272 882).

ten Dam, G., & Volman, M. (2004). Critical thinking as a citizenship competence: Teaching strategies. *Learning and Instruction, 14*, 359-379.

Thayer-Bacon, B. (2000). *Transforming critical thinking: Thinking constructively*. New York: Teachers College.

Tishman, S. & Andrade, A. (1996). Thinking Dispositions: A review of current theories, practices, and issues [Online]. Retrieved January 18, 2011, from http://learnweb.harvard.edu/alps/thinking/docs/Dispositions.pdf

Toulmin, S. (1990). *Cosmopolis: The hidden agenda of modernity*. Chicago: University of Chicago Press.

Trochim, W. M. (2006). The Research Methods Knowledge Base, 2nd Edition [Online]. Retrieved October 13, 2012, from http://www.socialresearchmethods.net/kb/

Vaughn, L. (2008). *The power of critical thinking: Effective reasoning about ordinary and extraordinary claims* (2nd ed.). London: Oxford University Press.

Wade, C., & Tavris, C. (1993). *Critical and creative thinking: The case of love and war.* New York: Harper Collins.

Walters, K. S. (1986). Critical thinking in liberal education: A case of overkill? *Liberal Education, 72*, 233-244.

Walters, K. S. (1994). *Re-thinking reason: New perspectives in critical thinking*. Albany, NY: State University of New York Press.

Watson, G., & Glaser, E. M. (1942). *Watson-Glaser test of critical thinking*. Oxford, England: World Book Co.

Weinberg, S. L., & Abramowitz, S. K. (2008). *Statistics using SPSS: An integrative approach* (2nd ed.). New York: Cambridge University Press.

Zemplén, G. (2007). Conflicting agendas: Critical thinking versus science education in the international baccalaureate theory of knowledge course. *Science & Education, 16*, 167-196.

Zohar, A., Weinberger, Y., & Tamir, P. (1994). The effect of the biology critical thinking project on the development of critical thinking. *Journal of Research in Science Teaching, 31*, 183-196.

APPENDIX A: AMERICAN INTERNATIONAL SCHOOL KUWAIT LETTER OF CONSENT

المدرسة الأمريكية الدولية
American International School

2 May 2011

Lehigh University
27 Memorial Drive West
Bethlehem PA 18015
USA

RE: Letter of Permission/Blair Lee

To Whom It May Concern:

As Superintendent of the American International School Kuwait (AIS), I have given Blair M. Lee permission to review and use certain student data collected by AIS for preparation of his Doctoral Dissertation. Specifically, I give him permission to use the Test of Everyday Reasoning test results and the PSAT results for students in Grades 11 and 12 from the years 2008, 2009, 2010, and 2011. Standard student profile data for these students may also be used (gender, enrollment in IB-DP, courses taken, grade point average, and nationality).

I have spoken with Mr. Lee and understand the scope of his research, and how he will be using this data. All information to be gathered will be done in a confidential and appropriate manner.

As Superintendent of AIS, I have the authority to release this data for research purposes. Should you have any questions, please feel free to contact me.

Sincerely,

Russell McLean
Superintendent

admin@aiskuwait.org بريد إلكتروني - 22255156 فاكس: 22255155 تلفون: الكويت - 22033 السالمية - 3267 ص.ب:
P.O.Box: 3267 - Salmiya 22033 - **Kuwait** - Tel.: 22255155 - Fax: 22255156 - E-mail: admin@aiskuwait.org
www.aiskuwait.org

APPENDIX B: THEORY OF KNOWLEDGE: AIMS AND OBJECTIVES

Aims and objectives

Aims

The aims of the TOK course are to:

- develop a fascination with the richness of knowledge as a human endeavour, and an understanding of the empowerment that follows from reflecting upon it
- develop an awareness of how knowledge is constructed, critically examined, evaluated and renewed, by communities and individuals
- encourage students to reflect on their experiences as learners, in everyday life and in the Diploma Programme, and to make connections between academic disciplines and between thoughts, feelings and actions
- encourage an interest in the diversity of ways of thinking and ways of living of individuals and communities, and an awareness of personal and ideological assumptions, including participants' own
- encourage consideration of the responsibilities originating from the relationship between knowledge, the community and the individual as citizen of the world.

Objectives

Having followed the TOK course, students should be able to:

1. analyse critically knowledge claims, their underlying assumptions and their implications
2. generate questions, explanations, conjectures, hypotheses, alternative ideas and possible solutions in response to knowledge issues concerning areas of knowledge, ways of knowing and students' own experience as learners
3. demonstrate an understanding of different perspectives on knowledge issues
4. draw links and make effective comparisons between different approaches to knowledge issues that derive from areas of knowledge, ways of knowing, theoretical positions and cultural values
5. demonstrate an ability to give a personal, self-aware response to a knowledge issue
6. formulate and communicate ideas clearly with due regard for accuracy and academic honesty.

APPENDIX C: IRB APPROVAL

LEHIGH UNIVERSITY

Office of Research and Sponsored Programs
[illegible]

DATE:	October 14, 2011
TO:	Blair Lee, MA
FROM:	Lehigh University IRB
STUDY TITLE:	[272222-1] Critical Thinking Skills Development in Secondary Students
IRB REFERENCE #:	12/036
SUBMISSION TYPE:	New Project
ACTION:	APPROVED
APPROVAL DATE:	October 14, 2011
PROJECT EXPIRATION DATE:	October 13, 2012
INITIAL APPROVAL DATE:	October 14, 2011
REVIEW TYPE:	Expedited Review

Thank you for your submission of materials for this research study. The Lehigh University IRB has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a study design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission. This approval is valid for one year.

This submission has received Expedited Review based on the Lehigh University Policy on the Protection of Human Subjects in Research.

Reapproval and Progress Report: The current approval will expire on October 13, 2012. If you wish to continue beyond that time, you must submit a renewal request and progress report on the Continuing Review form via IRBNet. This protocol will be due for continuing IRB review ***60 days*** before the expiration date of October 13, 2012.

Informed Consent: Please remember that INFORMED CONSENT is a process beginning with a description of the study and insurance of subject understanding followed by a signed consent form. Informed consent must continue throughout the study via a dialogue between the researcher and the research subject. The Lehigh University policy requires each subject receive a copy of the signed consent document.

Changes or Amendments: If during the year you propose significant changes in your approved protocol, please submit these changes for review using the amendment/modification form through IRBNet. The proposed changes may not be initiated without IRB approval (except when necessary to eliminate immediate hazards to subjects).

Adverse Events: All SERIOUS and UNEXPECTED adverse events must be reported to this office. Please use the appropriate adverse event forms through IRBNet for this procedure. All sponsor reporting requirements should also be followed. Any injuries or other unanticipated problems involving risks to research subjects and others resulting from this study must be reported promptly to the Lehigh University IRB. If the problem is serious, approval may be withdrawn pending further review by the committee.

Non-compliance or Complaints: Please report all NON-COMPLIANCE issues or COMPLAINTS regarding this study to this office.

Completion of Study and Record Retention: Please notify the Lehigh University IRB as soon as the research has been completed. Study records, including full protocols and signed consent forms (originals) for each subject, must be kept in a secured location by the investigator for 3 years following the study's completion.

If you have any questions, please contact Susan E. Disidore at 610-758-3020 (E-mail:sus5@lehigh.edu) or Troy Boni at 610-758-2985 (E-mail:tdb308@lehigh.edu). Please include your study title and reference number in all correspondence with this office.

APPENDIX D: EXAMPLE OF KUWAIT SCHOLARSHIP

The MBA Scholarships Program

KIA offers ten academic scholarships per year to Kuwaiti Nationals seeking to obtain their MBA (Masters in Business Administration) degree from a top ranked university in the world with high academic standards. The Cultural Division in the Kuwait Embassies supervise the Students during their courses of study.

Scholarship Terms and Conditions

- Maximum period of sponsorship: Two academic years
- Starting date of sponsorship: First day of classes

Scholarship Requirements

1. The applicant is required to have Kuwaiti Citizenship.
2. Unconditional admission to the MBA program at one of the universities approved by KIA.
3. Registration in the Academic MBA program and not Executive MBA (EMBA).
4. Dual Degree programs will not be accepted.
5. The applicant should not have a scholarship from another source.
6. The applicant must register as a full-time student.
7. Satisfactory KIA interview.
8. The applicant should not have an existing MBA degree.
9. GMAT score should be no less than 600.
10. Applicant must provide proof of corresponding with universities.

All scholarships applications should be submitted two months prior to the start of the first semester.

Financial Benefits in Kuwaiti Dinars (K.D.)

- Monthly stipend: o Twelve hundred (1200) K.D.
- Book allowance per year: o Five hundred (500) K.D.
- Computer allowance: o One time five hundred (500) K.D.
- Annual airline ticket: o Seven hundred fifty (750) K.D.

- Research and scientific journal subscription: Fifty (50) K.D. per academic year.
- The student may travel within 30 days of the program starting date and will be entitled to two months stipend and KD 750 airline ticket
- KIA will meet all expenses related to tuition, fees, examination charges and thesis costs. A copy of the thesis proposal must be submitted to KIA for approval.
- Attendance at one professional conference during the degree program. The

conference must be recommended by the program director of the university and approved by KIA. Students are required to submit a report on the conference to KIA and will be entitled to an economy airline ticket when participating in conferences related to the students major.

- Students are required to maintain full-time status during each academic term.

- **Upon graduation the student is entitled to:**
  - 2 month stipend.
  - KD750 airline ticket.
  - KD200 cargo shipment of personal items.

- Code of conduct: Students are required to comply with the laws of the host country; follow the rules and regulation of their schools; comply with the Kuwait Cultural Division's policies; maintain perfect attendance and graduate on time.
- KIA maintains the right to terminate or suspend a student's scholarship mission, based on the evaluation and recommendation of the KIA scholarship committee. This committee may also consider a maximum scholarship extension of six months. Financial entitlements during the period of scholarship suspension or extension will be determined by scholarship committees.
- A Student's resignation from scholarship becomes effective on the day the scholarship committee acknowledges the resignation.

Penalties

- KIA reserves the right to request reimbursement of its expenditure if the student fails to successfully complete the degree requirements according to the terms and conditions of the scholarship regulations.
- In addition to these rules and regulations, Civil Service Law No.15 of 1979 will be applied to all scholarship recipients.

KIA reserves the right to modify, add, delete or interpret the articles of the scholarship regulations in accordance with the committee's recommendations.

retrieved from:
http://www.kia.gov.kw/En/MD_Office/Training_Dep/MBA/Pages/default.aspx

APPENDIX E: TEST OF ASSUMPTIONS

For each ANCOVA model, there are assumptions that must be examined. The first assumption is that the variance of the observations around their respective group means (error variances) will be equal across groups. This is called the homogeneity of variance assumption and can be examined with a Levene's test. A Levene's test associated with a nonsignificant *F* statistic indicates the assumption was met. The results for Models A to F are presented in Table A. The dependent variable was Total score on the TER. In all but one model (Model C) the assumption was violated. However, I concluded that the ANOVA tests were robust to violations of the assumption due to the relatively large sample size. Typically, analyses of groups with more than 30 observations are robust to violations of the homogeneity of variance assumption. The homogeneity of variance assumption is also necessary for MANCOVA tests. The tests of this assumption for MANCOVA Models A1 to F2 are presented in Table B. In all but two cases (A1 for Evaluation and A2 for Deduction) the assumption was met. Based on my prior argument, I concluded that the MANCOVAs would be robust to violations of the assumption due to relatively large subsample sizes (n = 150).

Table A

Levene's Test of Equality of Error Variances for Models A to F

Model	Dependent Variable	*F*	$p < .05$	Conclusion
A	Total	7.33	Sig.	Assumption not met
B	Total	7.19	Sig.	Assumption not met
C	Total	2.54	NS	Assumption met
D	Total	5.10	Sig.	Assumption not met
E	Total	4.30	Sig.	Assumption not met
F	Total	4.30	Sig.	Assumption not met

Table B

Levene's Test of Equality of Error Variances for Models A1 to F2

Model	Dependent Variable	*F*	*p* < .05	Conclusion
A1	Analysis	0.69	NS	Assumption met
	Inference	0.69	NS	Assumption met
	Evaluation	5.37	Sig.	Assumption not met
B1	Analysis	0.22	NS	Assumption met
	Inference	0.45	NS	Assumption met
	Evaluation	2.13	NS	Assumption met
C1	Analysis	0.01	NS	Assumption met
	Inference	0.81	NS	Assumption met
	Evaluation	1.84	NS	Assumption met
D1	Analysis	0.27	NS	Assumption met
	Inference	0.53	NS	Assumption met
	Evaluation	2.86	NS	Assumption met
E1	Analysis	0.00	NS	Assumption met
	Inference	0.62	NS	Assumption met
	Evaluation	2.87	NS	Assumption met
F1	Analysis	0.00	NS	Assumption met
	Inference	0.62	NS	Assumption met
	Evaluation	2.87	NS	Assumption met

Table B (continued)

A2	Deduction	5.98	Sig.	Assumption not met
	Induction	3.17	NS	Assumption met
B2	Deduction	3.03	NS	Assumption met
	Induction	1.09	NS	Assumption met
C2	Deduction	0.96	NS	Assumption met
	Induction	1.06	NS	Assumption met
D2	Deduction	2.45	NS	Assumption met
	Induction	2.26	NS	Assumption met
E2	Deduction	1.88	NS	Assumption met
	Induction	1.55	NS	Assumption met
F2	Deduction	1.88	NS	Assumption met
	Induction	1.55	NS	Assumption met

The extension of the ANCOVA assumption of equal variances to the case of MANCOVA requires that the dependent variables have the same variance-covariance matrix in each group. Table C presents the results of Box's test of equality of variance-covariance matrices for Model A1 to F2. In every case, the assumption was met as indicated by a nonsignificant probability value associated with the *M* statistic.

Table C

Box's Test of Equality of Variance-Covariance Matrices

Model	Box's *M*	Significant	Conclusion
A1	3.72	NS	Assumption met
B1	14.50	NS	Assumption met
C1	6.20	NS	Assumption met
D1	2.58	NS	Assumption met
E1	3.61	NS	Assumption met
F1	3.61	NS	Assumption met
A2	3.94	NS	Assumption met
B2	13.53	NS	Assumption met
C2	2.38	NS	Assumption met
D2	2.55	NS	Assumption met
E2	2.88	NS	Assumption met
F2	2.88	NS	Assumption met

Table D shows the Chi-Square values associated with Bartlett's test of sphericity. The reason a researcher conducts a MANCOVA rather than an ANCOVA is he or she assumes the dependent variables in a given model are connected in some important ways. For example, in Model A1, there are three dependent variables, posttest scores on Analysis, Inference, and Evaluation. The researcher assumes these are three different dimensions of a common factor called intellectual skill. If the three variables *are* linked by a common factor, one must assume that the three variables are well-correlated with each other. In order to test whether these three variables are well-correlated SPSS's MANCOVA routine produces Bartlett's test of sphericity. Each Bartlett's test has a corresponding Chi-square value. If the Chi-square value is significant (i.e. significantly different than 0) we conclude that the dependent variables are sufficiently intercorrelated and this assumption of is met. The results of Bartlett's tests of sphericity for each MANCOVA are presented in Table D. The conclusion is that the assumption is met for all Models and that a multivariate test is reasonable for each one.

Table D

Chi-Square Values for Bartlett's Test of the Sphericity Assumption for Models A1 to F2

Model	Chi-Square	*df*	Significant	Conclusion
A1	86.41	5	Sig.	Assumption met
B1	86.14	5	Sig.	Assumption met
C1	19.82	5	Sig.	Assumption met
D1	78.16	5	Sig.	Assumption met
E1	61.52	5	Sig.	Assumption met
F1	61.52	5	Sig.	Assumption met
A2	73.72	2	Sig.	Assumption met
B2	76.51	2	Sig.	Assumption met
C2	7.59	2	Sig.	Assumption met
D2	67.72	2	Sig.	Assumption met
E2	54.48	2	Sig.	Assumption met
F2	54.48	2	Sig.	Assumption met

APPENDIX F: FURTHER ANALYSIS OF PERFORMANCE ON PSAT-CRITICAL READING AND TER GAIN SCORES

Analysis Comparing Effect of Treatment to PSAT CR Score

On average, students in the treatment group gained 3.39 points on the TER Total Score from pretest to posttest. In comparison, the students in the comparison group gained an average of .79 points from pretest to posttest. Thus the mean gain for the IB-DP students, or treatment group, is approximately 4.3 times greater than the mean gain for the non-IB-DP students, or comparison group. This is an impressive difference; however, the *ex-post facto* study design requires further analysis of possible third variable explanations prior to reaching the conclusion that treatment caused the greater gains on the TER-Total Score.

In order to contrast the effect of treatment to the PSAT CR skills effect, I ran another analysis in which I bisected the sample into two groups. Group 1 consisted of students that scored in the top 50% on the PSAT CR test. Group 2 consisted of students that scored in the bottom 50% on the PSAT CR test. Students in Group 1 gained, on average, 3.14 points on the TER from pretest to posttest. Those in Group 2 gained, on average, .70 points on the TER. Thus, the students who scored in the top 50% on PSAT CR gained, on average, 4.5 times more than those who scored in the bottom 50% on the PSAT CR.

This rudimentary analysis concludes that the advantage of scoring high on the PSAT CR (TER average gain scores that are 4.5 times higher) is greater than the advantage of being in the treatment group (TER gain scores that are only 4.3 times higher). This analysis demonstrates in a more simplistic method what the more refined ANCOVA analysis determined.

Four Quadrant Analysis (104 Comparison – 67 Treatment)

This analysis segmented the students into four categories. Students were first bisected into High Scorers on PSAT CR and Low Scorers on PSAT CR. Students were then divided again into either comparison or treatment groups. In Table ?, you can see that 29 Non-IB-DP students scored in the top 50% on the PSAT CR, whereas, the large majority of Non-IB-DP students (*n* = 71) scored in the bottom 50% on PSAT CR. Similarly, 45 IB-DP students scored in the top 50% on the PSAT CR and only 15 of the IB-DP students scored in the bottom 50% on the PSAT CR.

Table E

Quadrants: PSAT CR - High and Low score: Comparison and Treatment

	Comparison *Non-IB-DP*	*Treatment* *IB-DP*
Top 50% PSAT CR	Quadrant 1 *n* = 29	Quadrant 2 *n* = 45
Bottom 50% PSAT CR	Quadrant 3 *n* = 71	Quadrant 4 *n* = 15

We can gain some interesting insights by analyzing the average TER gain scores for students in each of these quadrants (see Table ?). IB-DP students who are in the top 50% of the PSAT-CR scores had a TER gain score of 3.38. This was not significantly different than the gain score for Non-IB-DP students who also scored high on the PSAT-CR (2.79). It did not matter

whether a student took the IB-DP course or not. Both groups of students that scored in the top 50% of the PSAT-CR showed impressive gains on the TER.

Non-IB-DP who scored low on the PSAT-CR showed marginal gains from pretest to posttest on the TER (only .08 points, on average). However, IB-DP students, who scored low on the PSAT-CR *did* show impressive gains from pretest to posttest on the TER (3.63 points, on average). The difficulty with this finding is that the sample size for this group (n = 15) was so small the effect was not significant due to low power.

Table F

Quadrants: TER Average Gain Scores

	Comparison *Non-IB-DP*	*Treatment* *IB-DP*
Top 50% PSAT CR	Quadrant 1 TER gain = 2.79	Quadrant 2 TER gain = 3.38
Bottom 50% PSAT CR	Quadrant 3 TER gain = .08	Quadrant 4 TER = 3.63

Vita

Dr. Blair Lee holds a Master's Degree in Education from the University of Regina, Canada and a Bachelor's Degree in Secondary Education, majoring in English, from the University of Saskatchewan, Canada.

In the spring of 2012, Blair assumed the role of Assistant Head of School – Academics at Seoul Foreign School in Seoul, South Korea. Blair spent nine years in a number of vice-principal and principal positions in Canada before heading into the international market in 2004. He worked at the American International School of Kuwait for seven years, serving initially as the Middle School Principal before assuming the position of High School Principal. Just prior to joining Seoul Foreign School, Blair served as the Superintendent of Mount Zaagkam International School on the island of Papua, Indonesia.